Evolution of the gene translation machinery and its applications to drug discovery

Eva Maria Novoa Pardo

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR

EVOLUTION OF THE GENE TRANSLATION MACHINERY AND ITS APPLICATIONS TO DRUG DISCOVERY

Eva Maria Novoa Pardo 2012

PROGRAMA DE DOCTORAT DE BIOMEDICINA

TESIS REALIZADA EN EL LABORATORIO DE TRADUCCIÓN GENÉTICA

INSTITUT DE RECERCA BIOMÈDICA - BARCELONA

EVOLUCIÓN DE LA MAQUINARIA DE TRADUCCIÓN GENÉTICA Y SUS APLICACIONES EN EL DESCUBRIMIENTO DE FÁRMACOS

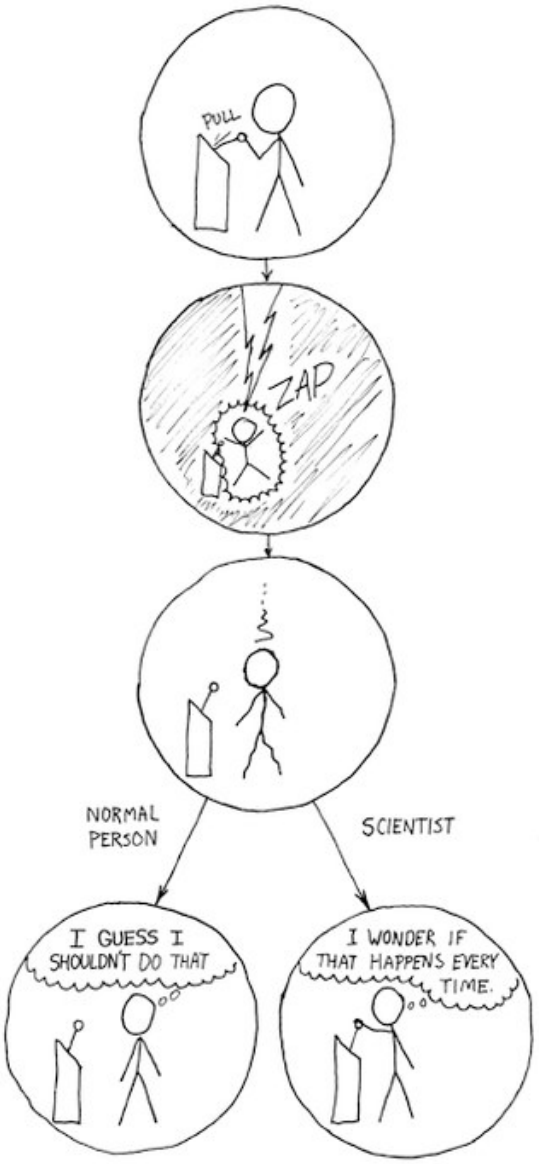
Memoria presentada por Eva Maria Novoa Pardo
para optar al grado de doctora por la Universitat de Barcelona

Director:	Tutor:	Doctoranda:
Lluís Ribas de Pouplana	**Modesto Orozco Lopez**	**Eva Maria Novoa Pardo**

A mi familia y a mi gente

A la meva família i a la meva gent

A miña familia e a miña xente



"The most beautiful thing we can experience is the mysterious."

-Albert Einstein-

-

Agradecimientos

¿Buf, por dónde empezar? A lo largo de una tesis son tantas las personas que te han ayudado, apoyado y marcado... Me gustaría aprovechar estas lineas para dar mi agradecimiento a todas aquellas personas que han hecho de esta etapa de mi vida personal y profesional una aventura maravillosa.

Me gustaría comenzar estas lineas dando las gracias a mi director de tesis, Lluis Ribas de Pouplana. No sólo por dirigirme la tesis, sino por tantas otras cosas... Por darme la oportunidad de hacer el doctorado contigo, por enseñarme a ser crítica, por formarme como estudiante. No sabes cuánto he aprendido estos años y cuan agradecida estoy. Por escucharme y apoyarme en mis momentos de dudas, tanto científicas como personales. Por las reuniones infinitas y "brain-stormings" que tanto echaré de menos. Por ser lo suficientemente cercano pero a la vez ejercer como jefe y director de tesis. Por la motivación que transmites cuando te hablo de proyectos e ideas que te parecen interesantes, pero también por tu sinceridad para decirme cuáles otros creías que no valían tanto la pena. Por tu capacidad para distinguirlos. Por enseñarme también a mí a distinguirlos poco a poco. Por darme alas y dejarme volar, y disfrutar de lo maravilloso de la curiosidad científica, pero a la vez, gracias por no dejarme que pierda el suelo de vista, y no perderme por el camino. Gracias por confiar en mi trabajo. Gracias por animarme a marchar fuera a hacer el postdoctorado. Sin tus consejos y ánimos, no tengo claro si me hubiera animado a tomar esta decisión. Gracias por ser mi director de tesis, no puedo imaginar ninguna otra forma mejor para haber llevado a cabo mi tesis que bajo tu dirección.

Tambien quiero dar las gracias a mi tutor, Modesto Orozco, por su tiempo y dedicación, y por dirigirme el proyecto de Máster. Gracias por darme la oportunidad de hacer el proyecto en tu grupo, y aprender de ti y de todos tus estudiantes las herramientas bioinformáticas que tanto me han ayudado a lo largo de mi tesis. Gracias por todo tu soporte científico y tus consejos incluso habiendo ya acabado mi proyecto de Máster.

Gracias de antemano a los miembros del tribunal, Manuel Santos, Baldomero Oliva y Lluís Ballell por acceder a evaluar y discutir este trabajo, y a Miquel Duran, Roderic Guigó y Xavier Daura por aceptar ser miembros suplentes de mi tribunal. También quiero agradecer a Xavier Barril, Alfred Cortés y Josep Farrera por formar parte de mi comité de evaluacion de tesis a lo largo de estos años. Gracias por vuestras críticas y sugerencias, han sido muy útiles y positivas para intentar mejorar mi trabajo a lo largo de estos años.

¿Qué puedo decir de mis compañeros de laboratorio? He sido una privilegiada por haber podido compartir la experiencia de estar en dos grupos distintos durante mi doctorado, primero en el de Modesto, y despues en el de Lluís. Nada más llegar, esa sonrisa profident que me dedicó Nacho, que se merece una mención especial aqui, jajaja. Y después, fueron llegando todos los demás: David, Jordi M, Sergi, Adam, Carles, Óscar, Guillem, Oliver, Rebeca, Nadine, Agustí, Jordi C, Antonella, Jose... Ay Jose, cuantas horas nos habremos estado peleando con el dichoso Maestro!!!!! Gracias por tu paciencia y tu ayuda, incluso cuando ya no estaba en el grupo, y por los cafés en la puerta del trabajo, que siempre se agradecían. Gracias Antonella por ayudarme con Maestro y Glide. Y también gracias Xavi, por tus consejos de docking y scripts de MOE, y por supervisar mi trabajo computacional estos años.

Y, casi dos añitos más tarde, me mudé de laboratorio. Y... ¡¡¡coger una pipeta otra vez!!!! Noelia, ¿qué habría hecho yo sin ti? No sólo me enseñaste cómo funcionaba el laboratorio, a hacer tRNAs, a transformar células, a hacer ensayos de aminoacilación.... todo eso es muchísimo, pero es lo de menos. Gracias por tu paciencia, por ser tan amable y a la vez tener energía para parar un tren. Yaiza, aunque compartimos muy poco tiempo el en labo, gracias por tus consejos antes de comenzar, y por dejar todo tan fantásticamente etiquetado y ordenado para los que hemos seguido usando tus cosas. Francesc, gracias por enseñarme a ser tan pulida haciendo Westerns, creo que es lo que mejor me sale de todo lo relacionado con experimental jejeje!! Tanit, por tu rigor científico, por ser una inspiración para mí en tu forma de trabajar, y por ser tan friki como para tener conversaciones sobre la iniciación de la traducción por el google-talk. Manuel, gracias por tu buen humor que siempre traias al labo -incluso cuando te peleabas con eternas transfecciones-, y por ser otro amante del “lado oscuro” de los bioinformáticos. Daria, por traer la misteriosa “masa madre” y tener que compartir nuestros amados seminarios de programa juntas. Anna, gracias por echarte el laboratorio a los hombros cuando Noelia no pudo estar. Laia, gracias por nuestras conversaciones intentando comprender cómo demonios funcionaban las sintetasas en Plasmodium. Paco, gracias por tus charlas sobre cáncer y por tu gran aportación de los “Juernes” al labo!! Alfred, Núria y Valerie, gracias por el soporte que me habéis dado en todo lo relacionado con Plasmodium, y por el buen clima que siempre creastéis en el labo. Y también a los newbies del labo (aunque ya no tanto, como pasa el tiempo!), David y Adé. A todos vosotros, gracias por los buenos ratos compartidos, por tantos ratos de desayunos y comidas, por soportar mis lab meetings bioinfórmaticos sin dormiros demasiado, por teñir de azul el laboratorio y ponernos mascarillas, por hacernos fotos de nuestra ropa, por las competiciones de moscas mutantes, y por hacer de cada día en el laboratorio un día agradable.

Y por supuesto, fuera del laboratorio hay grandes amigos a los que tengo mucho que agradecer. Barby y Gemma, mis nenas desde siempre, por ser como sois, por ser mis “amici prima”, y por estar siempre allí. Lore, mi amore, por tantos buenos ratos compartidos, tanto en la playita como con un buen vino. A todos mis compañeros de volei playa con los que hemos compartido tantos buenos momentos durante estos años: Saul, Xavi, Bea, Ana, Pau, Juju, Maribel, Berta, Juli, Toni.... gracias a todos!! Y por supuesto a mis compañeras de pista, mis Rolideras, por las cervecitas en Plaza Concordia y en el Quimet, por los desayunos post-fiesta, por hacer empanadillas gallegas a las seis de la mañana y por reirme tanto con todas vosotras. A Jordi, por su paciencia, comprensión y apoyo durante mi tesis, por dedicarme tanto tiempo y no reprocharme el que no pude dedicarte yo. A Óscar, gracias también, por ser el primer friki de mi mundillo bioinformático, y echarte tantas horas estudiando conmigo en el Jordi’s y en la Cerveceria Universidad. Y a los frikis que vinieron después: Meri, Alba, Cristian, Laia... A mis antiguos compis de piso, Albert, Gerard y Seán, por haber podido compartir con vosotros mucho más que piso y facturas. Y a mis nuevos compis, Nano, Pete y Ruso. Por las siestas comunes en el sofá, por las cenas caseras, por los encuentros nocturnos en el comedor a las 4 de la mañana, por ser tan ordenados y limpios...jajaja. Me ha encantado compartir estos meses con vosotros. Y en especial al nene, por convencerme con sus lentejas de que me fuera al piso. *Xa me dixo o meu pai que vixiase cos galegos!!!!* Pero como nunca se escucha a los padres, pues aquí estoy escribiendo mi tesis en Ribadavia...

Y para terminar, y en este caso está más que claro que *last is not least*, sino todo lo contrario.... mi familia. Por vuestro cariño, dedicación y todo tipo de mimos que me habéis dedicado a lo largo de mi vida. Por vuestro apoyo incondicional, por vuestras palabras siempre alentadoras. Por animarme siempre a seguir adelante. Caminante no hay camino, el camino se hace al andar!! Siempre faltarán palabras para decir todo lo que podría decir y agradeceros todo lo que habéis hecho por mi.

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

Contents

Abstract 1

Abbreviations 7

1. Introduction 11

1.1 Overview of the gene translation process 11
1.1.1 The central dogma 11
1.1.2 The two phases of gene translation 12
1.2 The genetic code 16
1.3 Transfer RNAs 18
1.3.1 tRNA structure 18
1.3.2 tRNA gene arrangement and transcription 19
1.3.3 tRNA gene copy number and abundance 20
1.4 tRNA modifications 22
1.4.1 Types of tRNA modifications 22
1.4.2 Functions of tRNA modifications 25
1.5 Aminoacyl-tRNA synthetases 26
1.5.1 Aminoacylation reaction 26
1.5.2 Classes of aminoacyl-tRNA synthetases 27
1.5.3 Evolution of aminoacyl-tRNA synthetases 29
1.5.4 Domains of aaRS and non-canonical functions 33

2. Objectives 37

3. PhD Advisor Report 41

4. Publications 47

4.1. Chapter 1: Genome-wide characterization of the gene translation machinery and its evolution across species

4.1.1. Introduction 49
4.1.1.1. Control of gene expression 49
4.1.1.2. Post-transcriptional regulation of gene expression 50
4.1.1.3. Translation efficiency 52
4.1.1.3.1. Codon usage bias 52
4.1.1.3.2. tRNA isoacceptor abundance 52
4.1.1.3.3. tRNA modifications 53

4.1.2. Publications
4.1.2.1. ***Publication 1: A role for tRNA modifications in genome structure and codon usage.*** ***55***

Novoa EM, Pavon-Eternod M. Pan T and Ribas de Pouplana L.
Cell 2012, 149: 202-213

4.1.2.2. ***Publication 2: Speeding with control: codon usage, tRNA and ribosomes*** ***77***
Novoa EM and Ribas de Pouplana L.
Trends in Genetics 2012 (in press)

4.2. Chapter 2: Aminoacyl-tRNA synthetases as antimalarial drug targets

4.2.1. Introduction to Plasmodium falciparum **87**
4.2.1.1. *Plasmodium falciparum* malaria 87
4.2.1.2. *P. falciparum* life cycle 88
4.2.1.3. *P. falciparum* genome 90
4.2.1.4. Transcription in *P. falciparum* 92
4.2.1.5. Protein translation in *P. falciparum* 93
4.2.1.6. Components of the *P. falciparum* translation machinery 95
4.2.1.6.1. *tRNA* *95*
4.2.1.6.2. *Aminoacyl-tRNA synthetases* *98*
4.2.1.7. *P. falciparum* aaRS as drug targets 101

4.2.2. Introduction: aaRS as drug targets **103**
4.2.2.1. Reaction intermediate analogues 104
4.2.2.2. Analogues of natural aaRS inhibitors 106
4.2.2.3. Drugs disrupting tRNA interaction 107
4.2.2.4. Inhibitors of the aaRS proofreading activity 111
4.2.2.5. Virtual screens and structure-based design 112
4.2.2.6. High-throughput screening programs 116

4.2.3. Publications
4.2.3.1. ***Publication 3: Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum*** ***119***

Hoen R*, Novoa EM*, López A, Camacho C, Cubells L, Martin P, Bautista JM, Vieira P, Santos M, Cortes A, Ribas de Pouplana L and Royo M.
J Med Chem (under review)

4.2.3.2. ***Publication 4: Systematic study on Plasmodium falciparum aminoacyl-tRNA synthetases as antimalarial drug targets*** ***191***
Camacho C, Novoa EM, Cubells L, Wilkinson B, Martin P, Bautista JM, Cortés A and Ribas de Pouplana L.
To be submitted

4.3. Chapter 3: Method development

4.3.1. Introduction **221**
4.3.1.1. A sequence-based prediction method to identify pathogenicity-related proteins 221

4.3.1.1.1. Use of sequence-based homology searches to predict protein function 221
4.3.1.1.2. Application of the method to functionally-related proteins 222
4.3.1.1.3. Application of the method to whole genomes 223
4.3.1.2. Ensemble docking from homology models 226

4.3.2. Publications

4.3.2.1. ***Publication 5: A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins*** **229**
Novoa EM, Castro de Moura M, Orozco M and Ribas de Pouplana L.
FEBS Lett 2010, 584 (2): 460-466.

4.3.2.2. ***Publication 6: Ensemble docking in homology models*** **239**
Novoa EM, Ribas de Pouplana L, Barril X and Orozco M.
J Chem Theory Comput 2010, 6 (8): 2547-2557

4.3.2.3. ***Publication 7: Small molecule docking from theoretical structural models*** **263**
Novoa EM, Ribas de Pouplana L and Orozco M.
In: "Computational Modelling of Biological Systems: From Molecules to Pathways". Ed Springer, New York (USA). Vol 4, pp. 75-96.

5. DISCUSSION AND CONCLUSIONS 289

5.1. Role of tRNA modifications in genome structure and codon usage 289
5.1.1. The appearance of two tRNA modification enzymes shaped the tRNA gene content and the codon usage bias 289
5.1.2. tRNA gene content within each kingdom does not follow the tree of life 291
5.1.3. tRNA modifications as a novel mechanism for post-transcriptional regulation 296

5.2. Aminoacyl-tRNA synthetases as antimalarial drug targets 303
5.2.1. Antimalarial drug discovery: old and new approaches 303
5.2.2. Our approach: aaRS as antimalarial drug targets 304
5.2.3. Future work on aaRS as antimalarial drug targets 307

6. SUMMARY (SPANISH) 311

7. REFERENCES 329

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

Abstract

Gene translation is a central process that occurs in all three domains of life, in which messenger RNA (mRNA) is decoded to produce a specific polypeptide according to the rules specified by the genetic code. Our research group studies gene translation, and more specifically, the mechanism of transfer RNA (tRNA) aminoacylation. In the aminoacylation reaction, a particular amino acid is transferred to its cognate tRNA. The enzymes catalyzing this highly specific reaction are the aminoacyl-tRNA synthetases (aaRS), which are responsible for establishing the genetic code. Aminoacyl-tRNA synthetases are the link between the worlds of protein and nucleic acids. It is not only the structure-function of these enzymes what has captured the biologist's imagination, but also the possibility that they could tell us the secrets of the genetic code. To understand these enzymes is to add a most important piece to the puzzle of what the cell is, and how it works.

This work focuses on the genome-wide study and characterization of the gene translation machinery using both *in silico* and *in vitro* approaches, with a special focus on the two major players of the aminoacylation reaction: aaRS and tRNAs. Indeed, in this work I have characterized with greater detail the gene translation machinery of *Plasmodium falciparum*, the most deadly species causing malaria, in order to design and screen inhibitors specifically targeting its gene translation machinery.

This PhD thesis has been structured into three different sections, corresponding to the different projects performed related to the characterization of the gene translation machinery:

1. Genome-wide characterization of the gene translation machinery and its evolution across species

Despite the central role of tRNAs in protein translation, the connections between tRNA gene population dynamics and genome evolution have rarely been explored. Indeed, we do not understand the reasons for the variability between tRNA pools of different species, nor the principles that determine tRNA gene abundances or genomic codon composition.

To understand the evolutionary pressures that shaped the gene translation machinery, we analyzed hundreds of genomes in terms of their tRNA gene contents and codon usage. Through our analysis we observe that tRNA pools have evolved in a kingdom-specific manner,

and that two kingdom-specific tRNA modifications greatly contributed to genome evolution and extant codon usage biases: tRNA-dependent adenosine deaminases (ADATs) in Eukarya, and uridine methylatransferases (UMs) in Bacteria. Our results suggest that these two tRNA modifications exerted a positive selection on their respective genomes, causing a bias towards specific codons that are read by these modified tRNAs in highly expressed genes. Therefore, the abundance of codons read by these modified tRNAs in a gene directly correlates with genome-wide expression levels. This suggests not only that codon usage bias is a strategy for regulating gene expression levels, but also that the modulation of translation efficiency is performed through the use of specific tRNA modifications.

The discovery of kingdom-specific strategies to optimize translation efficiency opens new possibilities to further improve heterologous gene expression systems. Indeed, preliminar results suggest that these modifications may also have potential roles in disease states. Thus, tRNA modifications may not be mere "decorations" of the function and structure of RNA molecules, but a whole layer of regulation of gene expression levels.

2. In silico and *in vitro* drug design targeting the *Plasmodium falciparum* gene translation machinery

The protein synthesis machinery represents one of the most useful targets for the development of new anti-infectives. Several families of broadly used antibiotics exert their function by blocking the protein synthesis machinery. And yet, very little is known about the specifics of the protein synthesis machinery in *Plasmodium*. In this work we aim to characterise the tRNA biology in *Plasmodium falciparum*, and to develop both *in silico* and *in vitro* screenings for the selection of new potential anti-malarial drugs targeting the plasmodial aminoacyl tRNA synthetases, which are essential enzymes and proven antimicrobial drug targets, and thus represent interesting novel targets for antimalarial drug discovery.

There are three different genomic reservoirs that can be translated in *P. falciparum*: the apicoplastic, the mitochondrial and the nuclear genome. Our results predict that there is a total of 37 nuclear-encoded aaRS genes, which are either targeted to the apicoplast or to the cytoplasm, obtaining a full set of aaRS in these two compartments. Amongst the 37 predicted ARS, we decided to focus on two of them as candidate antimalarial drug targets: the apicoplastic-targeted lysyl-tRNA synthetase (PfKRS-2) and the glutaminyl-tRNA synthetase (PfQRS).

Plasmodial proteins are difficult to characterize structurally using traditional *in vitro* approaches. However, these problems can be partially overcome using a number of *in silico* approaches. This work is a clear example showing that the combination of both *in silico* and *in vitro* procedures can facilitate and accelerate the discovery of candidate hits. Furthermore, we also show that plasmodial aminoacyl-tRNA synthetases are druggable enzymes that can be used as specific targets of antimalarials. Overall this work shows that it is worth to continue characterizing the protein synthesis machinery in *Plasmodium falciparum,* and use this knowledge for the development of new antimalarials.

3. Method development

To develop the projects mentioned above, two side-projects related to computational benchmarking and software development have been performed:

1. Development of a sequence-based method to predict pathogenicity-related proteins
2. Development of a docking strategy to decipher the reliability, enrichment ratios and docking binding pose accuracy when using homology models for docking purposes.

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

ABBREVIATIONS

3D three-dimensional
aa amino acid
aa-AMP aminoacyl-adenylate
aaRS aminoacyl-tRNA synthetase
ADAT tRNA-dependent adenosine deaminase
ADAR RNA-dependent adenosine deaminase
ADP adenosine diphosphate
ATP adenosine triphosphate
BLAST basic local alignment search tool
bp base pair
CDS coding sequence
C_{ter} carboxy-terminal
DAPI 4′,6-Diamidino-2-phenylindole dihydrochloride
DNA deoxyribonucleic acid
dNTP deoxyribonucleotide triphosphate
eEF eukaryotic elongation factor
EF elongation factor
eGFP enhanced green fluorescent protein
eIF eukaryotic initiation factor
EMAPII endothelial monocyte activating polypeptide II
ER enrichment ratio
eRF eukaryotic releasing factor
GFP green fluorescent protein
GTP guanosine triphosphate
GS GlideScore
hetADAT heterodimeric ADAT
HGT horizontal gene transfer
HTS high-throughput screening
Hs *Homo sapiens*
IC_{50} half maximal inhibitory concentration
IDC intraerythrocytic developmental cycle
IF initiation factor
IleRS isoleucyl-tRNA synthetase
iRBC infected red blood cell
Kb kilobase
kDa kilodaltons
K_i inhibition constant
K_M Michaelis constant
KRS lysyl-tRNA synthetase
LAD lysyl-adenylate intermediate
LC-MS/MS liquid chromatography-coupled tandem quadrupole mass spectrometry
LeuRS leucyl-tRNA synthetase
MARS multi-synthetase complex
MD molecular dynamics
MRSA methicillin-resistant *Staphylococcus aureus*
Mg magnesium
miRNA micro-RNA
mRNA messenger RNA
MS mass spectrometry
mTOR mammalian target of rapamycin
MW molecular weight
N_{ter} amino-terminal
ORF open reading frame
PCA principal component analysis
PCR polymerase chain reaction
PDB protein data bank
PI3K phosphoinositide 3-kinase
Pf *Plasmodium falciparum*
PfKRS-1 *Plasmodium falciparum* cytoplasmatic lysyl-tRNA synthetase
PfKRS-2 *Plasmodium falciparum* apicoplastic lysyl-tRNA synthetase
PfTRS *Plasmodium falciparum* threonyl-tRNA synthetase
PfQRS *Plasmodium falciparum* glutaminyl-tRNA synthetase
PheRS phenylalanyl-tRNA synthetase
PP_i inorganic pyrophosphate
ProRS prolyl-tRNA synthetase
RBC red blood cell
RF releasing factor
RGF relative gene frequency
RNA ribonucleic acid
RNAP RNA polymerase
RNase ribonuclease
RRF ribosome recycling factor
rRNA ribosomal RNA
RSCU relative synonymous codon usage
SBDD structure-based drug design
SD Shine-Dalgarno
siRNA small interference RNA
ThrRS threonyl-tRNA synthetase
tRNA transfer RNA
$tRNA^{aa}_{NNN}$ tRNA specific for aa bearing the codon NNN
TrpRS tryptophanyl-tRNA synthetase
UTR untranslated region
UM uridine methyltransferase
WB western blot
WHO World Health Organization
wt wild type
Zn zinc

Amino acid abbreviations

A	**Ala**	alanine		**fMet**	formyl-methionine
C	**Cys**	cysteine	**N**	**Asn**	asparagine
D	**Asp**	aspartic acid	**P**	**Pro**	proline
E	**Glu**	glutamic acid	**Q**	**Gln**	glutamine
F	**Phe**	phenylalanine	**R**	**Arg**	arginine
G	**Gly**	glycine	**S**	**Ser**	serine
H	**His**	histidine	**T**	**Thr**	threonine
I	**Ile**	isoleucine	**U**	**Sec**	selenocysteine
K	**Lys**	lysine	**V**	**Val**	valine
L	**Leu**	leucine	**W**	**Trp**	tryptophan
M	**Met**	methionine	**Y**	**Tyr**	tyrosine

Modified nucleotides abbreviations

mcm^{5}U	5-methoxycarbonylmethyluridine
mcm^{5}s^{2}U	5-methoxycarbonylmethyl-2-thiouridine
s^{4}U	4-thiouridine
t^{6}A	*N*^{6}-threonylcarbamoyladenosine
m^{7}G	7-methylguanosine
Cm	2'-*O*-methylcytidine
D	dihydrouridine
I	inosine
m^{2}G	*N*^{2}-methylguanosine
m$^2_2$G	*N*^{2},*N*^{2}-dimethylguanosine
mnm^{5}s^{2}U	5-methylaminomethyl-2-thiouridine
m^{5}U	5-methyluridine or ribothymidine
cmo^{5}U	uridine 5-oxyacetic acid
acp^{3}U	3-(3-amino-3-carboxypropyl)uridine
s^{2}C	2-thiocytidine
mcm^{5}Um	5-methoxycarbonylmethyl-2'-*O*-methyluridine
ms^{2}t^{6}A	2-methylthio-*N*^{6}-threonyl carbamoyladenosine
m^{5}C	5-methylcytidine
f^{5}C	5-formylcytidine
mnm^{5}U	5-methylaminomethyluridine
k^{2}C	lysidine
τm^{5}s^{2}U	5-taurinomethyl-2-thiouridine
Ψ	pseudouridine
xm^{5}s^2 U	5-methyl-2-thiouridine derivatives with any substitution at carbon 5 of the uracil
xmo^{5}U	5-methoxyuridine derivatives with any substitution at carbon 5 of the uracil

Contents

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

1. Introduction

1.1 Overview of the gene translation process

1.1.1 The central dogma

The cell is the functional unit of living organisms, and contains DNA, RNA and proteins (Crick 1958). Proteins comprise nearly 50% of the cellular mass and serve as enzymes, signalling molecules, structural, storage and mechanical components of the cell.

The sequential transition of information from DNA to mRNA to protein constitutes the central dogma of molecular biology (**Figure 1.1**). It states that such information cannot be transferred back from protein to either protein or nucleic acid (Crick, 1970).

The specific part of DNA that encodes for an mRNA sequence that will be transcribed into a protein is called a gene. For the information of DNA to be converted into protein, the gene has to be transcribed into mRNA by RNA polymerase and transcription factors. In eukaryotic cells the primary transcript (pre-mRNA) must be processed further in order to ensure translation. This normally includes a 5'-cap, poly-A tail and splicing. Eventually, this mature mRNA finds its way to a ribosome, where it is translated.

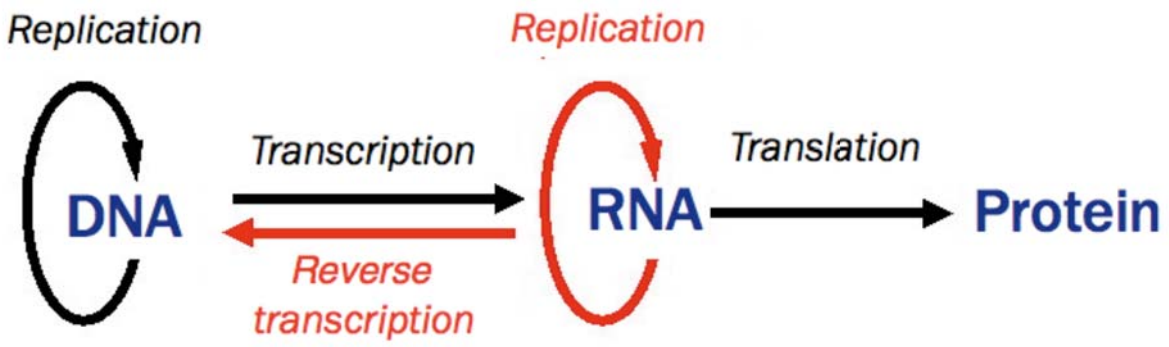


Figure 1.1. Central dogma of biology. The dogma is a framework for understanding the transfer of sequence information between sequential information-carrying biopolymers. Crick's initial proposal (1970) is shown in black, and later modifications are included in red.

1.1.2 The two phases of gene translation

Translating the 4-letter code of RNA into the 22-letter alphabet of proteins is a central feature of cellular life. In the gene translation process, messenger RNA (mRNA) is decoded by the ribosome to produce a specific amino acid chain that will later fold into an active protein. The translation machinery is dedicated to interpreting the nucleic acid code in a two-part process. First, amino acids are covalently linked to their cognate tRNAs via an aminoacylation reaction catalyzed by a diverse group of proteins, the aminoacyl-tRNA synthetases (aaRS). The aminoacyl-tRNAs (aa-tRNAs) are then delivered to the ribosome by elongation factors (EF-Tu in bacteria and EF-1A in archaea and eukaryotes) (Krab and Parmeggiani 2002; Hotokezaka et al. 2002). At the ribosome, the tRNA anticodon is matched to the mRNA codon and the charged tRNA delivers the next residue of a nascent protein chain (**Figure 1.2**).

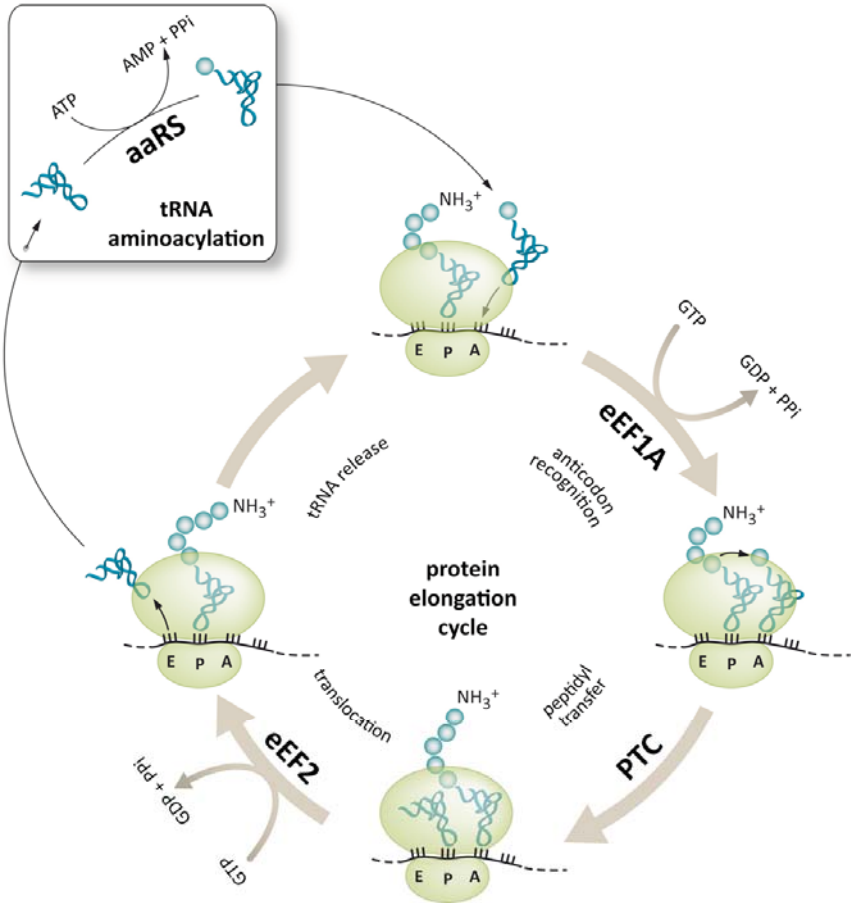


Figure 1.2. Two phases of protein synthesis. In the first phase, tRNAs are aminoacylated with their cognate tRNA by its specific aaRS (top left). Then, the aminoacylated tRNAs are delivered to the ribosome. The protein elongation cycle is also depicted.

The first phase, concerning tRNA aminoacylation by aaRS, will be covered in greater depth in section 1.5.1. Regarding the second phase, the ribosomal protein translation has been extensively reviewed (Dale and Uhlenbeck, 2005; Kapp and Lorsch, 2004; Jackson et al. 2010). Ribosomal translation occurs in four stages: initiation, elongation, termination and recycling (**Figure 1.3**).

- In the ***initiation step***, methionyl initiator tRNA (Met-tRNA^{Met-i}), GTP and eukaryotic initiator factor 2 (eIF2) are assembled in the eIF2-GTP-Met-tRNA^{Met-i} ternary complex, which binds to the P site of the small (40S) ribosomal subunit. The 43S complex begins to scan down the mRNA in the 5' to 3' direction, looking for the AUG initiation codon with the Kozak sequence (Kozak, 1986).

- During the ***elongation*** step, a new aminoacyl-tRNA is carried to the A site of the ribosome complexed with eukaryotic elongator factor 1A (eEF1A) and GTP in the ternary complex eEF1A·GTP·aa-tRNA. The ribosomal peptidyl transferase center (PTC) catalyzes the formation of a peptide bond between the incoming amino acid and the growing peptide, resulting in a deacylated tRNA with its acceptor end in the E (exit) site. The protein elongation cycle is repeated until a stop codon is encountered, in which case, the process of termination is triggered.

- At the ***termination*** stage, the eukaryotic releasing factor (eRF) 1 (eRF1) promotes, in response to any of the three eukaryotic stop codons UAA, UAG or UGA in the A site, the hydrolysis of the ester bond linking the polypeptide chain to the tRNA on the P site and, therefore, the release of the completed polypeptide from the ribosome.

- The ribosome ***recycling*** process is the less known of the four stages and, contrary to prokaryotes, no ribosome recycling factors (RRF) have been found in eukaryotes. Instead, eIF3 has been proposed as the principal factor that promotes recycling of the ribosomes after termination.

-

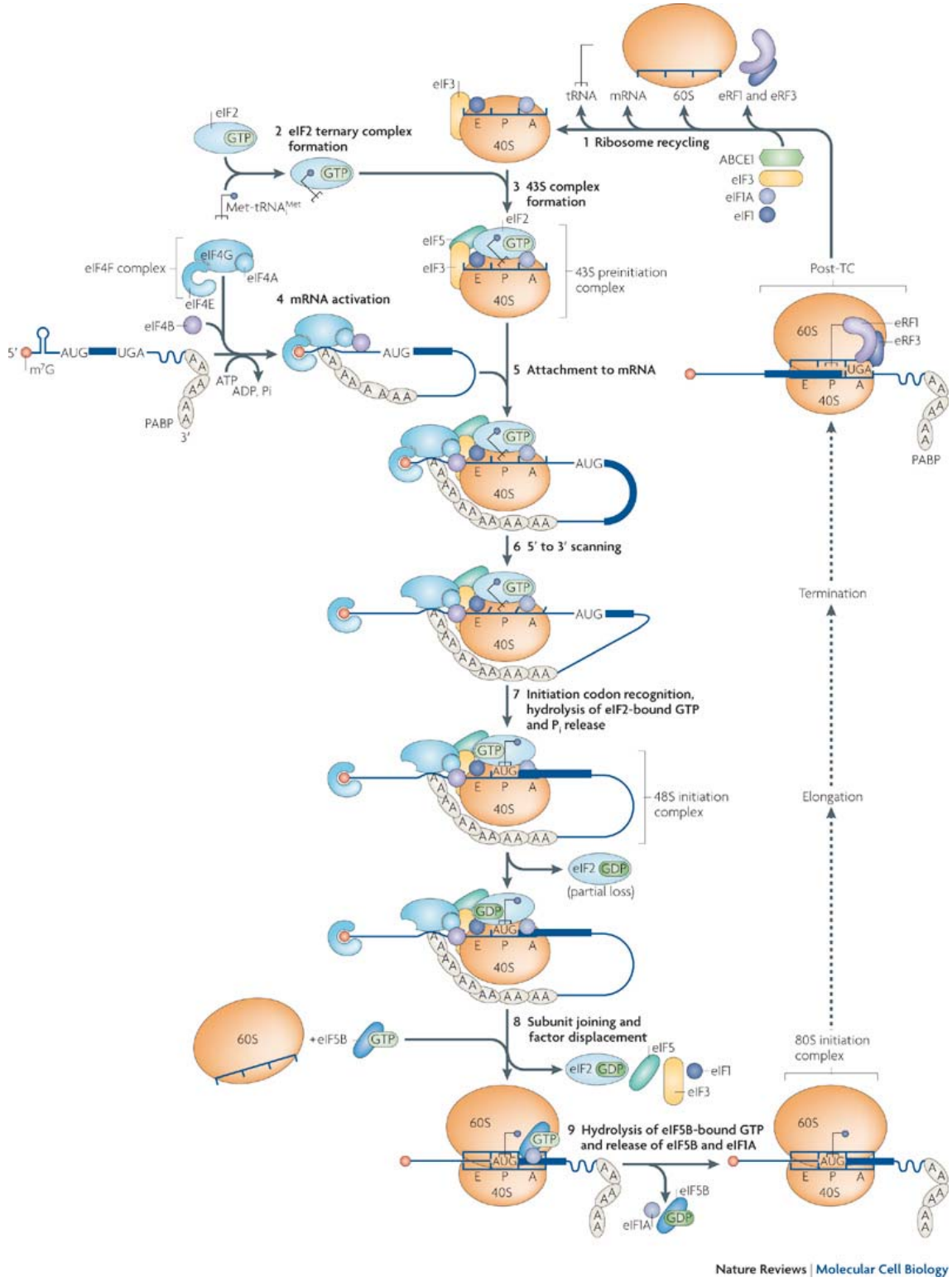


Figure 1.3. Model of the canonical pathway of eukaryotic translation initiation. The canonical pathway translation initiation is divided into eight stages (**2-9**), which are followed by the recycling of post-termination complexes (post-TCs; **1**). Adapted from Jackson et al., 2010.

1.2 The genetic code

The standard genetic code, with some exception, is found throughout all the kingdoms of life. It is composed of 64 different triplets (codons), with 61 of them encoding amino acids. As there are many more amino acid codons (64) than amino acids themselves (20), most amino acids are encoded by several related codons in what is referred to as degeneracy. The only exceptions to the degeneracy are methionine and tryptophan (**Figure 1.4**). There is a tendency for similar codons to specify for similar amino acids (Woese, 1965b). Mutation or misreading of the third base pair of a codon is therefore likely to preserve the amino acid specified, or switch it to an amino acid with similar properties. A certain resemblance is also observed between the amino acids specified by codons that share the same residue at the second position. Codons with U at the second position specify hydrophobic amino acids while those with an A in this position tend to code for strongly hydrophilic residues. These similarities seem to indicate that the genetic code has evolved to minimize the harm caused by error in the genes or in the translation process (Alff-Steinberger, 1969).

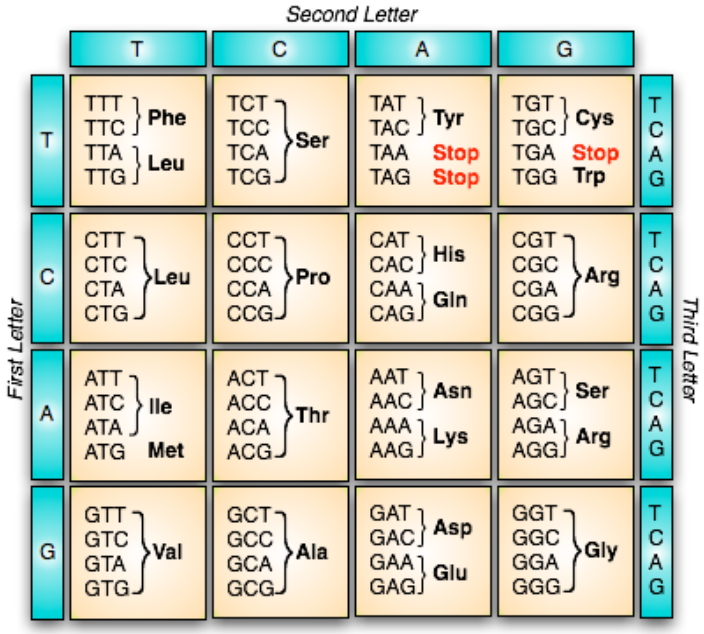


Figure 1.4. The genetic code. Correspondence between codons (DNA triplets) and its corresponding amino acid. The combination of the two first letters of the code creates 16 possible codon boxes, where each codon box is composed of 4 codons, and is differentiated by the third letter –also known as degenerate position, given that codon boxes this position is not important for the establishment of the identity of the correct amino acid-. Stop codons (TAA, TAG, TGA) are highlighted in red.

The mechanism through which an organism can read all 61 codons was first hypothesized by Francis Crick with his Wobble Hypothesis (Crick, 1958). Now it is widely known that a tRNA has the ability to decode multiple codons through the flexibility in base-pairing between the third position (3' position) of the mRNA codon and the first position (5' position) of the tRNA anticodon, also known as the wobble position (**Figure 1.5**).

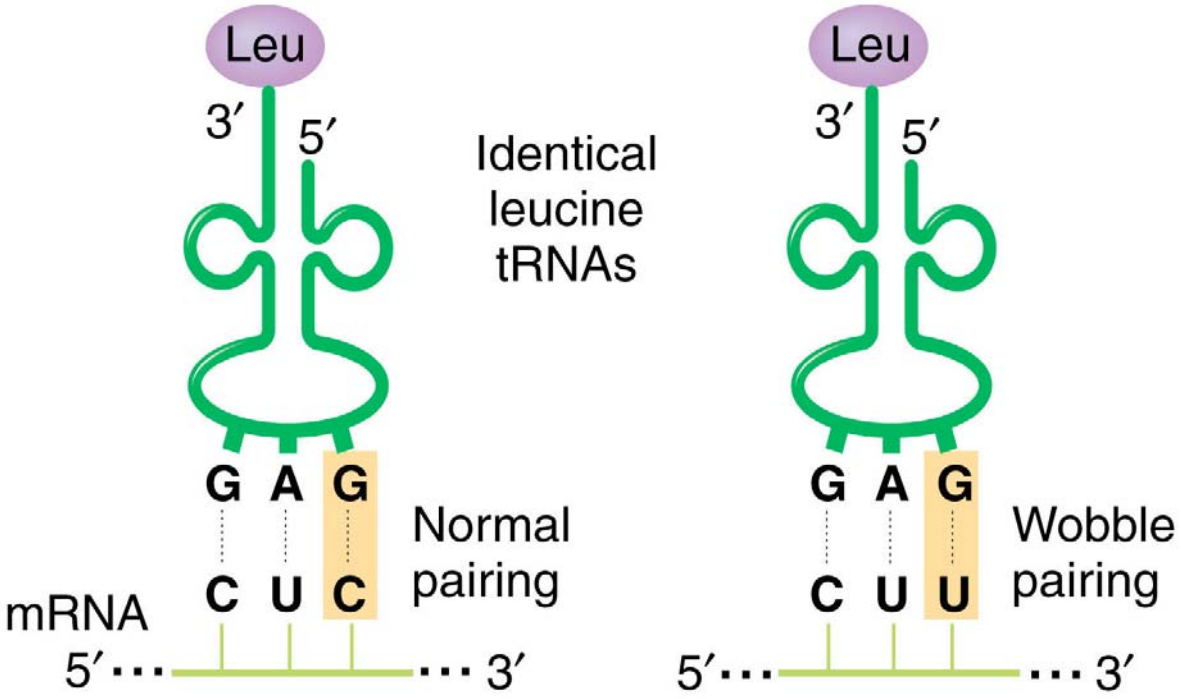


Figure 1.5. mRNA codon- tRNA anticodon base pairing. The base 34 of the tRNA, also known as wobble base, recognizes the base in the 3rd position of the mRNA, also known as the degenerate position. In the example above, the G can pair with either a U or a C. This allows mRNA to be translated with fewer than the 64 tRNAs that would be required without the wobble.

1.3 Transfer RNAs

Transfers RNAs (tRNAs) are the adaptor molecules first hypothesized by Crick over 50 years ago (Crick, 1958). As a general rule, there is at least one tRNA for each of the twenty amino acids used in the standard genetic code. In many cases, multiple tRNA isoacceptors exist for a given amino acid, with these isoacceptors recognizing different or overlapping sets of codons for that amino acid.

tRNAs carry amino acids to the ribosome and decode the genetic information of the mRNA. However, these ancient molecules have also been shown to participate in other cellular processes non-related to translation, such as control of their cognate aaRS expression (Ryckelynck et al. 2005) or a primer function in reverse transcription during retrovirus and retrotransposon replication (Mak and Kleiman, 1997).

1.3.1 tRNA structure

tRNA molecules are relatively short –typically 75 to 95 nucleotides long- that exhibit a strongly conserved secondary structure (Sprinzl et al. 1998). This secondary structure consists of a series of double-stranded stems and single stranded stems (**Figure 1.6**). The overall structure can be depicted in an unfolded cloverleaf form composed of an acceptor stem, D-arm, T-arm and T-loop (Holley et al., 1965).

The T and D tRNA loops owe their names to two strongly conserved modifications, ribothymidine (T) at position 54 and dihydrouridine (D) at position 16 (Bjork et al. 1999). Conserved G18 and G19 in the D-loop interact with conserved ψ55 and C56 of the T-loop to stabilize the structure, along with the Levitt base pair interaction between position 48 of the variable loop and position 15 of the D-loop (Kim et al. 1974; Levitt, 1969). The resulting bent hairpin places the acceptor stem and anticodon loop at opposite ends of the molecule.

The acceptor stem consists of seven base pairs followed by an unpaired discriminator base at position 73, which is followed by a conserved C74, C75 and A76 sequence. The amino acid is charged into the 3' terminal of the A76 residue. Bases 34, 35 and 36 of the anticodon loop constitute the anticodon that is used by the ribosome to recognize the mRNA codons. The first position of the anticodon (base 34) is named *wobble base*, as it allows non-Watson-Crick base pairing with the third position of the mRNA codon (Crick, 1970).

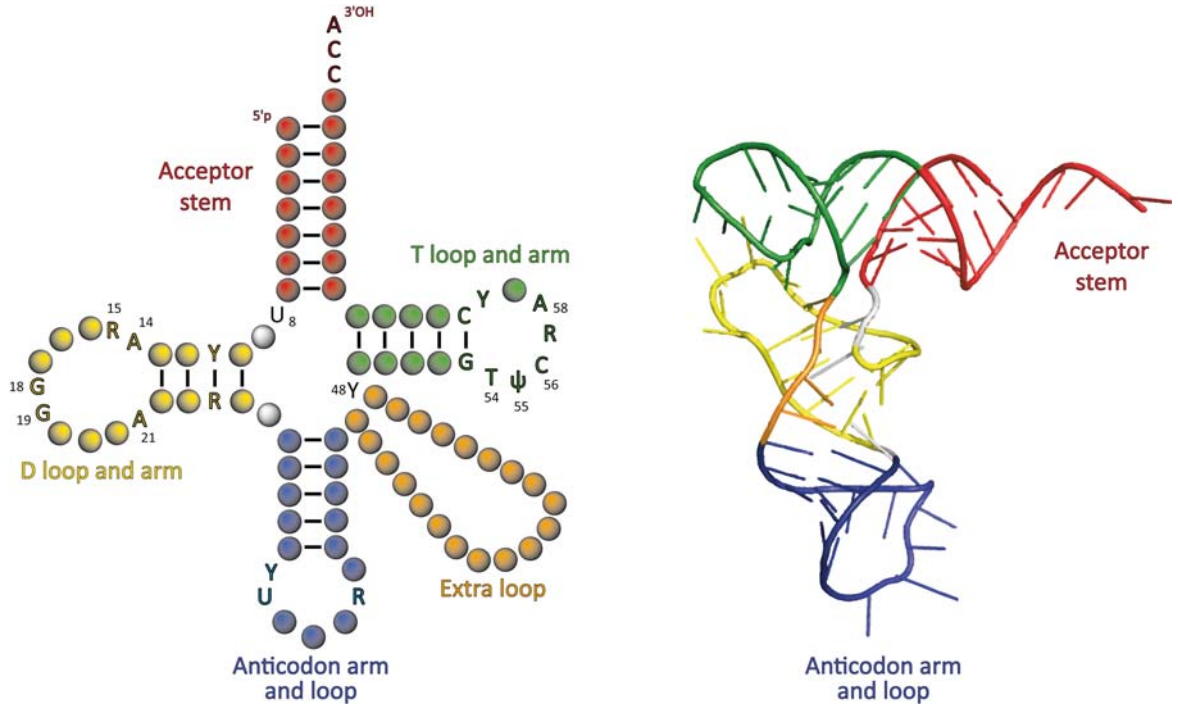


Figure 1.6. Structure of tRNA. On the left, cloverleaf representation of a tRNA with the key conserved residues indicated and each loop and stem highlighted in different colours. The extra loop or variable loop depicted is typically found in $tRNA^{Ser}$ and $tRNA^{Leu}$. On the right, structural 3D representation of a properly folded $tRNA^{Asp}$, with structures colour-coded as in the cloverleaf representation (Ruff et al., 1991).

1.3.2 tRNA gene arrangement and transcription

tRNAs tend to be transcribed into long RNA units that are enzymatically trimmed to yield a functional tRNA (Deutscher, 1984). In bacteria, polycistronic as well as monocistronic precursors are present, whereas in eukaryotes the majority of the primary transcripts are monocistronic. In eukaryotes, RNA polymerase III uses transcription factors to recognize two internal tRNA sequences, the A and B box, which are composed of parts of the T-arm and T-loop or D-arm and D-loop, respectively.

Following transcription and, if necessary, intron removal, the 5' end of the pre-tRNA transcripts are processed by RNAseP (Kole and Altman, 1979). In bacteria, exonucleases remove excess residues from the 3' end leaving a mature CCA 3' terminus (Reuven and Deutscher, 1993). However, eukaryotes lack this CCA sequence in the gene, and thus, after the processing of the 3' end by nucleases, a tRNA nucleotidyl transferase adds the CCA residues to the 3' terminus (Tomita and Weiner, 2001).

1.3.3 tRNA gene copy number and abundance

tRNA genes are often present in multiple copies, with higher copy number for tRNAs typically corresponding to more frequently used codons in the genome. There are 61 possible tRNA isoacceptors, each of them with a different anticodon. In all organisms, less than the possible 61 tRNA types carry out the decoding of all codons. For example, there are only 40 tRNA isoacceptors in *E. coli* K12, 44 tRNA isoacceptors in *D. melanogaster*, 48 in *C. elegans* and 51 in human (Lowe and Eddy, 1997). Each of these isoacceptors tends to be present in multiple copies, which are unequally distributed, with some tRNA isoacceptors that are over-represented compared to others, which are missing (**Figure 1.7**).

Organelles have suffered a drastic reduction of their contents. The number of tRNAs encoded by mitochondrial genomes is species-specific. For instance, the human genome encodes for 22 tRNA genes that are sufficient for mitochondrial protein synthesis. However, in other organisms such as plants, fungi and protozoa, the situation is different, and some tRNA genes are absent in the mitochondrial genome. Trypanosomatids (e.g. *Trypanosoma brucei*) and Apicomplexa (e.g. *Plasmodium falciparum*) are the most extreme situation with no tRNA genes encoded in their mitochondrial genomes. In these cases, nuclear-encoded tRNAs are imported into the mitochondria (Salinas et al. 2008).

Four Box tRNA Sets

Isotype	tRNA Count by Anticodon				Total
Ala	AGC 29	GGC	CGC 5	TGC 9	43
Gly	ACC	GCC 15	CCC 7	TCC 9	31
Pro	AGG 10	GGG	CGG 4	TGG 7	21
Thr	AGT 10	GGT	CGT 6	TGT 6	22
Val	AAC 11	GAC	CAC 16	TAC 5	32

Six Box tRNA Sets

Isotype	tRNA Count by Anticodon								Total
Ser	AGA 11	GGA	CGA 4	TGA 5	ACT	GCT 8			28
Arg	ACG 7	GCG	CCG 4	TCG 6			CCT 5	TCT 6	28
Leu	AAG 12	GAG	CAG 10	TAG 3			CAA 7	TAA 7	39

Two Box tRNA Sets

Isotype	tRNA Count by Anticodon				Total
Phe	AAA	GAA 12			12
Asn	ATT 2	GTT 32			34
Lys			CTT 17	TTT 16	33
Asp	ATC	GTC 19			19
Glu			CTC 13	TTC 13	26
His	ATG	GTG 11			11
Gln			CTG 20	TTG 11	31

Two Box & Other tRNA Sets

Isotype	tRNA Count by Anticodon				Total
Ile	AAT 14	GAT 3		TAT 5	22
Met			CAT 20		20
Tyr	ATA 1	GTA 14			15
Supres			CTA 1	TTA 2	3
Cys	ACA	GCA 30			30
Trp			CCA 9		9
SelCys				TCA 3	3

Figure 1.7. tRNA gene composition of *H. sapiens* build 37.1. Amino acids have been divided into different boxes depending on the number of isoaccepting tRNAs that encode for the same amino acid. The individual tRNA gene copies are shown in black, whereas the total tRNA gene copy number per amino acid is shown in green. The tRNA gene predictions have been performed using the tRNAscan-SE software (Lowe and Eddy, 1997).

In unicellular organisms such as bacteria and fungi, the genomic tRNA copy number correlates with the intracellular tRNA levels (Ikemura, 1981; Sorensen and Pedersen 1991; Kanaya et al. 1999; Tuller et al. 2010). Thus, the expression of a tRNA gene that is not subjected to regulation is expected to be similar to its respective copy number (**Figure 1.8**). However, in higher organisms such as human, several tRNA genes appear as outliers of the plot of tRNA levels versus tRNA gene copy number, suggesting that the epigenetic signature and chromatin state may play a role in tissue-specific tRNA expression levels (Ernst et al. 2011; Mahlab et al. 2012).

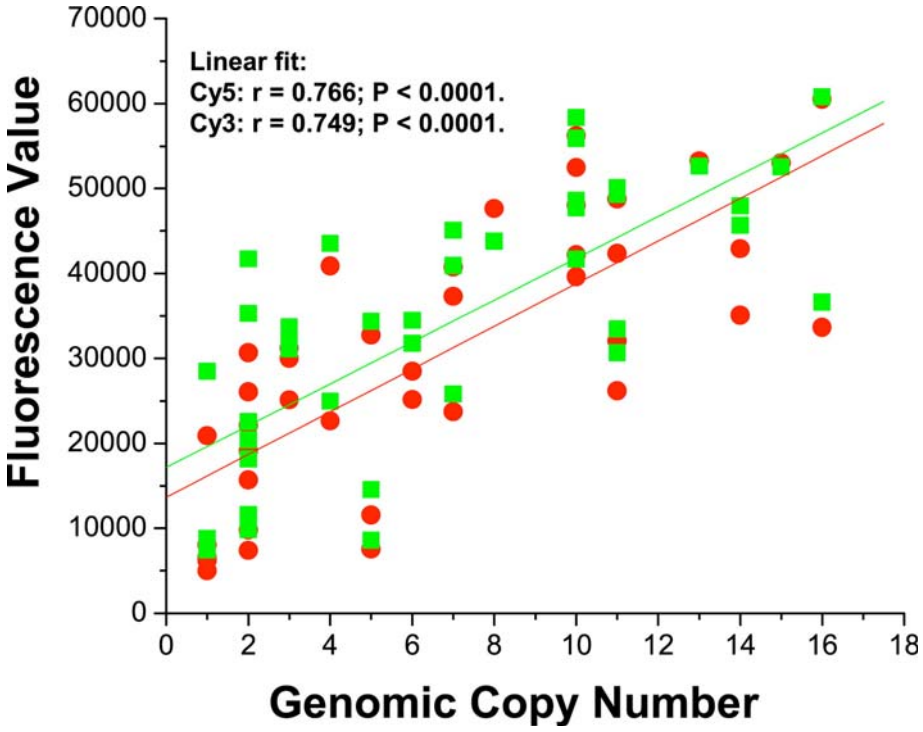


Figure 1.8. Correlation between tRNA gene copy number and tRNA abundances in *S. cerevisiae*. The tRNA abundances have been measured using specific tRNA microarray probes dedicated to this species (Dittmar et al. 2004). tRNA levels measured independently with two alternative dyes (Cy3 and Cy5). Adapted from Tuller et al. Cell 2010.

1.4 tRNA modifications

Unlike mRNA transcripts, transcripts of tRNA genes undergo extensive post-translational processing to become a fully functional and mature tRNA for protein translation. This process is known as tRNA editing, and is essential for cell survival (Döring et al. 2001; Nangle et al. 2006). tRNA modification is a function of the processing stage, the concentration of the substrate and the amount of activity of the tRNA-modifying enzyme.

1.4.1 Types of tRNA modifications

Currently, there are over 100 post-translational modifications that have been identified in tRNA (http://rna-mdb.cas.albany.edu/RNAmods/), some of which are shown in **Figure 1.9**. Some of these modifications are found in all three phylogenetic domains, whereas some others are domain-specific (Peterkofsky et al. 1971). The tRNA is modified post-transcriptionally by modifying enzymes, which are specific for their nucleoside substrate and its position in the tRNA. For instance, the pseudouridine residue (Ψ) found in the anticodon loop (Ψ38) and the one found in the TΨC loop (Ψ55) are synthesized by different enzymes.

Escherichia coli encodes 86 tRNA genes, which represent 40 different tRNA species (gtrnadb.ucsc.edu). Almost 30 different modified nucleotides have been identified in *E. coli* tRNAs (**Table 1**). All tRNA species contain Ψ55 and m^5U54, and modifications at positions 34 and 37 are frequent. Modifications at the wobble position (base 34) can directly affect translation by altering the pattern of hydrogen bond donors and acceptors. Some of these modifications increase its wobbling capacities, while others restrict it. Specific modifications can be either determinants or anti-determinants for the translation of specific codons.

Some tRNA modifications involve a complete base substitution, in which a tRNA nucleoside is post-transcriptionally modified into another nucleoside, with its respective consequences in its base-pairing capabilities. The two most common tRNA editing substitutions involve adenosine to inosine (A-to-I) deamination, and cytidine to uridine (C-to-U) deamination.

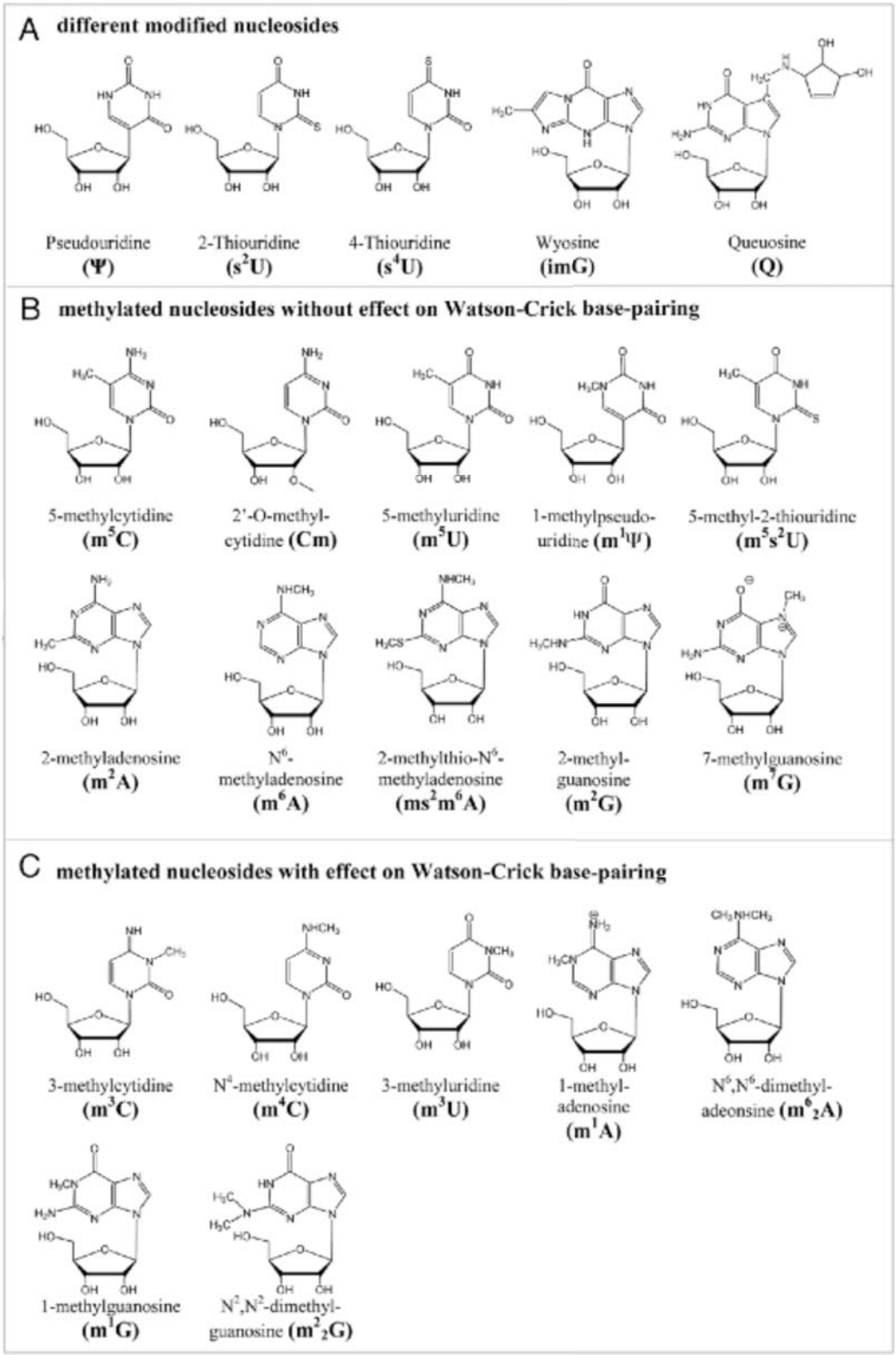


Figure 1.9. Common types and sites for modifications of the four major nucleotides. **A**) A choice of chemically unusual modified nucleosides. **B**) Nucleosides with modifications on the Hogsteen edge. **C**) Nucleosides carrying methyl groups on the Watson-Crick edge. Adapted from Kellner et al. 2010.

Table 1. tRNA modifications in *Escherichia coli*

Modification	Name	Function
m^2A	2-methyladenosine	
m^6A	N^6-methyladenosine	may prevent A36 from base-pairing other than U
i^6A	N^6-isopentenyladenosine	
ms^2i^6A	2-methylthio-N^6-pentenyladenosine	decodes UNN codons, stabilization of anticodon:codon interactions, effectively compensating for the weak A:U
t^6A	N^6-threonyl carbamoyladenosine	modification may stabilize U:A base pair at the first codon position, positive determinant for IleRS
m^6t^6A	N^6-methyl-N^6-threonylcarbamoyladenosine	may prevent misreading at the first position with a U:G base pair
s^2C	2-thiocytidine	may increase efficiency of codon:anticodon formation
ac^4C	N^4-acetylcytidine	reduces reading of AUG codons, decreases misreading of noncognate AUA codons
k^2C	lysidine	prevents misacylation, changes base-pairing ability of C to recognize only A
Cm	2'-O-methylcytidine	pos. 32: restricts nucleotide flexibility pos. 34: restricted wobbling with tRNAMet
D	dihydrouridine	establishing correct conformation for minoacylation
Gm	2'-O-methylguanosine	
m^1G	1-methylguanosine	methyl group prohibits base pairing with Watson-Crick geometry, might prevent out-of phase reading with shifted or expanded anticodon methyl group may increase base stacking
m7G	7-methylguanosine	
I	inosine	increase codon:anticodon pairing possibilities
Q	queuosine	minor effects on decoding of U and C
s^4U	4-thiouridine	pos. 8: acts as sensor for near-UV light and protects cells from such stress, prevents expression of SOS response and thus reduces mutagenesis
Y	pseudouridine	different functions depending on the position: pos. 32: ? pos. 34, 35: increases translational efficiency by stabilizing anticodon:codon pairing pos. 38-40: increases translational efficiency
Um	2'-O-methyluridine	
cmo^5 U	uridine 5-oxyacetic acid	enhance wobbling, tRNAs read A, G and U
$mcmo^5$ U	uridine 5-oxyacetic acid methyl ester	enhance wobbling, tRNAs read A, G and U
mnm^5 U	5-methylaminomethyluridine	
mnm^5 Um	5-methylaminomethyl-2-O-methyluridine	restricts wobbling, tRNAs read A > G
mnm^5 s^2 U	5-methylaminomethyl-2-thiouridine	restricts wobbling, tRNAs read A > G
mnm^5 Se^2 U	5-methylaminomethyl-2-selenouridine	
acp^3 U	3-(3-amino-3-carboxypropyl)-uridine	
m^5U	ribosylthymine	stabilizes tRNA structure, decreases errors and increase A-site binding

1.4.2 Functions of tRNA modifications

Why are RNA chains modified at all? The simplest answer may be that RNA, as opposed to proteins with 20 different amino acids, have only four nucleotides to use, and therefore, the evolution of modified nucleosides may have compensated for the shortcoming in flexibility and accuracy (Persson et al., 1993). However, most modified nucleotides may not be essential for the aminoacylation reaction (Chan et al. 2010), although they constitute important identity determinants for many aaRS.

Of special interest are those modifications that affect the functions of the translational machinery. In this regard, modifications at position 34 have an obvious effect on the decoding capacity by altering the array of H-bonding groups. The most well described biological functions for these tRNA modifications include:

i) Extension and restriction of base-pairing capacity
ii) Modification of the stability of codon-anticodon interaction
iii) Reading frame maintenance
iv) Modification in natural non-sense suppression
v) Effects on the efficiency of translation initiation

In vitro transcription of tRNAs allowed a direct comparison between native and unmodified tRNAs in *E. coli* (Sylvers et al. 1993; Tamura et al. 1992). Such comparisons of the kinetics of the aminoacylation reactions revealed that amongst 14 different unmodified tRNAs, all except 3 accepted the cognate amino acid. Thus, modified nucleotides are not a prerequisite for most aminoacylation reactions *in vitro*, but will affect the kinetics of the reaction.

The non-essentiality of many of these modifications parallels the fact that many DNA modifications are not essential for life. However, in a similar fashion to DNA modifications, increasing evidence indicates that tRNA modifications can play regulatory roles in cells, especially in response to stress conditions (Chan et al., 2010; Chan et al., 2012). Like the epigenetic states of histone protein modification and DNA methylation, the pattern and selectivity of tRNA modifications could also be regulated and maintained in distinct cell types and physiological states (Yi and Pan, 2011).

1.5 Aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases (aaRS) are, together with tRNA, the main players in the first step of the protein translation: the aminoacylation reaction. As a consequence of the aminoacylation reaction, a particular amino acid is specifically linked to its cognate tRNA. Once this reaction is completed, the tRNA is brought to the ribosome and participates in the second step of protein translation: the ribosomal peptide synthesis.

1.5.1 Aminoacylation reaction

Aminoacylation by aaRS occurs in two steps. First, the amino acid is adenylated or "activated" with ATP to form aminoacyl-adenylates (aa-AMPs), releasing pyrophosphate (PPi). Then the activated amino acid (aa-AMP), which remains complexed with the enzyme, is then transferred onto the 3' terminal nucleotide (A76) of the tRNA via covalent attachment, yielding free AMP and free aminoacyl-tRNA (**Figure 1.10**).

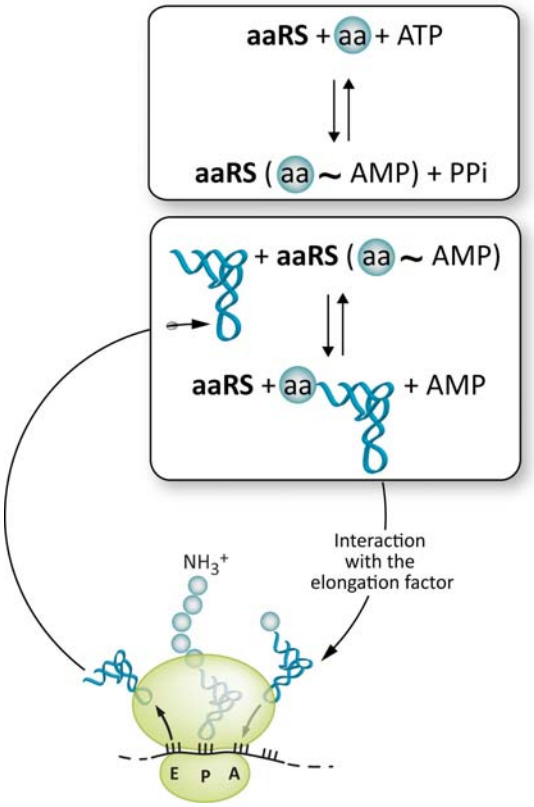


Figure 1.10. Aminoacylation reaction. Aminoacyl-tRNA synthetases catalyze the aminoacyl-tRNA (aa-tRNA) formation in two steps: i) activation of the amino acid and ii) transfer of the activated amino acid to its cognate tRNA.

1.5.2 Classes of aminoacyl-tRNA synthetases

With notable exceptions, there are 20 aaRS, one for each of the amino acids used in the genetic code. These aaRS are universally distributed across the tree of life (Nagel and Doolittle, 1991). Although the basic chemical reaction is the same in each case, the 20 aaRS fall into two classes containing distinct active site architectures (Cusack et al., 1990; Eriani et al., 1990). Class I and class II enzymes appear to have originated from two separate ancestral active site domains or catalytic cores, that contained both amino acid activation and tRNA aminoacylation activity (Schimmel and Ribas de Pouplana, 1995). With the exception of lysyl-tRNA synthetase, each of the 20 types of aaRS can be assigned to only one of these two classes (**Figure 1.11**).

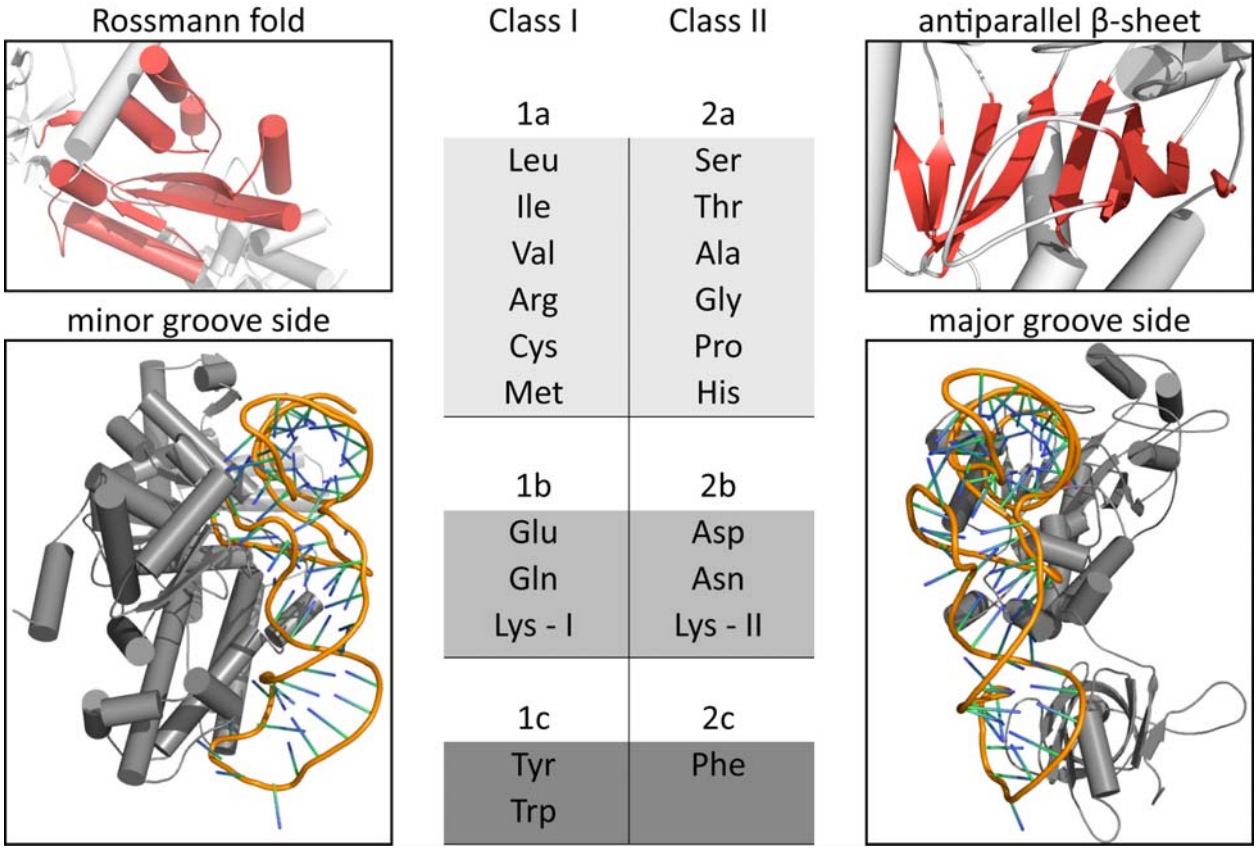


Figure 1.11. Classes of aaRS. Depending on the fold of the catalytic site, aaRS can be classified into two different classes: class I (Rossman fold) or class II (antiparallel ß-sheet). Each of the enzymes corresponding to a given class tend to recognize the tRNA from its minor or major groove side, respectively.

Class I enzymes have a catalytic core based on a nucleotide binding Rossman fold (consistent of a minimum of three parallel ß-strands connected by helices) and contain characteristic HIGH and KMSKS motifs for ATP and magnesium (Mg^{2+}) ion interaction (Eriani et al, 1990; Rould et al., 1989). In contrast, the catalytic core of class II aaRS is comprised of an antiparallel ß-sheet formation flanked by alpha helices. Class II enzymes contain three conserved motifs (Eriani et al., 1990; Leberman et al., 1991). Motif 1 forms part of the dimer interface whereas motifs 2 and 3, located near the active site, participate in the ATP, amino acid and tRNA acceptor stem binding.

The differences between the two classes extend beyond their active site structure. Class I aaRS affix the amino acid to the 2'-hydroxyl group of the 3' end of the tRNA while class II aaRS affix the amino acid to the 3'-hydroxyl group of the same residue (Fraser and Rich, 1975; Sprinzl and Cramer, 1975). In addition, class I aaRS approach the acceptor stem of the tRNA from the minor groove while class II aaRS approach tRNA from the major groove side (Sissler et al., 1997) (**Figure 1.11**).

Each of these two classes can be further subdivided into three subclasses based on sequence analysis of aaRS throughout the tree of life (Cusack, 1997; Nagel and Doolittle, 1991). Each of these six subclasses is believed to have evolved from a separate single common ancestor that had previously evolved from the common ancestor of the entire class (O'Donoghue and Luthey-Schulten, 2003).

Interestingly, aaRS of the same subgroup tend to recognize similar types of amino acids. For instance, class Ic aaRS recognize aromatic amino acids such as tyrosine and tryptophan while class Ib recognizes amino acids with charged side chains such as lysine, glutamate and its derivative glutamine. In addition, a certain loose symmetry in the type of amino acid recognized by aaRS of different class, but of corresponding subclass, appears to exist. For instance, class IIc enzymes recognize the aromatic amino acid phenylalanine, similar to the corresponding class Ic enzymes which recognize aromatic amino acids. This symmetry is intriguing given that aaRS from different classes approach the tRNA from opposite sides of the molecule and therefore might be able to bind the tRNA at the same time (Sissler et al., 1997). Molecular modelling studies have shown that the corresponding subclasses (Ia-IIa; Ib-IIb; Ic-IIc) can simultaneously fit on the same tRNA molecule (Ribas de Pouplana and Schimmel, 2001), suggesting that the progenitors of corresponding subclasses may have originally bound the same tRNA.

1.5.3 Evolution of aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases are among the oldest proteins. 17 of the aaRS are universally distributed across the tree of life and their subsequent evolution marks for the most part the evolution of life (Nagel and Doolittle, 1995). However, three other aaRS appear to have evolved clearly after the last common ancestor: glutaminyl-tRNA synthetase (QRS), asparaginyl-tRNA synthetase (NRS) and cysteinyl-tRNA synthetase (CRS). QRS and NRS are only found in some bacteria and some eukaryotes, while CRS is not found in some archaea (Pavlov et al. 1997). In species lacking QRS and NRS, $tRNA^{Gln}$ and $tRNA^{Asn}$ are first changed with glutamate and aspartate, which are afterward modified to glutamine and asparagine, respectively, by tRNA- dependent amidotransferases (Curnow et al. 1997) (**Figure 1.12**). This indirect aminoacylation pathway is also seen for cysteine in some archaea where $tRNA^{Cys}$ is first charged with O-phosphoserine, which is afterwards modified to cysteine by Sep-tRNA:Cys-tRNA synthase (Sauerwald et al. 2005).

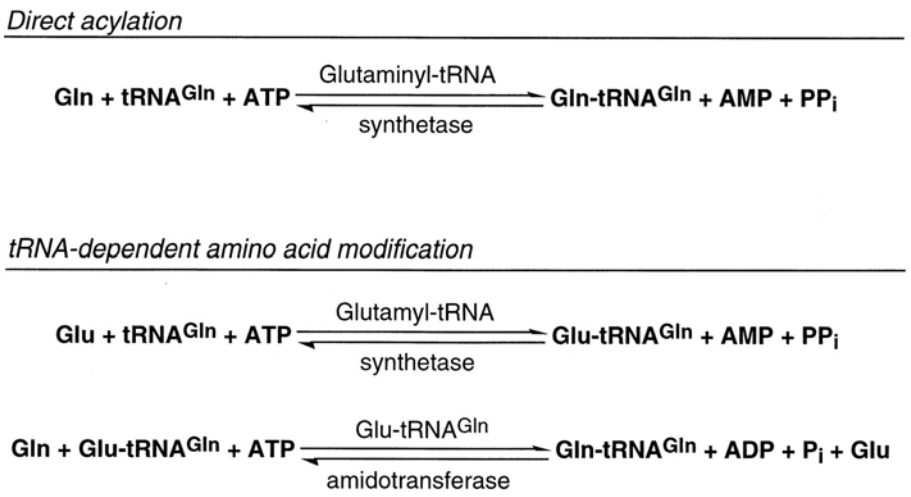


Figure 1.12. Mechanisms of aminoacyl-tRNA formation. Both pathways, direct acylation and tRNA-dependent amino acid modification, are depicted for glutaminyl-tRNA formation. For example, *E. coli* uses glutaminyl-tRNA synthetase (direct pathway), while *B. subtilis* employs $Glu\text{-}tRNA^{Gln}$ amidotransferase (indirect pathway) for $Gln\text{-}tRNA^{Gln}$ formation. Adapted from Woese et al. 2000.

Most of the aaRS phylogenies are often not consistent with accepted organismal phylogenies, *i.e.* they violate the so-called canonical phylogenetic pattern produced by 16s RNA sequences from the three domains of life: Archaea, Bacteria and Eukarya (**Figure 1.13**). Furthermore, the phylogenies inferred for aaRS of different amino acids often do not agree with another,

indicating that aaRS genes have undergone considerable horizontal gene transfer (HGT) across and within the three main branches of the tree of life (O'Donoghue and Luthey-Schulten, 2003). A clear HGT event from the endosymbiont (proto-mitochondria) to the nuclear genome of the host occurred at the origin of eukaryotes. The initial association of the ancestral alpha-proteobacteria and its host brought together two complete translation systems with a total of 40 different aminoacyl-tRNA synthetases. However, extant mitochondrial genomes encode no longer for aaRS.

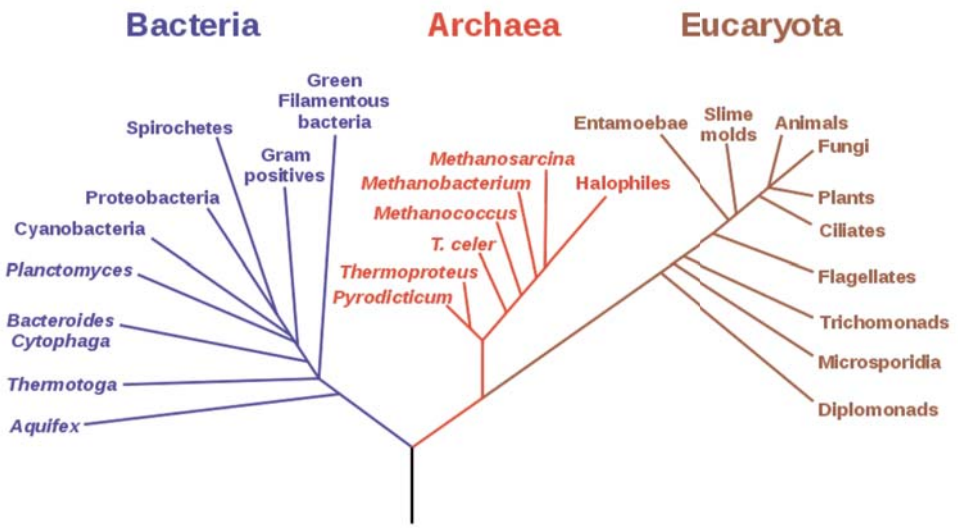


Figure 1.13. Canonical phylogenetic pattern of the tree of life based on 16s rRNA. Universal phylogenetic tree in rooted form, based on the work of Woese, 1977. Branching order and branch lengths are based upon rRNA sequence comparisons. Each kingdom has been coloured accordingly.

There are five possible fates of a single orthologous gene found in both the bacterial endosymbiont and the host after the endosymbiosis (Brown, 2003):

i) *gene retention*, when the gene is retained in the genomes of both organelle and the host (**Figure 1.14a**)
ii) *gene loss*, when the product encoded by a gene of the host genome functions now in two compartments: the organelle and the cytoplasm (**Figure 1.14b**)
iii) *gene co-existence*, when the organelle gene is transferred to the host genome, where it coexists with the host copy (**Figure 1.14c**)
iv) gene replacement, when the organelle gene is transferred to the eukaryotic host genome, where it substitutes an existing gene (**Figure 1.14d**)
v) gene function acquisition, when an unrelated nuclear gene encodes a protein that has acquired a new role in maintaining the organelle (**Figure 1.14e**)

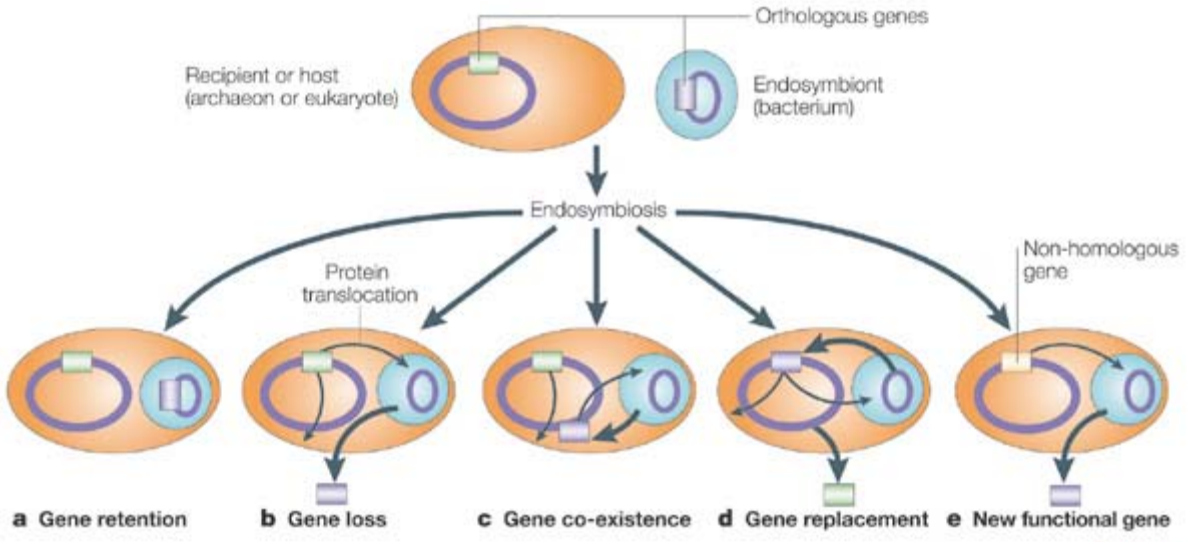


Figure 1.14. Possible fates of genes in endosymbiosis. A) *Gene retention*, where each orthologous gene is retained in its genome. This does not happen with aminoacyl-tRNA synthetases**. B)** *Gene loss* of the endosymbiotic gene, and dual targeting of the host gene. **C)** *Gene coexistence*, where the endosymbiont gene migrates to the nuclear genome, but each gene will be acting in its original compartment. **D)** *Gene replacement*, where the host gene is lost and the nuclear-encoded endosymbiont gene is dually targeted both to the cytosol and organelle. **E)** *New functional gene* targeted to the endosymbiont. Adapted from Brown, 2003.

1.5.4 Domains of aaRS and non-canonical functions

AaRS are multi-domain proteins. Their most conserved, and presumably oldest domains are the catalytic cores, which activate amino acids and transfer them to the 3' ends of tRNAs (**Figure 1.15**). Additional domains appended to or inserted in the body of aaRS increase efficiency and specificity of the aminoacylation process, either by providing additional tRNA contacts (e.g. anticodon binding domain), or by hydrolyzing non-cognate amino acid products (e.g. editing domains).

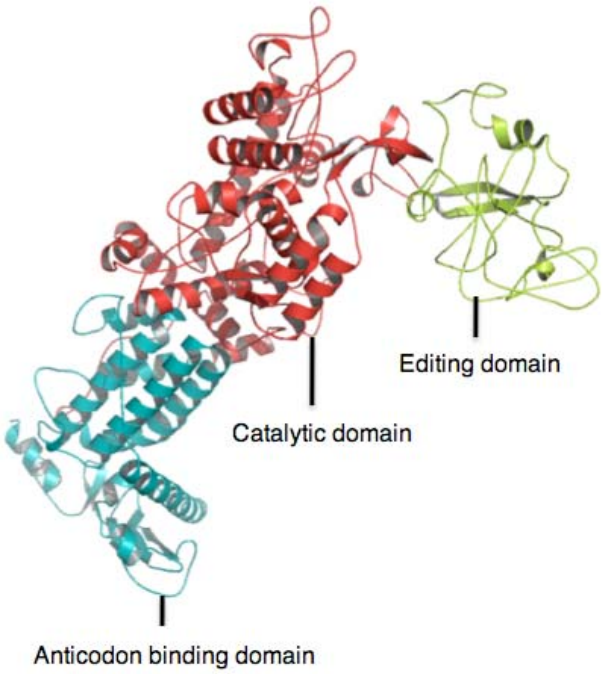


Figure 1.15. Basic domains of aminoacyl-tRNA synthetases. Three different domains found in *E. coli* isoleucyl-tRNA synthetase. Each domain has been coloured and is labelled accordingly.

Faithful translation of genetic information from mRNA to protein is critical for cellular function. Synthetases achieve the amino acid substrate specificity necessary to keep errors in translation to an acceptable level in two ways: preferential binding of the amino acid and selective editing of near-cognate amino acids. It has been postulated that error rates of >1 in 3.000 in the initial amino acid selection require correction mechanisms to increase the accuracy of aminoacylation and thereby reduce error in protein synthesis to a tolerable level (Fersht, 1981). When error rates exceed this threshold, the incorrect products are hydrolyzed at the secondary amino acid binding sites (editing sites), either by pre-transfer (hydrolysis of aminoacyl-adenylate) or post-transfer (hydrolysis of aminoacyl-tRNA) editing mechanisms (Jakubowski, 1981) (**Figure 1.16**).

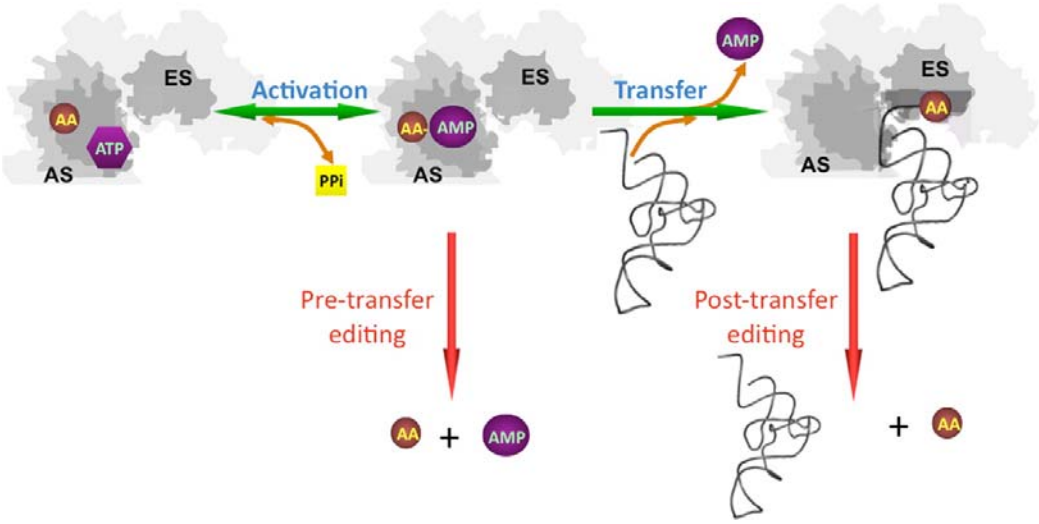


Figure 1.16. Pre-transfer and post-transfer editing of non-cognate amino acids by aaRS. The amino acid (AA) is activated at the active site (AS) to form aminoacyl-adenylate (AA-AMP). In pre-transfer editing, AA-AMP is hydrolyzed directly, whereas in post-transfer editing, the mischarged tRNA is translocated to the editing site, where the amino acid is removed. Adapted from Yadavalli et al. 2008.

Besides these basic domains (aminoacylation, anticodon binding and editing), new domains and motifs have been progressively added to aaRS to expand their functionalities (Brown et al. 2010; Park et al. 2008) (**Figure 1.17**). These appended domains, often dispensable for aminoacylation, are considered as markers for the aaRS-associated functions beyond translation (Guo et al. 2010).

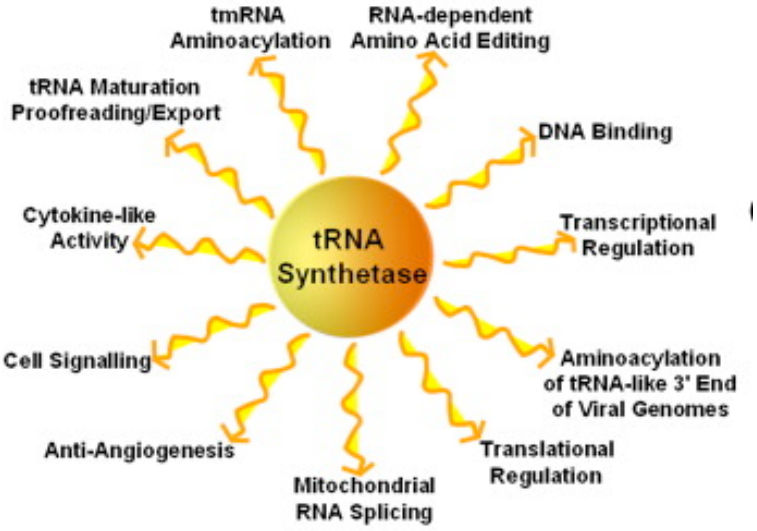


Figure 1.17. Non-canonical functions of aaRS. Adapted from Martinis and Pang, 2007

Indeed, during their extended evolution, aaRS have experienced numerous instances of duplication, insertion and deletion of domains. The aaRS-related proteins that have resulted from these genetic events are generally known as aaRS-like proteins. This heterogeneous group of polypeptides are paralogues of aaRS domains, and they carry out a varied number of functions that are not always related to gene translation (**Figure 1.18**).

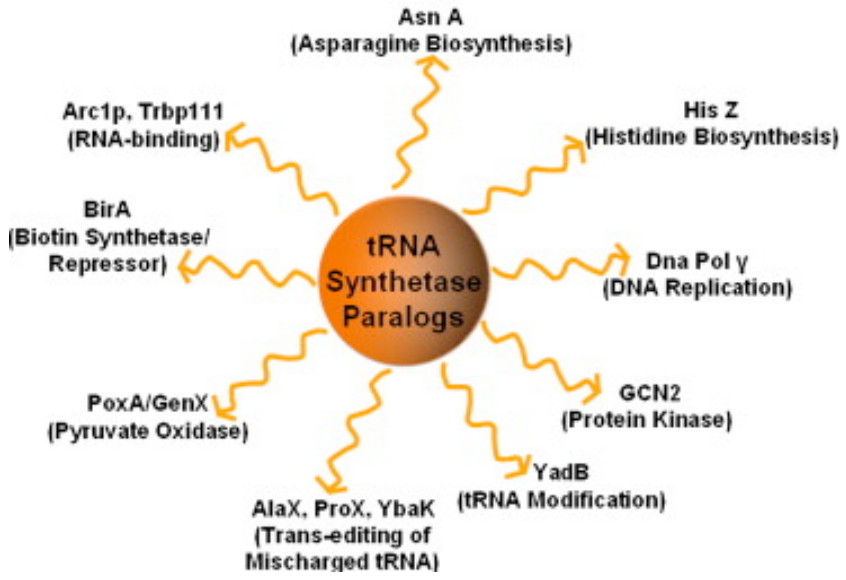


Figure 1.18. Alternate functions of aaRS-like proteins. Adapted from Martinis and Pang, 2007**.**

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

2. Objectives

Chapter 1: Genome-wide characterization of the gene translation machinery and its evolution across species

1.1 To characterize the evolution of tRNA genes across species

1.2 To identify the potential correlations between tRNA gene content and codon usage bias

1.3 To decipher the potential roles of codon usage bias across a species and between species

1.4 To apply the gained knowledge for biotechnological applications and understanding of translation defects in disease states

Chapter 2: *In silico* and *in vitro* drug design targeting the *Plasmodium falciparum* gene translation machinery

2.1 To study and characterize the translation machinery of *Plasmodium falciparum*, including the identification of its set of aaRS and tRNAs, genome-wide codon usage analysis and the identification of its strategy for maximizing its translation efficiency.

2.2 To identify and characterize novel drug targets in *Plasmodium falciparum*, including the phylogenetical and structural characterization and comparisons between plasmodial aaRS and its human homologues, as well as the i*n vitro* determination of the subcellular localisations of the candidate plasmodial drug targets.

2.3 To investigate known protein translation inhibitors, generate new lead compounds and test them, using diverse drug design strategies, which include structure-based drug design, high-throughput screening and screening of combinatorial libraries.

Chapter 3: Method development

3.1 To develop a pipeline and predict the pathogenicity of a protein based on its sequence

3.2 To quantify the reliability of homology models for docking purposes

3.3 To develop a workflow that maximizes the performance of homology models for docking purposes

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

3. PhD Advisor Report

The remarkable productivity achieved by Ms Novoa during her Ph.D. studies is partially reflected in the list of publications that she has, and will, obtain as a result of research during this period. It should be noted, however, that Eva started several other projects during these last years that have not been included in her thesis because they are still in earlier stages of development. I nevertheless expect two additional publications to appear as a direct result of these additional efforts.

Given the width of Eva's activity, I will now comment on her publications following the same structure that she has used for her thesis. In general it should be noted that she has managed to generate papers in first-rate journals, as well as technically specialized reports, and reviews in widely-read journals.

Chapter 1: Genome-wide characterization of the gene translation machinery and its evolution across species

***Publication 1**: A role for tRNA modifications in genome structure and codon usage.*
Novoa EM, Pavon-Eternod M. Pan T and Ribas de Pouplana L.
Cell 2012, 149: 202-213

In this paper, published in the most important biology journal, Eva reported that the emergence of two specific tRNA modifications shaped the structure and composition of all extant genomes. Through the analysis of more than 500 genomes, she identified two kingdom-specific tRNA modifications as major contributors that separated archaeal, bacterial, and eukaryal genomes in terms of their tRNA gene composition. We also experimentally demonstrated that human gene expression levels correlate well with genomic codon composition if these identified modifications are considered.

The relevance of this work cannot be understated. The realization that tRNA modifications may represent a new layer of gene translation regulation is completely new, and this paper by Eva puts her at the forefront of this new topic. The pioneering nature of the work is well reflected in the quality of the Journal that accepted to publish it.

Publication 2: *Speeding with control: codon usage, tRNA and ribosomes*
Novoa EM and Ribas de Pouplana L.
Trends in Genetics 2012 (in press)

As a result of our report in Cell we were invited to publish a review of the field in the broadly read journal Trends in Genetics. Here, we discussed the importance of codon-anticodon interactions in translation regulation and highlight the contribution of non-random codon distributions and post-transcriptional base modifications to this regulation. This article constitutes the most up-to-date review of this topic, and will serve as a reference for the whole RNA community.

Chapter 2: Aminoacyl-tRNA synthetases as antimalarial drug targets

Publication 3: *Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum*
Hoen R*, **Novoa EM***, López A, Camacho C, Cubells L, Martin P, Bautista JM, Vieira P, Santos M, Cortes A, Ribas de Pouplana L and Royo M. (*equal contributors)
J Med Chem (under review)

Eva's main interest during her thesis was combining *in silico* research with biochemical studies to develop new anti-malarial drugs. In this report she demonstrates the feasibility of coupling both strategies for the development of truly specific inhibitors. Indeed, she has demonstrated that selective inhibition of apicoplastic ARS is possible, and describes new compounds that show antimalarial activity and specifically inhibit *Plasmodium* apicoplastic lysyl-tRNA synthetase.

Publication 4: *Systematic study on Plasmodium falciparum aminoacyl-tRNA synthetases as antimalarial drug targets*
Camacho C, **Novoa EM**, Cubells L, Wilkinson B, Martin P, Bautista JM, Cortés A and Ribas de Pouplana L.
To be submitted

As a complementary work to the previous paper, Eva worked in collaboration with members of the lab to explore the potential of the aminoacyl-tRNA synthetase (ARS) family as source of antimalarial drug targets. The main conclusion of this work is that borrelidin, a natural inhibitor of threonyl-tRNA synthetase (ThrRS), stands out for its potent antimalarial effect. Moreover, we

found that certain borrelidin derivatives present higher selectivity towards the *P. falciparum* enzyme, thus revealing promising antimalarial scaffolds that should be further explored for the search of novel antimalarial drugs.

Chapter 3: Method development

Publication 5: *A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins.*
Novoa EM, Castro de Moura M, Orozco M and Ribas de Pouplana L.
FEBS Lett 2010, 584 (2): 460-466.

As part of her training in bioinformatics Eva developed several programs and approaches to improve our ability to identify relevant targets, and design active inhibitors against them. As part of this general goal she developed a new genomics method to determine the potential implication in pathogenicity of any given protein, and applied it in this paper to aminoacyl-tRNA synthetases.

Publication 6: *Ensemble docking in homology models.*
Novoa EM, Ribas de Pouplana L, Barril X and Orozco M.
J Chem Theory Comput 2010, 6 (8): 2547-2557

Following the same general objectives as in the previous article, Eva also dedicated time to the problem of inhibitor design against proteins of unknown three-dimensional structure. In this paper she described a systematic exploration of the quality of protein structures derived from homology modeling when used as templates for high-throughput docking. Remarkably, she found that structures derived from homology modeling are often similar in quality for docking purposes than real crystal structures, even in cases where the template used to create the structural model shows only a moderate sequence identity with the protein of interest. This work has the potential to greatly influence the way that researches approach the problem of inhibitor design against proteins of unknown structure but with close homologues in the PDB database.

Publication 7: *Small molecule docking from theoretical structural models*
Novoa EM, Ribas de Pouplana L and Orozco M.
In: “Computational Modelling of Biological Systems: From Molecules to Pathways”.
Ed Springer, New York (USA) Vol 4, pp 75-96.

As a result of her work on the problem of three-dimensional modelling of proteins, and her vast understanding of the literature on this topic, Eva was able to produce this extensive review chapter that provides readers with a comprehensive analysis of the current approaches to small molecule docking *in silico*. Together with publication 2, this work represents the contribution of Eva's thesis to literature review and topic analysis, and demonstrates the wide scope of contributions that her work towards her Ph.D. has produced.

Lluis Ribas de Pouplana
Gene Translation Laboratory
Institute for Research in Biomedicine

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

4. PUBLICATIONS

Chapter 1: Genome-wide characterization of the gene translation machinery and its evolution across species

***Publication 1**: A role for tRNA modifications in genome structure and codon usage.*
Novoa EM, Pavon-Eternod M. Pan T and Ribas de Pouplana L.
Cell 2012, 149: 202-213

***Publication 2**: Speeding with control: codon usage, tRNA and ribosomes*
Novoa EM and Ribas de Pouplana L.
Trends in Genetics 2012 (in press)

Chapter 2: Aminoacyl-tRNA synthetases as antimalarial drug targets

***Publication 3**: Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum*
Hoen R*, **Novoa EM***, López A, Camacho C, Cubells L, Martin P, Bautista JM, Vieira P, Santos M, Cortes A, Ribas de Pouplana L and Royo M. (*equal contributors)
J Med Chem (under review)

***Publication 4**: Systematic study on Plasmodium falciparum aminoacyl-tRNA synthetases as antimalarial drug targets*
Camacho C, **Novoa EM**, Cubells L, Wilkinson B, Martin P, Bautista JM, Cortés A and Ribas de Pouplana L.
To be submitted

Chapter 3: Method development

***Publication 5**: A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins.*
Novoa EM, Castro de Moura M, Orozco M and Ribas de Pouplana L.
FEBS Lett 2010, 584 (2): 460-466.

***Publication 6**: Ensemble docking in homology models.*
Novoa EM, Ribas de Pouplana L, Barril X and Orozco M.
J Chem Theory Comput 2010, 6 (8): 2547-2557

***Publication 7**: Small molecule docking from theoretical structural models*
Novoa EM, Ribas de Pouplana L and Orozco M.
In: "Computational Modelling of Biological Systems: From Molecules to Pathways".
Ed Springer, New York (USA) Vol 4, pp 75-96.

4.1. Chapter 1: Genome-wide characterization of the gene translation machinery and its evolution across species

4.1.1. Introduction

4.1.1.1 Control of gene expression

The control of gene expression is a fundamental process and its misregulation is usually associated with disease. It is now well established that gene expression is regulated at multiple levels. Gene regulation can be divided into transcriptional and post-transcriptional control (**Figure 4.1**). Furthermore, proteins themselves can be regulated by protein modifications and degradation.

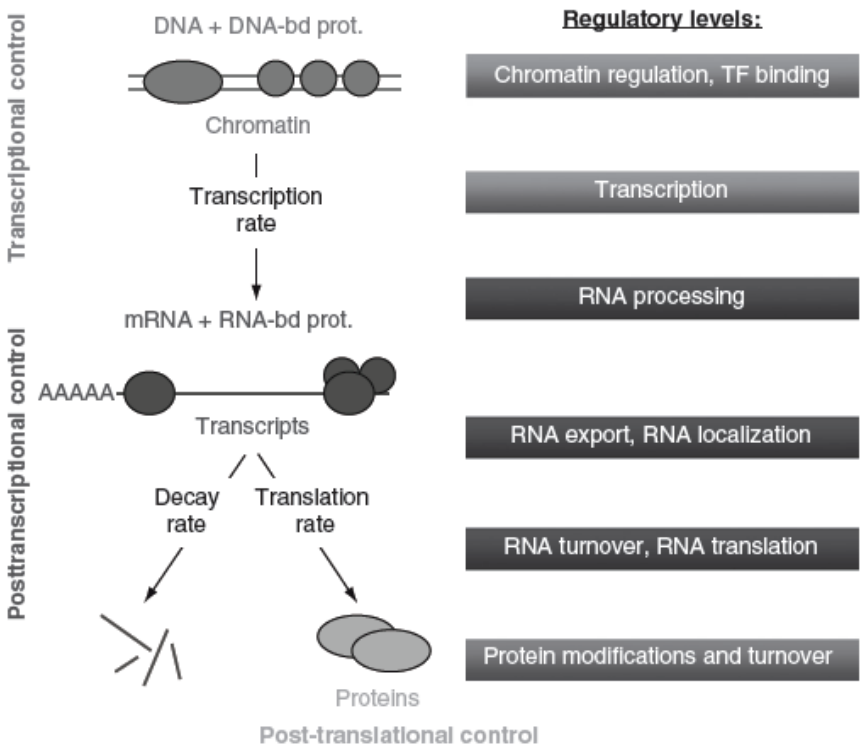


Figure 4.1. Scheme of different layers of gene regulation. The regulatory processes are listed according to their involvement in transcriptional, post-transcriptional or post-translational control. Adapted from Mata et al., 2005.

Transcriptional control has received much attention, through both traditional single-gene studies (Kadonaga, 2004) and genome-wide approaches, including expression profiling (Lockhart and Winzeler, 2000), transcription factor binding studies and identification of regulatory sequence elements (Sandelin et al, 2007) as well as chromatin remodelling and epigenetic analyses (Bernstein et al., 2007). In comparison, post-transcriptional control has been less extensively studied. Nevertheless, an increasing appreciation of the importance of post-transcriptional gene regulation is emerging.

4.1.1.2 Post-transcriptional regulation of gene expression

Post-transcriptional regulation mechanisms comprise various processes such as mRNA processing (polyadenylation, capping and splicing), mRNA export and localization, mRNA decay, and mRNA translation. Despite the variety of regulatory mechanisms, they all have one thing in common: they ultimately control if, where, and how efficiently a given mRNA is translated into protein. Consequently, translation and translational control are central to post-transcriptional regulation of gene expression.

Why do cells regulate translation and how do they benefit from it? There are several possible answers to this question. Regulation at the translational level can happen rapidly without the necessity of going through all the upstream processes of gene expression such as transcription, mRNA processing and mRNA export. Furthermore, translational regulation is usually reversible. Another reason for the regulation of translation is spatial control of gene expression within the cell (Schuman et al., 2006). Translational regulation also provides flexible control of gene expression, where translational efficiencies of few mRNAs can be affected selectively.

Although it is clear that translational regulation provides certain advantages in comparison to transcriptional regulation, the mechanisms through which it is accomplished have only recently started to be deciphered. Translation regulation can be performed both through external factors, such as protein factors and miRNAs (Gebauer and Hentze, 2009), although the latter remains widely unexplored. On the other hand, instrinsic factors also affect the mRNA translation rate. A major part of the control occurs at the stage of initiation, where ribosome recruitment takes place (Ingolia et al. 2009). The elongation phase has been shown to be governed by both mRNA secondary structure (Gray and Hentze, 1994) and the extent of adaptation of the coding sequence to the cellular tRNA pool (dos Reis et al. 2004; Sharp and Li, 1987).

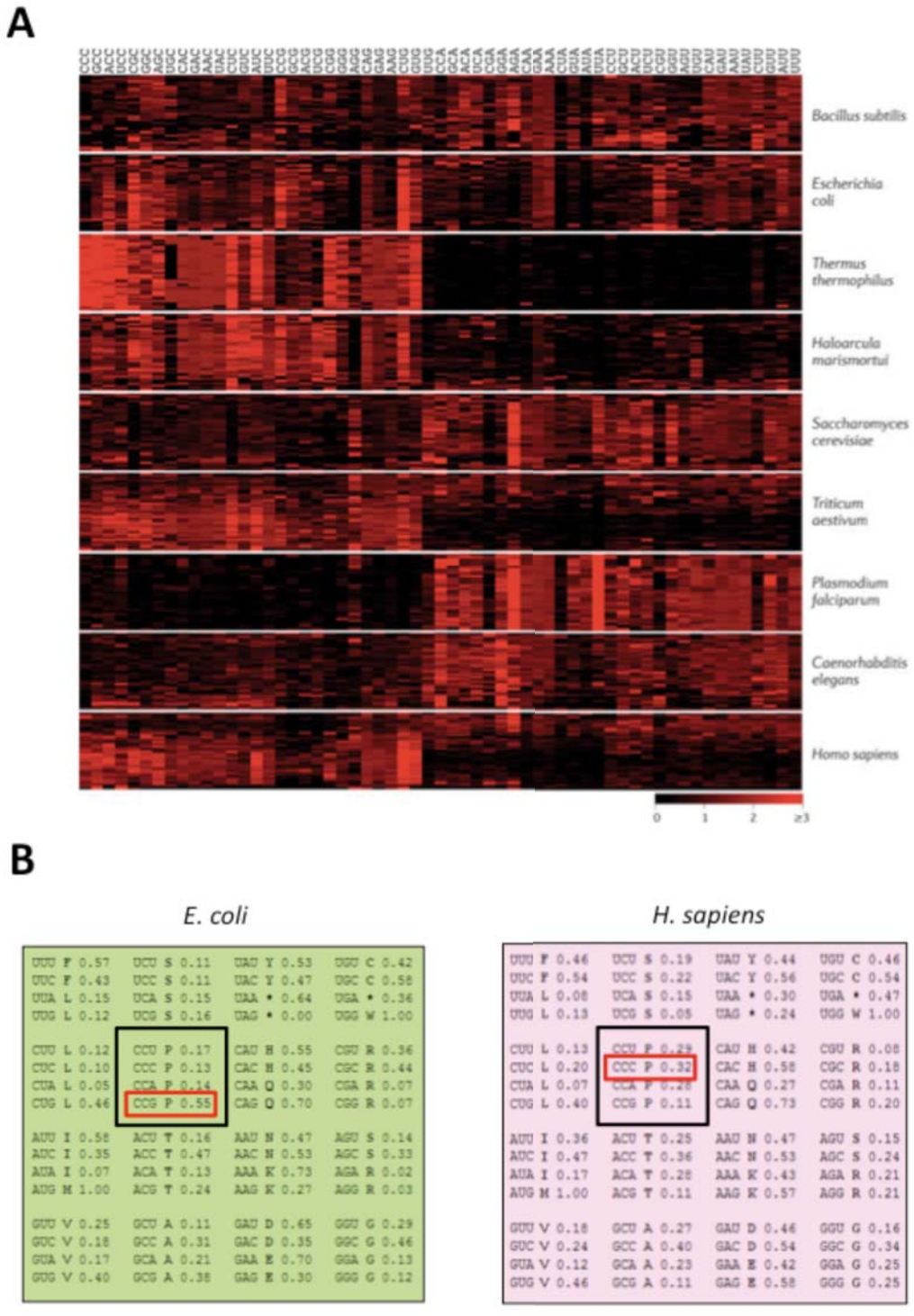


Figure 4.2. Codon bias within and between genomes. A) The relative synonymous codon usage (RSCU) is plotted for 50 randomly selected genes from each of nine species. RSCU ranges from 0 (when the codon is absent) through 1 (when there is no bias) to 6 (when a single codon is used exclusively in the case of a 6-codon family). Methionine, tryptophane and stop codons are omitted. Genes are in rows and codons are in columns. Adapted from Plotkin and Kudla, 2011. **B)** Codon usage data tables for *E. coli* and *H. sapiens*. Synonymous codons corresponding to proline are squared in black, with the preferred codon of the family box squared in red, respectively.

4.1.1.3 Translation efficiency

The final protein levels within a cell will depend on the rate or speed at which the mRNA is translated into proteins, on what is termed 'translation efficiency'. Several studies have shown that the choice of specific codons is important for determining the speed of translation (Arava et al. 2003; Tuller et al. 2007). Indeed, when designing heterologous genes, we intend to design the gene such that it matches the codon usage bias of the host species, because it will produce higher amounts of protein compared to genes containing 'rare' codons. But why would certain codons be 'preferred' in comparison to others?

4.1.1.3.1 Codon usage bias

As stated in the Introduction (section 1.2), the genetic code determines which of the 61 triplets or codons correspond to which of the amino acids. Because there are more codons than amino acids, the genetic code is necessarily redundant. While few amino acids are encoded by a single codon, most amino acids are encoded by two to six different codons. The different codons that encode the same amino acid are known as 'synonymous' codons. Changes in the DNA sequence of a protein between two synonymous codons are often assumed to have no effect and are thus called 'silent' changes caused by 'synonymous' mutations.

Even though synonymous codons encode for the same amino acid, it has been shown for a wide variety of organisms that they are not used with equal frequencies across different genes (**Figure 4.2a**). This phenomenon has been termed codon usage bias. Interestingly, the direction of codon bias shifts is different between organisms, *i.e.* the choice of the most abundant synonymous codon differs between organisms (Chen et al., 2004) (**Figure 4.2b**).

4.1.1.3.2 tRNA isoacceptor abundance

It is widely accepted that the reason why certain codons increase the rate of translation of a gene is because they are decoded by abundant tRNA species in the cell. Transcripts whose codons are biased toward the more abundant tRNAs are found to be more highly expressed (Man and Pilpel, 2007; Qin et al. 2004), whereas codons corresponding to rare tRNA species may induce long waiting time s and stall elongation at such positions, causing lower translation efficiencies.

The abundances of tRNA species tend to be proportional to the tRNA gene copy number (Tuller et al. 2010; see also Figure 1.8). However, why are certain tRNA isoacceptor abundances higher than others? And more importantly, why the choices of the most abundant tRNA isoacceptor change between organisms? (**Figure 4.3**) To understand the evolution that occurred across tRNA gene pools and to answer these questions, we performed a study on the evolution of tRNA gene contents throughout the 3 kingdoms of life. From our analysis, we find that the appearance of two different tRNA modification enzymes that increase the translation efficiency of certain codons explains the observed differences between the sets of 'preferred' tRNA isoacceptors across kingdoms. Indeed, the identified strategies to increase translation efficiency also explain the codon usage bias observed between high- and low-expressed genes in a given species (**Publication 1**).

AA	Species	tRNA isoacceptor			
		AGC	GGC	CGC	TGC
Ala	*Pyrococcus horikoshii*	-	1	1	1
	Escherichia coli	-	2	-	3
	Homo sapiens	29		5	9
		AGG	GGG	CGG	TGG
Pro	*Pyrococcus horikoshii*	-	1	1	1
	Escherichia coli	-	1	1	2
	Homo sapiens	10	-	4	7

Figure 4.3. tRNA isoacceptor gene copy numbers from 3 diverse species. The 'preferred' tRNA isoacceptor varies depending on the species. In this example, only the data tRNA isoacceptor gene copy numbers for the amino acids alanine (Ala) and proline (Pro) are shown. The tRNA gene copy numbers have been predicted using the tRNAscan-SE software (Lowe and Eddy, 1997). The species included belong to the following kingdoms: *P. horikoshii* (Archaea), *E. coli* (Bacteria), *H. sapiens* (Eukarya).

4.1.1.3.3 tRNA modifications

tRNA modifications, specially those affecting the wobble base position, can affect the translation efficiency of a given codon. A same tRNA isoacceptor can read more than one codon (e.g. $tRNA^{Phe}_{GAA}$ can read both UUU and UUC), but it may preferentially recognize one codon amongst the ones that it is capable to recognize. However, once this tRNA is modified, its pairing preferences may change, causing that another codon is preferentially recognized,

thus enhancing the translation efficiency of this codon. Therefore, the presence or absence of tRNA modifications can change the translation efficiencies of codons by its tRNAs, and consequently, the regulation of tRNA modifications constitutes a layer for post-transcriptional regulation of protein levels (**Publication 2**).

4.1.2. Publications

PUBLICATION 1:

A role for tRNA modifications in genome structure and codon usage.

Novoa EM, Pavon-Eternod M. Pan T and Ribas de Pouplana L.
Cell 2012, 149: 202-213

Cell

Theory

A Role for tRNA Modifications in Genome Structure and Codon Usage

Eva Maria Novoa,[1] Mariana Pavon-Eternod,[2] Tao Pan,[2] and Lluís Ribas de Pouplana[1,3,*]
[1]Institute for Research in Biomedicine, c/ Baldiri Reixac 15-21, 08028 Barcelona, Catalonia, Spain
[2]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA
[3]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain
*Correspondence: lluis.ribas@irbbarcelona.org
DOI 10.1016/j.cell.2012.01.050

SUMMARY

Transfer RNA (tRNA) gene content is a differentiating feature of genomes that contributes to the efficiency of the translational apparatus, but the principles shaping tRNA gene copy number and codon composition are poorly understood. Here, we report that the emergence of two specific tRNA modifications shaped the structure and composition of all extant genomes. Through the analysis of more than 500 genomes, we identify two kingdom-specific tRNA modifications as major contributors that separated archaeal, bacterial, and eukaryal genomes in terms of their tRNA gene composition. We show that, contrary to prior observations, genomic codon usage and tRNA gene frequencies correlate in all kingdoms if these two modifications are taken into account and that presence or absence of these modifications explains patterns of gene expression observed in previous studies. Finally, we experimentally demonstrate that human gene expression levels correlate well with genomic codon composition if these identified modifications are considered.

INTRODUCTION

Transfer RNAs (tRNAs) are present in all living organisms, acting as adaptors that link amino acids to codons in messenger RNAs (mRNA). Based on their aminoacylation identity, all tRNAs are subdivided into 20 accepting groups (alloacceptors). Each group comprises several tRNAs (isoacceptors) that translate synonymous codons with the same amino acid thanks to synonymous anticodons that vary mostly at the third position. The redundancy of the genetic code is due to synonymous codons, and solved by isoacceptor tRNAs.

tRNA genes tend to be present in multiple copies in the genomes of most organisms, from prokaryotes to eukaryotes, but the number of gene copies for each tRNA species (tRNAs with the same anticodon) varies widely from species to species (Marck and Grosjean, 2002). For any actively dividing cell, the translation efficiency of a given codon is determined by the amount of tRNA in the cell (Ikemura, 1981; Bennetzen and Hall, 1982; Sharp et al., 1988; Man and Pilpel, 2007; Akashi, 2003; Elf et al., 2003; Dittmar et al., 2005). The concentration of each tRNA is determined by its number of gene copies in the genome (Tuller et al., 2010a). Thus, tRNA gene content determines relative tRNA isoacceptor abundances that, in turn, determine codon translation efficiency. Therefore, the study of tRNA gene content bias may help explaining codon usage biases in extant genomes.

Previous reports have shown that the number of genes coding for each tRNA is not conserved between kingdoms (Gerber and Keller, 2001; Marck and Grosjean, 2002). The variability in tRNA gene number is extreme in some cases: certain tRNA species are absent in entire branches of the phylogenetic tree, whereas others are clearly predominant (e.g., in *Homo sapiens* 29 out of the 43 $tRNA^{Ala}$ genes (68%) correspond to the isoacceptor $tRNA^{Ala}_{AGC}$). The factors that influence tRNA gene copy number within genomes have been studied mostly in individual species (Withers et al., 2006; Gonos and Goddard, 1990; Kanaya et al., 1999; Dong et al., 1996), but the principles that govern the evolution of tRNA gene populations remain unknown.

In addition to the variability in tRNA gene content, the diversity of tRNA populations is further increased by species-specific base modifications. Thus, the tRNA signature of each species, defined as the total set of mature tRNAs that results from tRNA gene transcription, tRNA maturation, and the action of modification enzymes, is a complex evolutionary trait. Little is known about the parameters that shape the tRNA signature of species in evolution.

Two enzymes are known to cause modifications in base 34 of the anticodon that increase codon-pairing ability: tRNA-dependent adenosine deaminases (ADATs) and tRNA-dependent uridine methyltransferases (UMs) (Agris et al., 2007). tRNA-adenosine deaminases are essential enzymes found in Bacteria and Eukarya that catalyze the conversion of adenine-34 to inosine-34 (A-to-I editing) (Wolf et al., 2002; Gerber and Keller, 1999; Maas and Rich, 2000). I34 is able to wobble with adenine, cytosine, and uridine (Gerber and Keller, 2001). Thus, INN anticodons are capable of pairing with three different codons. Unlike in Bacteria, where ADAT only modifies $tRNA^{Arg}$, in Eukarya a heterodimeric form of this enzyme (hetADAT) formed by Tad2p and Tad3p deaminates several tRNAs (Gerber and Keller, 1999). On the other hand, bacterial UMs, modify uridine to xo^5U_{34}, enabling its pairing with adenine, guanosine and uridine (Yokoyama et al., 1985). Two enzymes have been identified as responsible for the last step of xo^5U modifications: CmoA and CmoB (Näsvall et al., 2004).

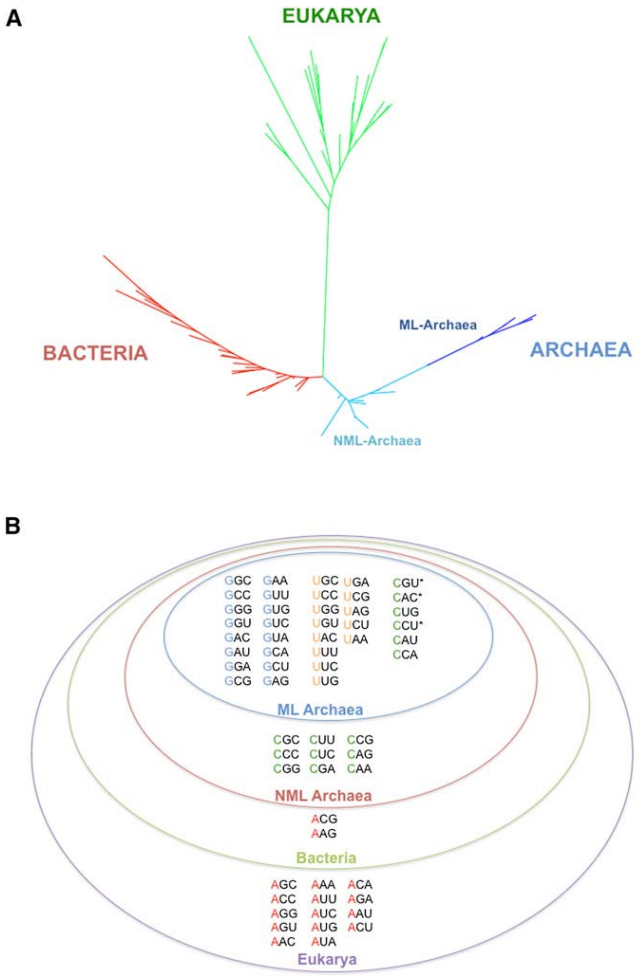


Figure 1. Genome Phylogeny Based on tRNA Gene Content

(A) Distance-based phylogeny based on tRNA gene content, performed with equal number of species of each kingdom. The four phylogenetic clusters have been labeled accordingly. The phylogeny performed with the whole set of 527 species is consistent with these results (see Figure S1).

(B) Diagram showing the increase in tRNA population complexity in the four main phylogenetic clusters found in this work (each tRNA is designated by its anticodon sequence). Each base at the wobble position is colored according to its chemical nature. Anticodons labeled with an asterisk (CGU, CAC, CCU) correspond to tRNA genes that are not found in all species comprising the ML-Archaea clade.

In this work, we have analyzed the distribution and abundance of all tRNA genes in more than 500 species across the three kingdoms of life. We first confirmed that tRNA gene composition can be considered a single trait that recapitulates the main evolutionary lines of the tree of life. Using principal component analysis, we identified those tRNA isoacceptors that became positively selected (increased in number) in Bacteria and Eukarya. Our results indicate that the appearance of UMs and hetADATs contributed to the divergence of eukaryal and bacterial genomes from their archaeal counterparts. The effect of the modifications caused by these enzymes increased the decoding capacity of modified tRNAs which, therefore, were positively selected during evolution. The diverse codon usage biases displayed by Bacteria and Eukarya are, at least partly, due to the different modification strategies used to improve translation efficiency, which are kingdom specific.

RESULTS

tRNA Gene Content as a Tool for Phylogenetic Analysis

The short sequence length of tRNAs, and their susceptibility to be transferred horizontally, limits the usefulness of their sequences for phylogenetic analysis. But tRNA gene content, defined as the set of tRNA genes used by a given organism to translate its genome, is unaffected by these limitations. In gene content-based phylogenies the evolutionary distance between species is calculated on the basis of acquisition or loss of genes. Gene content analyses using genome sequences (Snel et al., 1999; Iwasaki and Takagi, 2007; Fitz-Gibbon and House, 1999), protein domain content (Yang et al., 2005), and whole-proteome comparisons (Tekaia et al., 1999) have been previously reported.

Using tRNA gene content analysis, we have built a phylogenetic tree of more than 500 species that correctly identifies four known clades: (1) *Methanococcus*-like Archaea, (2) non-*Methanococcus*-like Archaea, (3) Bacteria, and (4) Eukarya (Figure 1A, see also Figure S1 available online). As can be seen in Figure 1A, tRNA gene content as a single trait follows the evolution of the whole tree of life, correctly clustering species into their corresponding kingdoms. Although this method is not powerful enough to correctly resolve the inner topology of individual clades, several outliers in tRNA signatures that have been previously reported (Man and Pilpel, 2007) are correctly identified by our approach. This indicates that kingdom-specific parameters drove the divergence of tRNA gene populations between the three kingdoms of life.

The four clades found in our gene-content analysis correspond to different levels of tRNA population complexity. Indeed, the tRNA gene populations of the clades vary from the relatively simple tRNA gene composition of Archaea, to an intermediate situation in Bacteria, and the most complex tRNA gene set found in Eukarya (Figure 1B). This increase in complexity implies that, along evolution, the number of tRNA species tended to increase through duplications or changes in anticodon specificity. Interestingly, the fact that *Methanococcus*-related species present

the simplest decoding strategy coincides with the proposed ancestral nature of this clade (Stetter, 1996; Brochier and Philippe, 2002).

To characterize the four identified phylogenetic clades, we quantified and analyzed the distribution of tRNA isoacceptor gene copy numbers within each of these four groups. As can be seen in Figure 2, each clade has different tRNA gene abundances and, more interestingly, unequal enrichment of certain tRNA isoacceptors. The archaeal clades are characterized by a relatively uniform distribution of tRNA gene copy numbers, with little variation between isoacceptors (e.g., all tRNA isoacceptors coding for alanine have similar gene frequencies). Thus, Archaea presents the simplest decoding scenario, with a minimal set of tRNA isoacceptors (Figure 1B) and uniform abundances of tRNA genes (Figure 2). In contrast Bacteria and Eukarya are more complex, both in terms of relative number of tRNA isoacceptors and in differences in the frequencies of tRNA gene copy numbers.

The loss of uniformity in tRNA gene abundances is not equivalent in Eukarya and Bacteria. For example, tRNAs with ANN anticodons tend to be absent both in archaeal and bacterial genomes, whereas in Eukarya they are the most abundant isoacceptors in four-codon (Pro, Ala, Val, Thr) and six-codon (Ser, Leu, Arg) tRNA sets (Figure 2). It is unclear, however, why should such selection act in a given kingdom and not in another. To try to answer this question, we first performed Principal Component Analysis (PCA) to statistically identify the tRNA isoacceptors that have been positively selected in each of the kingdoms.

Statistical Analysis of tRNA Gene Frequencies

PCA is a mathematical procedure that uses orthogonal transformation to reduce the dimensions of the data (correlated variables, in our case, tRNA gene frequencies), obtaining new variables (principal components, PCs) that are linear combinations of the original variables. Multivariate statistical analysis methods like PCA are particularly well adapted to the multidimensional nature of tRNA gene content data. If the original variables are correlated, most of the variance can be condensed in the two first PCs (PC1 and PC2). Analysis of our data shows that PC1 and PC2 account for 64.5% of the variance of tRNA gene content values, allowing us to analyze our results in two dimensions (Figure 3).

The *scores* plot—the transformed variable values (Figure 3A)—correctly clusters the species used in this analysis into their three respective kingdoms, and shows that PC1 is the principal component responsible for the separation of Bacteria, whereas PC2 is responsible for the separation of Eukarya (confirmed by t test, p values of 1e-5 and 2e-16, respectively). On the other hand, the *loadings* plot (Figure 3B) identifies which variables (tRNA isoacceptors) are contributing most to the differences between clusters. Top-ranked tRNA isoacceptors that are significantly associated to Bacteria and Eukarya are included inside an ellipse. The individual correlation values are listed in Table S1. Our data shows that eukaryal species present a positive selection of tRNA(ANN) isoacceptors belonging to four-codon families (Val, Pro, Ala, Thr), six-codon families (Leu, Ser) and split tRNA sets (Ile). On the other hand, bacterial species positively selected tRNA(UNN) isoacceptors for the same codon families.

The analysis of additional PCs was also performed to identify minor contributors to the differences between kingdom-specific tRNA gene populations (Figure S2). Interestingly, PC3 separates both Bacteria and Eukarya from Archaea due to the contribution of $tRNA^{Arg}$(ACG), confirming the importance of ANN isoacceptor tRNAs in the divergence of tRNA gene populations in the three kingdoms of life (r = 0.44, p value = 5.6e-27).

tRNA Modification as a Factor in Translational Efficiency

Translational efficiency is increased by optimized codons, i.e., those codons that correspond to the most abundant tRNA species (Hershberg and Petrov, 2008). Therefore, the positive selection of tRNA isoacceptors that we observe in our data could be due to the increased translational efficiency allowed by these tRNAs. As mentioned previously, kingdom-specific modifying enzymes exist that can increase the translational efficiency of tRNAs through modifications of the anticodon wobble base. We hypothesized that the selection of certain tRNAs over other isoacceptors, i.e., those identified in our analysis, may be due to their ability to incorporate anticodon modifications that increase their pairing repertoire (Figure S3).

If base modifications in the anticodon increase translational efficiency then those anticodons capable of accepting I_{34} and xo^5U_{34} modifications should be positively selected in the species where the corresponding modification enzymes exist. We first checked whether genes coding for tRNA(ANN) isoacceptors capable of being modified by hetADATs are overrepresented (Table 1) in species that contain these enzymes. This is exactly the case, indicating that the activity of hetADATs is exerting a selective force on the tRNA pool. We then checked whether genes coding for tRNA(UNN) isoacceptors modifiable by UMs are enriched among Bacteria. Indeed, UNN anticodons that are modified by UMs are enriched in bacterial genomes, indicating that the activity of UMs is associated with the tRNA composition of bacterial species toward U34 tRNAs (Table 1).

The analysis of further PCs supports the role of these two tRNA modifications in the divergence of tRNA gene populations. As mentioned above, PC3 clusters the bacterial and eukaryal kingdoms, and separates them from the archaeal species, mainly due to the contribution of $tRNA^{Arg}$(ACG). This tRNA isoacceptor is the only tRNA species deaminated by ADATs both in Bacteria (through TadA) and Eukarya (through Tad2/Tad3). Thus, our analysis indicates that the vast majority of the contributions to the segregation of extant tRNA gene populations are related to the activity of anticodon-modifying enzymes.

It should be noted that sequence modifications outside the anticodon can also have effects on codon:anticodon interactions (Geslain and Pan, 2010; Ledoux et al., 2009). However, to our knowledge, tRNA modifications outside the anticodon have not been found to expand the decoding capacity of tRNAs. The analysis of the full set of known tRNA anticodon modification enzymes (Table S2) reveals that only bacterial UMs and eukaryal hetADATs display phylogenetic distributions and sets of tRNA substrates fully compatible with the families of tRNAs found to be enriched in our study.

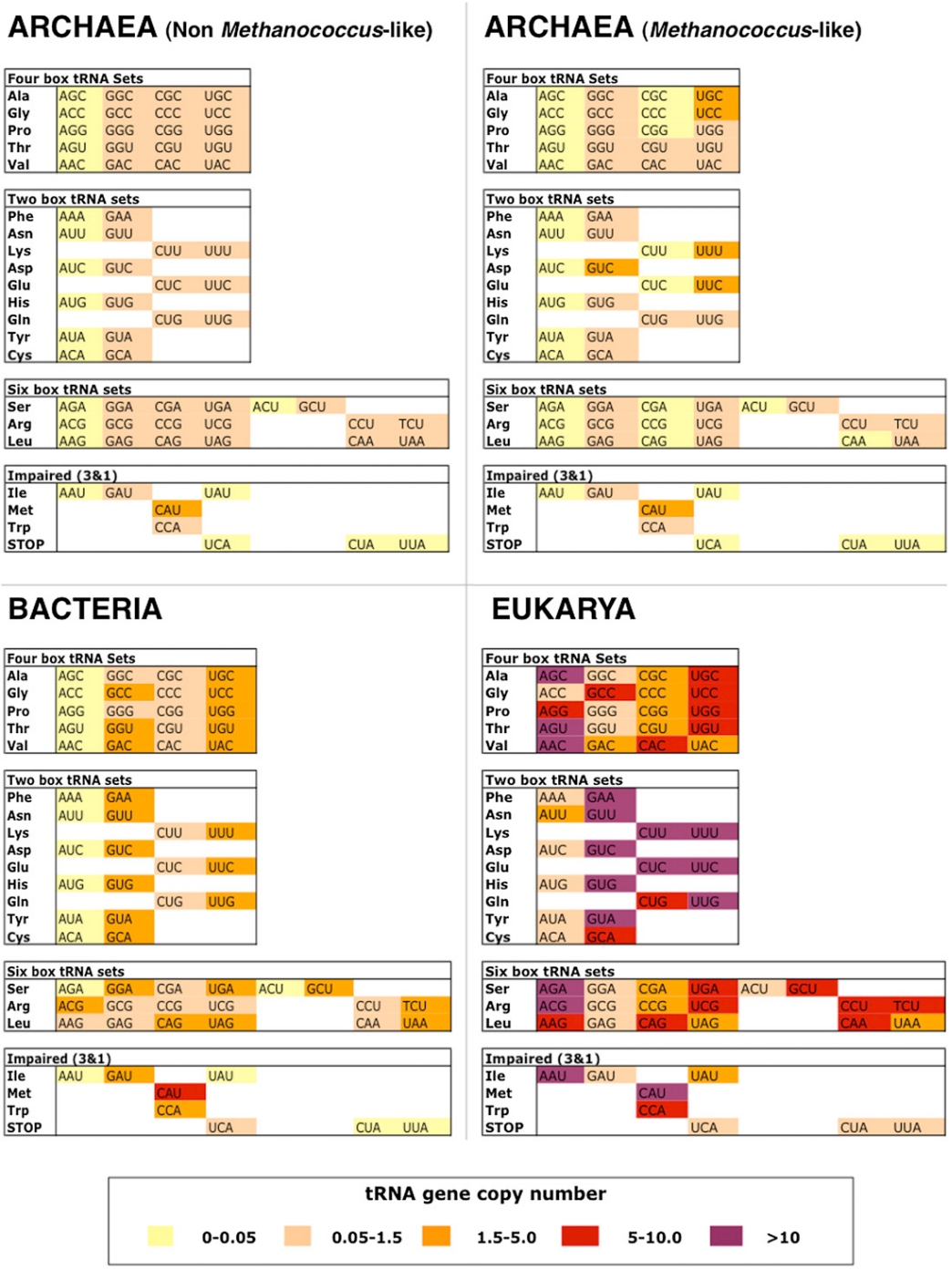


Figure 2. Unequal Enrichment of tRNA Isoacceptors Is Kingdom Specific

Mean tRNA abundances in the four phylogenetic clusters identified by gene content analysis: (1) *Methanococcus*-like Archaea, (2) non-*Methanococcus*-like Archaea, (3) Bacteria, and (4) Eukarya. Each tRNA anticodon is colored according to its average number of encoding tRNA genes. To deal with exceptional cases such as *Ferroplasma acidarmanus*, which is the sole archaea with a $tRNA^{Leu}$(AAG) gene (Marck and Grosjean, 2002), we have considered as absent those tRNA isoacceptors whose average tRNA gene copy number is between 0 and 0.05 (shown in yellow).

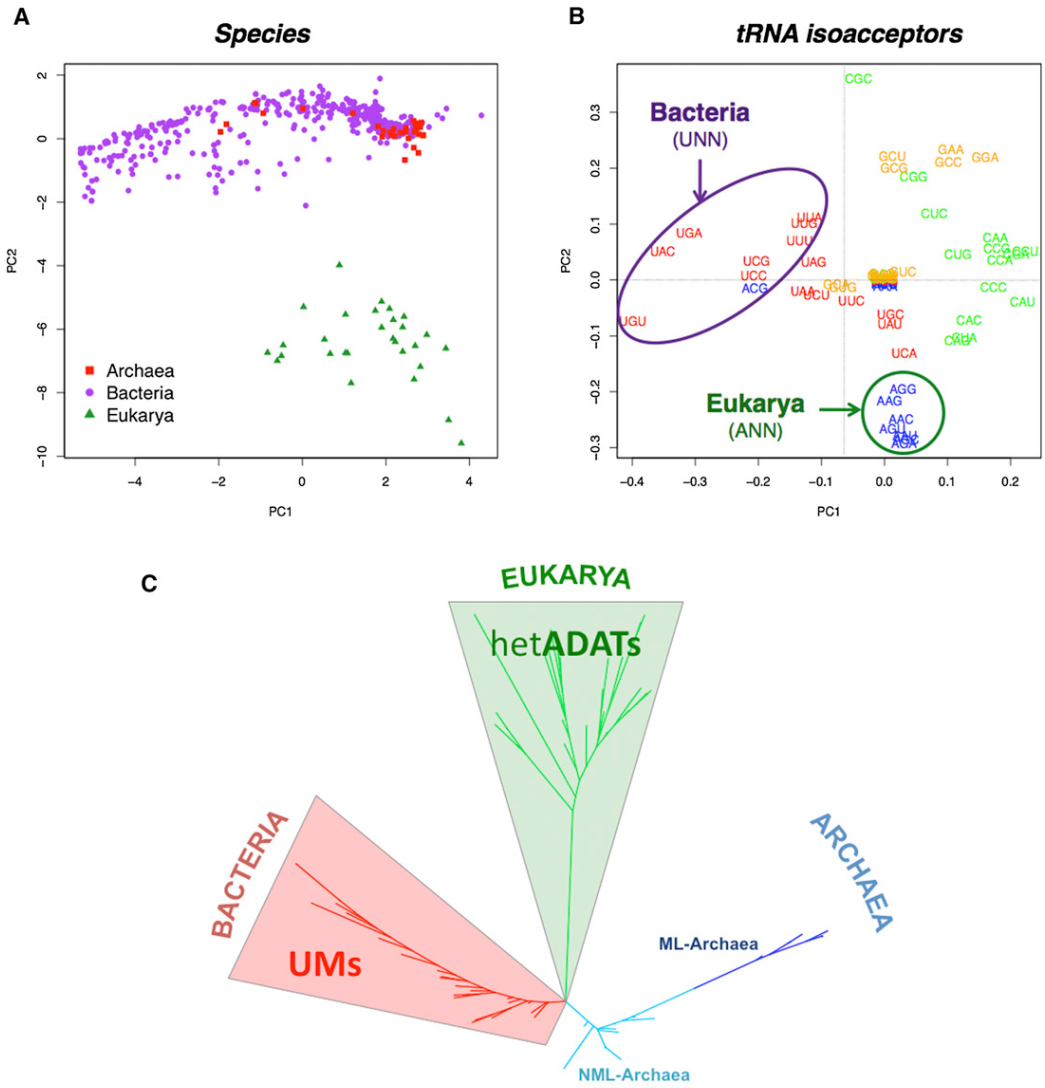


Figure 3. Identification and Quantification of Overrepresented tRNA Isoacceptors

(A) Biplot of the scores after performing Principal Component Analysis (PCA). Archaea (red), Bacteria (purple) and Eukarya (green) are distinguishable clusters using this analysis. The archaeal outliers correspond to *Methanococcus* species, which were already identified as a separate cluster using the tRNA gene content analysis.

(B) Biplot of the loadings, indicating the tRNA isoacceptors whose frequencies contribute the most to each of the clusters. Each anticodon has been colored depending on its wobble base. The ellipses surround those anticodons that are significantly associated to the PCs, either with PC1 negative values, which correspond to Bacteria (purple), or with PC2 negative values, which correspond to Eukarya (green) (see Table S1 for the individual correlation values). See also Figure S2 and Table S2.

(C) Genome phylogeny based on tRNA-gene content. The distributions of the two wobble base modification enzymes that act upon the tRNA isoacceptors identified in the PCA are shown. Uridine methyltransferases (UMs, labeled in red) are exclusively distributed across the bacterial kingdom. Heterodimeric adenosine deaminases (ADATs, labeled in green) are exclusively distributed in eukaryotes. Homodimeric forms of ADATs (TadA) are found in bacteria, but they only increase the decoding capacity of $tRNA^{Arg}$, and for simplicity, are not shown in the phylogeny.

Correlation between tRNA Gene Abundances and Codon Usage

Several studies performed on unicellular species have shown a correlation between tRNA abundance and codon usage (Ikemura, 1981; Ran and Higgs, 2010; Kanaya et al., 2001; Dong et al., 1996). In higher eukaryotes the search for this correlation has been less successful (Kanaya et al., 2001; dos Reis et al., 2004), and it has been proposed that in these species translation efficiency might not be the primary factor influencing codon usage (Kanaya et al., 2001). Studies in *Drosophila melanogaster* have concluded that in this organism selection acts to increase translation accuracy (Akashi, 1994; Moriyama and Powell, 1998), whereas other authors have linked codon usage in metazoans to several parameters, including average gene length

Table 1. Overrepresented tRNA Genes Correspond Exactly to Those Isoacceptors Modifiable at the Wobble Position by UMs and ADATs

	ADAT Gene	Anticodons Modified by ADATs	A34 Anticodons with RGF > 1.6[a]
Archaea			
Any species	—	—	none
Bacteria			
E. coli	*tadA*	ACG	ACG
Eukarya			
S. cerevisiae	*tad2p/tad3p*	AGA, AGG, AGU, AAC, AGA, ACG, AAU	AGA, AGG, AGU, AAC, AGA, ACG, AAU
H. sapiens	*tad2/tad3*	AGA, AGG, AGU, AAC, AGA, ACG, AAU, AAG	AGA, AGG, AGU, AAC, AGA, ACG, AAU, AAG
	UM Gene	Anticodons Modified by UMs	U34 Anticodons with RGF > 1.6[a]
Archaea			
Any species	—	—	none
Bacteria			
S. enterica	*cmoA/cmoB*	UGC, UGG, UGU, UAC, UGA, UAG	UGC, UGG, UGU, UAC, UGA, UAG
Eukarya			
Any species	—	—	none

[a]The RGF threshold was chosen such that the overrepresented tRNA isoacceptors also correspond to the most abundant isoacceptor among its tRNA codon family.

(Duret and Mouchiroud, 1999), cost of proofreading, or translation efficiency (Duret and Mouchiroud, 1999; Duret, 2000; Tuller et al., 2007, 2010b).

We analyzed the correlation between tRNA gene copy number and codon usage in more than 500 genomes using previously reported approaches. We first determined the set of *highly adapted* codons (those recognized by tRNAs coded by the most abundant tRNA genes) and compared them to the set of *highly abundant codons* (those with high relative synonymous codons usage [RSCU], determined from gene sequences of ribosomal proteins). Our results confirm that the most abundant codons (highest RSCU) in general correspond to the most adapted codons (61% match) (for four- and six- codon families, the two most abundant codons are included in the analysis). However, as previously reported, this correlation is not perfect, and it is poor in eukaryotic genomes. Indeed, when considering the top two tRNA isoacceptors, archaeal species present the best match (75%), whereas Bacteria and Eukarya show matches of 59% and 41%, respectively (Table S3).

Strikingly, the codons whose frequencies do not correlate well with tRNA gene content values are precisely those codons corresponding to tRNAs susceptible to be modified either by adenosine deaminases or uridine methyltransferases (Figure 4A, see also Figure S4). It is worth noting that hetADATs and UMs exclusively modify those previously nonmatching codons (Figure S4). We reclassified those codons in the correlation analysis to account for the increased pairing ability of anticodons modified by UMs and hetADATs. This new analysis provided quasiperfect correlations between RSCU values and tRNA gene copy numbers in Bacteria and Eukarya (95% match) (Figure 4A). Therefore, tRNA gene copy number is almost perfectly correlated with codon usage in all kingdoms, provided that tRNA modifications caused by hetADATs and UMs are considered. This implies that, in all kingdoms of life, translational efficiency seems to be a primary factor influencing codon usage.

To experimentally confirm that association between codon usage and tRNA abundance is enhanced by the inclusion of modification enzymes, we determined $tRNA^{Arg}$ isoacceptor concentrations in HeLa and Hek 293T cell lines. We chose $tRNA^{Arg}$ for this analysis because all five human arginine isoacceptors can be individually quantified thanks to isoacceptor-specific probes. We performed an association analysis for $tRNA^{Arg}$ expression and codon usage in the absence or presence of modification information. Only after the inclusion of hetADAT modification information in the calculations could a good correlation be found between tRNA abundance and codon usage (Pearson correlation: 0.86 and 0.81 for HeLa and 293T, respectively) (Figure 4B).

To further confirm these results we also analyzed published data on gene expression levels in other species. In a recent study, Kudla et al. synthesized a library of 154 genes coding for green fluorescent protein (GFP) that varied randomly at synonymous sites (Kudla et al., 2009). These genes were expressed in *Escherichia coli*, and GFP expression levels were obtained that varied 250-fold across the library. The initial analysis of this data failed to find a correlation between codon composition and gene expression (however, see Supek and Smuc, 2010; Navon and Pilpel, 2011). We wondered whether the inclusion of the activity of UMs in the model would improve the correlation between translation efficiency and codon composition. Thus, we tested whether codon composition correlated with protein production when the frequencies of UM- and hetADAT-modifiable anticodons (hereinafter named "preferred codons") and nonmodifiable anticodons (hereinafter named "nonpreferred codons") were taken into account. This was indeed the case, and we obtained quasiperfect correlations in the set of highly expressed GFP genes (94% match) (Figure S4).

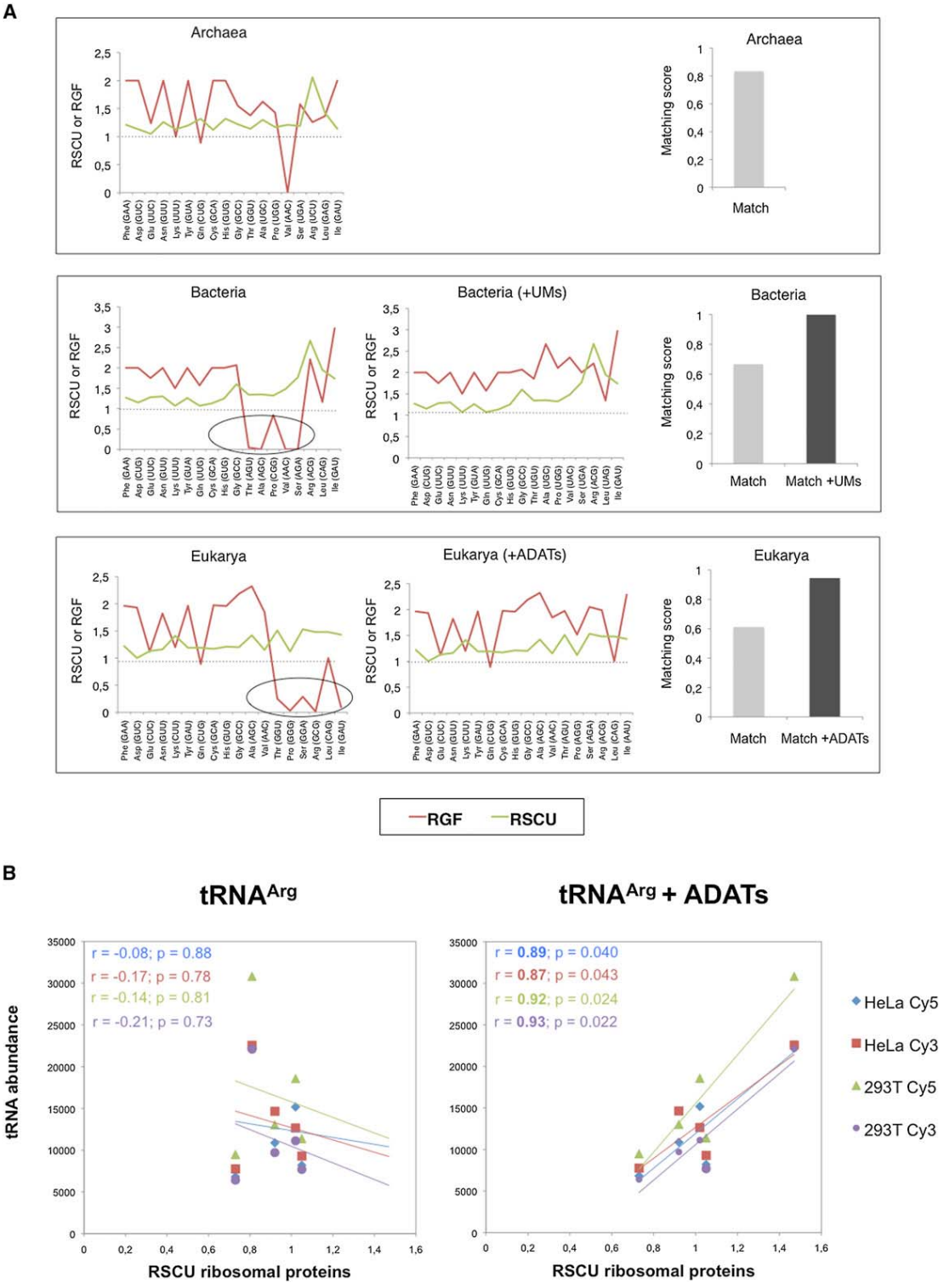


Figure 4. Match between Most Adapted Codons and Most Abundant Codons

(A) The match between the highest RSCU codon (green, most abundant codons) and the RGF value of its decoding tRNA (red, most adapted codons) is shown, for each kingdom, in the left column. The match after correcting the RGF values to account for the activity of UMs and ADATs is shown in the middle column. Archaea present neither ADATs nor UMs, and therefore the middle column is missing for this kingdom. The increase in the match score between RSCU and RGF after the correction is shown for each kingdom in the right histogram (except for Archaea).

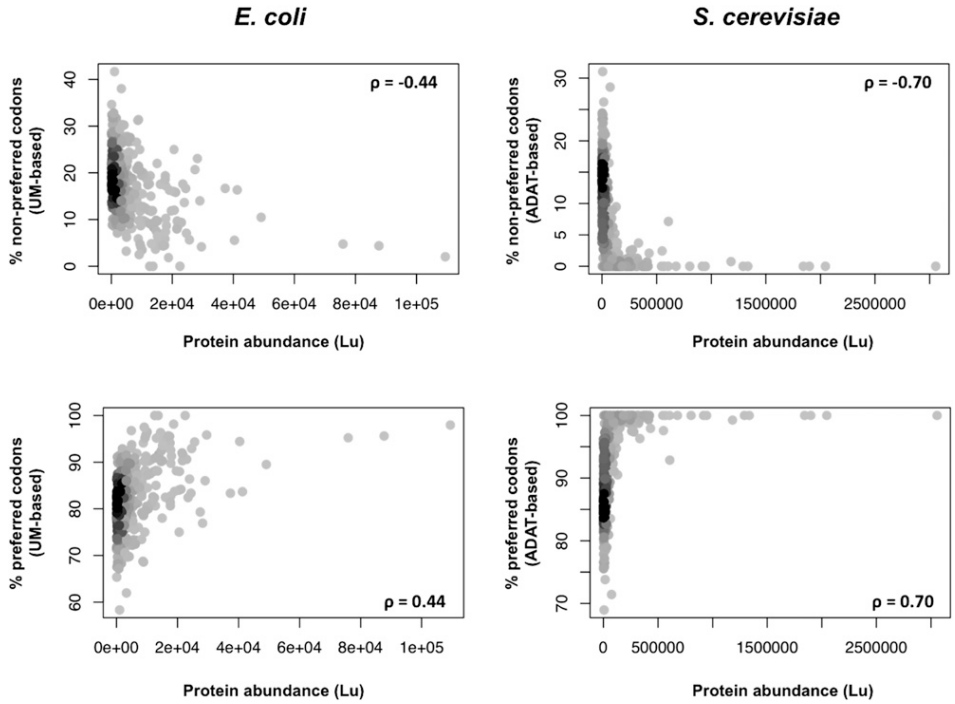


Figure 5. Correlation between Preferred Codons and Protein Abundance

In both *E. coli* and *S. cerevisiae*, the abundance of preferred codons in a gene correlates with protein abundance (Spearman correlation: 0.44 and 0.70, with p values of 9.7e-20 and 5.1e-52, respectively). Complementarily, the frequency of nonpreferred codons in genes decreases proportionally to protein abundance. The local density of data points in the graph is signified by their color (darker corresponding to more populated areas of the plot). See also Figure S5.

Analysis of the Influence of "Preferred Codons" in Protein Synthesis

Our results indicate that those transcripts whose codon composition is best adapted to anticodons modified by ADATs and UMs are the most efficiently translated. We therefore checked whether the relative abundance of preferred codons correlates with expression levels of any given gene. In this regard, genome-wide expression analyses (Lu et al., 2007; Ingolia et al., 2009; Ishihama et al., 2008; Ghaemmaghami et al., 2003; Taniguchi et al., 2010) provide experimental quantification of translational efficiency across a whole genome.

We examined the effect of UM and hetADAT modifications in published whole genome expression data obtained through the analysis of the *E. coli* and *Saccharomyces cerevisiae* transcriptomes. We found a good correlation between relative abundance of "preferred codons" of any given gene and its protein abundance in *E. coli* and *S. cerevisiae* ($\rho = -0.44$ and -0.70, respectively) (Figure 5, see also Figure S5). Different genome-wide expression data sets (Lu et al., 2007; Ishihama et al., 2008; Newman et al., 2006) produced similar correlations for both species ($\rho = -0.27$ and -0.74, respectively) (Figure S5). Moreover, an inverse correlation between protein abundance and nonpreferred codons was also detected, suggesting the existence of an upper maximum limit of nonpreferred codons per gene. Thus, the abundance of preferred codons possibly represents an additional level of translation control that needs to be considered in addition other mechanisms of posttranscriptional regulation (Mata et al., 2005).

DISCUSSION

Despite the central role of tRNAs in protein translation, the connections between tRNA gene population dynamics and genome evolution have rarely been explored. It is known that in unicellular organisms the most abundant codons are recognized by the most abundant tRNAs in the cell (Withers et al., 2006; Tuller et al., 2010a). However, we do not understand the reasons for the variability between tRNA pools of different species, nor the principles that determine tRNA gene abundances or genomic codon composition.

Our tRNA gene content analysis shows that genomic tRNA gene composition is an evolutionary trait that separates the main kingdoms of life. This separation is mainly due to the selection of tRNA genes containing anticodons modifiable by

(B) Correlation between human $tRNA^{Arg}$ isoacceptor abundance determined using tRNA microarrays and codon usage of ribosomal proteins (shown as RSCU), both for HeLa and HEK293T cell lines. The lack of correlation between these two parameters in the left plot is corrected in the right plot by the inclusion of the activity of ADATs.

See also Figure S4 and Tables S3–S6.

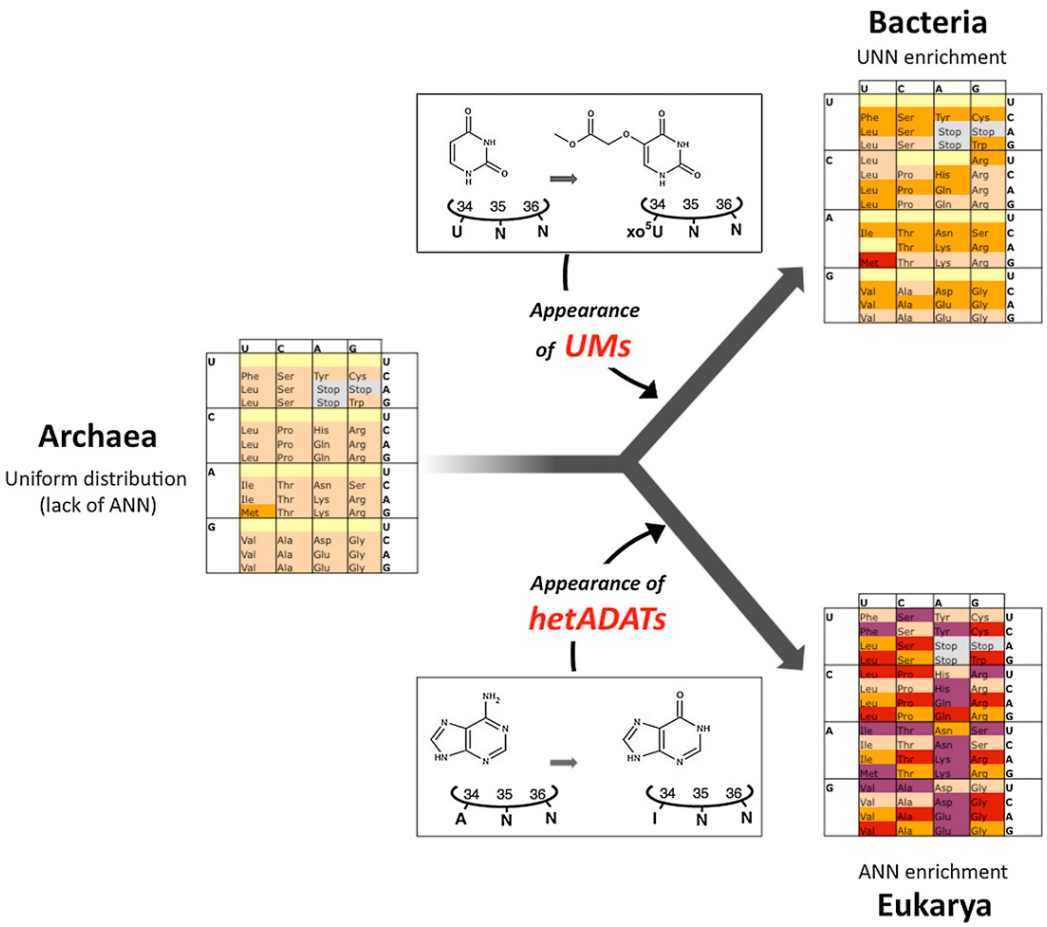


Figure 6. Model for the Role of Modification Enzymes in the Evolution of Genome Compositions

The emergence of the two tRNA modification enzymes (heterodimeric ADATs and UMs) was the main factor causing the divergence of decoding strategies between kingdoms. Archaea represents the most ancestral decoding strategy, where all isoacceptors are equally represented (and ANN anticodons are missing). ANN anticodons became overrepresented in eukaryotes due to the emergence of heterodimeric ADATs. Similarly, UNN anticodons became overrepresented in bacteria due to the appearance of UMs. Modification of the wobble position increased the decoding capacity of tRNAs, and consequently, translation efficiency. Thus, modifiable tRNAs were positively selected, causing a bias in tRNA gene content distribution which, in turn, caused the codon usage bias characteristic of the three main kingdoms.

kingdom-specific enzymes. This selection is likely driven by the improved decoding capacity that these modifications instill upon the modified tRNAs. A different solution to maximize tRNA decoding capacity was applied by Bacteria and Eukarya, thus contributing to the extant differences in tRNA pools and genome compositions.

Archaea would be the most ancestral kingdom in terms of decoding complexity (Figure 6). In Archaea neither ANN anticodons (Marck and Grosjean, 2002) nor ADATs are found (Mian et al., 1998). Therefore, the emergence of ADATs might be responsible for the appearance and selection of ANN-containing tRNAs that increased translation efficiency. In a similar fashion, the emergence of bacterial UMs would have driven the enrichment of tRNA genes with UNN anticodons in these organisms.

Several groups have demonstrated that preferred codon frequencies in highly expressed genes correlate with tRNA abundances within the cell (Withers et al., 2006; Tuller et al., 2010a). However, whether codon usage bias is caused by mutational bias or by natural selection has been a matter of controversy (Yang and Nielsen, 2008; Duret, 2002). In fast-growing organisms such as *E. coli* or *S. cerevisiae*, codon usage is generally thought to be under selective pressure (Sharp et al., 2005, 2010; Dong et al., 1996). On the other hand, in slowly growing organisms such as vertebrates, the existence of this selective pressure is controversial.

We have shown that the inclusion of modification data caused by ADATs and UMs in the definition of tRNA populations improves the codon usage-tRNA gene content correlation in Bacteria and Eukarya. Likely, the emergence of UMs and hetADATs in Bacteria and Eukarya allowed for the selection of new tRNAs that improved translation efficiency, and thus contributed to the evolution of genomic codon composition and tRNA gene content differences. Using published experimental data, we have shown that codons recognized by UM- and

hetADAT-modifiable anticodons are significantly enriched in highly expressed genes. Conversely, lowly expressed genes are enriched in codons recognized by nonmodifiable anticodons.

We have also shown that $tRNA^{Arg}$ populations in human cells do correlate well with genomic codon composition provided that anticodon modifications caused by hetADATs are considered in the definition of the different $tRNA^{Arg}$ isoacceptor concentrations. Thus, as previous studies have proposed for limited sets of species (Supek et al., 2010; Hershberg and Petrov, 2009; Drummond and Wilke, 2008), we conclude that translation efficiency influences tRNA gene populations in all kingdoms of life.

Several studies claim that the most significant parameter explaining codon bias differences among organisms is the level of GC content (Chen et al., 2004; Knight et al., 2001). Nevertheless, this observation does not explain codon bias variations within genomes, nor its correlation with gene expression levels. Anticodon modification strategies designed to improve translational efficiency could have evolved in parallel to the establishment of species-specific GC contents to ensure that tRNA gene populations were adapted to optimize translation. It should be noted that the triplet decoding strategies used by individual organisms have been determined (Marck and Grosjean, 2002; Grosjean et al., 2010). Each decoding strategy defines the minimum set of tRNAs needed to read all codons, and ranges from 25 up to 46 tRNAs. Interestingly, the defined minimal sets of eukaryotic and bacterial tRNAs conserve tRNA(ANN) and tRNA(UNN) isoacceptors respectively.

To summarize, Bacteria and Eukarya used two different tRNA modifications to increase the translational efficiency of their respective genomes. This phenomenon, in turn, contributed to the extant differences in tRNA gene populations and codon compositions of the main kingdoms of life. The discovery of kingdom-specific strategies to optimize translation efficiency opens new possibilities to further improve heterologous gene expression systems. Indeed, heterologous protein expression may be further improved if gene compositions are designed to match the mature tRNA gene population of the host species. In this regard, recent studies have started to analyze the potential of codon selection to tune translation efficiency (Cannarozzi et al., 2010; Tuller et al., 2010b) or protein folding (Zhang et al., 2009).

EXPERIMENTAL PROCEDURES

tRNA Sequence Retrieval

We have extracted, analyzed and compared over 53,000 sequences corresponding to cytoplasmatic nonorganellar tRNAs from 527 genomes distributed throughout the three kingdoms of life. All tRNA sequences have been downloaded from the GtRNAdb (http://gtrnadb.ucsc.edu), which uses the predictions made by the program tRNAscan-SE (Lowe and Eddy, 1997). Given that our analysis is based on average tRNA abundances, minor misannotations that may happen in tRNA genes using this prediction program are not statistically significant and thus should not affect the final results of this work.

Gene Content Analysis

Using the complete set of tRNA sequences we have built a distance-based phylogeny constructed on the basis of gene content. The similarity between two species is determined by the number resulting from dividing the number of tRNA genes that they have in common by the total number of gene types (isoacceptors). Using this method we have calculated a distance matrix that contains all pairwise distance values between the species analyzed. The distance matrix obtained has been used to cluster the sequences and build the phylogenetic tree, using the neighbor-joining method implemented in the program PHYLIP (Felsenstein, 1989). The program iTOL (Letunic and Bork, 2007) has been used for the visualization of the resulting phylogenetic tree.

Principal Component Analysis

A matrix consisting of the tRNA relative gene frequencies (RGF) for each anticodon and for all the analyzed species was used as input to perform PCA analysis (Jolliffe, 2002) using the program R (Team RDC, 2008, R: A Language and Environment for Statistical Computing, Vienna Austria R Foundation for Statistical Computing). The same software was used to obtain the resulting plots and to perform the t test and Wilcoxon test on the results. The significance of the association of the loadings with each principal component was computed using the FactoMineR package for R (Lê et al., 2008).

Retrieval of Coding Sequences and Codon Usage Estimation

All complete protein-coding sequences (CDS) for each of the selected 107 species were downloaded from the EMBLCDS database (http://www.ebi.ac.uk/embl/cds). For each species, a subset corresponding to ribosomal proteins was selected and visually inspected, and finally used as input to estimate the codon usage of highly expressed proteins using the GCUA software (McInerney, 1998).

Correlation between Codon Usage and tRNA Gene Content

For each species analyzed, the set of 18 preferred codons and preferred tRNA isoacceptors was computed (one for each amino acid, excluding Met and Trp). Initial correlations were computed by using the Watson-Crick base pairing rules (U:A; A:U; C:G; G:C), and extended correlations were computed including the extended wobble base pairing that result from the activities of ADATs (I:A; I:C; I:U) and UMs (xo^5U:A; xo^5U:G; xo^5U:U).

Correlation coefficients were computed as: $C = (\Sigma M / N) * 100$, where M is the number of codon-anticodon pairs for which there is a match (using Watson-Crick or extended wobble base pairing rules), and N is the number of codon-anticodon pairs considered in the analysis. We considered three different sets of matching codon-anticodon pairs. The simplest set (N = 8) includes the major tRNA isoacceptors with modifiable anticodons. A second set (N = 18) includes all major tRNA isoacceptors with the exception of methionine and tryptophan. Finally, a larger set (N = 27) was built by also considering the second most abundant tRNA isoacceptor from all four-, six-, and split (Ile) codon families.

The inclusion of modification data in our correlation analysis increases the number of acceptable codon-anticodon pairs, which could artificially increase correlation coefficients. To discard the possibility that the correlations that we obtain are simply the result of the increased number of acceptable pairs we tested the statistical significance of our data in both scenarios, i.e., with and without the inclusion of modification data. To that end, we approximated our data to a binomial distribution, computing for each set of data the expected distribution of random matches (Table S4). Our results show that the significance of our data is not due to the increased number of acceptable pairs caused by the inclusion of modification data (Tables S5 and S6). Using the same approach we confirmed that the statistical significance of our results is independent of the subset of tRNA isoacceptors analyzed.

tRNA Microarrays

tRNA abundance from HeLa and HEK293T cells was measured using a tRNA specific microarray method described previously (Dittmar et al., 2006; Pavon-Eternod et al., 2010). The standard tRNA microarray experiment consists of four steps starting from total RNA: (1) deacylation to remove remaining amino acids attached to the tRNA, (2) selective Cy3/Cy5 labeling of tRNA, (3) array hybridization, and (4) data analysis. The relative Cy3 or Cy5 fluorescent values from each tRNA probe of the same sample are used to determine the relative abundance of each tRNA in this sample, as described previously (Pavon-Eternod et al., 2010; Tuller et al., 2010a).

Protein Abundance and mRNA Levels

Protein abundance values and mRNA measurements of *E. coli* were taken from the work of Lu et al. (2007) and Ishihama et al. (2008); protein abundance values and mRNA levels of *S. cerevisiae* were taken from the work of Lu et al. (2007) and Newman et al. (2006). Correlation between protein expression levels and the abundance of preferred codons is shown in Figure 5 and Figures S4 and S5, and has been quantified using the Spearman's rank correlation coefficient.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at doi:10.1016/j.cell.2012.01.050.

ACKNOWLEDGMENTS

We thank Dr. M. Santos and Dr. V. de Crécy-Lagard for their critical analysis of the manuscript. We also thank E. Planet and D. Rossell for their help with the statistical analysis of the data. This work has been supported by grant BIO2009-09776 from the Spanish Ministry of Education and Science, and by grant MEPHITIS-223024 from the European Union. E.M.N. is supported by a La Caixa/IRB International Ph.D. Programme Fellowship. M.P.-E. was supported by a Ruth Kirshstein Pre-doctoral Fellowship from the NIH (1F31CA139968).

Received: September 20, 2011
Revised: November 23, 2011
Accepted: January 12, 2012
Published: March 29, 2012

REFERENCES

Agris, P.F., Vendeix, F.A., and Graham, W.D. (2007). tRNA's wobble decoding of the genome: 40 years of modification. J. Mol. Biol. *366*, 1–13.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics *136*, 927–935.

Akashi, H. (2003). Translational selection and yeast proteome evolution. Genetics *164*, 1291–1303.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. J. Biol. Chem. *257*, 3026–3031.

Brochier, C., and Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature *417*, 244.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. Cell *141*, 355–367.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA *101*, 3480–3485.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep. *6*, 151–157.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet. *2*, e221.

Dong, H., Nilsson, L., and Kurland, C.G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. *260*, 649–663.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. *16*, 287–289.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. *12*, 640–649.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA *96*, 4482–4487.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. Science *300*, 1718–1722.

Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164–166.

Fitz-Gibbon, S.T., and House, C.H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. *27*, 4218–4222.

Gerber, A.P., and Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. Science *286*, 1146–1149.

Gerber, A.P., and Keller, W. (2001). RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem. Sci. *26*, 376–384.

Geslain, R., and Pan, T. (2010). Functional analysis of human tRNA isodecoders. J. Mol. Biol. *396*, 821–831.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gonos, E.S., and Goddard, J.P. (1990). Human tRNAGlu genes: their copy number and organisation. FEBS Lett. *276*, 138–142.

Grosjean, H., de Crecy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett. *584*, 252–264.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu. Rev. Genet. *42*, 287–299.

Hershberg, R., and Petrov, D.A. (2009). General rules for optimal codon choice. PLoS Genet. *5*, e1000556.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. *146*, 1–21.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics *9*, 102.

Iwasaki, W., and Takagi, T. (2007). Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics *23*, i230–i239.

Jolliffe, I.T. (2002). Principal Component Analysis (New York: Springer Series in Statistics).

Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene *238*, 143–155.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., and Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. *53*, 290–298.

Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. *2*, RESEARCH0010.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science *324*, 255–258.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. J. Stat. Softw. *25*, 1–18.

Ledoux, S., Olejniczak, M., and Uhlenbeck, O.C. (2009). A sequence element that tunes Escherichia coli tRNA(Ala)(GGC) to ensure accurate decoding. Nat. Struct. Mol. Biol. *16*, 359–364.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. *25*, 117–124.

Maas, S., and Rich, A. (2000). Changing genetic information through RNA editing. Bioessays *22*, 790–802.

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat. Genet. *39*, 415–421.

Marck, C., and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA *8*, 1189–1232.

Mata, J., Marguerat, S., and Bähler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. Trends Biochem. Sci. *30*, 506–514.

McInerney, J.O. (1998). GCUA: general codon usage analysis. Bioinformatics *14*, 372–373.

Mian, I.S., Moser, M.J., Holley, W.R., and Chatterjee, A. (1998). Statistical modelling and phylogenetic analysis of a deaminase domain. J. Comput. Biol. *5*, 57–72.

Moriyama, E.N., and Powell, J.R. (1998). Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. *26*, 3188–3193.

Navon, S., and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biol. *12*, R12.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Näsvall, S.J., Chen, P., and Bjork, G.R. (2004). The modified wobble nucleoside uridine-5-oxyacetic acid in tRNAPro(cmo5UGG) promotes reading of all four proline codons in vivo. RNA *10*, 1662–1673.

Pavon-Eternod, M., Wei, M., Pan, T., and Kleiman, L. (2010). Profiling non-lysyl tRNAs in HIV-1. RNA *16*, 267–273.

Ran, W., and Higgs, P.G. (2010). The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol. Biol. Evol. *27*, 2129–2140.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., and Wright, F. (1988). Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. *16*, 8207–8211.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., and Sockett, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. *33*, 1141–1153.

Sharp, P.M., Emery, L.R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 1203–1212.

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. Nat. Genet. *21*, 108–110.

Stetter, K.O. (1996). Hyperthermophiles in the history of life. Ciba Found. Symp. *202*, 1–10, discussion 11–18.

Supek, F., and Smuc, T. (2010). On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. Genetics *185*, 1129–1134.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., and Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. PLoS Genet. *6*, e1001004.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Tekaia, F., Lazcano, A., and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res. *9*, 550–557.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput. Biol. *3*, e248.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. USA *107*, 3645–3650.

Withers, M., Wernisch, L., and dos Reis, M. (2006). Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. RNA *12*, 933–942.

Wolf, J., Gerber, A.P., and Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. EMBO J. *21*, 3841–3851.

Yang, S., Doolittle, R.F., and Bourne, P.E. (2005). Phylogeny determined by protein domain content. Proc. Natl. Acad. Sci. USA *102*, 373–378.

Yang, Z., and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. *25*, 568–579.

Yokoyama, S., Watanabe, T., Murao, K., Ishikura, H., Yamaizumi, Z., Nishimura, S., and Miyazawa, T. (1985). Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc. Natl. Acad. Sci. USA *82*, 4905–4909.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. *16*, 274–280.

Cell

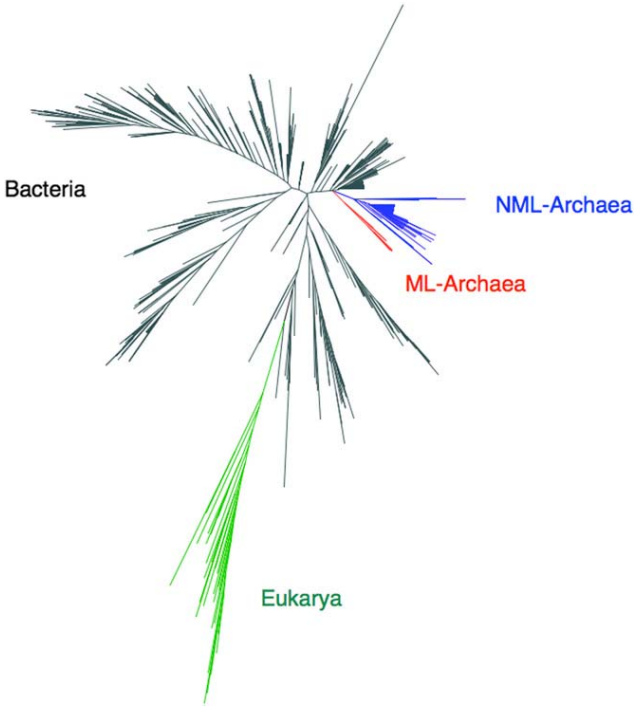


Figure S1. Genome Phylogeny of the 527 Species Based on tRNA Gene Content, Related to Figure 1
Each identified phylogenetic cluster has been labeled accordingly, and is shown in green (Eukarya), black (Bacteria), red (ML-Archaea) and blue (NML-Archaea).

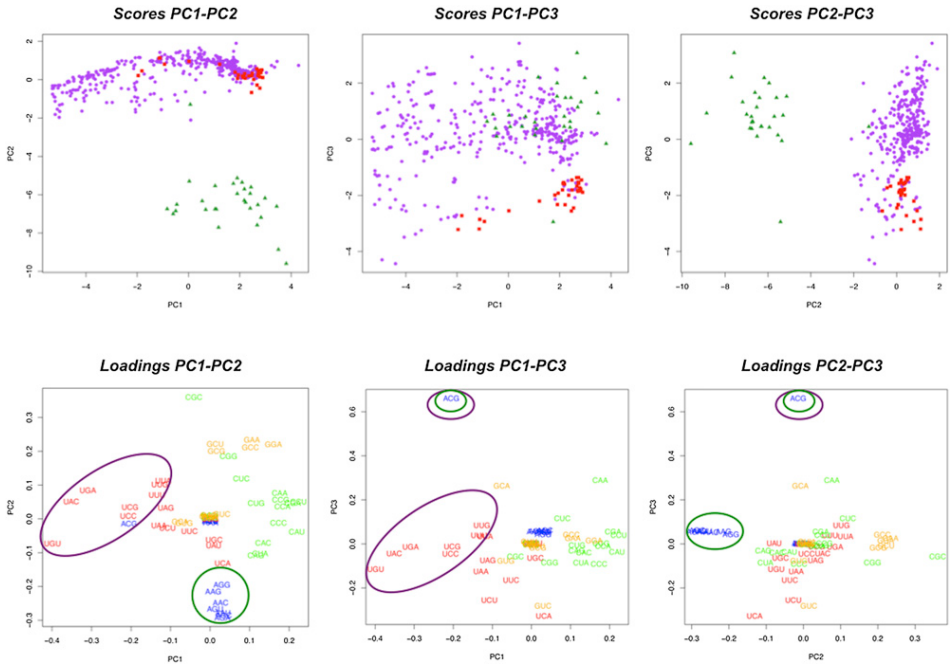


Figure S2. Identification and Quantification of Overrepresented tRNA Isoacceptors, Related to Figure 3

(A) Biplot of the scores after performing Principal Component Analysis. The left plot shows the PC1 versus PC2 loadings, the middle plot shows PC1 versus PC3, and the right plot shows PC2 versus PC3. The species have been colored according to their kingdom: Archaea (red), Bacteria (purple) and Eukarya (green). (B) Biplot of the loadings, indicating the tRNA isoacceptors whose frequencies contribute most to each of the clusters. The tRNA isoacceptors that are significantly associated to the PCs are circled, and colored according to the kingdom in which they are enriched: Eukarya (green) and Bacteria (purple).

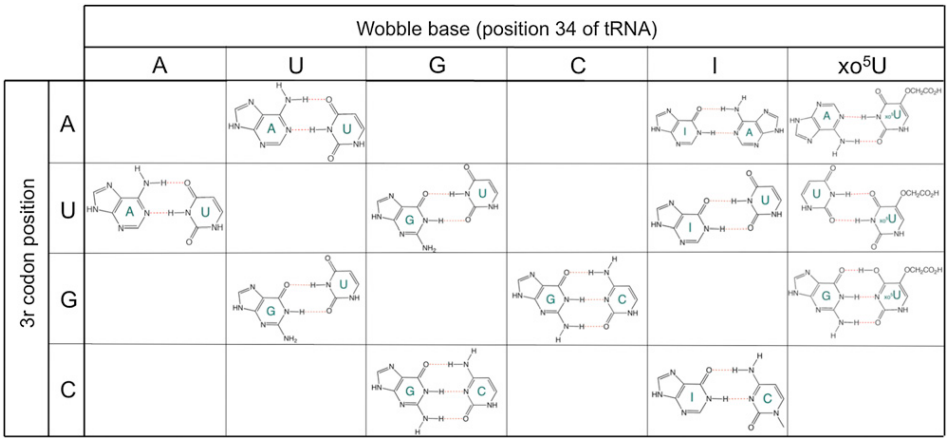


Figure S3. Codon-Anticodon Recognition at the Wobble Position, Related to Table 1

Representation of all possible codon:anticodon pairings according to the extended wobble base pairing rules. The decoding capacity of both xo5U and I is increased in comparison to the other bases that can found at the wobble position of the anticodon.

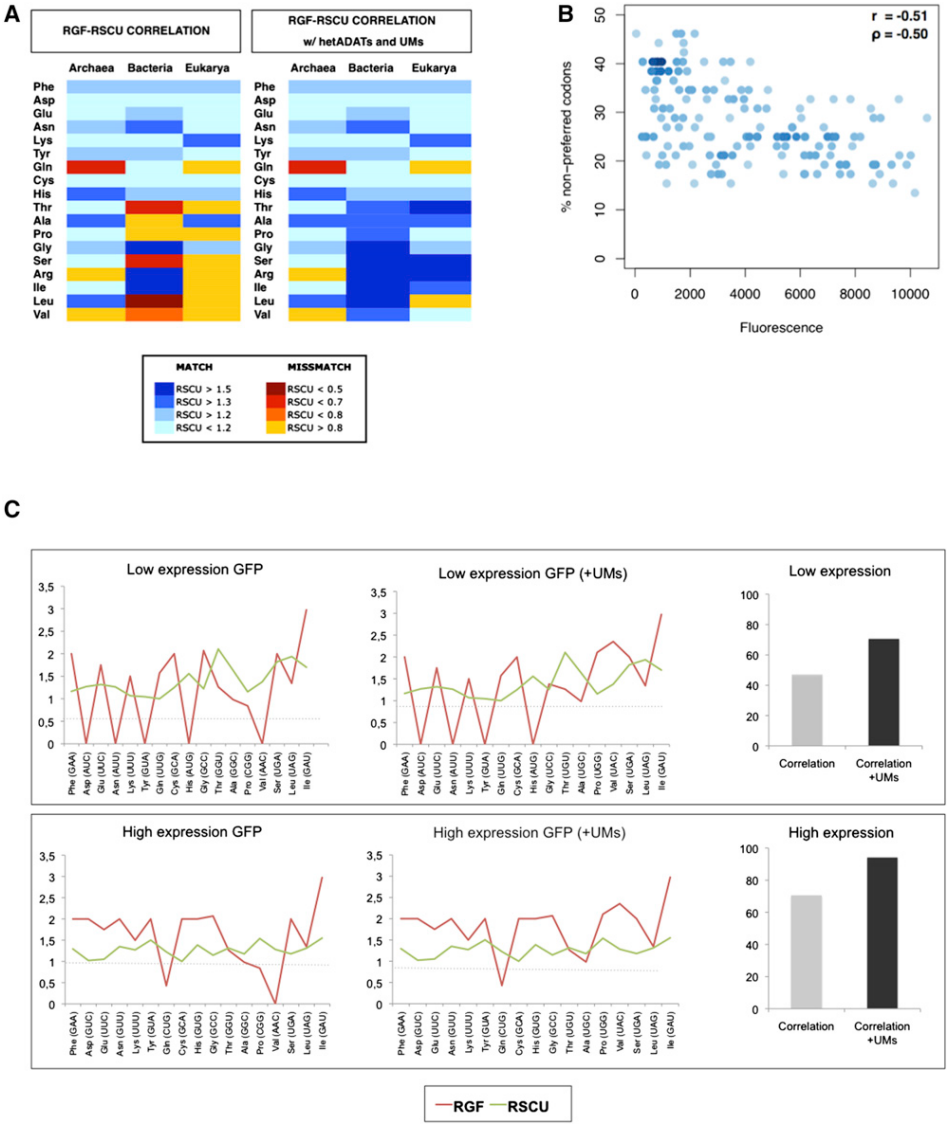


Figure S4. Correlation between Codon Usage and tRNA Gene Copy Number, Related to Figure 4

(A) For each amino acid and each kingdom, the *highest adapted codon*, i.e., that corresponding to the highest tRNA gene copy number, has been computed and compared to the codon with highest RSCU. Correlation results have been presented as heat map, either as match (blue) or mismatch (orange-red). Interestingly, the majority of nonmatching codons are precisely those susceptible of being recognized by modified tRNAs. When taking into account the wobbling (right) facilitated by uridine methyltransferases (UMs) or adenosine deaminases (ADATs) we can observe that most nonmatching codons are now matching.

(B) Correlation between the relative abundance of "nonpreferred" codons and GFP fluorescence values (Pearson's r: −0.51, p-value = 2.5e-11; Spearman's rho: −0.50, p-value = 6.2e-11). The local density of data points in the graph is signified by color (darker corresponding to more populated areas of the plot).

(C) Correlation between the most adapted codon (highest RGF) and the most abundant codon (highest RSCU) for low and high expressed GFP sequences. Low expressed and high expressed GFP sequences were chosen as those having the top or bottom 5% fluorescence among the 154 GFP set. In the left, the direct correlation between most adapted codon and most frequent codon is shown; in the middle graph the increase in correlation due to the incorporation of UMs is shown. The histograms represent the percentage of match between RGF and RSCU with and without the inclusion of the activity of UMs in the calculation of RGF values. Arginine codons have been excluded from the analysis because all synthetic GFP sequences present extremely high RSCU values for $tRNA^{Arg}$(AGA) (ranging from 1.71 to 4.29), and therefore cannot be used to measure codon preferences or correlations. Similar results were obtained when using larger datasets (i.e., the top and bottom 10% fluorescence values in the GFP set).

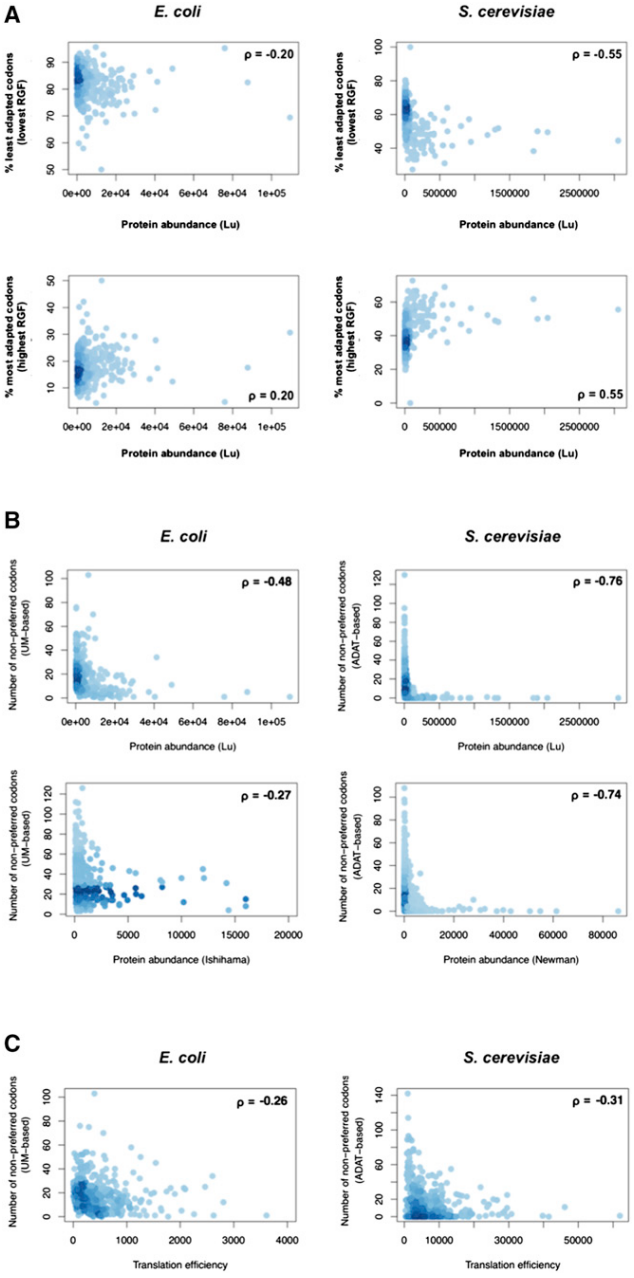


Figure S5. Correlations between Codon Usage and Protein Expression Levels, Related to Figure 5

(A) Correlations between the most adapted codons (highest RGF) and protein expression levels without considering modifications are shown both for *E. coli* and *S. cerevisiae* (Spearman's rho: 0.20 and 0.55, respectively). Similarly, the correlations between the least adapted codons (all those not included before) and protein expression levels are shown. Only amino acids with modifiable tRNA isoacceptors have been considered in this data to make them comparable with those results shown in Figure 5. The significance of the differences between Figure 5 (with modifications) and this figure is p = 9.4e-5 for *E. coli*, and p = 6.3e-4 for *S. cerevisiae* data, respectively.

(B) Correlation between the number of nonpreferred codons and protein abundance using diverse sources of experimentally determined protein abundances. For *E. coli*, the Spearman correlations are −0.48 (p-value = 8.5e-24) and −0.27 (p value = 6.5e-10) using experimental data from Lu et al. (2007) and Ishihama et al. (2008), respectively. For *S. cerevisiae*, the Spearman correlations are −0.76 (p value = 4.0e-68) and −0.74 (p-value = 1.2e-69) using data obtained by Lu et al. (2005) and Newman et al. (2006), respectively.

(C) Correlation between "nonpreferred" codons and translation efficiency. The number of non-UM-preferred and non-ADAT-preferred codons has been computed for each of the genes whose mRNA and protein abundance data was available (Lu et al., 2007), both for *E. coli* and *S. cerevisiae*, respectively. Translation efficiency has been defined as the ratio between mRNA and protein abundance levels.

SUPPLEMENTAL TABLES

Table S1. List of Anticodons that Are Significantly Associated to Each Kingdom, Related to Figure 3

BACTERIA				EUKARYA			
Anticodon	Correlation	P-value	Amino acid	Anticodon	Correlation	P-value	Amino acid
UGG	0 .769	9 .47E-105	Pro	**AGC**	0 .784	2 .74E-105	Ala
UAC	0 .765	6 .41E-103	Val	**AAU**	0 .782	2 .28E-104	Ile
UGC	0 .745	5 .06E-95	Ala	**AGU**	0 .771	1 .58E-99	Thr
UUG	0 .743	5 .09E-94	Gln	**AAC**	0 .764	2 .07E-96	Val
UUC	0 .742	1 .05E-93	Glu	**AGA**	0 .757	1 .01E-93	Ser
UGA	0 .740	4 .36E-93	Ser	**AGG**	0 .708	7 .49E-76	Pro
UGU	0 .723	5 .50E-87	Thr	**AAG**	0 .692	8 .65E-71	Leu
UUU	0 .709	5 .53E-82	Lys				
UAG	0 .619	2 .08E-57	Leu				

* In bold are shown those anticodons susceptible to be modified by UMs and ADATs.

Table S2. tRNA Modifications at Base 34, Related to Figure 3

Symbol	Common name	Distribution	3rd codon read	Usage
Gm	2'-O-methylguanosine	B, E	C	2-codon sets
I	**inosine**	**B, E**	**U, C, A**	**Family boxes**
Q	queuosine	B, E, M	U, C	2-codon sets
Y	pseudouridine	E	A	Ile AUA
cmnm5U	5-carboxymethylaminomethyluridine	B, M	A, G	2-codon sets
cmnm5Um	5-carboxymethylaminomethyl-2-O-methyluridine	B, E	A, G	2-codon sets
cmnm5s2U	5-carboxymethylaminomethyl-2-thiouridine	B, M	A, G	2-codon sets
cmo5U	**5-methoxyuridine**	**B**	**A, G, U**	**Family boxes**
k2C	lysidine	B, M	A	Ile AUA
mcm5s2U	5-methoxycarbonylmethyl-2-thiouridine	E	A, G	2-codon sets
mnm5s2U	5-methylaminomethyl-2-thiouridine	B, E	A, G	2-codon sets
mo5U	**uridine 5-oxyacetic acid**	**B**	**A, G, U**	**Family boxes**
m7G	2-O-methylguanosine	M	A, G, U, C	Family boxes
s2U	2-thiouridine	E	A, G	2-codon sets
f5C	5-formylcytidine	M	G	Met AUR
f5Cm	5-formyl-2-O-methylcytidine	E	A, G	Leu UUR
ac4C	N-acethylcytidine	A, B	G	Met AUG
Cm	2-O-methylcytidine	A, B, E	G	Met AUG, Trp UGG
m5C	5-methylcytidine	E	G	Leu UUG

* In bold are shown those tRNA modifications whose phylogenetic distributions and sets of tRNA substrates match with the families of tRNAs found to be enriched.

Table S3. Correlation Coefficients between RGF and RSCU Values, Related to Figure 3

		Archaea	Bacteria	Eukarya	Global
1 isoacceptor	Correlation	83 .3	66 .7	55 .6	68 .5
	Correlation including tRNA modifications	83 .3	100	88 .9	90 .7
2 isoacceptors	Correlation	75 .1	59 .3	40 .7	60 .5
	Correlation including tRNA modifications	75 .1	100	88 .9	88 .0

Table S4. Table of Random Probabilities of Match for Each Amino Acid, Used to Assess the Statistical Significance of the Correlation Coefficients between Codon Usage and tRNA Gene Content, Related to Figure 4

	without modification enzymes	with modification enzymes
2box codon families	0 .5	0 .5
4box codon families (except Gly)	0 .25	0 .4375
Gly	0 .25	0 .25
6box codon families	0 .17	0 .25
Split codon families (Ile)	0 .33	0 .56

Table S5. Correlation Coefficients between RGF and RSCU Values and Statistical Significances, Related to Figure 4

Only amino acids with modifiable isoacceptors have been considered in these data.

		Archaea		Bacteria		Eukarya	
		w/o modifying enzymes	with modifying enzymes	w/o modifying enzymes	with modifying enzymes	w/o modifying enzymes	with modifying enzymes
Nº isoacceptors							
1	Probability of random match	0 .229	0 .229	0 .229	0 .323	0 .229	0 .382
	Correlation	0 .75	0 .75	0 .375	1	0 .25	0 .875
	p-value	2 .6e-3	2 .6e-3	0 .395	1 .2e-4	1	6 .3e-3
2	Probability of random match	0 .229	0 .229	0 .229	0 .323	0 .229	0 .382
	Correlation	0 .81	0 .81	0 .38	1	0 .56	0 .94
	p-value	1 .3e-6	1 .3e-6	0 .228	1 .4e-8	3 .9e-3	5 .5e-6

Table S6. Correlation Coefficients between RGF and RSCU Values and Statistical Significances, Related to Figure 4

All amino acids have been considered in these data.

		Archaea		Bacteria		Eukarya	
		w/o modifying enzymes	with modifying enzymes	w/o modifying enzymes	with modifying enzymes	w/o modifying enzymes	with modifying enzymes
Nº isoacceptors							
1	Probability of random match	0.365	0.365	0.365	0.416	0.365	0.433
	Correlation	0.833	0.833	0.667	0.94	0.556	0.889
	p-value	5.6e-5	5.6e-5	6.8e-3	1.4e-7	0.048	7.5e-5
2	Probability of random match	0.321	0.321	0.321	0,388	0.321	0.412
	Correlation	0.751	0.751	0.593	0.96	0.407	0.89
	p-value	7.9e-6	7.9e-6	2.3e-3	3.9e-8	0.154	3.6e-5

PUBLICATION 2:

Speeding with control: codon usage, tRNA and ribosomes.

Novoa EM and Ribas de Pouplana L.

Trends in Genetics 2012 (in press)

Review

Cell PRESS

Speeding with control: codon usage, tRNAs, and ribosomes

Eva Maria Novoa[1] **and Lluís Ribas de Pouplana**[1,2]

[1] Institute for Research in Biomedicine (IRB), c/Baldiri Reixac 15-21 08028, Barcelona, Catalonia, Spain
[2] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain

Codon usage and tRNA abundance are critical parameters for gene synthesis. However, the forces determining codon usage bias within genomes and between organisms, as well as the functional roles of biased codon compositions, remain poorly understood. Similarly, the composition and dynamics of mature tRNA populations in cells in terms of isoacceptor abundances, and the prevalence and function of base modifications are not well understood. As we begin to decipher some of the rules that govern codon usage and tRNA abundances, it is becoming clear that these parameters are a way to not only increase gene expression, but also regulate the speed of ribosomal translation, the efficiency of protein folding, and the coordinated expression of functionally related gene families. Here, we discuss the importance of codon–anticodon interactions in translation regulation and highlight the contribution of non-random codon distributions and post-transcriptional base modifications to this regulation.

Codon usage bias

What is codon usage bias?

Due to the degeneracy of the genetic code, several codons ('synonymous' codons; see Glossary) are translated into the same amino acid. Synonymous codons are used with different frequencies, a phenomenon known as codon bias. Codon bias is a defining characteristic of each genome and is maintained by a balance between selection, mutation, and genetic drift [1–3]. Despite the relative universality of the genetic code and the conservation of the translation machinery across species, codon biases vary dramatically between organisms. Thus, the most frequent or most rare codon in a gene varies both between and within species depending on the gene [1,4].

It is generally accepted that the speed at which ribosomes decode a codon depends on the cellular concentration of the tRNA that recognize it [5–8], although there is some debate about this assumption [9]. Nevertheless, the most abundant codons pair with the most abundant tRNAs and vice versa. As a result, gene codon bias strongly correlates with gene expression levels in organisms as diverse as *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Drosophila melanogaster* [10–15]. It has been shown that the use of particular codons can increase the expression of a gene by more than 1000-fold [16].

Corresponding author: Ribas de Pouplana, L. (lluis.ribas@irbbarcelona.org).
Keywords: codon usage; tRNA; translation efficiency; tRNA modifications.

Why does codon usage bias exist?

The existence of codon bias is explained by two different lines of thought [1]. According to 'selectionists', codon bias contributes to the efficiency and accuracy of amino acid sequence, and this bias is maintained by selection [2,17]. By contrast, 'mutationalists' suggest that codon bias exists because of non-randomness in the mutational patterns, whereby some codons would be more mutable and, therefore, would have lower equilibrium frequencies [18,19]. According to this latter theory, genomic G+C composition is thought to be a major factor affecting codon usage variation [20], given that G+C frequencies can range from <20% to >90% in the third position of codons. These two explanations are not mutually exclusive, and both are supported by several studies (*vide supra*).

A clear association exists between the expression level of a gene and its codon composition, an observation that holds for organisms ranging from bacteria to mammals. It is generally accepted that the variation of codon usage

Glossary

Anticodon: sequence of three nucleotides of a tRNA that is complementary to a given codon.
Codon: sequence of three nucleotides of an mRNA that specifies the amino acid that will be added next during protein synthesis.
Mistranslation: phenomenon that occurs when an amino acid is attached to the wrong tRNA and subsequently misincorporated into the nascent protein.
Preferred codons: subsets of rapidly translated codons that are expected to increase translation efficiency and, therefore, to be over-represented in highly expressed proteins.
Ribosome stalling: ribosome pausing, which is thought to happen for several reasons, including the presence of rare codons, which are decoded more slowly.
Shine-Dalgarno sequence: ribosomal binding site that exists in the mRNA of Bacteria and Archaea, generally located eight base pairs upstream of the start codon AUG.
Translation efficiency: rate of mRNA translation into proteins within cells.
tRNA channeling: direct transfer of tRNAs from the aminoacyl-tRNA synthetases to the elongation factor and ribosomes without dissociation. It also includes the transfer of tRNAs leaving the ribosome to their cognate aminoacyl-tRNA synthetases, which will regenerate newly charged tRNAs ready to use again in protein synthesis.
tRNA decoding capacity: ability of a tRNA to recognize more than one codon from a subset of codons that encode the same amino acid, including both Watson-Crick and wobble base pairings.
tRNA isoacceptors: tRNA molecules that bind to alternate codons encoding the same amino acid residue.
tRNA microarray: specific microarray method to quantify tRNAs based on a fluorescent dye-labeling technique.
tRNA modifications: nucleotide modifications that alter the biophysical and biochemical properties of a tRNA, causing changes in the structure and dynamics of the tRNA to fine-tune its function.
Wobble base-pair: non-Watson-Crick base-pairing between two nucleotides in RNA molecules, but the thermodynamic stability is comparable to that of Watson-Crick base pairs.
Wobble position: third position in the codon, or first position of the tRNA anticodon (base 34).

0168-9525/$ – see front matter http://dx.doi.org/10.1016/j.tig.2012.07.006 Trends in Genetics xx (2012) 1–8

between genes of the same genome is a product of selection, based on the observation that codon bias is more extreme in highly expressed genes, which are enriched in those codons that match the most abundant cognate tRNAs [15,18,21]. However, whether the codon bias found in highly expressed genes serves to optimize translational efficiency or improve codon reading accuracy has been a topic of active debate [18,19,22,23].

A priori, both translation efficiency and accuracy should be under positive selection. On the one hand, efficient elongation of a transcript might increase its protein yield [16] or provide a global benefit to the cell by freeing up ribosomes that can then translate other messages [24]. On the other hand, accurate elongation would benefit the cell by reducing the costs of useless mistranslation products [22].

Beyond this direct relation between codon composition and translation speed lies a more complex set of parameters that link codon usage and tRNA abundance to gene expression regulation. Factors such as codon autocorrelation [25], clustering of rare codons [26], mRNA secondary structure [24], ribosomal density [6], relative abundance of wobble base pairs [27], presence of Shine-Dalgarno-like features in coding sequences [9], or interactions with modified tRNAs [28] can further contribute to the regulation of gene expression through the phenomenon of synonymous codon bias and tRNA dynamics (Figure 1).

For example, codon usage bias has been linked to the control of cell cycle development [29] and stress-mediated specific responses [30]. Specific tRNAs and, consequently, certain codon compositions are a crucial component in the activation of some genetic programs [31], suggesting a novel layer of genomic regulation that is only now starting to be explored. Similarly, it has been recently shown that the emergence of certain anticodon modification enzymes during evolution has shaped the structure of genomes, contributing to the regulation of the speed of gene translation [28] (Figure 2). In this review, we discuss the latest advances leading to current understanding of how codon usage and tRNA populations evolved not only to optimize gene expression, but also to regulate it.

Codon usage and tRNA

Codon frequencies and tRNA abundances

tRNAs translate codons into amino acids during protein synthesis. Every organism has multiple tRNA species that read the codons for the same amino acid (tRNA isoacceptors). Several reports have shown that synonymous triplet variation across species is driven by the adaptation of codon usage to tRNA abundances or vice versa [15,16,28,32,33]. However, the search for a correlation between tRNA abundance and codon usage has been successful only in some organisms [5,34]. In several species, including many bacteria and eukaryotes, this search has failed [35,36], prompting the proposal that, in the latter organisms, translation efficiency might not be the primary factor influencing codon usage [36,37]. However, it was recently reported that two distinct modifications at the wobble position of certain anticodons are at the core of this apparent lack of correlation [28]. These modifications 'extend' the wobble pairing ability of anticodons and influence the codon usage bias in bacteria and eukaryotes, ultimately affecting codon usage and genomic tRNA compositions. The inclusion of these modifications corrects previously reported discrepancies between codon usage and tRNA abundance across all extant major phylogenetic groups, These results suggest not only that codon usage and tRNA abundances coevolve, but also that the diversification of the genetic code usage in evolution was at least partially driven by the appearance of certain tRNA modification enzymes [28].

Recently, *in vivo* translational speeds for all sense codons from *S. cerevisiae* were determined [38] using genome-wide ribosome profiling data. Surprisingly, similar translational speeds among synonymous codons were found, suggesting that preferentially used codons in highly expressed proteins are not translated faster than non-preferred ones. However, a correlation between codon usage bias and cognate tRNA abundances was indeed observed. These findings suggest that codon usage bias found in highly expressed genes is a product of natural selection for an overall cellular efficiency, rather than a product of stronger selection for translation efficiency in more highly expressed genes.

Variability in tRNA pools

tRNA gene copy number has often been used as a proxy for tRNA abundance in the cell [26,34,35,39]. This approximation has been validated for some unicellular organisms, such as yeast [6] and *E. coli* [5], but recent studies have demonstrated that tissue-specific differences in the expression of tRNA genes exist in more complex organisms [40]. Indeed, microarray-based quantification of cellular tRNAs shows significant variation in their levels among different tissues, both in terms of relative enrichments of specific tRNA isoacceptors and in total tRNA concentration [40]. Importantly, the correlation between relative tRNA abundances and the codon composition of highly expressed, tissue-specific genes was also observed in the different tissues analyzed.

Because mature tRNAs in humans are thought to be very stable, cellular tRNA levels are mostly determined by tRNA transcription rates [41]. tRNAs are transcribed by a multisubunit complex of RNA polymerase III (Pol III), TFIIIB, and TFIIIC [42,43] and are negatively regulated by Maf1, a protein under the control of the mammalian target of rapamycin (mTOR) pathway [44]. Thus, the regulation of tRNA transcript levels is closely linked to cellular conditions, such as nutrient availability and genome integrity.

To explore the evolutionary dynamics of tRNA gene transcription and the variation across different tissues in mammals, Pol III occupancy has been experimentally determined in several tissues from six mammalian species [45]. Pol III binding to different tRNA genes varies substantially in strength and genomic location for different species. However, there is a strong conservation of Pol III occupancy at the genes of grouped tRNA isoacceptor families [45]. These results suggest that, although the usage of individual tRNA genes has evolved rapidly, functional tRNA isoacceptor families have been maintained throughout evolution. This indicates that the major evolutionary

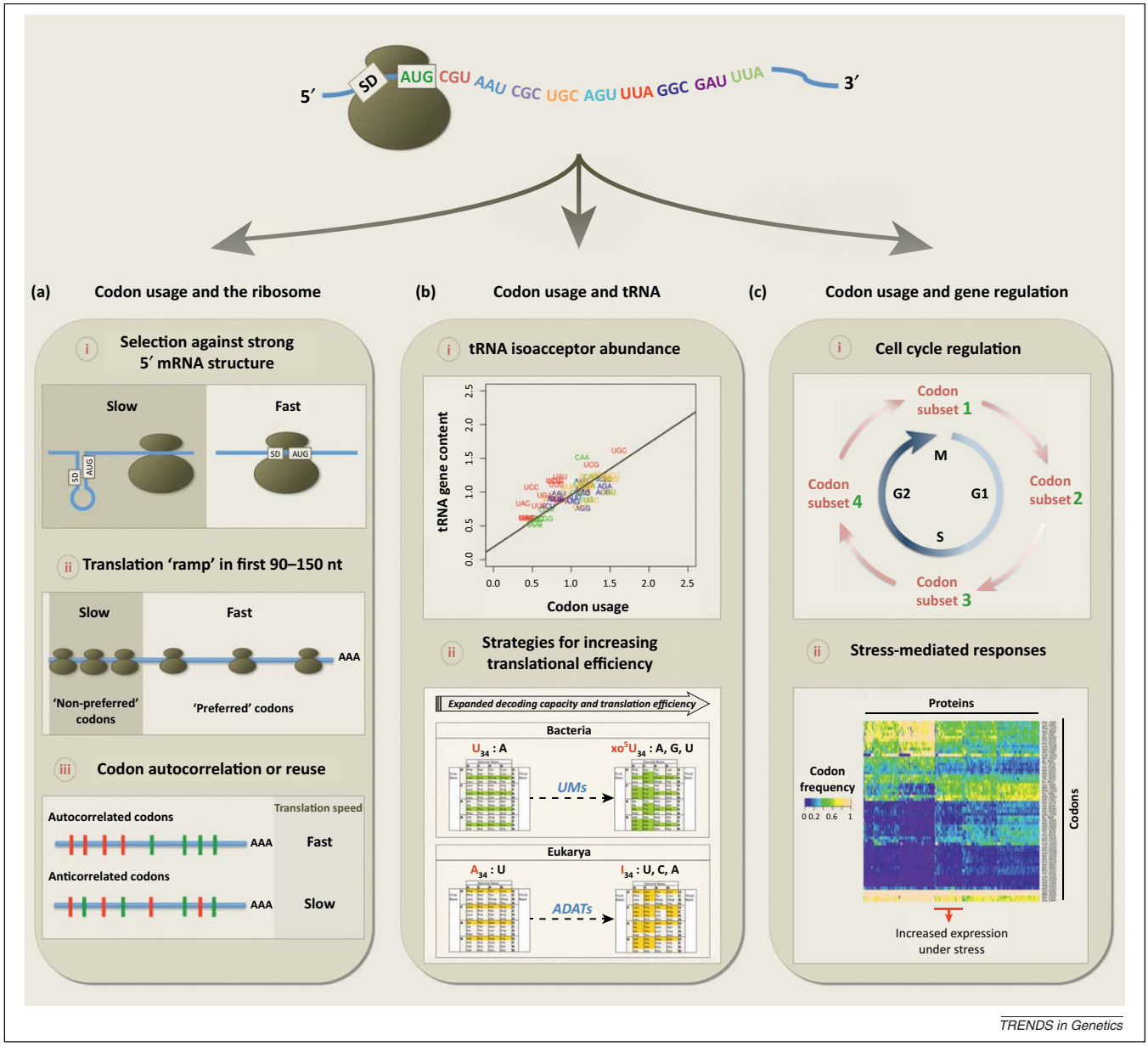


Figure 1. Relevant mechanisms involving the unequal use of synonymous codons and their effect on translation efficiency. **(a)** The distribution of synonymous codons along the gene affects the speed of the ribosome and, consequently, the translation efficiency. As general rules, mRNA transcripts lack strong 5′ secondary structure (i) [24], 'non-preferred' codons cluster at the beginning of the transcript (ii) [6], and autocorrelated codons, which allow tRNA recycling, increase the speed of translation (iii) [25]. **(b)** tRNA gene content tends to correlate with the codon usage bias of highly expressed genes (i) [34]. tRNA gene content biases appear to increase protein translation efficiency by increasing the number of tRNA isoacceptors that are capable of being modified by tRNA modification enzymes, which expands their wobbling capacity. These tRNA modification enzymes differ between bacterial [uridine methyltransferases (UMs)] and eukaryal [adenosine deaminases (ADATs)] species, which in turn have caused differential increases of specific tRNA isoacceptors between kingdoms (ii) [28]. **(c)** The sets of genes that are expressed in each stage of the cell cycle present similar codon covariations, and these differ from those found in other stages, suggesting that the codon preferences change during the cell cycle (i) [29]. Codon preferences may change due to the activity of specific tRNA modification enzymes [e.g., tRNA methyltransferase 9 (Trm9) in *Saccharomyces cerevisiae*]. Under stress conditions, Trm9 modifies a subset of tRNAs and, consequently, their decoding capacities and codon preferences, thus enhancing the expression of a subset of codons that is enriched in proteins that respond to stress (ii) [30].

forces driving relative tRNA abundance and codon composition are conserved across mammals.

Codon usage and the ribosome

Codon distribution and local enrichment

Traditionally, analyses of codon usage for individual genes have only considered the overall codon composition of transcripts. However, patterns of unequal codon distribution along genes exist, and these are thought to be important for the control of ribosome speed and translation stability [6,25]. Indeed, the notion that translation rates can change across different regions of an mRNA transcript has been known for some time [46,47] and has recently gained additional experimental support [48,49].

Recently, a study of the translation efficiency of codons as a function of their location on the transcript [6] reported that, for most genes, the speed of translation is reduced during the first 30–50 codons (known as the 'translation ramp') and then increases for the remainder of the gene. This ramp of poorly adapted codons (i.e., those read by low

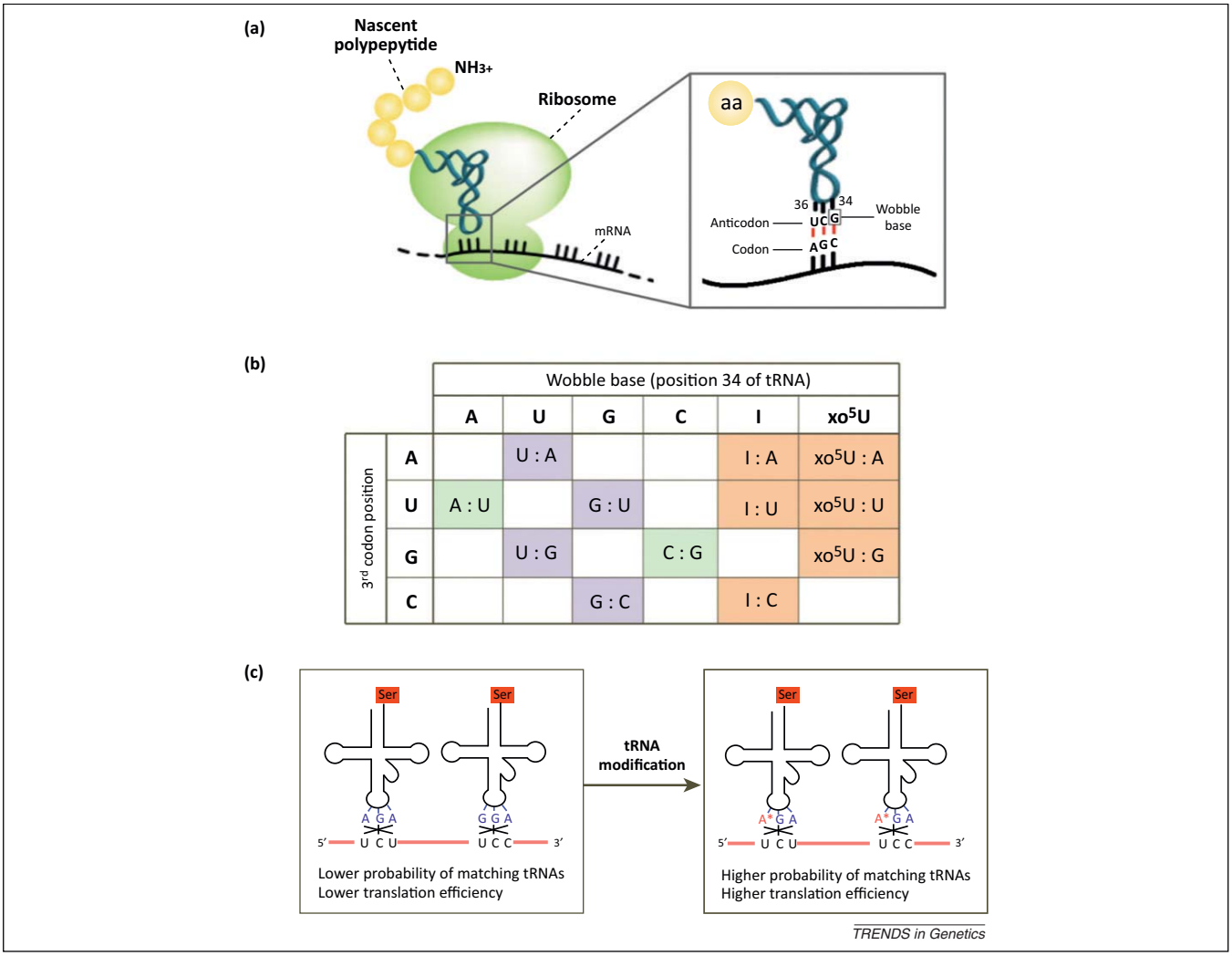


Figure 2. Increased decoding capacity of modified tRNAs at the wobble position. **(a)** The aminoacylated tRNA enters the ribosome and is selected based on the correct pairing of its anticodon bases (positions 34, 35, and 36 of the tRNA) with the respective codon. The wobble base (position 34) recognizes the third position of the mRNA codon, the degenerate codon position. **(b)** Representation of all possible codon–anticodon pairings according to the extended wobble base-pairing rules. A and C in position 34 can only recognize one base (shown in green), whereas G and U can recognize two different bases (shown in purple): one through Watson-Crick pairing and the other through the G:U wobbling. The activities of adenosine deaminases (ADATs; A-to-I conversion) and uridine methyltransferases (UMs; U-to-xo^{5}U conversion) expand the wobbling capacities of base 34, allowing them to pair with three different codon bases (shown in orange). **(c)** Proposed model for the effect of tRNA modification enzymes upon translation efficiency.

abundant tRNAs) would presumably slow elongation at the beginning of the gene, reducing the frequency of ribosomal stalling. This model is further supported by experimentally determined ribosome profile densities along mRNAs [50] and has now been proposed as a general feature of gene translation in both prokaryotic and eukaryotic species.

It has also been shown that once a particular synonymous codon has been used in a transcript, other codons recognized by the same tRNA isoacceptor will be favored in that gene [25]. This observation holds for both frequent and rare codons, and the observed enrichment diminishes as a function of the distance between subsequent synonymous codons. This indicates that sequences optimized for tRNA reuse are expressed more efficiently than are sequences that require different tRNA isoacceptors. In accordance with this model, previous studies have proposed that tRNA diffusion away from the ribosome is slower than translation, and that some tRNA channeling takes place to optimize ribosome function [51]. Specifically, it has been suggested that, after release from the ribosome, tRNAs remain bound to the multi tRNA–synthetase complex [52] or to elongation factors [53], which might themselves be associated with the ribosome. Such a mechanism would effectively raise the local concentration of tRNAs that recognize codons that have already appeared in given transcript. Thus, genes that reuse the same codons, even rare codons, may be more efficiently transcribed.

Codon composition and RNA secondary structure

At the gene level, it is well known that mRNA structure influences translational efficiency. In bacteria, the formation of strong hairpin loops around the Shine-Dalgarno ribosomal binding site and the initiation codon can significantly reduce expression levels [54]. Therefore, strong mRNA structure near the 5′ end of a transcript is generally

thought of as disadvantageous and can inhibit ribosomal translation initiation [55,56]. This is further supported by a study that looked at the expression levels of a synthetic library of GFPs with random synonymous changes in *E. coli* and found that upstream sequence composition influenced mRNA folding near the ribosomal binding site. In fact, it was estimated that the mRNA sequence composition in this region explained more than half of the variation that was found in protein levels [24].

Protein folding regulated by codon composition

Numerous reports indicate that the speed of translation along some transcripts may be critical to the formation of the native structure of a protein. Pausing has been identified during the translation of certain proteins [57–59] and, in many cases, it appears to be caused by local mRNA structures [60], which may be required for the correct folding of the nascent polypeptide [49,61]. These translation effects support the theory of co-translational protein folding and highlight the importance of mRNA sequence and codon usage bias in protein structure formation [62,63]. Indeed, synonymous mutations can have significant consequences in the folding process of the nascent protein and even change the substrate specificity of enzymes [64].

Codon usage and gene regulation

Codon usage and cell cycle control

It is becoming increasingly evident that the use of specific subsets of codons can be a strategy to optimize parameters other than protein synthesis efficiency. For example, previous works in bacteria and fungi demonstrated that functionally related genes that probably need to be expressed at similar levels tend to have similar patterns of codon bias [65,66]. In a recent study, it was shown that certain non-optimal codon compositions were related to cell cycle-dependent oscillations in protein levels [29]. Indeed, cell cycle-regulated genes display different codon preferences, suggesting that codon usage has a role in cell cycle regulation.

The same study also concluded that cell cycle-regulated genes have a strong preference for codons with low codon–anticodon binding affinity [29], based on published thermodynamic data for binding affinities of several possible base pairings [67]. If subsets of functionally related genes exhibit specific biases towards particular codons, then the regulation of the expression of these genes may also be linked to specific codon usage patterns. Thus, it appears that subsets of 'preferentially expressed' genes form coherent groups in terms of codon usage, and that these codon composition 'preferences' change throughout the life cycle of a cell. Similarly, these codon preferences might also be capable of responding to a variety of external stimuli, such as stress [30].

Proteome regulation through modulation of codon-anticodon pairings

As stated above, genes that need to be expressed at similar levels tend to have similar codon biases [65,66]. Importantly, anticodon bases can be customized by tRNA modification enzymes to alter their translation decoding capacity, potentially impacting the subset of 'preferred' codons in the genome. This potential variability in the sets of 'preferred' codons implies that modulating the activity of modification enzymes may be an avenue for regulating the composition of the proteome when needed (Figure 3).

The relation between codon frequency and tRNA abundance is further confounded by the existence of post-transcriptional modifications in tRNA nucleotides, over 100 of which have been recognized and described to date (http://rna-mdb.cas.albany.edu/RNAmods/). Modifications contribute to tRNA folding, structure, and stability, as well as to translation efficiency and amino acid substitution rates [68–70]. Many of the known base modifications are not essential for life and have often been characterized as a mere expansion of the repertoire of the nucleotide bases. Nevertheless, increasing evidence indicates that tRNA modifications can have regulatory roles in cells, especially in response to stress conditions [30,71].

The function of many tRNA modifications, particularly with regards to gene expression regulation, remains unclear. To approach this problem, novel mass spectrometric methods to quantify tRNA modifications with high precision are being used [72]. One such study exposed *S. cerevisiae* to various environmental stresses and analyzed the resulting changes in tRNA modification levels. Interestingly, the prevalence of several tRNA modifications changed as a function of the stress response being activated, suggesting

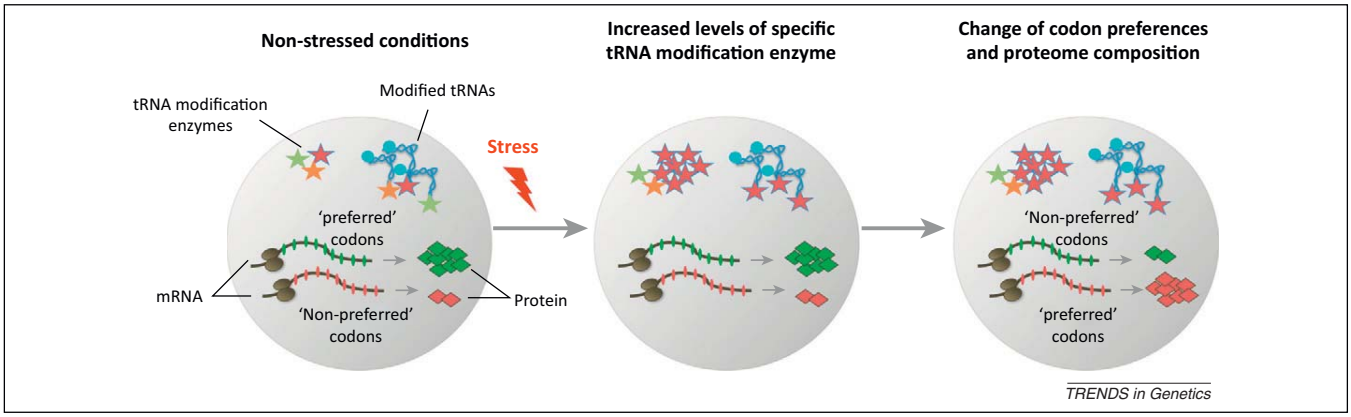


Figure 3. Codon usage bias as a mechanism for tuning the proteome. In response to a particular signal, such as an environmental stress, the levels of a given tRNA modification enzyme change, altering the codon preferences of the tRNA. These changes in turn cause an increase in the protein expression levels of those mRNAs that are found to be enriched in that specific subset of newly 'preferred' codons. Such a mechanism may operate on a set of proteins involved in the specific response to the signal or stress.

Table 1. Characterization of tRNA modifications in *Saccharomyces cerevisiae*[a]

Modification enzyme	tRNA modification	Target tRNAs[c]	tRNA position[c]	Changes in modifications in enzyme-lacking mutants[b]		Function[c]
				Decreased[d]	Increased	
Trm1	m_2^2G	Several	G26	m_2^2G	–	
Trm2	m^5U	All	U54	m^5U	–	Suggested role in tRNA stabilization and maturation
Trm3	Gm	Several	G18	–	–	
Trm4	m^5C	Several	C34, C40, C48, C49	m^5C	–	Suggested role in ribosome biogenesis
Trm5	m^1G/yW	Several	G37	yW	Y, Gm, Um, Am, m_2^2G	Required for yW modification
Trm7	Cm	Several	C32, G34	ncm^5U, yW	–	
Trm8	m^7G	Several	G46	m^7G, (yW)	–	Required to maintain tRNA stability
Trm9	mcm^5U/mcm^5s^2U	Arg(UCU), Glu(UUC)	U34	mcm^5U, mcm^5s^2U	–	Role in stress response; interacts with Trm112
Trm10	m^1G	Several	G9	m^1G	ncm^5U	
Trm11	m^2G	Several	G10	m^2G, (yW)	–	Interacts with Trm112
Trm12	yW	Phe	G14	yW	–	Not methyltransferase
Trm13	Cm	Gly, His, Pro	G4	(Cm)	–	
Trm44	Um	Ser	U44	(Um)	––	
Trm82	m^7G	Several	G46	m^7G, (yW)	m^1G, m^3C, t^6A, m_2^2G, m^2G, m^1I, mcm^5U, mcm^5s^2U	Required to maintain tRNA stability; complexes with Trm8
Tad1	I	Ala	A37	m^1I, yW, (D), (Y)	–	
Mod5	i^6A	Several	A37	i^6A, yW, (D), (Y)	–	
Tan1	ac^4C	Leu, Ser	C12	ac^4C, (yW)	m^1G, m^3C,m^1A, m^2G, m^1I, mcm^5U, m^7G	Acetyltransferase

[a]Abbreviations: A, adenosine; ac, acetyl; C, cytidine; D, dihydrouridine; G, guanosine; I, inosine; m, methyl; mcm, methoxycarbonylmethyl; s, thio; t, threonyl; U, uridine; Y, pseudouridine; yW, wybutosine.

[b]Data from [72].

[c]Data from public databases.

[d]Changes shown in parenthesis correspond to more subtle changes in modification levels compared to those shown without parenthesis.

that the control of tRNA modifications in cellular response pathways is a dynamic process.

Similarly, it has been shown that certain clusters of yeast mRNAs enriched in AGA codons are differentially translated under stress due to an increase in anticodon modifications mediated by tRNA methyltransferase 9 (Trm9) [30]. Therefore, similar transcriptomes may result in different proteome compositions as a consequence of changes in the activity of anticodon modification enzymes [30]. This mechanism is probably not limited to Trm9, and it is possible that other responses are linked to the activity of other tRNA modification enzymes (Figure 3, Table 1).

The redundancy of the genetic code offers an opportunity to fine-tune gene expression levels depending on the usage of synonymous codons. In this regard, functionally related genes (i.e., cell cycle-related genes or stress-response genes) seem to have similar codon usage profiles, suggesting that their translation is somehow favored under certain conditions. Whether this regulation is achieved through changes in tRNA abundance or through the regulation of modifications is something that must be further studied, although supporting evidence for both regulatory mechanisms exists [30,40,72,73].

Two tRNA modification enzymes are known to increase codon-pairing ability: tRNA-dependent adenosine deaminases (ADATs) and tRNA-dependent uridine methyltransferases (UMs) [74,75]. These enzymes expand the wobbling capacity of tRNAs and increase the translation efficiency of the codons recognized by the modified tRNAs. Indeed, highly expressed genes (e.g., ribosomal genes) are found to be most enriched in 'preferred' codons (in this case, those read by tRNAs with modified anticodons), again supporting the possibility that the activity of tRNA modification enzymes constitutes a novel mechanism for post-transcriptional regulation of protein abundance [28]. However, it is an open question whether the activity of these enzymes is regulated in response to specific conditions and, if so, by what means this regulation is accomplished.

Future directions

mRNA sequences contain far more information than just the encoded amino acids. Although the multiple regulatory layers that result from modification of DNA and proteins have been extensively studied, RNA modifications still remain an unexplored territory [76]. In this regard, the complexity of cellular tRNA populations holds great potential for the discovery of new cellular regulatory mechanisms.

It has been shown that certain post-transcriptional RNA modifications can be dynamic and reversible, suggesting that some modifications have functions beyond fine-tuning the structure and function of the RNA [77]. Similarly, tRNA modifications can also be regulated and maintained in distinct cell types and physiological states [76,78].

Information about the range of biological functions of tRNA modifications has only recently begun to emerge [30,72]. Due to their complex nature, post-transcriptional RNA modifications are difficult to study, and understanding of these modifications is sorely lacking compared with

other areas in cell biology. New technologies capable of systems-level analyses of RNA modification changes under diverse cellular conditions will surely bring novel insights into the biosynthesis of tRNA modifications and their role in cellular responses [72,79]. We expect that, during the coming years, these types of approach will shed light on the roles of tRNA modification enzymes in the proteomic changes that accompany transitions in the cell cycle, stress responses, and cell differentiation.

In addition to affecting translation efficiency, codon choice has also been shown to govern translation fidelity by influencing the rate of mistranslation [18,19,22]. Indeed, in the search for the right tRNA, the ribosome might incorrectly bind to a near-cognate tRNA (i.e., a tRNA with one base mismatch relative to the codon), causing the incorporation of a different amino acid. The frequency of this type of mistranslation error has been estimated *in vivo* to range from 10^{-2} in *Bacillus subtilis* [80] to 10^{-5} in yeast cells [81]. The fact that there is a significant correlation between codon conservation and conserved amino acid position suggests that translation accuracy has been under positive selection [22]. Importantly, however, mistranslation might also be beneficial. Recent work has shown that, under certain stress conditions, mistranslation rates increase, leading to increased misincorporation of methionine residues into the mammalian proteome [82]. Moreover, in certain organisms, proteome-wide mistranslation has been shown to increase their fitness under particular environmental conditions [83]. These novel observations suggest that mistranslation evolved as a cellular strategy to adapt to environmental changes and that the codon choice have evolved such that errors can be introduced in nonessential positions of proteins. Exploring the biological significance of mistranslation represents one of the most exciting new directions in this field.

Concluding remarks

Although any given amino acid can be encoded by multiple codons, these 'synonymous' codons are not equally used across genes or genomes. Codon usage has been shown to influence gene expression levels, but the precise rules that govern codon composition remain unclear. Recent efforts have started to uncover specific parameters that affect codon choice, such as codon autocorrelation, codon order, tRNA isoacceptor abundance, or gene coregulation.

In this already complex scenario, tRNA modifications emerge as novel players that can modulate the translation efficiency of codons and, consequently, the expression levels of specific subsets of genes. Indeed, recent studies suggest that tRNA modifications have an important role in genome regulation by specifically enhancing the expression levels of those genes involved in a cellular response. This indicates that the complexity of mature cellular tRNA populations, which has only recently started to be appreciated, holds great potential for the discovery of new cellular regulatory mechanisms.

Acknowledgments

This work has been supported by grant BIO2009-09776 from the Spanish Ministry of Education and Science, and by grant MEPHITIS-223024 from the European Union. EMN is supported by a La Caixa/IRB International PhD Programme Fellowship.

References

1 Hershberg, R. and Petrov, D.A. (2008) Selection on codon bias. *Annu. Rev. Genet.* 42, 287–299
2 Bulmer, M. (1991) The selection–mutation–drift theory of synonymous codon usage. *Genetics* 129, 897–907
3 Rocha, E.P. (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* 14, 2279–2286
4 Plotkin, J.B. and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32–42
5 Dong, H. *et al.* (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* 260, 649–663
6 Tuller, T. *et al.* (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141, 344–354
7 Sorensen, M.A. *et al.* (1989) Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* 207, 365–377
8 Varenne, S. *et al.* (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.* 180, 549–576
9 Li, G.W. *et al.* (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538–541
10 Goetz, R.M. and Fuglsang, A. (2005) Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 327, 4–7
11 Ghaemmaghami, S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425, 737–741
12 Castillo-Davis, C.I. and Hartl, D.L. (2002) Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* 19, 728–735
13 Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12, 640–649
14 Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10, 7055–7074
15 Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13–34
16 Gustafsson, C. *et al.* (2004) Codon bias and heterologous protein expression. *Trends Biotechnol.* 22, 346–353
17 Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.* 15, 8023–8040
18 Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136, 927–935
19 Stoletzki, N. and Eyre-Walker, A. (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* 24, 374–381
20 Chen, S.L. *et al.* (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3480–3485
21 Sharp, P.M. and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 4, 222–230
22 Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134, 341–352
23 Zhou, T. *et al.* (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.* 26, 1571–1580
24 Kudla, G. *et al.* (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324, 255–258
25 Cannarozzi, G. *et al.* (2010) A role for codon order in translation dynamics. *Cell* 141, 355–367
26 Parmley, J.L. and Huynen, M.A. (2009) Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genet.* 5, e1000548
27 Stadler, M. and Fire, A. (2011) Wobble base-pairing slows *in vivo* translation elongation in metazoans. *RNA* 17, 2063–2073
28 Novoa, E.M. *et al.* (2012) A role for tRNA modifications in genome structure and codon usage. *Cell* 149, 202–213
29 Frenkel-Morgenstern, M. *et al.* (2012) Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Mol. Syst. Biol.* 8, 572
30 Begley, U. *et al.* (2007) Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Mol. Cell* 28, 860–870

31 Maraia, R.J. *et al.* (2008) It's a mod mod tRNA world. *Nat. Chem. Biol.* 4, 162–164
32 Kanaya, S. *et al.* (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238, 143–155
33 Percudani, R. *et al.* (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 268, 322–330
34 Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 151, 389–409
35 Kanaya, S. *et al.* (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* 53, 290–298
36 dos Reis, M. *et al.* (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32, 5036–5044
37 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4482–4487
38 Qian, W. *et al.* (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* 8, e1002603
39 Duret, L. (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16, 287–289
40 Dittmar, K.A. *et al.* (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* 2, e221
41 Lin, K. *et al.* (2002) Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. *Nucleic Acids Res.* 30, 2599–2607
42 Paule, M.R. and White, R.J. (2000) Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* 28, 1283–1298
43 Geiduschek, E.P. and Kassavetis, G.A. (2001) The RNA polymerase III transcription apparatus. *J. Mol. Biol.* 310, 1–26
44 Phizicky, E.M. and Hopper, A.K. (2010) tRNA biology charges to the front. *Genes Dev.* 24, 1832–1860
45 Kutter, C. *et al.* (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43, 948–955
46 Arava, Y. *et al.* (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3889–3894
47 O'Brien, T. and Lis, J.T. (1991) RNA polymerase II pauses at the 5′ end of the transcriptionally induced *Drosophila hsp70* gene. *Mol. Cell. Biol.* 11, 5285–5290
48 Siller, E. *et al.* (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* 396, 1310–1318
49 Zhang, G. *et al.* (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* 16, 274–280
50 Ingolia, N.T. *et al.* (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223
51 Stapulionis, R. and Deutscher, M.P. (1995) A channeled tRNA cycle during mammalian protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 92, 7158–7161
52 Petrushenko, Z.M. *et al.* (2002) Novel complexes of mammalian translation elongation factor eEF1A.GDP with uncharged tRNA and aminoacyl-tRNA synthetase. Implications for tRNA channeling. *Eur. J. Biochem.* 269, 4811–4818
53 Gaucher, E.A. *et al.* (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl. Acad. Sci. U.S.A.* 98, 548–552
54 Kubo, M. and Imanaka, T. (1989) mRNA secondary structure in an open reading frame reduces translation efficiency in *Bacillus subtilis*. *J. Bacteriol.* 171, 4080–4082
55 de Smit, M.H. and van Duin, J. (1990) Control of prokaryotic translational initiation by mRNA secondary structure. *Prog. Nucleic Acid Res. Mol. Biol.* 38, 1–35
56 Gu, W. *et al.* (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* 6, e1000664
57 Kim, J. *et al.* (1991) Ribosomes pause at specific sites during synthesis of membrane-bound chloroplast reaction center protein D1. *J. Biol. Chem.* 266, 14931–14938
58 Makhoul, C.H. and Trifonov, E.N. (2002) Distribution of rare triplets along mRNA and their relation to protein folding. *J. Biomol. Struct. Dyn.* 20, 413–420
59 Yanagitani, K. *et al.* (2011) Translational pausing ensures membrane targeting and cytoplasmic splicing of XBP1u mRNA. *Science* 331, 586–589
60 Zama, M. (1995) Discontinuous translation and mRNA secondary structure. *Nucleic Acids Symp. Ser.* 34, 97–98
61 Saunders, R. and Deane, C.M. (2010) Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.* 38, 6719–6728
62 Cortazzo, P. *et al.* (2002) Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 293, 537–541
63 Komar, A.A. (2007) Genetics. SNPs, silent but not invisible. *Science* 315, 466–467
64 Kimchi-Sarfaty, C. *et al.* (2007) A 'silent' polymorphism in the *MDR1* gene changes substrate specificity. *Science* 315, 525–528
65 Fraser, H.B. *et al.* (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9033–9038
66 Lithwick, G. and Margalit, H. (2005) Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res.* 33, 1051–1057
67 Watkins, N.E. and SantaLucia, J. (2005) Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.* 33, 6258–6267
68 Agris, P.F. (2004) Decoding the genome: a modified view. *Nucleic Acids Res.* 32, 223–238
69 Alexandrov, A. *et al.* (2006) Rapid tRNA decay can result from lack of nonessential modifications. *Mol. Cell* 21, 87–96
70 Urbonavicius, J. *et al.* (2001) Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J.* 20, 4863–4873
71 Kramer, G.F. and Ames, B.N. (1988) Isolation and characterization of a selenium metabolism mutant of *Salmonella typhimurium*. *J. Bacteriol.* 170, 736–743
72 Chan, C.T. *et al.* (2010) A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet.* 6, e1001247
73 Pavon-Eternod, M. *et al.* (2009) tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res.* 37, 7268–7280
74 Gerber, A.P. and Keller, W. (1999) An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science* 286, 1146–1149
75 Nasvall, S.J. *et al.* (2004) The modified wobble nucleoside uridine-5-oxyacetic acid in tRNAPro(cmo5UGG) promotes reading of all four proline codons *in vivo*. *RNA* 10, 1662–1673
76 Yi, C. and Pan, T. (2011) Cellular dynamics of RNA modification. *Acc. Chem. Res.* 44, 1380–1388
77 Jia, G. *et al.* (2011) N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* 7, 885–887
78 He, C. (2010) Grand challenge commentary: RNA epigenetics? *Nat. Chem. Biol.* 6, 863–865
79 Globisch, D. *et al.* (2011) Systems-based analysis of modified tRNA bases. *Angew. Chem. Int. Ed. Engl.* 50, 9739–9742
80 Meyerovich, M. *et al.* (2010) Visualizing high error levels during gene expression in living bacterial cells. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11543–11548
81 Stansfield, I. *et al.* (1998) Missense translation errors in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 282, 13–24
82 Netzer, N. *et al.* (2009) Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature* 462, 522–526
83 Moura, G.R. *et al.* (2009) Genetic code ambiguity: an unexpected source of proteome innovation and phenotypic diversity. *Curr. Opin. Microbiol.* 12, 631–637

4.2. Chapter 2: Aminoacyl-tRNA synthetases as antimalarial drug targets

4.2.1. Introduction to *Plasmodium falciparum*

4.2.1.1. *Plasmodium falciparum* malaria

Apicomplexan protozoans form a phylum of obligate intracellular parasites that comprises important human pathogens, including *Plasmodium sp*., the causative agent of malaria. Malaria is a mosquito-borne infectious disease that can result in symptoms such as coma, respiratory distress, severe anaemia and death. Approximately half of the world's population is at risk of contracting malaria; however most malaria cases and deaths occur in sub-Saharan Africa (**Figure 4.4**).

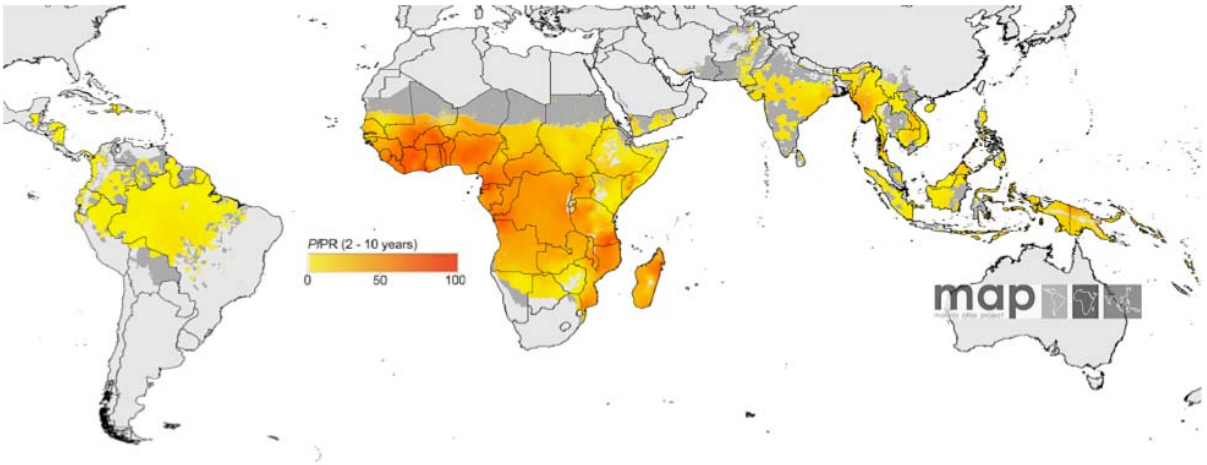


Figure 4.4. Spatial distribution of *Plasmodium falciparum* endemicity. The annual mean of P. falciparum parasite rate (PfPR) is the proportion of blood samples showing detectable parasites between 2 and 10 years of age, as a measure of endemicity. Adapted from Hay et al. 2009.

It has been estimated that malaria causes 225 million cases of clinical malaria and approximately one million deaths per year (WHO, 2010). Most of these deaths are caused by *Plasmodium falciparum*, one of the four distinct *Plasmodium* species that affect humans,

whereas the high morbidity of *P. vivax*, which can cause dormant liver-stage infections, also accounts for the enormous economic burden.

While isolated efforts to curb malaria with combinations of vector control, education, and drugs have proven successful, a global situation has not been reached. Current control strategies include both therapeutic and prophylactic chemotherapy and the blocking of transmissin using vector control including insectide-impregnated bed nets. Achieving eradication would be eased by the availability of effective drugs that would target both liver and blood stages of the parasite.

4.2.1.2. *P. falciparum* life cycle

P. falciparum has a complex life cycle involving several differentiated forms and two different hosts (**Figure 4.5**). Within the human host, *Plasmodium* has two major phases: liver and blood. Subsequent to a bite from an infected mosquito, the *Plasmodium* sporozoites migrate to the liver and infect hepatocytes. After replicating within the hepatocyte, the parasites rupture the cell to release merozoites, a stage specialized to infect erythrocytes. However, two species (*P. vivax* and *P. ovale*) can assume a dormant state (hypnozoites) within the liver that commonly lasts for months to years (White, 2011). After exiting the liver, the parasite establishes a recurring life cycle in the erythrocytes, named asexual intraerythrocytic cycle (IEC), which lasts approximately 48h. All clinical symptoms of malaria are associated with the IEC, and it is the target of most antimalarial drug and vaccine strategies (**Table 2**). During the blood stage, some parasites will differentiate into sexual forms that will be transmitted back to the mosquito.

Each devlopmental stage is characterised by distinct physiology, and each has varying sensitivity to most drugs, making the discovery of a single drug active against all stages challenging (Wells et al., 2009). To date, antimalarial chemotherapy has primarily targeted the blood stages (**Table 3**). Various drugs including atoquavone, primaquine and artemisin derivatives are effective to varying degrees against liver-stage parasites. However, primaquine is the only drug effective against the dormant hypnozoites.

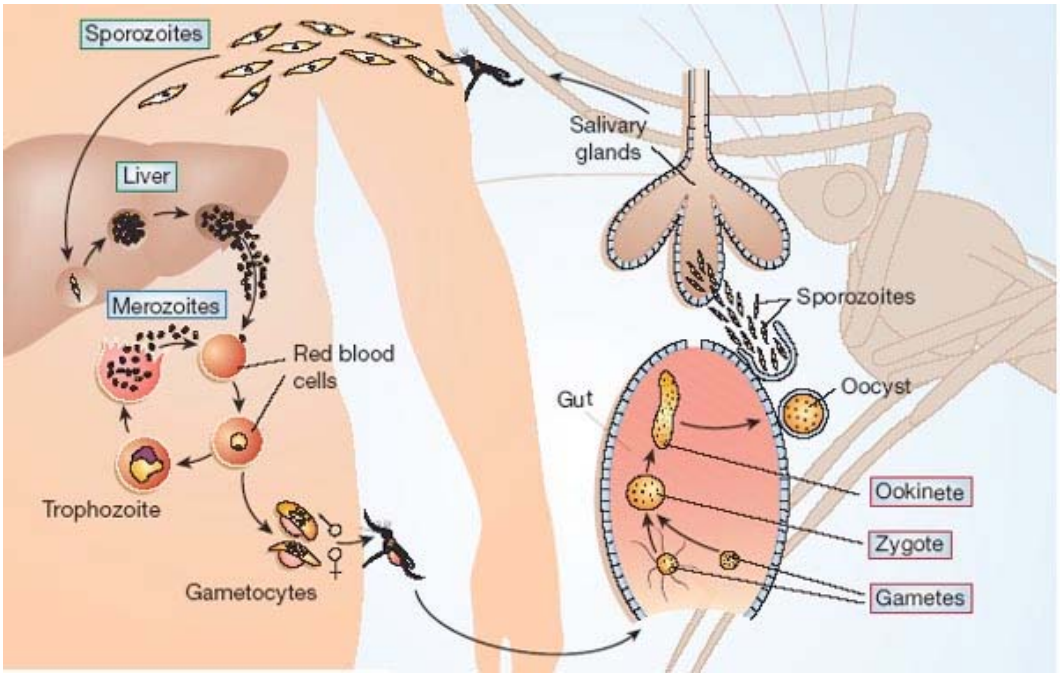


Figure 4.5. Complex life cycle of *Plasmodium falciparum*. The life cycle includes two hosts: the mosquito *Anopheles*, and human. The sexual stage occurs in the mosquito, whereas the asexual stages occurring in human erythrocytes cause the symptoms typical of malaria. Adapted from Menard, 2005.

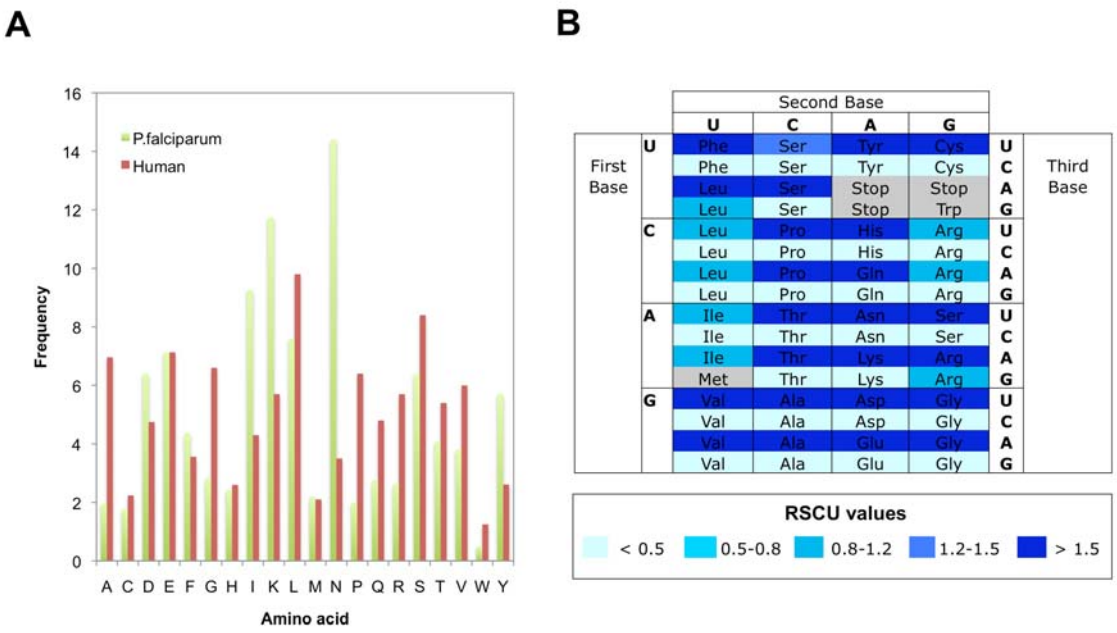
Table 2. Current antimalarial drugs, showing their targets and stage of action

		Stage specificity			
		Liver		Blood	
Drug	**Target**	**Hypnozoite**	**Schizont**	**IEC parasites**	**Gametocytes**
Artemisins	Heme? ATP6?	No	No	Yes	Yes
Atovaquone	Cytochrome bc1	No	Yes	Yes	Yes
Chloroquine	Hemozoin	No	No	Yes	No
Doxycycline	Ribosomal RNA (apicoplast)	No	Yes	Yes	No
Primaquine	Unknown	Yes	Yes	Yes	Yes
Pyrimethanime	Dihydrofolate reductase	No	Yes	Yes	Yes
Proguanil	Dihydrofolate reductase	No	Yes	Yes	Yes
Quinine	Hemozoin	No	No	Yes	Yes
Tetracycline	Ribosomal RNA (apicoplast)	No	Yes	Yes	No

4.2.1.3. *P. falciparum* genome

The 22.8Mb genome of *P. falciparum* is comprised of 14 linear chromosomes, a circular plastid-like genome, and a linear mitochondrial genome. The malaria genome-sequencing consortium estimates that more than 60% of the 5772 predicted open reading frames (ORFs) lack sequence similarity to genes from any other known organism (Gardner et al. 2002). Although ascribing putative roles for these ORFs in the absence of sequence similarity remains challenging, their unique nature may be key to identifying *Plasmodium*-specific candidates for antimalarial strategies.

Plasmodium falciparum presents unique genetic peculiarities compared to other genome. Firstly, its nuclear genome is extremely rich in A and T (over 80% AT-rich). This extreme AT bias causes, in turn, an extreme codon usage bias towards those codons that have an A or T nucleoside in its degenerate position (**Figure 4.6**). Indeed, the amino acid composition is also biased towards those amino acids encoded by AT-rich codons, such as lysine –encoded by AAA- and aspararagine -encoded by AAT-. Secondly, protein-encoding genes in *P. falciparum* tend to be longer than its homologues in other species –up to 50% longer than their yeast homologues- (Frugier et al., 2010). These insertions correspond in most cases to Low Complexity Regions (LCRs), which are characterized by single amino acid repeat sequences (**Figure 4.6c**). Examples of LCRs can be found for all 20 amino acids in the three domains of life. However, *Plasmodium* proteins possess many unique insertions not found in homologous proteins from other species. These insertions often take the form of non-globular segments that are integrated directly into structured domains that, in theory, bulge out of the protein core (Feng et al. 2006).



C

GluRS
(1) PNNNNNNNNNNNNNNNNNNNNNNNI
(2) LKTDLYNISEKIDPKNNKVVNIYVQTQYEEWLDFFRSRNVNKDIVFIC
(3) WKKNLKCLICDNKMNFDDRK

IleRS
(1) LCSDFPKVKNVYVIQDDIQPLVQKYSVLYDCYMKSEKCDTSNEKHADEDIYSDHINVNKNNNKNNNSNNNNNNNTNCNDNNNINLNYNNNCYA
(2) NIKYTNNQNMEHNENNETTNIKEEDLNRNDIHK

LeuRS
(1) EDDYSLCRSISDDKNNVDKN
(2) FMNSIRSSQSVLNACIEKKEAPNNNNNKINNNKINNNKINNNKINNNKSNNNKSNNNNNCGSSANISNTFFTDFEKGEEDLKNKIWNE
(3) LHYGDKSYVNNIMTKEEIINSC
(4) PLLKNKNIKNVQIPKSEDNTNDDDTLEKKNNDVITKNTNNNNNENNNNMNNVYNLD
(5) QNNQTNKNVDNYQNNNEYVIPSNDDNNQINNHHINVVKEEEEKNVHIKNNSRKE
(6) EKYDRNKEDVEERSILDRQTEILNELENIDLSSYEK
(7) NYSIVMHKQNNNLLNVVEIFRKSYDKVINKCNKQKVV
(8) EKENNNDNLKKNIQQIKQINHISNNNNNKDNICDGHNNDIDEDDDEERDDEEGANYYKEDEEE

MetRS
(1) ELYLKENKNKNKNFINNDDEKD

GlnRS
(1) NYDSSSGTCNVNNINQPNNNNNNKENINDNNNNNKENINDNNNNNKENNINPCSTNSFNINSFSTNNYNNFNNPCCSFSCNIYDYYDI
(2) SKSGLCTNQKKDNVNRQGGDGDNK

ValRS
(1) INHNDDNNDDNNNNNNNNNNL
(2) YELIKPRLNTDSHDNFVSGLKEKSDMNNIEKREVVEQNDNNIESVNTNHINEINKCDNYILDGYG

TrpRS
(1) KKKKDDNGKSNSTFDH

Figure 4.6. Biased composition of the *P. falciparum* nuclear genome. **A**) Amino acid composition of the *P. falciparum* protein coding genes (green), compared to the human genome (red). **B**) Codon usage of the protein coding genes, measured as relative synonymous codon usage (RSCU). The stop codons, Met and Trp not been included in the analysis. **C**) Examples of LCRs found in some cytosolic plasmodial aaRS. Lysine residues (K) have been coloured in yellow, and asparagine residues (N) have been coloured in red. Panel C is adapted from Frugier et al. 2010.

4.2.1.4. Transcription in *P. falciparum*

The transcription in *P. falciparum* has evolved in an extremely specialized mode of transcriptional regulation such that it produces a continuous cascade of gene expression that is unprecedented in eukaryotic biology (Bozdech et al. 2003). It starts with genes corresponding to general cellular processes, such as protein synthesis, and ending with *Plasmodium*-specific functionalities, such as erythrocyte invasion. The transcriptome of the IDC resembles a "just-in-time" manufacturing process whereby induction of any gene occurs once per cycle, and only at a time when it is required (**Figure 4.7**).

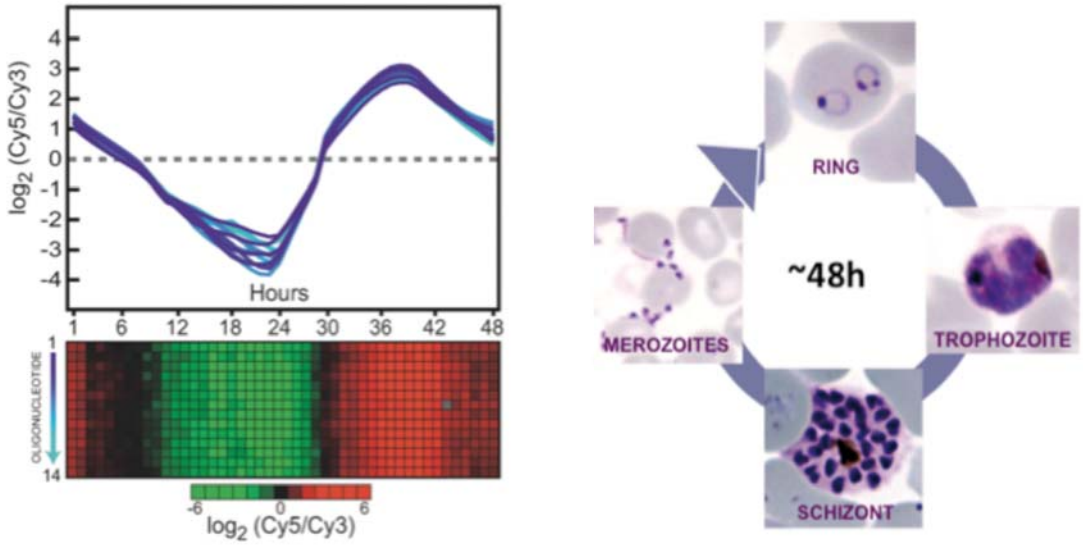


Figure 4.7. Transcriptional regulation during IDC. A) Example expression profile of MAL6P1.147 gene during the IDC, showing the corresponding individual expression profiles for the 14 oligonucleotides used to measure the gene's expression levels (Bozdech et al. 2003). This gene is expressed during the schizont stage. A red/green colorimetric representation of the gene expression ratios for each oligonucleotide is shown. B) Giemsa-staining of the IDC stages.

4.2.1.5. Protein translation in *P. falciparum*

The protein translation machinery of the *P. falciparum* parasite is the target of important antimalarial drugs, and encompasses many promising targets for future drugs. Nevertheless, its protein translation machinery remains poorly characterized. *Plasmodium* parasites have three subcellular compartments that house genomes: the nucleus, the mitochondrion and the apicoplast (an essential organelle that is present in all Apicomplexa and is thought to be derived from a secondary endosymbiosis of a red algae). Each of these three compartments requires its own compartmentalized transcription and translation apparatus for its survival (Jackson et al., 2011).

The apicoplastic machinery translates less than 50 genes encoded by its 35kb circular apicoplast genome, which include rRNAs, tRNAs, ribosomal proteins, the translation elongation factor Tu (EF-Tu), and also a handful of other poorly characterized proteins unrelated to translation (**Figure 4.8**). However, it also imports many other nuclear-encoded proteins needed for its translation (Roy et al., 1999).

The mitochondrion of *Plasmodium* is amongst those with highest size reduction (6kb), and encodes only three proteins: cytochrome c oxidase subunits I and III (Cox1 and Cox3), and cytochrome b (Cytb) (Feagin, 1992). The mitochondrial genome does not encode tRNA genes, but encodes small fragmented rRNAs, although it is unclear if they are functional. Therefore, the mitochondrion must import all proteins and tRNAs needed for the translation of its three mitochondrial-encoded transcripts. Despite the apparent incompleteness of the mitochondrial machinery and lack of direct evidence for translation in mitochondria, indirect evidence suggests that mitochondrial translation is active and is essential. A major antimalarial drug (atovaquone) targets the mitochondrial cytochrome bc1 complex, and point mutations in the *cytb* gene correlate with drug resistance (Afonso et al., 2010). Indeed, cyanide, a drug that inhibits mitochondrial electron transport, is also active against *Plasmodium*, suggesting that the mitochondrial *cox* genes found in the mitochondrion are translated (Painter et al., 2007).

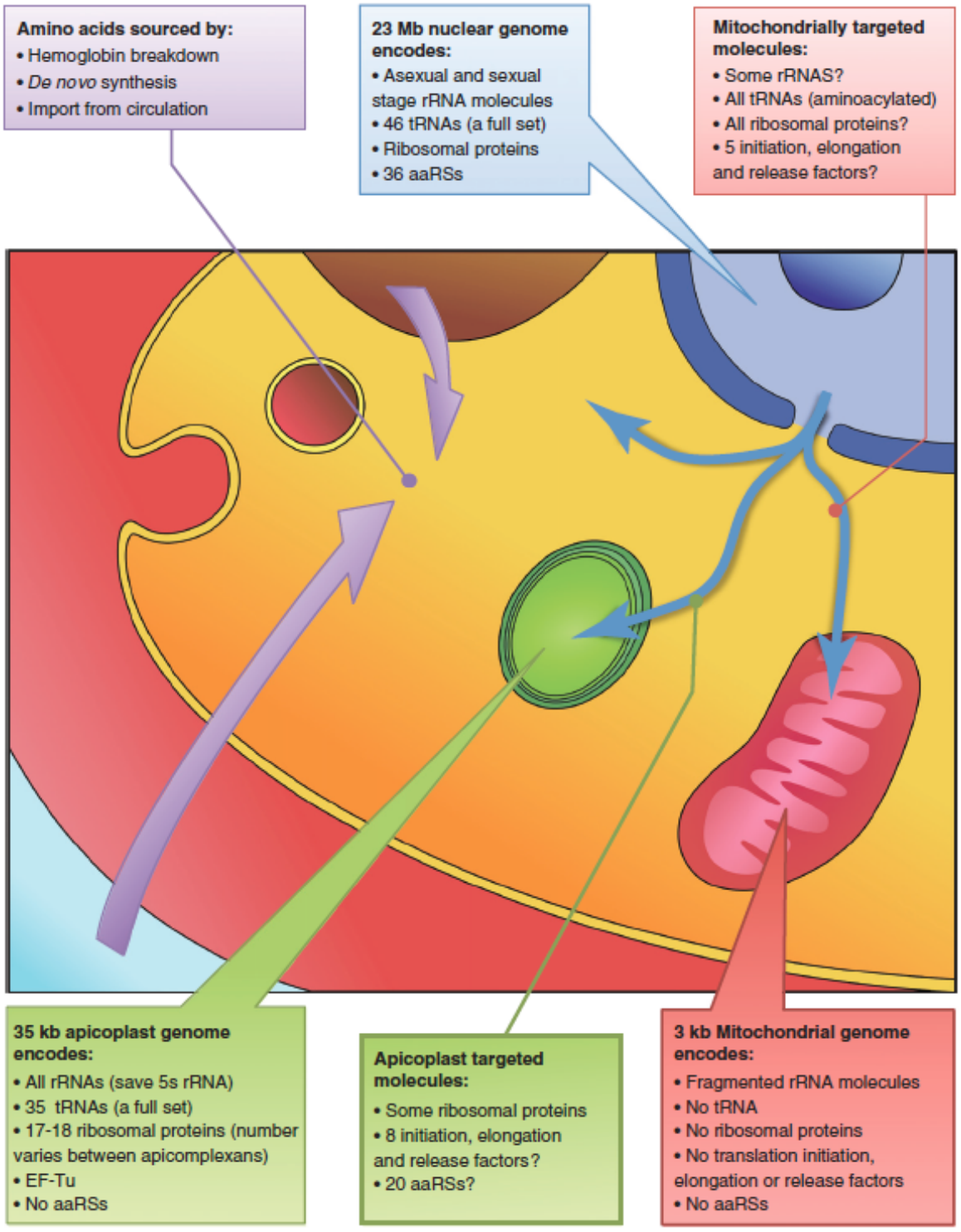


Figure 4.8. Protein translation compartments in *Plasmodium falciparum*. Translation takes place in the *Plasmodium* apicoplast (green), mitochondrion (red) and cytosol (orange). Adapted from Jackson et al. 2011.

4.2.1.6. Components of the *P. falciparum* translation machinery

4.2.1.6.1. tRNA

A total of 46 tRNA genes, coding for 44 different tRNA isoacceptors, are found in the nuclear genome, whereas the apicoplast genome contains 35 genes encoding 26 tRNA isoacceptors (**Figure 4.9**). With the exception of the apicoplastic initiator $tRNA^{Met}$, characterized by a unique variable region (11 nucleotides), they all resemble other eukaryotic tRNAs (Pütz et al., 2010), and have the potential to adopt a canonical tertiary fold.

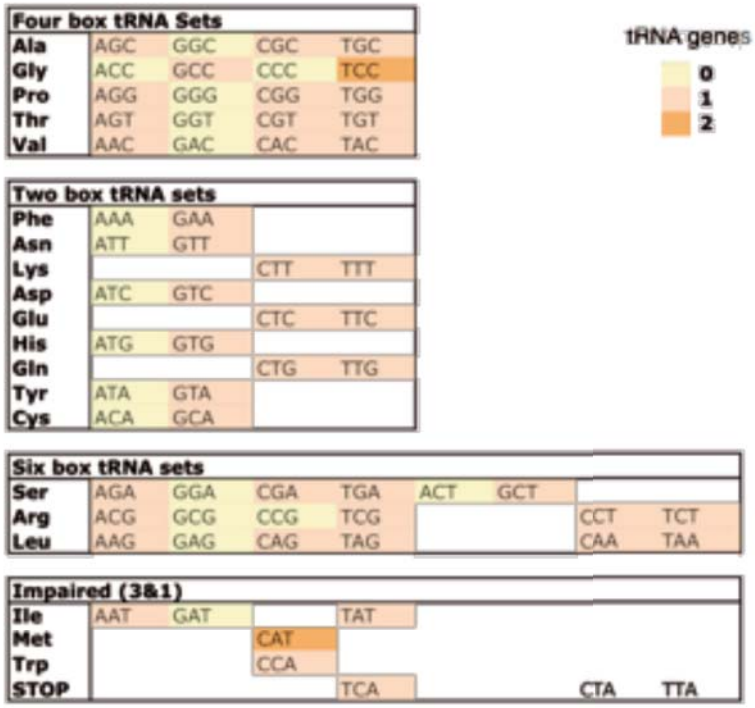


Figure 4.9. Distribution of nuclear-encoded tRNA genes in *Plasmodium falciparum.* Each tRNA isoacceptor is represented by its anticodon, and has been coloured depending on its tRNA gene copy number. As a general trait, Plasmodium falciparum has one single tRNA gene copy per tRNA isoacceptor.

The most striking observation is that there is only one copy per tRNA isoacceptor in the nuclear genome, and therefore *Plasmodium* has the fewest known tRNA genes of any eukaryote. Whether this limits the translation efficiency in the parasite or whether its RNA

polymerase III machinery is regulated by unknown factors is still unclear. If the different tRNAs are equally abundant (tRNA abundance is thought to be proportional to the relative tRNA gene copy number), the most used codons in the genome will have relatively fewer tRNA molecules available per codon, which should limit the translation efficiency.

It is unknown if the expression of plasmodial tRNA genes is under some type of regulation, or if other mechanisms compensate such a uniform tRNA gene distriution –e.g. differential codon-anticodon affinities, different tRNA half lifes- in order to match its AT-biased genome. To solve this intriguing question, members of our lab analyzed the tRNA expression levels using published *P. falciparum* microarray data (Rovira-Graells et al. 2012). The tRNA levels seem to vary across the IDC life cycle –and are similar across different parasite clones-, in a similar fashion as plasmodial aaRS (**Figure 4.10**) Interestingly, the absolute levels of tRNA seem to be dispare between different tRNA isoacceptors, suggesting that the tRNA abundances in the cell are not reflecting the uniform tRNA gene copy number but instead, that there might be a regulation upon tRNA expression levels.

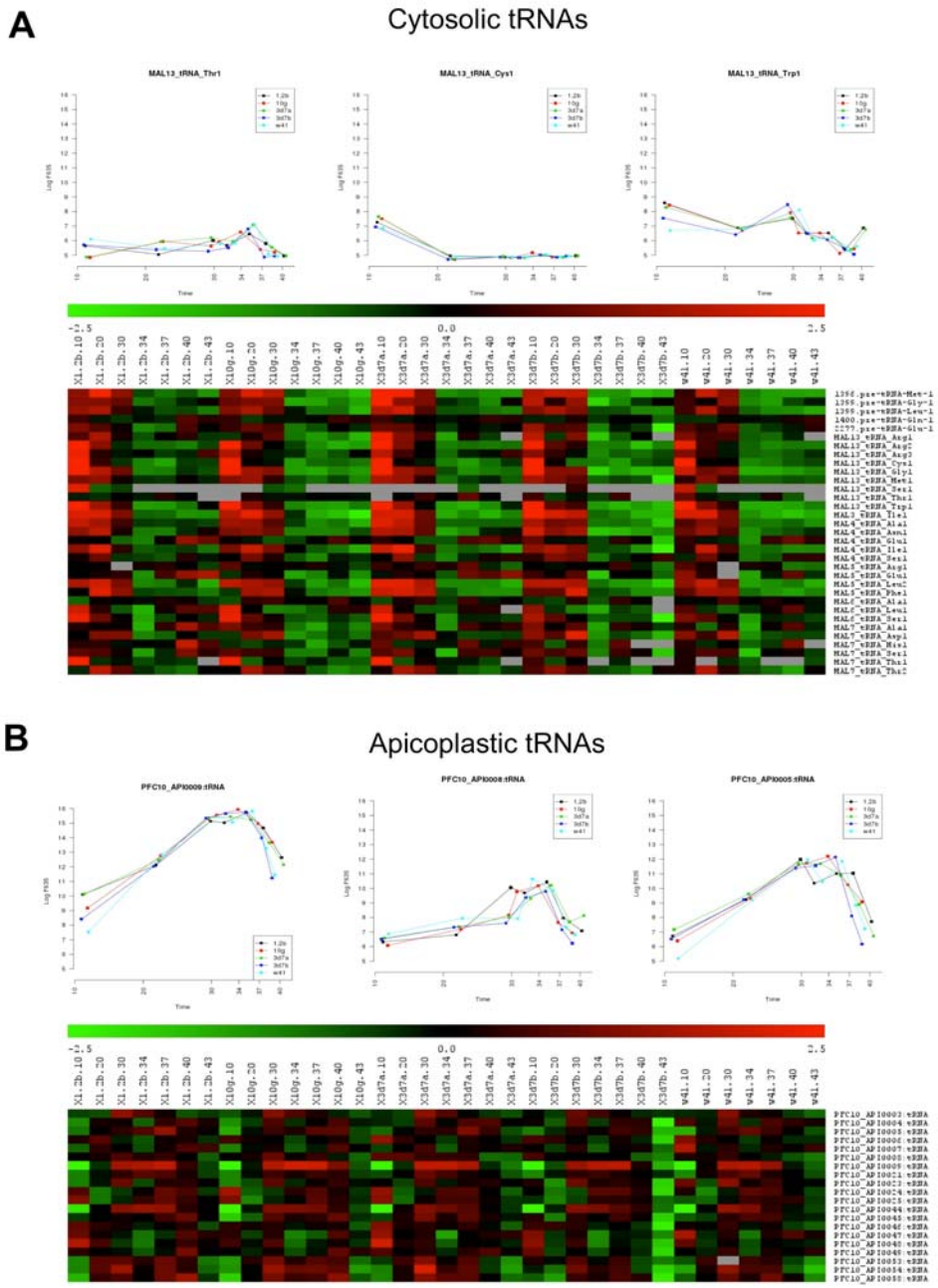


Figure 4.10. Microarray-based *P. falciparum* tRNA expression levels. Differential expression of nuclear-encoded tRNA genes (A) and apicoplast-encoded tRNA genes (B) along the IDC life cycle. For each tRNA gene (columns), 7 different time points across the life cycle have been taken (t=10,20,30,34,37,40 and 43h), and from 5 different parasite lines (1.2B, 10G, 3D7A, 3D7B and w41). Three tRNA isoacceptors have been randomly chosen for both cytosolic and apicoplastic tRNA genes (upper part of each subfigure), and the absolute levels of each tRNA isoacceptor across the IDC and for each parasite line (coloured differently) are plotted. Raw data taken from Rovira-Graells et al. 2012.

4.2.1.6.2. Aminoacyl-tRNA synthetases

Nuclear, plastid and mitochondrial genomes show differences in codon usage, but every amino acid is used at least once in each of the three compartments, and therefore a full complement of tRNAs and aaRS should be necessary for the translation of each genome.

Plasmodium has genes only for 37 aaRS, and these are apparently sufficient to translate the nuclear, apicoplast and mitochondrial genomes. The reduction from a theoretical maximum of 60 to 37 genes implies that several *Plasmodium* aaRS aminoacylate tRNAs that are active in more than one subcellular compartment. Our initial computational studies suggested that most of these proteins are either targeted to the apicoplast or the cytosol, but not to the mitochondria (**Table 3**). In accordance with our predictions, the expression peaks of the predicted subsets of cytosolic and apicoplast-targeted aaRS seemed to be coordinated, with the maximal levels of expression of cytosolic aaRS during the ring stage and of apicoplastic aaRS during the trophozoite stage, also correlating with the in life cycle times of cytosolic and apicoplastic translation, respectively (**Figure 4.11a**).

Several *P. falciparum* aaRS (GlyRS, AlaRS, CysRS and ThrRS) are present only once in the genome, and thus should be dually targeted to both compartments. Some of these dual-targeting predictions have been recently verified by some of our collaborators (Jackson et al., 2012). In each of these cases, the gene models start with a predicted apicoplast targeting sequence (Foth et al. 2003), which is either alternatively spliced or has two initiation start sites, generating two possible isoforms that are either targeted to the cytosol or to the apicoplast, respectively. The other enzyme that is found once is GlnRS, and is expected to be located in the cytosol, while the Gln-tRNA^{Gln} in the apicoplast is thought to beformed through the indirect pathway using apicoplast-targeted amidotransferases.

Table 3. List of predicted *P. falciparum* aaRS

	Mitochondria	Apicoplast	Cytosol
Class I			
Arg		1	1
Cys		1	
Glu		1	1
Gln	?	*	1
Ile		1	1
Leu		1	1
Met		1	1
Tyr		1	1
Trp		1	1
Val		1	1
Class II			
Ala		1	
Asp		1	1
Asn		1	1
Gly		1	
His		1	1
Lys		1	1
Phe		2	2
Pro		1	1
Ser		1	1
Thr		1	
Sum	1 ?	20	16

* Amidotransferases predicted to have apicoplastic localization

Recent works have started to experimentally demonstrate the specific targeting of the two isoleucyl-tRNA synthetases, which are targeted to the apicoplast and cytosol, respectively (Istvan et al., 2011). In addition, in this work we have demonstrated the localisation of the apicoplast-targeted lysyl-tRNA synthetase (PfKRS-2) and the cytosolic localisation of glutaminyl-tRNA synthetase (PfQRS) (**Figure 4.11b and 4.11c**).

With respect to mitochondrial protein translation, it is probable that in *P. falciparum* all mitochondrial tRNAs are aminoacylated in the cytosol and transported into the mitochondria for use in protein synthesis, in a similar fashion to what has been suggested for *Toxoplasma gondiii* –a closely related apicomplexan protozoa- (Pino et al., 2010; Esseiva et al., 2004). Such a mechanism presumably suffers from the deficiency that imported tRNAs would be unable to be recharged readily. It is currently unknown if these tRNAs would in some way be recycled or degraded after a single use.

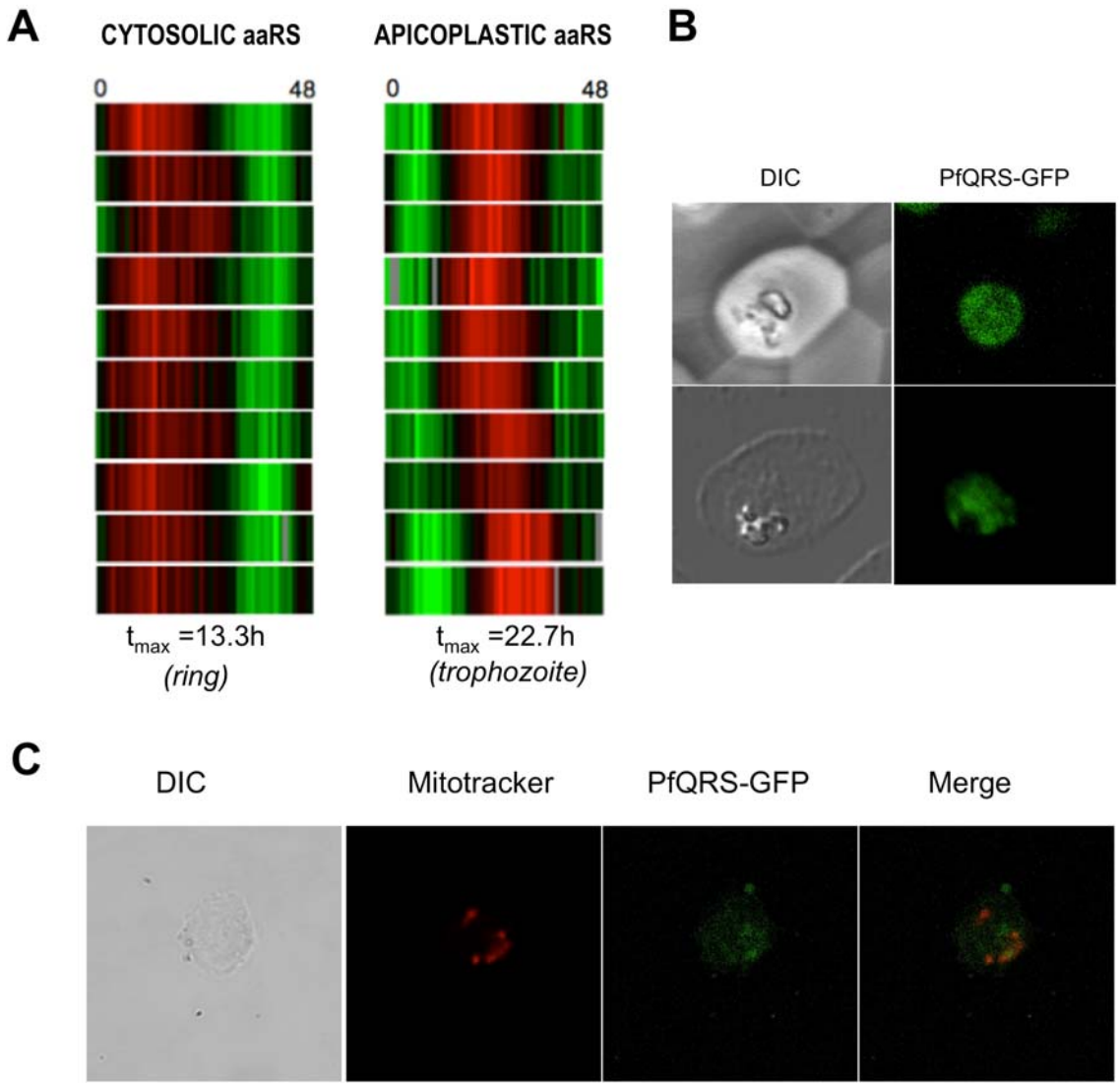


Figure 4.11. Predicted set of *P. falciparum* aaRS. A) Expression times of compartment-specific aaRS during the IDC. Cytosolic aaRS are maximally expressed (red) in the ring stage, while apicoplastic aaRS are maximally expressed in the trophozoite stage. Transcriptomic data has been taken from the work published by Bozdech et al. 2003. **B**) Immunofluorescence assays on transfected *P. falciparum*-infected red blood cells (iRBCs). *P. falciparum* parasites have been transfected with a vector containing PfQRS-GFP, which localizes in the cytosol. **C**) Mitotracker has been used to specifically label the mitochondria. To check whether PfQRS-GFP also localizes in the mitochondria -as suggested by the PlasMit subcellular prediction software-, iRBCs have been treated with saponin 0.10% to break the parasite membrane -but not the organellar membranes- and remove the cytosolic fluorescence (Jackson et al., 2012). However, as can be seen in the figure, there is no colocalization between PfQRS-GFP and Mitotracker, indicating that PfQRS is exclusively located in the cytosol. Immunofluorescence images obtained with Leica SP2 confocal software.

4.2.1.7. *P. falciparum* aaRS as drug targets

Once the set of plasmodial aaRS has been identified, which ones would be best to choose for antimalarial drug design? The TDR target priorization gives a list of desirable characteristics than should be accomplished by a drug target (Aguero et al., 2008):

i. **Essentiality**: the target must be an essential gene required for growth and viability, whose inhibition kills the pathogen
ii. **Druggability**: the target should present an ability to produce a drug against it, i.e. a protein that favours interactions with drug-like chemical compounds
iii. **Selectivity**: the target should not have human homologues, or drugs should be able to avoid cross-reactivity with its human homologues
iv. **Stage-specificity**: the target should be highly expressed in stages affecting the human host
v. **Feasability for structure-based drug design (SBDD)**: the target should be available as an X-ray crystal, or a good template should be available to build a reliable homology model for SBDD purposes.
vi. **Assayability**: the target should be easy to produce as a recombinant protein in order to perform an enzyme-based assay, which will be necessary to validate the target.

Taking into account these criteria, all plasmodial aaRS present the same profiles with respect to essentiality –they are all essential-, druggability –they are *a priori* all druggable-, stage-specificity –they are all highly expressed in human stages-, and feasability for SBDD –there is no X-ray for any of them, and similar templates are available for all aaRS to build homology models-. Thus, the two characteristics that can help us choose our targets amongst the set of aaRS are selectivity and assayability.

With respect to ***selectivity***, all plasmodial aaRS have human homologues. However, we can choose those that are more distantly related to the human counterparts through the use of phylogenetic analyses. We find that the apicoplastic lysyl-tRNA synthetase (PfKRS-2) is an interesting drug target because it has a bacterial origin, whereas the human counterpart has an eukaryotic origin. Indeed, lysyl-tRNA synthetase is the unique human aaRS that does not have a mitochondrial specific gene –which would be more closely related to the bacterial-like gene- because it is produced from alternative splicing from the eukaryotic human KRS, thus

being a very good target in terms of selectivity (see **Publication 3** for further details). On the other hand, we also find plasmodial glutaminyl-tRNA synthetase (PfQRS) as a candidate drug target for its selectivity, because it has a different phylogenetic origin than its human counterpart (**Figure 4.12**). The human enzyme clusters with the eukaryotic QRS, whereas all the Apicomplexan QRS cluster with bacterial QRS.

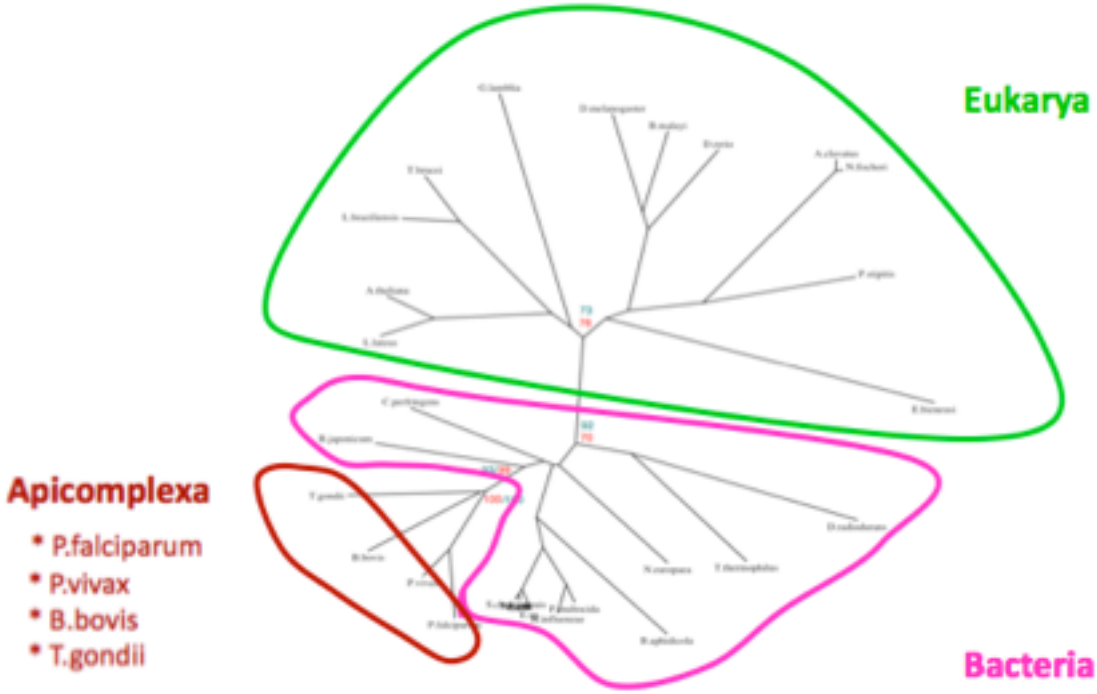


Figure 4.12. Phylogenetic analysis of glutaminyl-tRNA synthetases. The consense distance matrix is shown. Bootstrapping has been performed using 100 replicates, using the neighbour-joining method (green) and maximum-likelihood method (red).

On the other hand, ***assayability*** has always been an issue for plasmodial proteins in general. Its AT-rich genome and consequently its usual codon usage bias causes that heterologous protein expression of *P. falciparum* genes is in general unsuccessful, obtaining either unsoluble proteins or inactive proteins, probably due to incorrect folding (Mehlin et al., 2006). In this regard, once our initial trials for obtaining soluble and active proteins were unsuccessful, we decided to build codon-optimized synthetic genes both for PfQRS and PfKRS-2, which were expressed under a wide range of expression conditions –by the IRB protein expression facility-, finally obtaining soluble and active protein for *in vitro* testing.

4.2.2. Introduction: *aaRS as drug targets*

The emergence of resistance to existing antibiotics demands the development of novel antimicrobial agents directed against novel targets. Historically, bacterial cell wall synthesis, protein, DNA and RNA synthesis have been major targets of very successful classes of antibiotics such as beta-lactams, glycopeptides, macrolides, aminoglycosides, tetracyclines, rifampicins and quinolones. Amongst the less exploded targets is the family of aaRS, which are ancestral enzymes and essential for protein synthesis.

AaRS play essential roles not only in bacteria but also in eukaryotic cells and in mitochondria. Therefore, a major prerequisite for any potential new drug is a high selectivity for inhibition of the bacterial aaRS over inhibition of their eukaryotic and mitochondrial counterparts, given that Inhibition of the host enzymes would have major toxicological implications. From a phylogenetical point of view, mitochondrial aaRS are more closely related to bacterial aaRS than mammalian cytoplasmatic aaRS. However, there are still sufficient structural differences between mitochondrial and bacterial aaRS to allow the development of selective antibiotics. As a rule of thumb, selectivity of greater than 100-fold is desirable (Schimmel et al. 1998).

An important example of the clinical application of an aaRS inhibitor is provided by the antibiotic mupirocin (marketed as Bactroban), which selectively inhibits bacterial isoleucyl-tRNA synthetase (IleRS). This product is currently the world's most widely used topical antibiotic for the control of methicillin-resistant *Staphylococcus aureus* infections (MRSA) (Boyce, 2001).

The drug design strategies regarding the discovery of aaRS inhibitors can be classified into:

1. Reaction-intermediate mimics
2. Analogues of natural aaRS inhibitors
3. Drugs disrupting tRNA interaction
4. Inhibitors of the aaRS proofreading activity
5. Virtual screens and structure-based design aaRS inhibitors
6. High-throughput screening programs

4.2.2.1 Reaction intermediate mimics

The structures of the reaction intermediates have been the focus for the development of novel synthetic compounds that target aaRS. Aminoacyl-adenylates (AA-AMP) have lower dissociation constants than the amino acids and ATP, and thus, the choice of the intermediate is advantageous in the design of novel synthetic compounds with high affinity (Kim et al. 2003). In the reaction intermediate mimics, the acylphosphate linker of the adenylate is generally substituted for another chemically stable group, such as sulfonamides, phosphonates or phosphonamides (**Figure 4.13**). Other compounds replace the adenine with a tetrazole that is linked to one or two additional five- or six-member aromatic or heterocyclic rings.

The main issue with these compounds is whether they will be sufficiently selective for pathogen enzymes, *i.e.* they must not interfere with their human counterparts. Initial efforts to produce intermediate mimics produced non-selective inhibitors (Heacock et al. 1996) (**Figure 4.13a and 4.13b**). Further efforts produced mimics showing up to 250-fold selectivity towards bacterial enzymes versus its human counterpart (Yu et al. 1999). Replacement of the adenine moiety in a series of glutamyl-adenylate analogues by other bases resulted in more than 1000-fold loss of activity, suggesting that the contribution of the adenine ring is important for the binding (Desjardins et al. 1998) (**Figure 4.13c and 4.13d**). Contrarily, in a set of isoleucyl-adenylate mimics, substitutions in the adenine ring produced strong *in vitro* inhibitors (**Figure 4.13e, 4.13f and 4.13g**).

As stated above, previous studies have shown that some adenylate mimics can present *in vitro* species-specificity and antibacterial activity without inhibiting their human counterparts. However, these compounds generally show limited whole cell activity –probably due to poor penetration through the cell wall- (Kim et al. 2003; Schimmel et al. 1998), and low bioavailability –due to binding to serum albumin- (Cubist Pharmaceuticals Inc. 1998).

In this work, we have designed and screened a virtual library of ~1800 compounds mimetizing the structure of the lysyl-adenylate intermediate. The library was docked both to the homology-based model of the apicoplastic PfKRS-2 and to its human homologue. The positive predictions were used to synthesize a solid-phase combinatorial library consisting of 48 compounds, which were *in vitro* tested using *P. falciparum* cell cultures. From these, the

two best-performing compounds, M-26 and M-37, were chosen as drug candidates for further *in vitro* and *in vivo* analyses. We confirmed by aminoacylation assays that the inhibitory activity of the compounds was due to the specific inhibition of the apicoplastic PfKRS-2. Indeed, enzymatic assays indicate that the aminoacylation activity of the human KRS is being minimally affected, showing that these compounds selectively inhibit *Plasmodium* apicoplastic KRS (**Publication 3**).

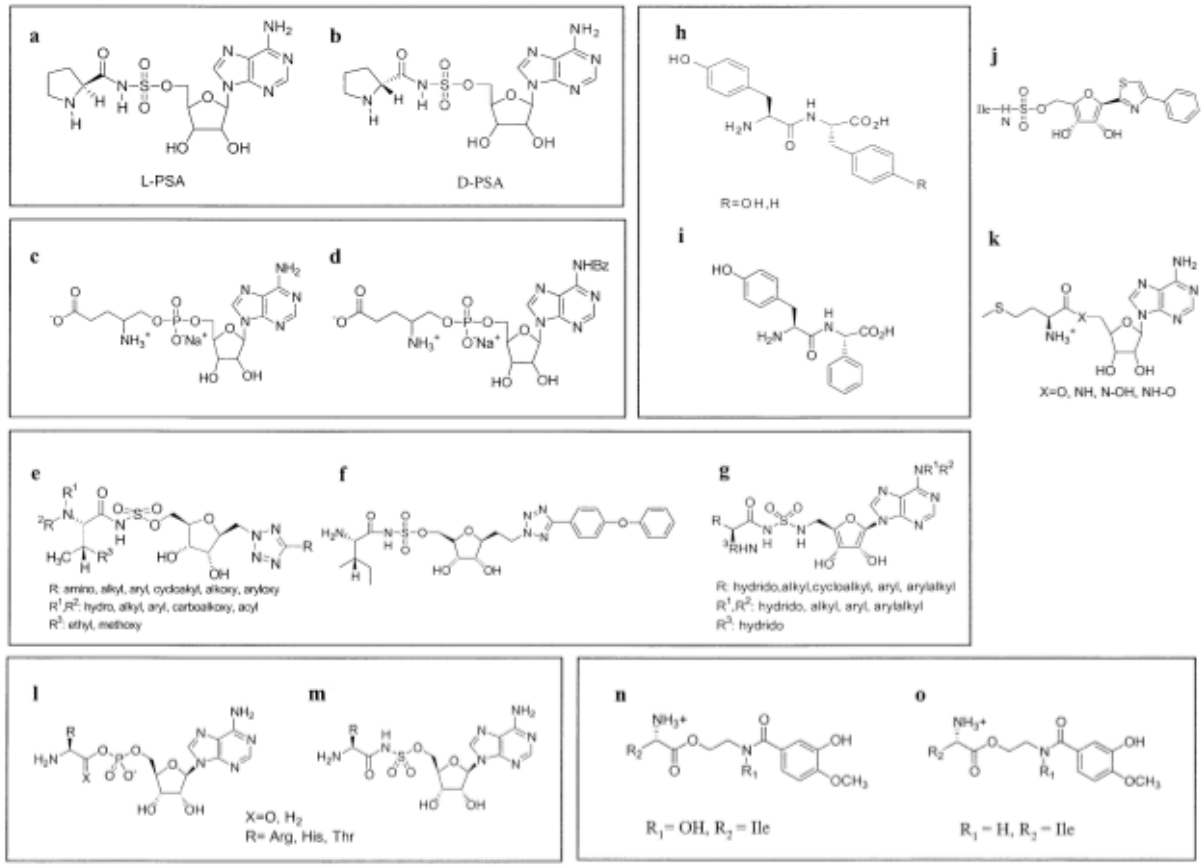


Figure 4.13. Analogues of the reaction intermediate. **A**) and **B**) Prolyl-adenylate analogues inhibiting both bacterial and mammalian ProRS (Heacock et al. 1996). **C**) and **D**) Glutamyl-adenylate analogues (Desjardins et al. 1998). The replacement of the N6 position of the adenine reduced the inhibitory activity of the compound. **E**) **F**) and **G**) Isoleucyl-adenylate analogues (Hill et al. 1998). **H**) and **I**) Tyrosyl-aryl dipeptides (Jarvest et al. 1999). **J**) Thiazole adenylate mimics inhibiting LeuRS (Yu et al. 1999). **K**) Methionyl-adnylate analogues (Lee et al. 1999). **L**) and **M**) Aminoalkyl- and sulfamoyl-adenylates of arginine, histidine and threonine (Forrest et al. 2000). **N**) and **O**) Hydroxamate derivatives of the isoleucyl- and methionyl-adenylate intermediates (Lee et al. 2001).

4.2.2.2 Analogues of natural aaRS inhibitors

A number of natural products have been discovered that inhibit the activities of aaRS (**Figure 4.14**). Besides the activity of pseudomonic acid –or mupirocin-, which inhibits IleRS, other known natural-product inhibitors are directed against aaRS, such as borrelidin (Nass et al. 1969); furanomycin (Tanaka et a. 1969); granaticin (Ogilvie et al. 1975); indolmycin (Werner et al. 1976); ochratoxin A (Konrad and Roschenthaler, 1977); cispentacin (Konishi et al. 1989) and purpuromycin (Kirillov et al. 1997). However, none of these aaRS inhibitors have been yet commercially developed for different reasons.

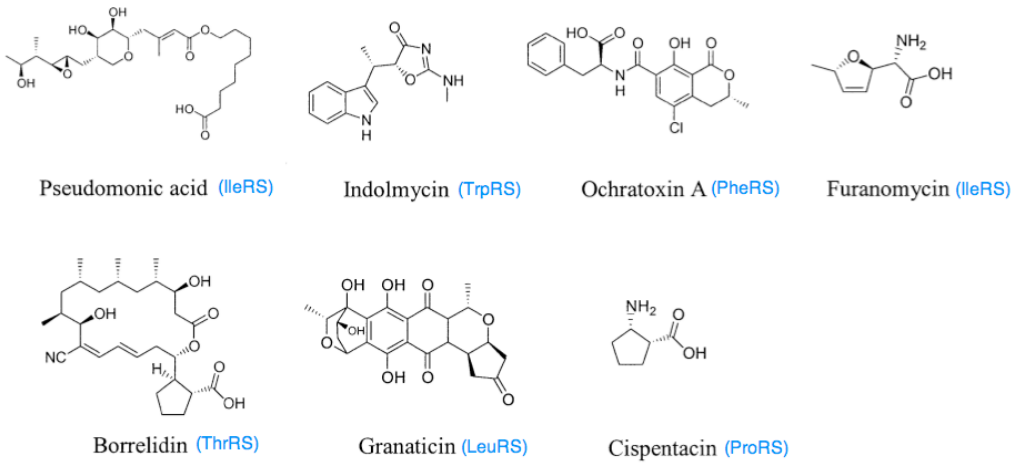


Figure 4.14. Natural aaRS inhibitors. In blue is shown the specific aaRS that is inhibited by each compound. Adapted from Kim et al. 2003.

In vivo, the ester bond of mupirocin is rapidly hydrolyzed, resulting in monic acid, which is inactive, and therefore it is only employed for topical use. Many attempts have been made to develop mupirocin derivatives with desirable properties for systemic use. Several laboratories designed mupirocin analogues with other substituents replacing the ester bond, but the resulting analogues presented higher MIC values, probably due to poor penetration into the bacterial cell (**Figure 4.15**) (Brown et al. 1997). Additional analogues were synthesized, which resulted in a lower potency but better pharmacokinetic properties. At GlaxoSmithKline (GSK), a docking model based on a crystal structure of IleRS with mupirocin was used for the rational design of new inhibitors. The introduction of an Ile side chain with appropriate spacing to the monate ring yielded femtomolar inhibitors due to a gain in binding energy in the Ile-binding pocket (Brown et al. 2000).

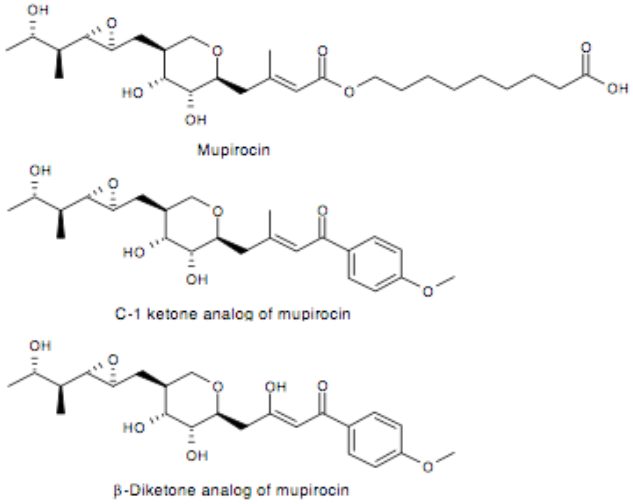


Figure 4.15. Molecular structures of mupirocin and some of its analogues. On the top, structure of mupirocin, which is the only commercially available aaRS inhibitor. In the middle and bottom, mupirocin analogue structures are shown, taken from the work published by Brown et al.1997.

In order to explore the potential of the aminoacyl-tRNA synthetase family as a source of antimalarial drug targets, we have treated *Plasmodium falciparum* cultures with a battery of known and novel aaRS inhibitors, and compared their activities. Amongst the compounds tested, borrelidin, a natural inhibitor of threonyl-tRNA synthetase (TRS) stands out for its potent antimalarial effect. Despite its promising antimalarial activity, borrelidin also inhibits human TRS, and is highly toxic to human cells. To circumvent this problem, we have explored the antimalarial activities of a library of borrelidin derivatives, and evaluated their cytotoxicity in human cells. We find that some of these compounds present higher selectivity towards the *P. falciparum* enzyme, whilst maintaining their antiparasitic activity both *in vitro* and *in vivo*. We propose that borrelidin is a promising antimalarial scaffold that should be further explored for the search of novel antimalarial drugs (**Publication 4**).

4.2.2.3 Drugs disrupting the tRNA interaction

The examples above follow the tradition of targeting drugs to the catalytic sites of enzymes. For many metabolic enzymes, the binding substrate and product is imbedded within the catalytic site, because the ligands themselves are small. In the case of aaRS, the amino acids and ATP binding sites are found in the catalytic site, but only the 3' end of the tRNA is

accommodated into the active site. Thus, a compound that blocks these protein-RNA interactions that occur out of the catalytic site would be also a potential drug.

Interestingly, in many cases eukaryotic and prokaryotic aaRS have different recognition mechanisms for the tRNA molecule. This fact is reflected in the kingdom-specific aminoacylation of tRNAs. For example, eukaryotic tyrosyl-tRNA synthetase does not acylate bacterial $tRNA^{Tyr}$, and vice versa. The reason for this species specificity is due in large part to the difference of a single base pair near the acceptor end of $tRNA^{Tyr}$: C:G for eukaryote and G:C for bacterial $tRNA^{Tyr}$. Interchange of this base pair switches the species specificity of acylation, so that the eukaryote enzyme is now capable of charging the bacterial substrate and vice versa (Quinn et al. 1995).

Similarly it has been proposed that drugs blocking the hinge movement occurring between the anticodon binding domain and the catalytic domain upon tRNA binding would also inhibit the aminoacylation reaction by restricting the inter-domain movement required for tRNA binding. The lower level of conservation between pathogenic aaRS and its human homologues out of the catalytic site make these drugs have lower chances of cross-reactivity.

In this regard, we attempted to find compounds that could inhibit the hinge movement between the catalytic and the anticodon-binding domain of PfKRS-2. First, we predicted druggable regions of the protein (SiteMap, Schrödinger), finding that the hinge was amongst the top-ranked druggable regions of the protein. Then, to explore the conformational variability of the PfKRS-2 hinge, we performed molecular dynamics (MD) using GROMACS 3.3. The MD simulation included: i) preparation of the structure, ii) MD running (100ns), iii) quality assurance of the MD run, and iv) structural analysis of the results. A principal component analysis of the MD simulation allowed us to conclude that the residues involved in the hinge movement were amongst the most fluctuating of the protein (**Figure 4.16**).

Figure 4.16. PCA analysis confirms the hinge movement as principal movement of the protein. Each residue has been heat-palette coloured according to its root mean square fluctuation (RMSF). Besides the N-ter and C-ter ending residues, the two regions with highest movement are a loop found in the catalytic domain, and the hinge found between the catalytic and the anticodon binding domain.

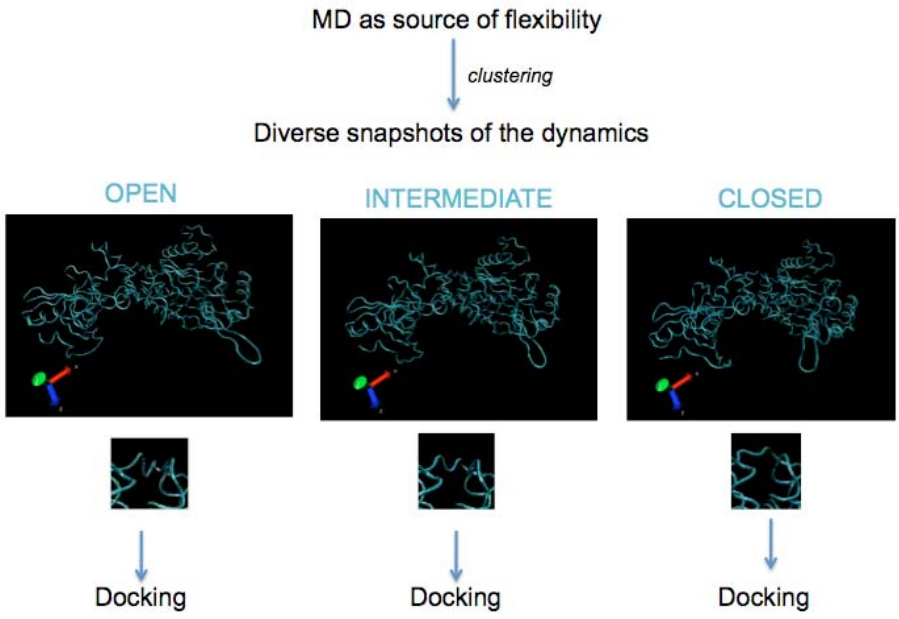


Figure 4.17. Docking workflow used to include the hinge flexibility in the docking run.

In order to include flexibility in our dockings, we selected a diverse subset of snapshots to be docked. For this, a clustering method was performed on the MD simulation, and the centroid of the 6 most populated clusters was selected as a member for the protein ensemble (**Figure 4.17**). Each protein conformation of the ensemble was docked against a virtual library of ligands that was built using different commercially available databases, which included the Prestwick Chemical Library, FDA drugs repository, the E-molecules dataset, and ZINC drug-like library (Irwin and Shoichet, 2005). We selected the 25 top-ranked candidate inhibitors for further *in vitro* testing, although this part has not been finished yet (**Figure 4.18**). To confirm the mode of action a compound *in vitro,* it should inhibit the aminoacylation reaction, but not the amino acid activation -which can be specifically measured using a PPi exchange assay- (Fersht et al., 1976).

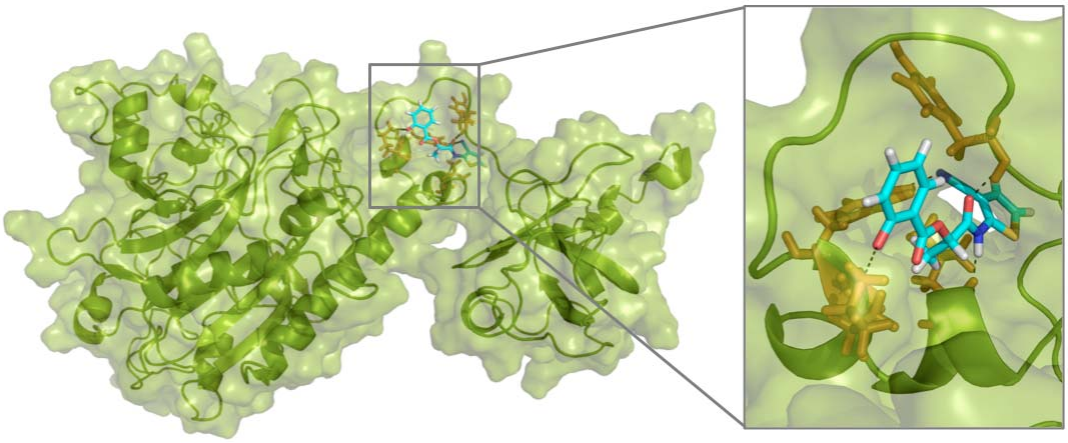


Figure 4.18. Predicted binding mode of a potential hinge inhibitor, as determined by docking. Cartoon representation of PfLysRS-2 (green), highlighting the hinge residues involved in the ligand recognition (orange). The ligand is shown in cyan.

4.2.2.4 Inhibitors of the aaRS proofreading activity

Many aaRS enzymes possess a proofreading (editing) mechanism that hydrolyzes tRNAs aminoacylated with the incorrect amino acid (Schimmel and Schimdt, 1995). These editing domains, which are separated of more than 30Å of the catalytic (aminoacylation) site, can be specifically inhibited, causing the death of the pathogen.

Amongst the few compounds described to specifically target aaRS editing domains are the benzoxaboroles, such as the well-described AN2690 (Rock et al. 2007) (**Figure 4.19**). These compounds have an usual chemical attribute: a boron atom. They specifically inhibit the leucyl-tRNA synthetase (LeuRS) editing site by forming an adduct with the terminal adenosine (A76) of $tRNA^{Leu}$ in the editing active site, thus trapping $tRNA^{Leu}$ in the editing active site

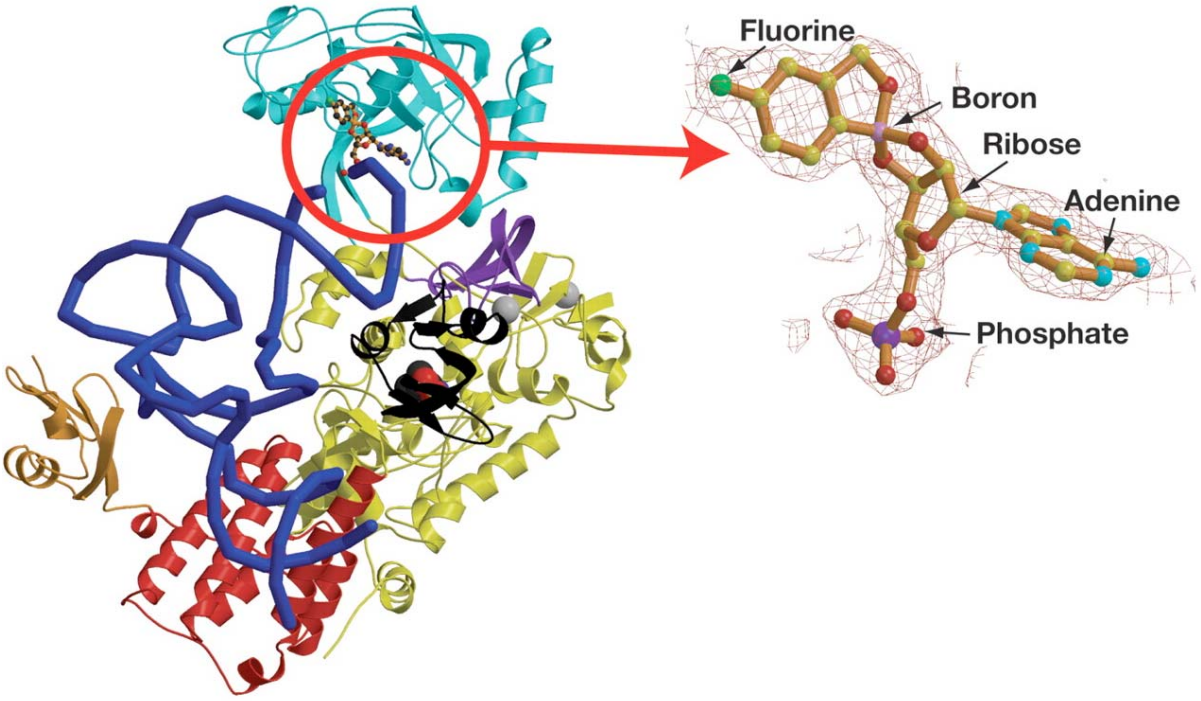


Figure 4.19. Binding mode of AN2690, as shown by X-ray crystallography. The compound forms a covalent bond with the terminal adenosine nucleoside of $tRNA^{Leu}$ in the editing active site of LeuRS. Each domain of the LeuRS has been coloured independently: editing domain (cyan) catalytic domain (yellow), Zn-domain (purple), LeuRS specific insertion (black), anticodon-binding domain (red) and C-terminal domain (gold). The tRNA structure is shown in blue tube. Adapted from Rock et al. 2007.

4.2.2.5 Virtual screens and structure-based design

Crystal structures have been determined for many aaRS, often complexed with substrates inhibitors of aminoacyl-adenylate intermediate analogues (Cusack, 1997). Such structural information has been used to screen for or optimize compounds that exhibit potent aaRS inhibitory activity (**Table 4**). For instance, a virtual screening of 500.000 compounds yielded 91 potential MetRS inhibitors with novel scaffold, and the most potent of them had an IC_{50} of 237nM (Kim et al. 2006).

Table 4. aaRS inhibitors identified through virtual screening or structure-based drug design

Screening/compound	Target	Inhibitor characteristics	Reference
Virtual screen hits	MetRS (*Escherichia coli*)	IC_{50}: 237 nM	Kim et al., 2006
Virtual screen hits	AsnRS (*Brugia malayi*)	Micromolar inhibition	Sukuru et al., 2006
Molecular modeling	LysRS (*Treponema pallidum*)	Selective for class I	Rao et al., 2006
Molecular modeling	LysRS (*Borrelia burgdorferi*)	Selective for class I	Ambrogelly et al., 2005
Molecular modeling	TyrRS (bacterial)	Selective over eukaryotic	Austin and First, 2002
Glutamyl-sulfamoyl-adenosine	GluRS (*E. coli*)	Ki: 2.8 nM (*E. coli* GluRS) Ki: 70 nM (mammalian enzyme)	Bernier et al., 2005
Tyrosinyl-adenylate	TyrRS (*S. aureus*)	IC_{50}: 11 nM	Brown et al., 1999
Isovanillate-hydroxamate	IleRS (*E. coli*)	IC_{50}: 4.5uM	Lee et al., 2001
Methionyl-adenylate analogues	MetRS (*E. coli*)	IC_{50}: 0.4-2.4nM	Vaughan et al., 2005

As previously stated, drug resistance to available drugs is a major problem causing that many antimalarials end losing efficacy. That is why, in a second part of this project, we tried to target this problem by designing dual inhibitors that target two different aaRS (multisynthetase inhibitors). The existence of structurally conserved residues across related

aaRS provides a realistic opportunity for the discovery of a single molecule that simultaneously inhibitrs multiple enzymes. Such molecules could be of major clinical importance, since the pathogen would require simultaneous point mutations within each drug target to become resistant, which is an unlikely event. However, we must consider that multisynthetase inhibitors may be susceptible to other strategies of antimicrobial drug resistance, including target up-regulation, reduced permeability, drug efflux or drug modification systems.

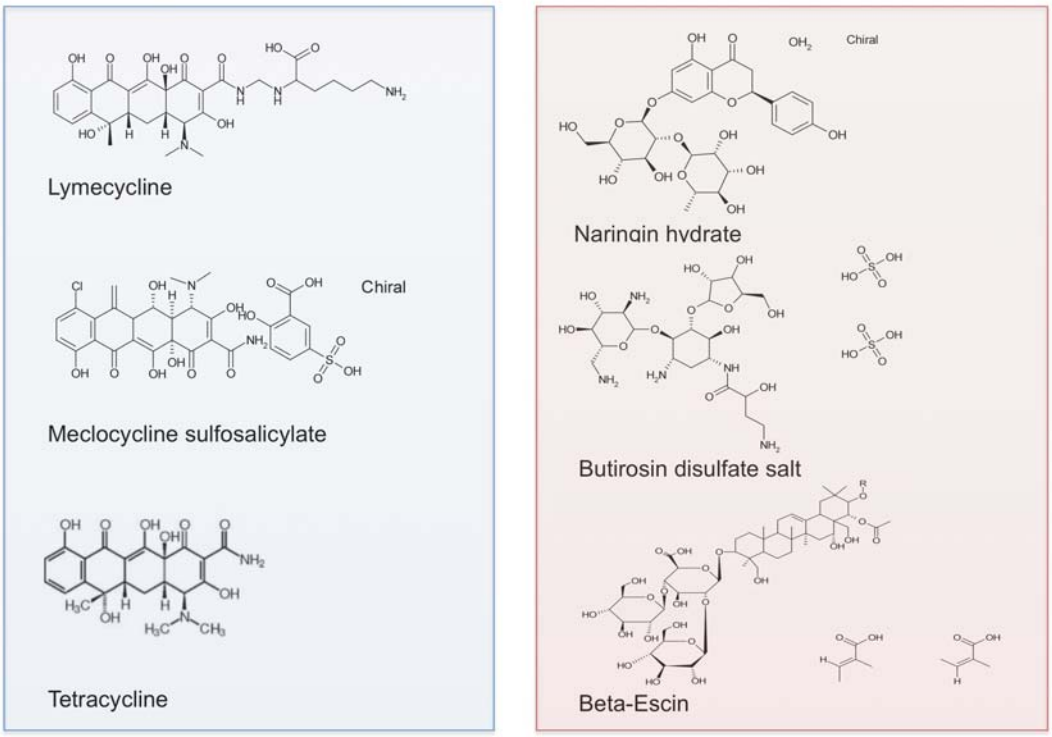


Figure 4.20. Dual hits targeting aaRS, as predicted from high-throughput docking

Using a high-throughput docking strategy, more than 300 molecules were predicted to be candidate inhibitors of the two aaRS docked (PfQRS and PfKRS-2). From these, nine showed dual specificity for the two enzymes. Interestingly, half of these dual hits were tetracycline derivatives (**Figure 4.20**). Tetracyclines are commonly used wide-spectrum antibiotics that bind to the 30s ribosomal subunit of bacteria, by preventing the docking of aminoacylated tRNA to the ribosome. Thus, it is not surprising that aminoacyl-tRNA

synthetases, which also bind the aminoacylated tRNA could be also the target of these drugs. We initially tested a battery of tetracyclines on *P. falciparum* cultures, finding that these compounds inhibited the parasite with a delayed death, probably due to the inhibition of the apicoplastic 30s ribosomal subunit (**Figure 4.21**).

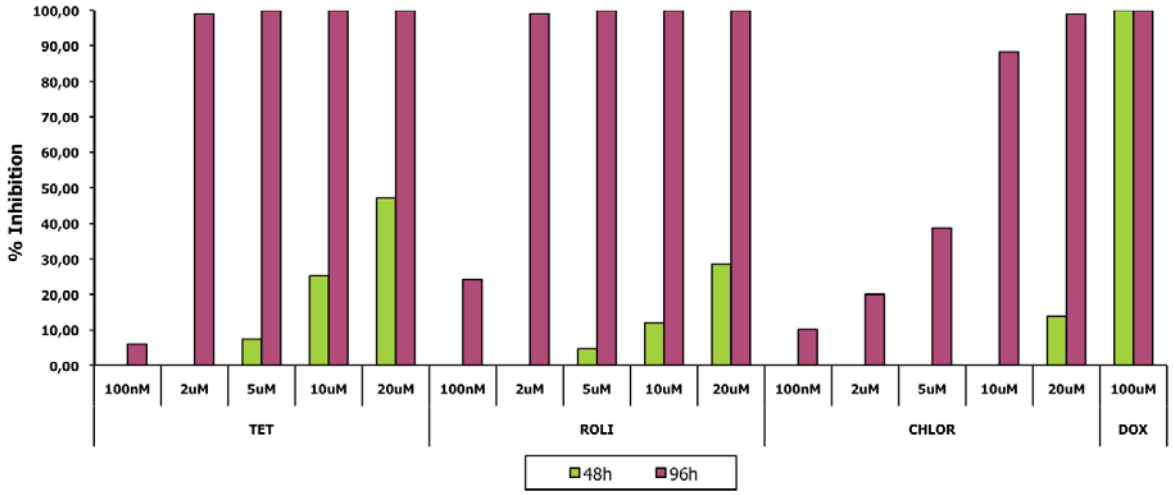


Figure 4.21. ***In vitro*** **activities of a battery of tetracyclines tested** ***in vitro*** **on** ***P. falciparum*****-iRBCs.** The inhibition has been measured both at 48h first life cycle, inhibition of the cytosolic translation machinery- (shown in green) and at 98h -second life cycle, inhibition of the apicoplastic translation machinery- (shown in red). Inhibition measured from Giemsa-stained smears counting. Abbreviations: TET, tetracycline; ROLI, rolitetracycline; CHLOR, chlortetracycline; DOX-doxycycline.

Guided by the predicted binding modes of the docked tetracycline molecules, we have designed modifications on the tetracycline scaffold in order to increase both the selectivity and the binding affinity of these molecules towards the aaRS (**Figure 4.22a**). These modifications mainly consist in the inclusion of aminoacyl side chains, such that they specifically enter the amino acid pocket of the catalytic site.

Interestingly, the aminoacyl-adenylate intermediates present two different binding modes depending on whether they correspond to class I or class II. Class I aminoacyl-adenylates place the amino acid side chain pointing down with respect to the N6 position of the adenine ring, whereas class II aminoacyl-adenylates place the amino acid side chain pointing left with respect to the same position (**Figure 4.22b**). Thus, these differences in binding modes allow designing tetracyclines derivatives that target both class I and class II aaRS simultaneously. In our case, PfQRS is a class I enzyme, whereas PfLysRS-2 is a class II enzyme.

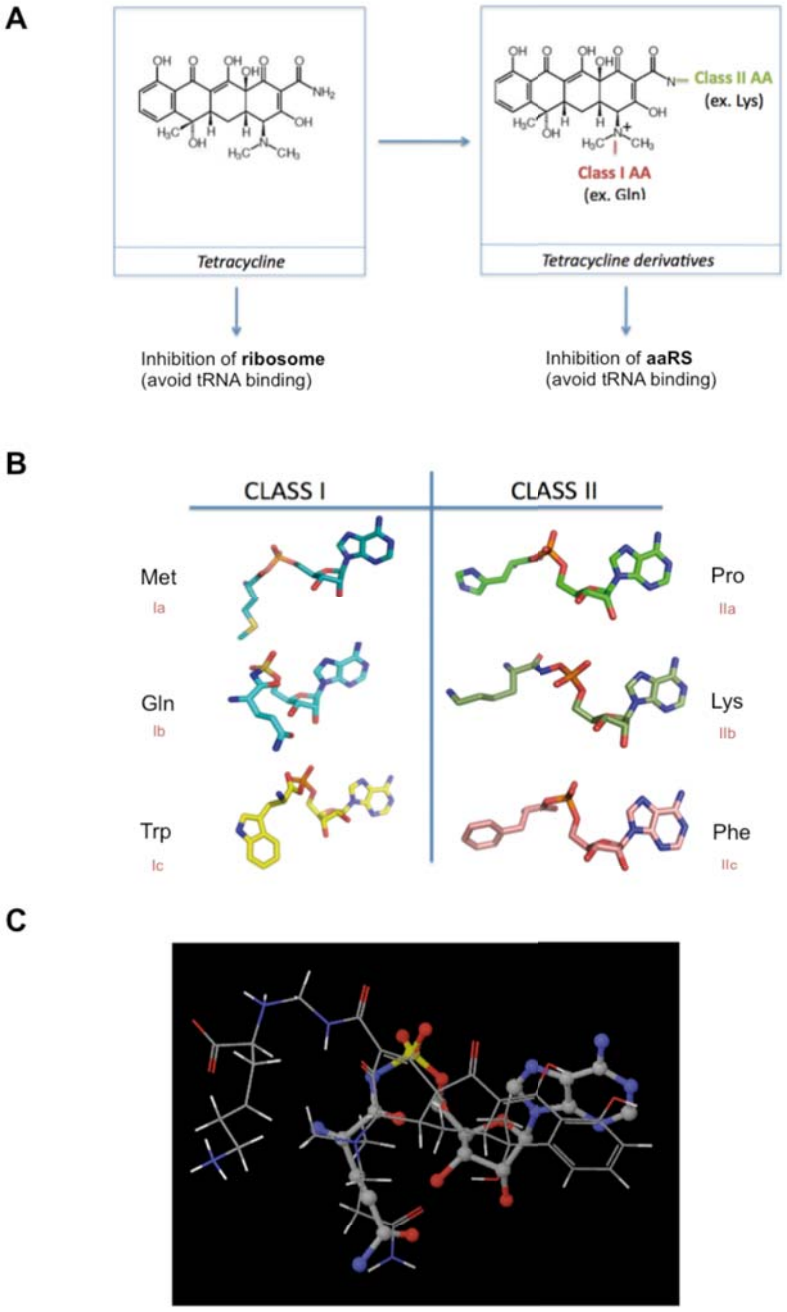


Figure 4.22. Structure-based drug design strategy to increase the binding affinity of tetracycline derivatives towards aaRS. **A**) Modifications added to the tetracycline derivatives. **B**) Binding modes of class I and class II aminoacyl-adenylate analogues. **C**) Predicted binding mode of the natural ligand -glutaminyl-adenylate- (shown in sticks) on the PfQRS binding pocket, and predicted binding mode of the proposed tetracycline derivative (shown in lines) also on the PfQRS binding pocket. The binding score of the modified tetracycline (GS=-12.65) is increased with respect to the binding score of the initial tetracycline (GS=-10.55).

The binding mode of the tetracycline derivative is very similar to the predicted binding mode of the initial tetracycline, but the addition of the amino acid side chain -which enters the amino acid side chain such as in the case of the aminoacyl-adenylate- (**Figure 4.22c**) increases the binding energy of the ligand, thus potentially changing the binding preferences of this compound towards aaRS. However, the synthetic chemistry to build these molecules has been problematic, and these molecules are still not available for testing.

4.2.2.6 High-throughput screening programs

Recently, many efforts have focused on the target-based approach that utilizes high-throughput screening assays (HTS). This technology has been applied to identify novel synthetase inhibitors from large compound libraries. In fact, the similar activities shared by the tRNA synthetases allow the utilisation of one kind of *in vitro* assay to screen all 20 synthetases (Tao and Schimmel, 2000).

For instance, using this strategy GSK scientists identified antibacterial pyridones and pyrimidones that specifically inhibit methionyl-tRNA synthetase (MetRS) and were selective against the mammalian enzyme (SmithKline Beecham PLC 2000a); and also a series of benzimidazole derivatives with antibacterial activity (SmithKline Beecham PLC 2000b), all of them with IC_{50} values in the nanomolar range. Furthermore, they also disclosed a novel class of substituted quinolones that are potent inhibitors of bacterial MetRS (SmithKline Beecham PLC 1999). Although quinolone derivatives are not structurally related to methionine, 3D quantitative structure-activity relationships (3D-QSAR) have shown that these molecules compete with methionine for important binding interactions in the amino acid binding pocket of MetRS (Kim and Lee, 2003).

Cubist Pharmaceuticals reported other novel pyrazoles (Finn et al. 2003) and proline derivatives (Finn et al. 2001), which were potent and selective inhibitors of *S.aureus* MetRS. Other scaffolds of aaRS inhibitors described include compounds as diverse as spirocyclic tetrahydrofurans (Hill et al. 2001) or thiazolidinones.

This type of strategy presents several advantages: i) it does not need *a priori* knowledge on any known drugs inhibiting the pathogen; ii) it does not require any knowledge on the structure of the drug target; iii) it is an efficient manner to discover novel drug scaffolds

inhibiting a given pathogen; and iv) it allows to screen thousands to millions of compound libraries in an automated manner, covering a wide range of the chemical space compared to low-throughput strategies. The main drawback, however, is that once an active compound is found, the target of the drug remains unknown, which makes further hit-to-lead optimisation a difficult task for the medicinal chemists.

To experimentally uncover the target of a given compound, assays such as the haploinsufficiency-profiling assay (HIP) (Giaever et al. 1999; Hoepfner et al. 2012) can be used. The HIP target discovery assay is based on a genome-wide collection of heterozygous knockout yeast strains, each of which contains a marked gene deletion (Winzeler et al. 1999). It has been shown that heterozygous diploid strains that bear a deletion in one copy show increased sensitivity to a drug compared to those strains that have two copies of the gene (Giaever et al. 1999). However, the drawback of this technique is that the given drug must inhibit the same target in yeast, and at the same time, if the yeast target is being inhibited, there are higher possibilities that the human counterpart is also inhibited.

4.2.2. Publications

PUBLICATION 3:

***Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum*.**

Hoen R*, **Novoa EM***, López A, Camacho C, Cubells L, Martin P, Bautista JM, Vieira P, Santos M, Cortes A, Ribas de Pouplana L and Royo M.

** equal contributors*

J Med Chem (under review)

Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from *Plasmodium falciparum*

Rob Hoen[†,#], Eva Maria Novoa[§,#], Alba López[†], Noelia Camacho[§], Laia Cubells[§], Pedro Vieira[&], Manuel Santos[&], Patricia Marin-Garcia[⊥], Jose Maria Bautista[⊥], Alfred Cortés[§], Lluís Ribas de Pouplana[*,§,±] and Miriam Royo[*,†]

[†] Combinatorial Chemistry Unit, Barcelona Science Park, University of Barcelona, C/ Baldiri Reixac 10, 08028 Barcelona, Catalonia, Spain.

[§] Institute for Research in Biomedicine, C/ Baldiri Reixac 10, Barcelona 08028, Catalonia, Spain.

[±]ICREA. Passeig Lluís Companys 1, Barcelona 08010, Catalonia Spain.

[&]RNA Biology Laboratory, Department of Biology and Centre for Environmental and Marine Studies (CESAM), University of Aveiro, Aveiro, Portugal.

[⊥] Department of Biochemistry and Molecular Biology IV, Complutense University of Madrid, Madrid, Spain.

[#] These authors have contributed equally to this paper.

*Corresponding authors

Contact Information:

For L.RdP.: telephone, +34 934034868; e-mail, lluis.ribas@irbbarcelona.org. For M.R.: telephone, +34 934037120; e-mail, mroyo@pcb.ub.cat.

ABSTRACT

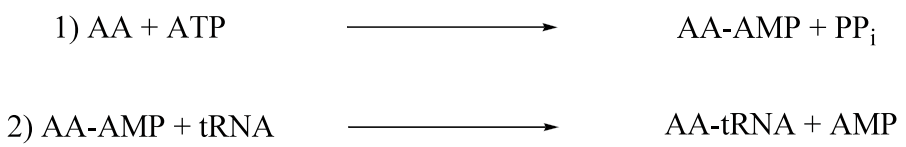
Resistance of malaria parasites to available drugs continues to grow, making the need for new antimalarial therapies pressing. In this regard aminoacyl-tRNA synthetases (ARS) constitute a promising set of targets to develop novel antimalarials. ARS are essential enzymes and proven antibacterial targets whose ancestral nature facilitates the development of specific inhibitors. The cyanobacterial origin of the apicoplast, a relict plastid common to all Apicomplexa that is essential for *Plasmodium*, is reflected in its bacterial-like enzymes (including the ARS). Despite their potential as drug targets, apicoplastic ARS remain unexplored. Here we demonstrate that selective inhibition of apicoplastic ARS is feasible, and describe new compounds that show antimalarial activity and specifically inhibit *Plasmodium* apicoplastic lysyl-tRNA synthetase.

INTRODUCTION

Malaria remains one of the most important infectious diseases in the world, causing acute illness on more than 100 million people and leading to approximately 1 million deaths annually[1]. In addition to its human cost, malaria causes a massive economic burden, contributing substantially to poverty in the developing world. Effective antimalarial drugs are available, but their efficacy is compromised by emerging resistance[2]. Thus, there is a broad consensus about the need to develop new antimalarial drugs. Malaria is caused by *Plasmodium*, a genus of parasitic protists. At the moment there are over 200 species known of this genus, of which at least 11 can infect humans. Amongst them, *Plasmodium falciparum* causes the most severe form of malaria, being responsible for 90% of the deaths[1].

The *P. falciparum* genome project revealed many new potential drug targets[3-9], of which several are enzymes acting in the apicoplast, a relict plastid derived from secondary endosymbiosis of cyanobacteria[10] which is essential for the parasite's survival[11,12] . Many of its bacterial-like enzymes are substantially different from its mammalian homologues[13-15], making them excellent drug target candidates. Several antibacterial drugs that are clinically used for the treatment of malaria and toxoplasmosis (e.g. doxycycline, clindamycin and spiramycin) act upon apicoplastic targets. These drugs typically display a "delayed death" phenotype, which is characterized by the inhibition of parasite growth on the second erythrocytic cycle after the drug treatment[16-20].

Aminoacyl-tRNA synthetases (ARS) are essential enzymes and proven antimicrobial drug targets[21,22], and thus represent interesting novel targets for antimalarial drug discovery[23]. They perform a central role in the translation of the genetic code by catalyzing the attachment of each amino acid to its cognate transfer RNA (tRNA). Although these enzymes differ widely in size, sequence, and oligomeric state, they all carry out a similar two-step reaction[22]. In a first step, the ARS catalyzes the activation of the amino acid, and in a second step, the aminoacyl-adenylate intermediate (AA-AMP) is transferred to the tRNA.



Currently, ARS inhibition is the mechanism of action of one commercial antibiotic, *i.e.* pseudomonic acid or mupirocin (GSK), a natural product that inhibits bacterial isoleucyl-

tRNA synthetases with an 8000-fold selectivity over their mammalian homologues. Mupirocin has also been shown to inhibit apicoplastic isoleucyl-tRNA synthetase of *Plasmodium*[24]. Other ARS inhibitors described to date include natural products, such as borrelidin[25,26], granaticin[27], indolmycin[28] furanomycin[29], ochratoxin A[30], cispentacin[31], and several semisynthetic products[32-34]. Most efforts on the design of new synthetic drugs targeting ARS have focused on mimicking the aminoacyl adenylate intermediate (AA-AMP)[21,35-37]. Finally, it has recently been reported that cladosporin, a fungal secondary metabolite, targets *Plasmodium falciparum* cytosolic lysyl-tRNA synthetase (PfKRS-1) with a selectivity of 100-fold with respect to its human homologue[38].

Indeed, the main challenge in using ARS as drug targets is to avoid cross-reactivity with their human homologues. In this regard, apicoplast-specific *P. falciparum* lysyl-tRNA synthetase (PfKRS-2) is interesting because its cyanobacterial origin makes it evolutionarily distant from human lysyl-tRNA synthetase (HsKRS). In this work we present a new series of compounds which selectively inhibit apicoplastic PfKRS-2, thus validating its potential as antimalarial drug target, and demonstrating that specific inhibition of apicoplastic ARS is feasible.

RESULTS

Characterization of the lysylation system in *Plasmodium falciparum*

Malaria parasites possess two distinct lysyl-tRNA synthetases, PfKRS-1 (PF13_0262), and PfKRS-2 (PF14_0166). Based on subcellular localization prediction software[11], PfKRS-1 is expected to be cytosolic, whereas PfKRS-2 is expected to be targeted to the apicoplast (Figure 1A). Immunofluorescence assays on PfKRS-2_{leader}-GFP transfected *P. falciparum* parasites (Figure1B) indicate that, as expected, PfKRS-2 is exclusively located in the apicoplast.

In general, apicoplastic-targeted enzymes tend to be of bacterial origin[39]. To confirm the bacterial origin of PfKRS-2 and evaluate its evolutionary distance with its human homologue, we performed a structure-based phylogenetic analysis on class II lysyl-tRNA synthetases from all kingdoms (Figure 1C). Our results show that apicoplastic lysyl-tRNA synthetases cluster with bacterial enzymes and are distantly related to HsKRS.

Using a manually curated homology model of PfKRS-2 (Figure S1) we noticed that those residues involved in the recognition of lysine in the bacterial, human, and *P. falciparum* enzymes are conserved across species. Importantly, however, other residues in the active site cavity that are not involved in substrate recognition are not so well conserved, and the sizes of the catalytic cavities are significantly different (Figure 1D). This suggests that PfKRS-2 might be able to accommodate ligands in the active site that may not be able to bind in the HsKRS cavity due to sterical restrictions. Altogether, our analyses suggest that specific design of inhibitors targeting the active site of PfKRS-2 is feasible.

Design, selection and synthesis of a library of lysyl-adenylate analogues

A compound virtual library was designed to identify molecules that may mimic the lysyl-adenylate intermediate (Figure 2). To construct the library, a proline derivative was used as a ribose mimetic, and an heterocycle as an adenylate substitute, as previously described[36]. In addition, four more points of chemical diversity were explored: i) both lysine and thialysine derivatives were used as lysine analogues; ii) the phosphate linker was replaced by other types of chemical linkers; iii) heterocyclic groups were used as substituents of the adenylate moiety, and iv) the stereochemistry of the proline and the lysine derivatives was varied. With this approach a library of 1764 compounds was designed and evaluated by docking the molecules against the 3D structures of both PfKRS-2 and HsKRS.

All 1764 compounds were docked both to PfKRS-2 and HsKRS, and the different ligand poses obtained were ranked using GlideScore[40]. The compounds to be synthesized for experimental testing were selected on the basis of their selectivity towards the PfKRS-2 enzyme. By this criterion we selected 36 compounds for further analysis (Table S1).

Amongst the 36 compounds selected, 70% contained an hydroxymate group as phosphate analogue, and a proline ring with an (*S*,*S*) configuration. Thus, a library of 50 lysyl-adenylate analogs based on the (*S*,*S*)-4-amino proline scaffold was designed using a hydroxymate group as a phosphate linker mimic. 25 compounds contained lysine, while the other 25 carried thialysine (Table 1, see also Figure 2). In addition to the predicted hits, a number of predicted non-selective and non-active compounds were also synthesized to evaluate the performance of the docking calculations (Figure S2B).

The library of potential PfKRS-2 inhibitors was synthesized employing an Alloc/Boc strategy based solid-phase synthesis[36]. Coupling of the Alloc-protected hydroxyproline to the resin was followed by the introduction of the protected lysine or thialysine hydroxamic acid moiety under Mitsunobu conditions. After removal of the Alloc group with $Pd(PPh_3)_4$ and $PhSiH_3$ introduction of the different carboxylic acids was carried out under standard peptide coupling conditions. Subsequent cleavage and protecting group elimination of the products under strong acidic condition produced the crude inhibitors. Purification by preparative HPLC yielded the desired products with purities of ≥ 85% (Figure 3A). A number of thialysine-derived products were not obtained due to degradation of the products during purification. The obtained products were subjected to biological evaluation.

In vitro testing of the compounds

All synthesized compounds were initially tested for their ability to kill *P. falciparum* parasites using the pLDH assay[41]. Inhibitors of apicoplastic protein synthesis kill the parasite in a retarded manner[16-18], and therefore, we used the "delayed death" phenotype as an initial indication that a compound in our library may be preferentially targeting apicoplastic lysyl-tRNA synthetase. Our initial screening allowed us to select five compounds that presented a clear delayed inhibitory effect (Table 2). The activity of these compounds was further confirmed by visual inspection of smears.

The five most active compounds from the library (**M-12**, **M-24**, **M-26**, **M-33** and **M-37**) were re-synthesized. Solution synthesis was used to improve purity and yields, and to minimize possible side-reactions (Figure 3B). All products were obtained in purities of >98.5%. The

antimalarial activity of the re-synthesized compounds was evaluated by visual analysis of *P. falciparum* smears. Highest inhibition rates were observed for compounds **M-12**, **M-33** and **M-37** (Table 2). In order to select specific inhibitors of the apicoplastic translation machinery it was decided to focus on those compounds causing a clear delayed effect phenomenon. Thus, **M-26** and **M-37** were chosen as drug candidates for further *in vitro* and *in vivo* analyses, given that these compounds show maximal difference between the inhibitory rates of these compounds at 48 and 96h.

In order to investigate the selectivity and specificity of **M-26** and **M-37**, we first verified their ability to inhibit HsKRS, PfKRS-1, and PfKRS-2. *In vitro* aminoacylation assays were performed using radiolabelled lysine and *in vitro* transcribed $tRNA^{Lys}$, and the effect of the compounds upon these aminoacylation reactions was quantified. Both **M-26** and **M-37** were found to inhibit PfKRS-2, but were not active against HsKRS or PfKRS-1 (Figure 4), which is in accordance with the docking predictions. Thus, we can conclude that **M-26** and **M-37** are selective inhibitors of apicoplastic PfKRS-2.

Structural basis for selectivity

Through the analysis of the binding mode of the natural ligand (lysyl-adenylate; LAD), it was observed that both the adenine and lysine moiety of LAD are being recognized at the binding site, and are major contributors to the free energy of binding of the reaction intermediate. In agreement with this observation, analogues showing a lysyl-adenylate-like binding mode (Figure 5, see also Figure S2A) tend to present higher docking scores. Interestingly, both **M-26** and **M-37** present a LAD-like binding mode in PfKRS-2, whereas in HsKRS they present either an adenine-like or a lysine-like binding mode, respectively (Figure S2A). These diverse binding modes are due to differences in size of the active site cavities of the two enzymes. Whereas the PfKRS-2 cavity can accommodate the inhibitors maintaining the recognition of both the lysine and adenine moieties, the catalytic cavity of HsKRS cannot accommodate both moieties of these compounds at the same time.

DISCUSSION AND CONCLUSIONS

During the last years, cell-based screening has been presented as an attractive way to find new leads for malaria drug development. However, although these approaches are capable of identifying large numbers of hits, they also present serious limitations. For instance, if an initial hit is chemically untreatable, or its target is not known, there may be no possibility to proceed to hit-to-lead optimization. In this regard, the initial validation of targets based on chemoinformatic predictions can be a useful approach.

Aminoacyl-tRNA synthetases (ARS) have been recognized for decades as useful targets for drug design[42,43]. Indeed, ARS continue to be used as targets in antibacterial and antiparasitic drug discovery programs[34,44,45]. However, targeting the ARS of a microorganism without inhibiting the human counterpart remains a major challenge. The use of methods for phylogenetic inference helps to recognize targets whose evolutionary history may favor the identification of selective compounds.

Among the twenty-odd aminoacyl-tRNA synthetases LysRS represents an evolutionary exception, because this enzyme exists with a class I and a class II fold[46]. Previous analyses have proposed that these two enzyme forms co-existed before the emergence of the last universal common ancestor[47]. Interestingly, the endosymbiotic theory of the origin of eukaryotes implies that LysRS of two distinct folds perhaps coexisted in the first eukaryote ancestor.

Our phylogenetic results indicate that two class II LysRS evolved during the early maturation of the eukaryotic protein synthesis machinery. Our data (Figure 1C) shows that all eukaryotic cytosolic LysRS form a sister clade with all the mitochondria-specific enzymes. This supports the idea that a bacterial lysS gene (class II fold) functionally replaced its archaeal equivalent U (class I fold) in early eukaryotes, and duplicated to produce cytosolic- and mitochondrial-specific genes. This ancestral composition would then be inherited by all eukaryotic clades, evolving differently in each of them. Diplomonads, plants, and metazoans all lost the mitochondrial lysS gene, which was replaced in plants and metazoans by the cytosolic isozyme. By contrast, some protozoa and fungi retained distinct mitochondrial and cytosolic forms of lysS.

Remarkably, apicomplexan protozoa incorporated a second form of LysRS during the endosymbiotic event that gave rise to apicoplasts. This is evident from the phylogenetic

position assigned in our trees to the apicoplast LysRSs of *P. falciparum* and *P. yoelii* (Figure 1C). Whether the apicoplast ancestor was a green or a red algae is still a matter of debate[11,15], but our data indicates that some of the genes acquired from the apicoplast genome by *Plasmodium* could be direct descendants of bacterial symbionts.

Structural and evolutionary data strongly point at PfKRS-2 as a promising target for the development of inhibitors of the apicoplast metabolism in apicomplexan organisms. Here, we demonstrate that this enzyme can indeed be specifically inhibited. Using both computational and experimental approaches, we have built a series of lysyl-adenylate analogue inhibitors designed against apicoplastic lysyl-tRNA synthetases. We show that some of these molecules do inhibit PfKRS-2, and that the activity of its human homologue remains unaffected by these compounds. Based on *in silico* predictions, we identify PfKRS-2-specific features of the active site that can explain the selectivity of these compounds. Unfortunately, the analogues presented in this work do not possess *in vivo* antimalarial activity in *P. yoelii*-infected mice. Given the chemical nature of the compounds, their lack of activity is likely due to *in vivo* instability or degradation in the blood stream (Figure S3). Nevertheless, our work validates the apicoplastic lysyl-tRNA synthetase as a druggable enzyme that can be selectively inhibited, and therefore could be further explored for the development of novel antimalarial chemotherapies.

EXPERIMENTAL SECTION

Homology modeling of PfKRS-2

The sequence of the apicoplastic lysyl-tRNA synthetase of *P. falciparum* was retrieved from the PlasmoDB database (http://PlasmoDB.org). A PSI-BLAST[48] search was performed against the Uniprot database (http://www.uniprot.org) in order to obtain a Position-Specific Scoring Matrix, which was used as input to perform a new BLAST search against the PDB database (http://www.rscb.org), obtaining a list of candidate templates to build the model (1lyl, 1bbu, 1e1o). The templates were structurally aligned using STAMP[49] to create a profile using HMMER[50], which was introduced as meta-template for alignment with the target sequence. Finally, the 9v5 version of MODELLER was employed to create structural models using default options[51]. The models generated were manually refined using several methods, including corrections of the alignment using the PSI-PRED[52] secondary structure predictions as guideline, and then followed by a new rebuilding of the model. The final model was analyzed with ProSA[53], and validated using PROCHECK[54].

Virtual Screening and docking

Ligand screening and docking was performed with Glide 5.0[40]. Ligands were prepared such that several conformations were generated for each input ligand, using the LigPrep[55], facility of MAESTRO[56], while the set-up of the proteins (PfKRS-2 and HsKRS) was done with the Protein Preparation Wizard facility. The receptor grid defining the docking universe was defined by defining a cubic box centered on the lysyl-adenylate. Schrödinger's GlideScore scoring function was used to score the poses.

Solid-phase synthesis

All solid-phase syntheses were carried out manually in a polypropylene syringe fitted with a polyethylene porous disk. Solvents and soluble reagents were removed by suction. Peptide synthesis for this work employed a combined Boc/Alloc solid phase strategy on a Fmoc-Rink-Amide-MBHA resin. Washings between deprotection, coupling, and subsequent deprotection steps were carried out with DMF (5×1min) and DCM (5×1 min) using 10 mL of solvent/g of resin each time. All the couplings and Fmoc removal were monitored using the Kaiser test. See also Figure S4 and S5 and Supplementary Methods.

Synthesis in solution

Compared to the solid phase synthesis, the protecting group of the amine was changed to the UV-active *p*-nitrobenzyl carbamate (PNZ), which is more stable under acidic conditions and is more readily cleaved by hydrogenolysis than the related benzyl carbamate (Z). Straightforward introduction of the PNZ protecting group followed by the conversion of the carboxylic acid into the corresponding primary amide, using standard peptide coupling conditions, gave **60** in good yield. Subsequently, lysine or thialysine were introduced by a Mitsonubu reaction. Separation of the formed triphenylphosphine oxide from the corresponding products was laborious. In spite of intensive attempts, complete removal of the triphenylphosphine oxide from the desired product could not be achieved. It was decided to continue the synthesis using the mixture, since the triphenylphosphine oxide does not interfere with the outcome of the following reactions. Deprotection of the proline derivative by hydrogenolysis, using Pd/C under an H_2-atmosphere, followed by coupling of the corresponding carboxylic acids under standard peptide coupling conditions (WSC·HCl/HOBt·H_2O) gave products **65**-**69** in good yields. Full deprotection of **65**-**69** by 40% TFA in DCM and immediate purification of the crudes by preparative HPLC gave the final products (**M-12**, **M-24**, **M-26**, **M-33** and **M-37**) in good yields. The final products were characterized by standard techniques such as ^{1}H and ^{13}C NMR (Figure S5), exact mass and HPLC-MS.

Phylogenetic analysis of lysyl-tRNA synthetases

The sequences of lysyl-tRNA synthetases reported here are available in Uniprot[57]. We applied the method of structure-based alignment of the active sites of the enzymes, as described elsewhere[58]. Archaeal LysRS-II sequences were initially included in our analysis, but were later dropped for two reasons: a) they are generally believed to have emerged through lateral gene transfer events[59-62], b) in our initial analyses they clustered consistently within the bacterial clade. Phylogenetic distributions were calculated by distance and maximum likelihood methods using PHYLIP 3.68 package[63], using 1000 bootstrap replicates in the distance calculations and 100 bootstrap replicates for the maximum likelihood trees.

Cloning and expression of P. falciparum PfKRS-2

Soluble and active PfKRS-2 was obtained with the following procedure: the nucleotide sequence of the gene coding for PfKRS-2 without the predicted bipartite signal sequence (ΔPfKRS-2) was codon optimized for *E. coli*, synthesized (MrGene), and inserted into the plasmid pQE70. Expression screening was done using the In-Fusion™ based Vector Suite at IRB Protein Expression Core Facility. Different tagged-ΔPfKRS-2 constructions were built and inserted into both *E. coli* Rosetta and B834 strains. Three soluble proteins were finally selected for activity assays, namely ΔPfKRS-2-sumo, ΔPfKRS-2-His and ΔPfKRS-2-Z, which included a C-terminal sumo-tag, N-terminal His-tag and a C-terminal Z-tag, respectively. Final concentrations of the His-tag, Sumo-His-tag and Z-tag enzyme were 3.2 μM, 2.7 μM, 3.4μM, respectively. Final aminoacylation assays were performed with ΔPfKRS-2-His.

In vitro aminoacylation assays

To characterize the activity of our compounds, their effect upon the lysylation of $tRNA^{Lys}$ by lysyl-tRNA synthetase was tested. *In vitro* transcribed *P. falciparum* $tRNA^{Lys}$ was prepared as described[64]. Aminoacylation was performed at 37ºC in 100 mM Hepes-KOH (pH = 7.2), 20 mM $KCl_{(aq)}$, 30 mM $MgCl_{2(aq)}$, 0.5 mM $DTT_{(aq)}$, 5 mM ATP, 0.1 mg / ml BSA, 20 μM [^{3}H]lysine (500 Ci / mol) (Perkin Elmer) and 5 μM *in vitro* transcribed tRNA. Reaction aliquots were spotted on 3 mm filter disks and washed in 5% trichloroacetic $acid_{(aq)}$ with 100 μM $lysine_{(aq)}$. Radioactivity was determined by liquid scintillation counting. Aminoacylation rates of human lysyl-tRNA synthetase (HsKRS) were also determined to check for possible cross-reactivity of PfKRS-2 inhibitors towards the human homologue. The reaction was performed using *in vitro* transcribed *H. sapiens* $tRNA^{Lys}$ added to 5 μL of human HEK 293T cell extracts, at same conditions as for PfKRS-2. Similarly, aminoacylation rates of *P.*

falciparum cytosolic lysyl-tRNA synthetase (PfKRS-1) were determined using *in vitro* transcribed nuclear-encoded $tRNA^{Lys}$ of *P. falciparum* added to 5 μL of plasmodial extracts.

Cell-based drug inhibition assays

a) LDH activity assay

Initial screens to test the activity of the compounds were done through the LDH activity assay, as previously described[65]. Smears were also prepared for each drug assay to visually confirm the absorbance results. For each tested compound, and LDH activity was measured both at 48 and 96h in order to check for delayed effect. Mupirocin[24] and borrelidin[25] were used as positive controls of inhibition[23,24,65].

b) Fluorescence-assisted cell sorting (FACS)

FACS was used to calculate the IC_{50} of the most active compounds. For FACS analysis, Syto-11 was used to discriminate parasitized from non-parasitized RBCs. Each sample was diluted at 1:100 in PBS and 0.5 mM Syto-11 in DMSO was added to a final concentration of 0.5 μM. Samples were excited at 488 nm and analyzed using an FC500 flow cytometer.

Subcellular localization of PfKRS-2 by immunofluorescence

Apicoplastic susbcellular localization of PfKRS-2 was predicted using different algorithms, including PlasmoAP[11], PATS[66], PlasMit[67], PSORT[68] and SignalP[69]. To experimentally prove this prediction, the PfKRS-2 leader sequence was inserted into the XhoI/XmaI digested pGlux.1 vector[70] (kind gift from Alan Cowman), to generate a C-terminal GFP fusion to the N-terminal region of PfKRS-2 that contains the predicted apicoplastic localization signal of the protein[11]. Synchronized cultures of *P. falciparum* 3D7A were electroporated and transfected with the PfKRS-2_{leader}-GFP-containing vector. After 24h of growth, WR99210 was added to a final concentration of 10nM to select for transfected parasites. RBCs containing transfected parasites expressing PfKRS-2-GFP were washed and fixed for 5 min in 90:10 methanol-acetone, and incubated with anti-ACP primary antibody (kindly provided by Dr. McFadden), which was used to check for colocalization in the apicoplast. Anti-ACP antibodies were detected with a secondary fluorescent antibody AlexaFluor 555 mouse anti-rabbit (Invitrogen). The samples were mounted with Mowiol (Calbiochem, Merck Chemicals), and analyzed with a Leica SP2 confocal microscope.

Zebrafish assays for drug toxicity screening

Compounds **M-26** and **M-37** were tested for toxicity on zebrafish embryos[71] to anticipate possible undesired toxic effects of the compounds. **M-26** and **M-37** were added at their

respective IC_{50} concentrations to zebrafish embryos (20 embryos per compound, independent duplicates), and the toxicity was evaluated at 24, 48 and 72h.

In vivo antimalarial activity of the compounds in mice

The rodent malaria parasite *Plasmodium yoelii 17XL* (*Py17XL*) MRA-267 was obtained from the Malaria Research and Reference Resource Center, and was maintained by serial blood passage in mice and stored in liquid nitrogen. Inbred BALB/cAnNHsd female, 6-8 weeks aged, were purchased from Harlan Laboratories and housed under standard conditions of light and temperature in the Animal Housing Facility at Complutense University. All mice were fed *ad libitum* on a commercial diet. *In vivo* experiments were carried out in accordance with national and international guidelines for Animal Care. The *in vivo* antimalarial activity of **M-26** and **M-37** was analyzed by using a 4-day-blood suppressive test as previously described[72,73]. Briefly, mice were inoculated $2x10^7$ red blood cells from *Py17XL*-infected mice by intraperitoneal injection. The chemotherapy treatment started 2h later (day 0) with a single dose of **M-37** stock1 and stock2 (9.75 mg kg-1 day-1) or **M-26** stock1 and stock2 (25 mg kg-1 day-1) by an intraperitoneal injection followed by identical dose administration for the following 3 days. Tested drugs were prepared at appropriate doses in aqueous vehicle containing 7% Tween-80 and 3% ethanol. The control groups received aqueous vehicle. The parasitemia was monitored daily by microscopic examination of Wright's-stained thin-blood smears.

SUPPORTING INFORMATION AVAILABILITY

Supplemental material includes figures S1-S5, table S1 and supplementary methods. This material is available free of charge via the Internet at http://pubs.acs.org

ACKNOWLEDGMENTS

This work has been supported by the EU FP7 Grant HEALTH-F3-2009-223024 – Mephitis, and by grant BIO2009-09776 (to L.R. d. P.) and CTQ2008-00177 and SAF2011-30508-C02-01(to M.R.) from the Spanish Ministry of Education and Science. E.M.N. is supported by a La Caixa/IRB International Ph.D. Program Fellowship. R.H., A.L., L.C. and N.C. were supported by the EU FP7 project grant.

REFERENCES

1. World Malaria Report - 2010. In *World Malaria Report - 2010*, 2010.

2. Petersen, I.; Eastman, R.; Lanzer, M., Drug-resistant malaria: Molecular mechanisms and implications for public health. *FEBS Letters* **2011,** *585* (11), 1551-1562.

3. Carlton, J. M.; Angiuoli, S. V.; Suh, B. B.; Kooij, T. W.; Pertea, M.; Silva, J. C.; Ermolaeva, M. D.; Allen, J. E.; Selengut, J. D.; Koo, H. L.; Peterson, J. D.; Pop, M.; Kosack, D. S.; Shumway, M. F.; Bidwell, S. L.; Shallom, S. J.; van Aken, S. E.; Riedmuller, S. B.; Feldblyum, T. V.; Cho, J. K.; Quackenbush, J.; Sedegah, M.; Shoaibi, A.; Cummings, L. M.; Florens, L.; Yates, J. R.; Raine, J. D.; Sinden, R. E.; Harris, M. A.; Cunningham, D. A.; Preiser, P. R.; Bergman, L. W.; Vaidya, A. B.; van Lin, L. H.; Janse, C. J.; Waters, A. P.; Smith, H. O.; White, O. R.; Salzberg, S. L.; Venter, J. C.; Fraser, C. M.; Hoffman, S. L.; Gardner, M. J.; Carucci, D. J., Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **2002,** *419* (6906), 512-519.

4. Gardner, M. J.; Shallom, S. J.; Carlton, J. M.; Salzberg, S. L.; Nene, V.; Shoaibi, A.; Ciecko, A.; Lynn, J.; Rizzo, M.; Weaver, B.; Jarrahi, B.; Brenner, M.; Parvizi, B.; Tallon, L.; Moazzez, A.; Granger, D.; Fujii, C.; Hansen, C.; Pederson, J.; Feldblyum, T.; Peterson, J.; Suh, B.; Angiuoli, S.; Pertea, M.; Allen, J.; Selengut, J.; White, O.; Cummings, L. M.; Smith, H. O.; Adams, M. D.; Venter, J. C.; Carucci, D. J.; Hoffman, S. L.; Fraser, C. M., Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. *Nature* **2002,** *419* (6906), 531-534.

5. Gardner, M. J.; Hall, N.; Fung, E.; White, O.; Berriman, M.; Hyman, R. W.; Carlton, J. M.; Pain, A.; Nelson, K. E.; Bowman, S.; Paulsen, I. T.; James, K.; Eisen, J. A.; Rutherford, K.; Salzberg, S. L.; Craig, A.; Kyes, S.; Chan, M.-S.; Nene, V.; Shallom, S. J.; Suh, B.; Peterson, J.; Angiuoli, S.; Pertea, M.; Allen, J.; Selengut, J.; Haft, D.; Mather, M. W.; Vaidya, A. B.; Martin, D. M. A.; Fairlamb, A. H.; Fraunholz, M. J.; Roos, D. S.; Ralph, S. A.; McFadden, G. I.; Cummings, L. M.; Subramanian, G. M.; Mungall, C.; Venter, J. C.; Carucci, D. J.; Hoffman, S. L.; Newbold, C.; Davis, R. W.; Fraser, C. M.; Barrell, B., Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **2002,** *419* (6906), 498-511.

6. Hall, N.; Pain, A.; Berriman, M.; Churcher, C.; Harris, B.; Harris, D.; Mungall, K.; Bowman, S.; Atkin, R.; Baker, S.; Barron, A.; Brooks, K.; Buckee, C. O.; Burrows, C.; Cherevach, I.; Chillingworth, C.; Chillingworth, T.; Christodoulou, Z.; Clark, L.; Clark, R.; Corton, C.; Cronin, A.; Davies, R.; Davis, P.; Dear, P.; Dearden, F.; Doggett, J.; Feltwell, T.; Goble, A.; Goodhead, I.; Gwilliam, R.; Hamlin, N.; Hance, Z.; Harper, D.; Hauser, H.; Hornsby, T.; Holroyd, S.; Horrocks, P.; Humphray, S.; Jagels, K.; James, K. D.; Johnson, D.; Kerhornou, A.; Knights, A.; Konfortov, B.; Kyes, S.; Larke, N.; Lawson, D.; Lennard, N.; Line, A.; Maddison, M.; McLean, J.; Mooney, P.;

Moule, S.; Murphy, L.; Oliver, K.; Ormond, D.; Price, C.; Quail, M. A.; Rabbinowitsch, E.; Rajandream, M. A.; Rutter, S.; Rutherford, K. M.; Sanders, M.; Simmonds, M.; Seeger, K.; Sharp, S.; Smith, R.; Squares, R.; Squares, S.; Stevens, K.; Taylor, K.; Tivey, A.; Unwin, L.; Whitehead, S.; Woodward, J.; Sulston, J. E.; Craig, A.; Newbold, C.; Barrell, B. G., Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. *Nature* **2002,** *419* (6906), 527-531.

7. Hyman, R. W.; Fung, E.; Conway, A.; Kurdi, O.; Mao, J.; Miranda, M.; Nakao, B.; Rowley, D.; Tamaki, T.; Wang, F.; Davis, R. W., Sequence of Plasmodium falciparum chromosome 12. *Nature* **2002,** *419* (6906), 534-537.

8. Lasonder, E.; Ishihama, Y.; Andersen, J. S.; Vermunt, A. M. W.; Pain, A.; Sauerwein, R. W.; Eling, W. M. C.; Hall, N.; Waters, A. P.; Stunnenberg, H. G.; Mann, M., Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **2002,** *419* (6906), 537-542.

9. Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J., A proteomic view of the Plasmodium falciparum life cycle. *Nature* **2002,** *419* (6906), 520-526.

10. Lim, L.; McFadden, G. I., The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society B-Biological Sciences* **2010,** *365* (1541), 749-763.

11. Foth, B. J.; Ralph, S. A.; Tonkin, C. J.; Struck, N. S.; Fraunholz, M.; Roos, D. S.; Cowman, A. F.; McFadden, G. I., Dissecting Apicoplast Targeting in the Malaria Parasite Plasmodium falciparum. *Science* **2003,** *299* (5607), 705-708.

12. Waller, R. F.; McFadden, G. I., The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol* **2005,** *7* (1), 57-79.

13. Dahl, E. L.; Rosenthal, P. J., Apicoplast translation, transcription and genome replication: targets for antimalarial antibiotics. *Trends in Parasitology* **2008,** *24* (6), 279-284.

14. Fichera, M. E.; Roos, D. S., A plastid organelle as a drug target in apicomplexan parasites. *Nature* **1997,** *390* (6658), 407-409.

15. Ralph, S. A.; D'Ombrain, M. C.; McFadden, G. I., The apicoplast as an antimalarial drug target. *Drug Resistance Updates* **2001,** *4* (3), 145-151.

16. Dahl, E. L.; Rosenthal, P. J., Multiple Antibiotics Exert Delayed Effects against the Plasmodium falciparum Apicoplast. *Antimicrob. Agents Chemother.* **2007,** *51* (10), 3485-3490.

17. Dahl, E. L.; Shock, J. L.; Shenai, B. R.; Gut, J.; DeRisi, J. L.; Rosenthal, P. J., Tetracyclines Specifically Target the Apicoplast of the Malaria Parasite Plasmodium falciparum. *Antimicrob. Agents Chemother.* **2006,** *50* (9), 3124-3131.

18. Goodman, C. D.; Su, V.; McFadden, G. I., The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. *Molecular and Biochemical Parasitology* **2007,** *152* (2), 181-191.

19. Ramya, T. N. C.; Mishra, S.; Karmodiya, K.; Surolia, N.; Surolia, A., Inhibitors of Nonhousekeeping Functions of the Apicoplast Defy Delayed Death in Plasmodium falciparum. *Antimicrob. Agents Chemother.* **2007,** *51* (1), 307-316.

20. Sidhu, A. B. S.; Sun, Q.; Nkrumah, L. J.; Dunne, M. W.; Sacchettini, J. C.; Fidock, D. A., In Vitro Efficacy, Resistance Selection, and Structural Modeling Studies Implicate the Malarial Parasite Apicoplast as the Target of Azithromycin. *Journal of Biological Chemistry* **2007,** *282* (4), 2494-2504.

21. Kim, S.; Lee, S. W.; Choi, E. C.; Choi, S. Y., Aminoacyl-tRNA synthetases and their inhibitors as a novel family of antibiotics. *Applied Microbiology and Biotechnology* **2003,** *61* (4), 278-288.

22. Schimmel, P. R.; Soll, D., Aminoacyl Transfer Rna-Synthetases - General Features and Recognition of Transfer-Rnas. *Annual Review of Biochemistry* **1979,** *48*, 601-648.

23. Ishiyama, A.; Iwatsuki, M.; Namatame, M.; Nishihara-Tsukashima, A.; Sunazuka, T.; Takahashi, Y.; Omura, S.; Otoguro, K., Borrelidin, a potent antimalarial: stage-specific inhibition profile of synchronized cultures of Plasmodium falciparum. *J Antibiot* **2011,** *64* (5), 381-384.

24. Istvan, E. S.; Dharia, N. V.; Bopp, S. E.; Gluzman, I.; Winzeler, E. A.; Goldberg, D. E., Validation of isoleucine utilization targets in Plasmodium falciparum. *Proceedings of the National Academy of Sciences* **2011,** *108* (4), 1627-1632.

25. Nass, G.; Poralla, K.; Zahner, H., Effect of Antibiotic Borrelidin on Regulation of Threonine Biosynthetic Enzymes in E Coli. *Biochemical and Biophysical Research Communications* **1969,** *34* (1), 84-&.

26. Paetz, W.; Nass, G., Biochemical and Immunological Characterization of Threonyl-Transfer-Rna Synthetase of 2 Borrelidin-Resistant Mutants of Escherichia-Coli K-12. *European Journal of Biochemistry* **1973,** *35* (2), 331-337.

27. Ogilvie, A.; Wiebauer, K.; Kersten, W., Inhibition of Leucyl-Transfer Ribonucleic-Acid Synthetase in Bacillus-Subtilis by Granaticin. *Biochemical Journal* **1975,** *152* (3), 511-515.

28. Werner, R. G.; Thorpe, L. F.; Reuter, W.; Nierhaus, K. H., Indolmycin Inhibits Prokaryotic Tryptophanyl-Transfer-Rna Ligase. *European Journal of Biochemistry* **1976,** *68* (1), 1-3.

29. Tanaka, K.; Tamaki, M.; Watanabe, S., Effect of furanomycin on the synthesis of isoleucyl-tRNA. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* **1969,** *195* (1), 244-245.

30. Konrad, I.; Roschenthaler, R., Inhibition of Phenylalanine Transfer-Rna Synthetase from Bacillus-Subtilis by Ochratoxin-A. *Febs Letters* **1977,** *83* (2), 341-347.

31. Konishi, M.; Nishio, M.; Saitoh, K.; Miyaki, T.; Oki, T.; Kawaguchi, H., Cispentacin, a New Antifungal Antibiotic .1. Production, Isolation, Physicochemical Properties and Structure. *Journal of Antibiotics* **1989,** *42* (12), 1749-1755.

32. Bennett, I.; Broom, N. J.; Cassels, R.; Elder, J. S.; Masson, N. D.; O'Hanlon, P. J., Synthesis and antibacterial properties of beta-diketone acrylate bioisosteres of pseudomonic acid A. *Bioorg Med Chem Lett* **1999,** *9* (13), 1847-52.

33. Broom, N. J.; Cassels, R.; Cheng, H. Y.; Elder, J. S.; Hannan, P. C.; Masson, N.; O'Hanlon, P. J.; Pope, A.; Wilson, J. M., The chemistry of pseudomonic acid. 17. Dual-action C-1 oxazole derivatives of pseudomonic acid having an extended spectrum of antibacterial activity. *J Med Chem* **1996,** *39* (18), 3596-600.

34. Hurdle, J. G.; O'Neill, A. J.; Chopra, I., Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents. *Antimicrobial Agents and Chemotherapy* **2005,** *49* (12), 4821-4833.

35. Ding, D.; Meng, Q.; Gao, G.; Zhao, Y.; Wang, Q.; Nare, B.; Jacobs, R.; Rock, F.; Alley, M. R.; Plattner, J. J.; Chen, G.; Li, D.; Zhou, H., Design, synthesis, and structure-activity relationship of Trypanosoma brucei leucyl-tRNA synthetase inhibitors as antitrypanosomal agents. *J Med Chem* **2011,** *54* (5), 1276-87.

36. Farrera-Sinfreu, J.; Espanol, Y.; Geslain, R.; Guitart, T.; Albericio, F.; de Pouplana, L. R.; Royo, M., Solid-phase combinatorial synthesis of a lysyl-tRNA synthetase (LysRS) inhibitory library. *Journal of Combinatorial Chemistry* **2008,** *10* (3), 391-400.

37. Van de Vijver, P.; Ostrowski, T.; Sproat, B.; Goebels, J.; Rutgeerts, O.; Van Aerschot, A.; Waer, M.; Herdewijn, P., Aminoacyl-tRNA synthetase inhibitors as potent and synergistic immunosuppressants. *Journal of Medicinal Chemistry* **2008,** *51* (10), 3020-3029.

38. Hoepfner, D.; McNamara, C. W.; Lim, C. S.; Studer, C.; Riedl, R.; Aust, T.; McCormack, S. L.; Plouffe, D. M.; Meister, S.; Schuierer, S.; Plikat, U.; Hartmann, N.; Staedtler, F.; Cotesta, S.; Schmitt, E. K.; Petersen, F.; Supek, F.; Glynne, R. J.; Tallarico, J. A.; Porter, J. A.; Fishman, M. C.; Bodenreider, C.; Diagana, T. T.; Movva, N. R.; Winzeler, E. A., Selective and Specific Inhibition of the Plasmodium falciparum Lysyl-tRNA Synthetase by the Fungal Secondary Metabolite Cladosporin. *Cell Host Microbe* **2012,** *11* (6), 654-63.

39. Brown, J. R., Ancient horizontal gene transfer. *Nat Rev Genet* **2003,** *4* (2), 121-32.

40. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **2004,** *47* (7), 1739-49.

41. Makler, M. T.; Hinrichs, D. J., Measurement of the Lactate Dehydrogenase Activity of Plasmodium falciparum as an Assessment of Parasitemia. *The American Journal of Tropical Medicine and Hygiene* **1993,** *48* (2), 205-210.

42. Hughes, J.; Mellows, G., Interaction of pseudomonic acid A with Escherichia coli B isoleucyl-tRNA synthetase. *Biochem J* **1980,** *191* (1), 209-19.

43. Sheng, C.; Zhang, W., New lead structures in antifungal drug discovery. *Curr Med Chem* **2011,** *18* (5), 733-66.

44. Shibata, S.; Gillespie, J. R.; Kelley, A. M.; Napuli, A. J.; Zhang, Z.; Kovzun, K. V.; Pefley, R. M.; Lam, J.; Zucker, F. H.; Van Voorhis, W. C.; Merritt, E. A.; Hol, W. G.; Verlinde, C. L.; Fan, E.; Buckner, F. S., Selective inhibitors of methionyl-tRNA synthetase have potent activity against Trypanosoma brucei Infection in Mice. *Antimicrob Agents Chemother* **2011,** *55* (5), 1982-9.

45. Ziegelbauer, K.; Babczinski, P.; Schonfeld, W., Molecular mode of action of the antifungal beta-amino acid BAY 10-8888. *Antimicrob Agents Chemother* **1998,** *42* (9), 2197-205.

46. Ibba, M.; Morgan, S.; Curnow, A. W.; Pridmore, D. R.; Vothknecht, U. C.; Gardner, W.; Lin, W.; Woese, C. R.; Soll, D., A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **1997,** *278* (5340), 1119-22.

47. Ribas de Pouplana, L.; Turner, R. J.; Steer, B. A.; Schimmel, P., Genetic code origins: tRNAs older than their synthetases? *Proc Natl Acad Sci U S A* **1998,** *95* (19), 11295-300.

48. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** *25* (17), 3389-402.

49. Russell, R. B.; Barton, G. J., Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **1992,** *14* (2), 309-23.

50. Eddy, S. R., Profile hidden Markov models. *Bioinformatics* **1998,** *14* (9), 755-63.

51. Sali, A.; Blundell, T. L., Comparative Protein Modeling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **1993,** *234* (3), 779-815.

52. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999,** *292* (2), 195-202.

53. Wiederstein, M.; Sippl, M. J., ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W407-10.

54. Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M., Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography* **1993,** *26*, 283-291.

55. LigPrep, v., Schrödinger, LLC, New York, NY, 2008.

56. Maestro, v., Schrödinger, LLC, New York, NY, 2008.

57. Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O'Donovan, C.; Redaschi, N.; Suzek, B., The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **2006,** *34* (Database issue), D187-91.

58. Ribas de Pouplana, L.; Brown, J. R.; Schimmel, P., Structure-based phylogeny of class IIa tRNA synthetases in relation to an unusual biochemistry. *J Mol Evol* **2001,** *53* (4-5), 261-8.

59. Diaz-Lazcoz, Y.; Aude, J. C.; Nitschke, P.; Chiapello, H.; Landes-Devauchelle, C.; Risler, J. L., Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. *Mol Biol Evol* **1998,** *15* (11), 1548-61.

60. O'Donoghue, P.; Luthey-Schulten, Z., On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol Mol Biol Rev* **2003,** *67* (4), 550-73.

61. Woese, C. R.; Olsen, G. J.; Ibba, M.; Soll, D., Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* **2000,** *64* (1), 202-36.

62. Wolf, Y. I.; Aravind, L.; Grishin, N. V.; Koonin, E. V., Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* **1999,** *9* (8), 689-710.

63. Felsenstein, J., Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* **1988,** *22*, 521-65.

64. Geslain, R.; Aeby, E.; Guitart, T.; Jones, T. E.; Castro de Moura, M.; Charriere, F.; Schneider, A.; Ribas de Pouplana, L., Trypanosoma seryl-tRNA synthetase is a metazoan-like enzyme with high affinity for tRNASec. *J Biol Chem* **2006,** *281* (50), 38217-25.

65. Nkhoma, S.; Molyneux, M.; Ward, S., In vitro antimalarial susceptibility profile and prcrt/pfmdr-1 genotypes of Plasmodium falciparum field isolates from Malawi. *Am J Trop Med Hyg* **2007,** *76* (6), 1107-12.

66. Zuegge, J.; Ralph, S.; Schmuker, M.; McFadden, G. I.; Schneider, G., Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. *Gene* **2001,** *280* (1-2), 19-26.

67. Bender, A.; van Dooren, G. G.; Ralph, S. A.; McFadden, G. I.; Schneider, G., Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. *Mol Biochem Parasitol* **2003,** *132* (2), 59-66.

68. Nakai, K.; Horton, P., PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **1999,** *24* (1), 34-6.

69. Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S., Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **2004,** *340* (4), 783-95.

70. Fidock, D. A.; Wellems, T. E., Transformation with human dihydrofolate reductase renders malaria parasites insensitive to WR99210 but does not affect the intrinsic activity of proguanil. *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94* (20), 10931-10936.

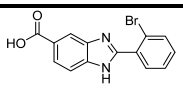
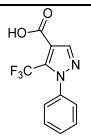
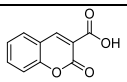
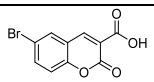
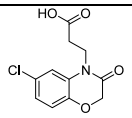
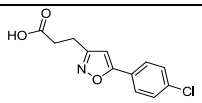
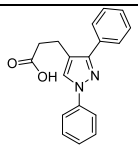
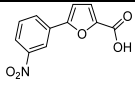
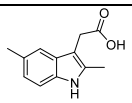
71. Rubinstein, A. L., Zebrafish assays for drug toxicity screening. *Expert Opin Drug Metab Toxicol* **2006,** *2* (2), 231-40.

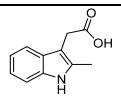
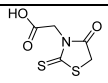
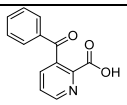
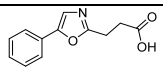
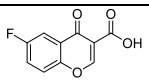
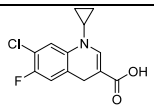
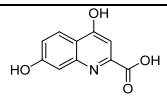
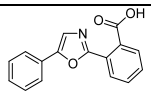
72. Moneriz, C.; Marin-Garcia, P.; Bautista, J. M.; Diez, A.; Puyet, A., Parasitostatic effect of maslinic acid. II. Survival increase and immune protection in lethal Plasmodium yoelii-infected mice. *Malar J* **2011,** *10*, 103.

73. Moneriz, C.; Marin-Garcia, P.; Garcia-Granados, A.; Bautista, J. M.; Diez, A.; Puyet, A., Parasitostatic effect of maslinic acid. I. Growth arrest of Plasmodium falciparum intraerythrocytic stages. *Malar J* **2011,** *10*, 82.

TABLES

Table 1: List of 46 synthesized compounds, ranked by its GlideScore when docked into the PfKRS-2 homology model (Pf-Gscore). The GlideScore on the human homologue is also shown (Hs-Gscore). See also Figure S2 and Table S1.

	LYSINE DERIVATIVES					THIALYSINE DERIVATIVES				
R-group	**Code[1]**	**Pf-Gscore**	**Hs-Gscore**	**Yield (mg)**	**Purity (%)**	**Code**	**Pf-Gscore**	**Hs-Gscore**	**Yield (mg)**	**Purity (%)**
	M-01	-12.94	-9.63	3.5	84.4	**M-26**	-11.99	-10.85	3.2	96.0
	M-02	-12.23	-8.85	9.3	95.0	M-27	-11.06	-7.94	8.1	90.9
	M-03	-10.76	-10.32	1.9	53.0	M-28	-9.28	-8.40	-	-
	M-04	-11.46	-9.66	1.6	90.2	M-29	-10.88	-10.60	-	-
	M-06	-11.22	-9.91	3.9	98.7	M-31	-9.68	-9.73	-	-
	M-07	-11.74	-9.40	4.0	93.9	M-32	-10.43	-8.40	11.2	98.5
	M-08	-12.70	-8.86	1.1	90.2	**M-33**	-12.00	-9.89	6.5	92.0
	M-09	-9.38	-8.63	3.4	91.4	M-34	-9.25	-10.89	5.1	>99.9
	M-11	-10.50	-8.68	12.0	85.8	M-36	-9.75	-9.50	10.1	96.9
	**M-12**	-10.97	-9.12	6.3	89.4	**M-37**	-10.54	-8.23	8.3	88.2
	M-13	-11.00	-8.88	10.4	92.7	M-38	-10.54	-8.23	20.1	86.8
	M-14	-11.76	-8.23	4.2	86.5	M-39	-9.84	-9.37	12.6	84.9
	M-15	-9.78	-10.68	1.4	99.0	M-40	-10.42	-10.66	12.7	90.0

	M-16	-10.43	-11.62	0.9	93.9	M-41	-10.11	-8.38	11.4	88.1
	M-17	-9.89	-9.42	1.0	95.3	M-42	-9.89	-10.21	4.9	92.8
	M-18	-9.59	-10.55	4.2	86.7	M-43	-9.66	-9.83	8.6	92.6
	M-19	-10.26	-8.12	2.0	86.5	M-44	-11.17	-9.54	3.9	96.5
	M-20	-10.22	-8.86	1.3	95.3	M-45	-10.49	-9.10	-	-
	M-21	-9.75	-11.09	1.7	76.5	M-46	-9.99	-8.54	6.5	91.3
	M-22	-10.62	-10.69	2.0	89.5	M-47	-12.49	-9.85	9.0	92.7
	M-23	-11.19	-10.45	3.0	95.9	M-48	-9.94	-11.59	3.5	93.1
	**M-24**	-9.52	-11.16	0.5	99.0	M-49	-9.79	-10.53	9.0	95.8
	M-25	-9.59	-10.77	4.7	92.0	M-50	-10.08	-8.90	-	-

[1] The subset of selected compounds which were re-synthesized for further *in vitro* analyses are shown in bold.

Table 2: *In vitro* inhibition of the five re-synthesized compounds. See also Figure S5

	SCREENING (purity > 85%, 150 μM)					**RE-SYNTHESIZED HITS (purity >95%, 50μM)**					
	Purity (%)	LDH 48h (% inhib.)	LDH 96h (% inhib.)	Smears 48h (% inhib.)	Smears 96h (% inhib.)	Purity (%)	Smears 48h (% inhib.)	Smears 96h (% inhib.)	IC_{50} 48h (μM)	IC_{50} 96h (μM)	Selectivity (fold)
M-12	93	0	45.9	23.2	58.0	99.5	31.7	70.4	172	83.2	2.2
M-24	92	4.42	58.7	0	58.3	99.5	32.6	42.4	518	427.1	1.2
M-26	96	0	35.5	0.18	66.0	100	3.90	41.6	551	84.7	6.5
M-33	85	24.7	100	76.6	98.9	98.6	43.5	77.9	48.1	29.5	1.4
M-37	88	15.7	65.5	49.7	100	100	6.20	72.7	151	38.4	3.9

* Inhibition of *P. falciparum* cultures has been measured using: i) the LDH (lactate dehydrogenase) assay, and ii) visual inspection of Giemsa-stained smears.

FIGURES

Figure 1. Characterization of the lysylation system in *P. falciparum*. A) Domain structure of *P. falciparum* KRS and KRS-2. The bipartite apicoplastic-targeting signal, consisting of signal peptide (SP) followed by a transit peptide (TP) is only found shown in PF14_0166, suggesting that PF13_0262 corresponds to the cytosolic enzyme, whereas PF14_0166 corresponds to the apicoplastic enzyme. B). Immunofluorescence assays of infected red blood cells (iRBCs) containing PfKRS-2_{leader}-GFP transfected *P. falciparum* parasites. The GFP-tagged sequence is colocalizing with ACP and not with Mitotracker, indicating that it is being specifically targeted to the apicoplast, in agreement with the bioinformatic predictions. C) Phylogenetic analysis of class II lysyl-tRNA synthetases. The plasmodial PfKRS-1 and PfKRS-2 are boxed in red, whereas the human HsKRS is boxed in blue. PfKRS-2 clusters with bacterial sequences, whereas PfKRS-1 and HsKRS cluster with eukaryal sequences. D) Active site comparison between PfKRS2 and its human homologue. The active site is defined as those residues having at least one of their atoms at less than 4Å from the ligand. Differing residues are highlighted. See also Figure S1.

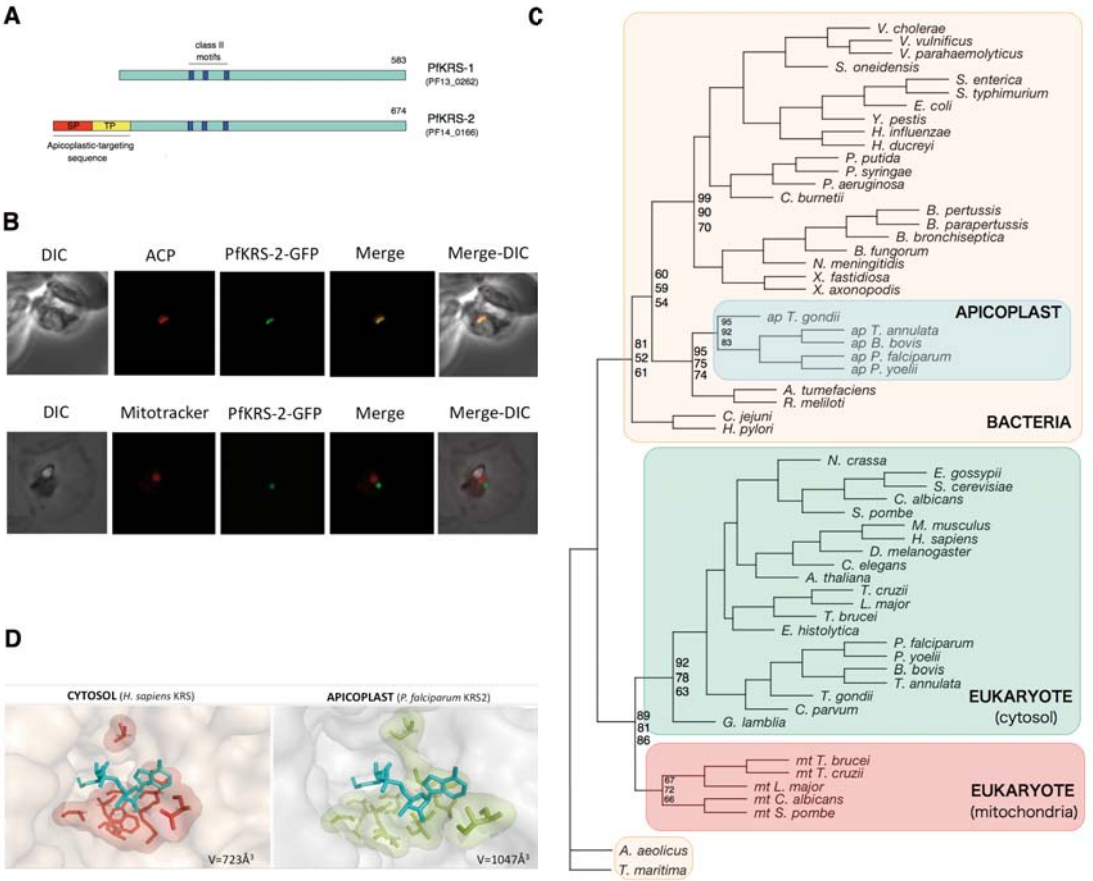


Figure 2. **Design of the virtual library and synthetic library.** The reaction intermediate lysyl-adenylate (left panel) has been subdivided in four parts, which have been colored accordingly. A virtual library of 1764 compounds (middle panel) was constructed based on the structure of the lysyl-adenylate complex. Based on the docking predictions, a library of 48 compounds was built (right panel).

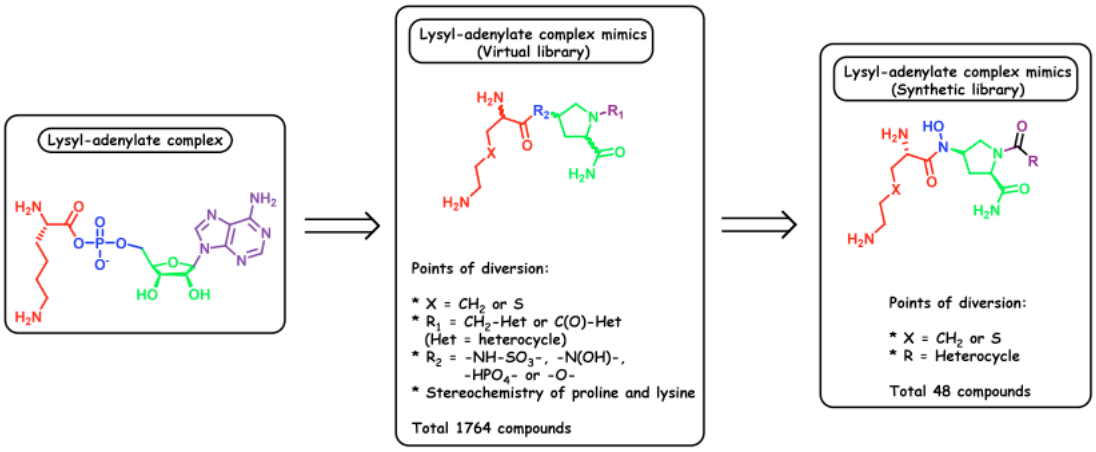


Figure 3. Synthesis of the library of lysyl-adenylate analogues. A) Solid-phase synthesis of the library. B) Synthesis of the active compounds in solution. See also Figure S3.

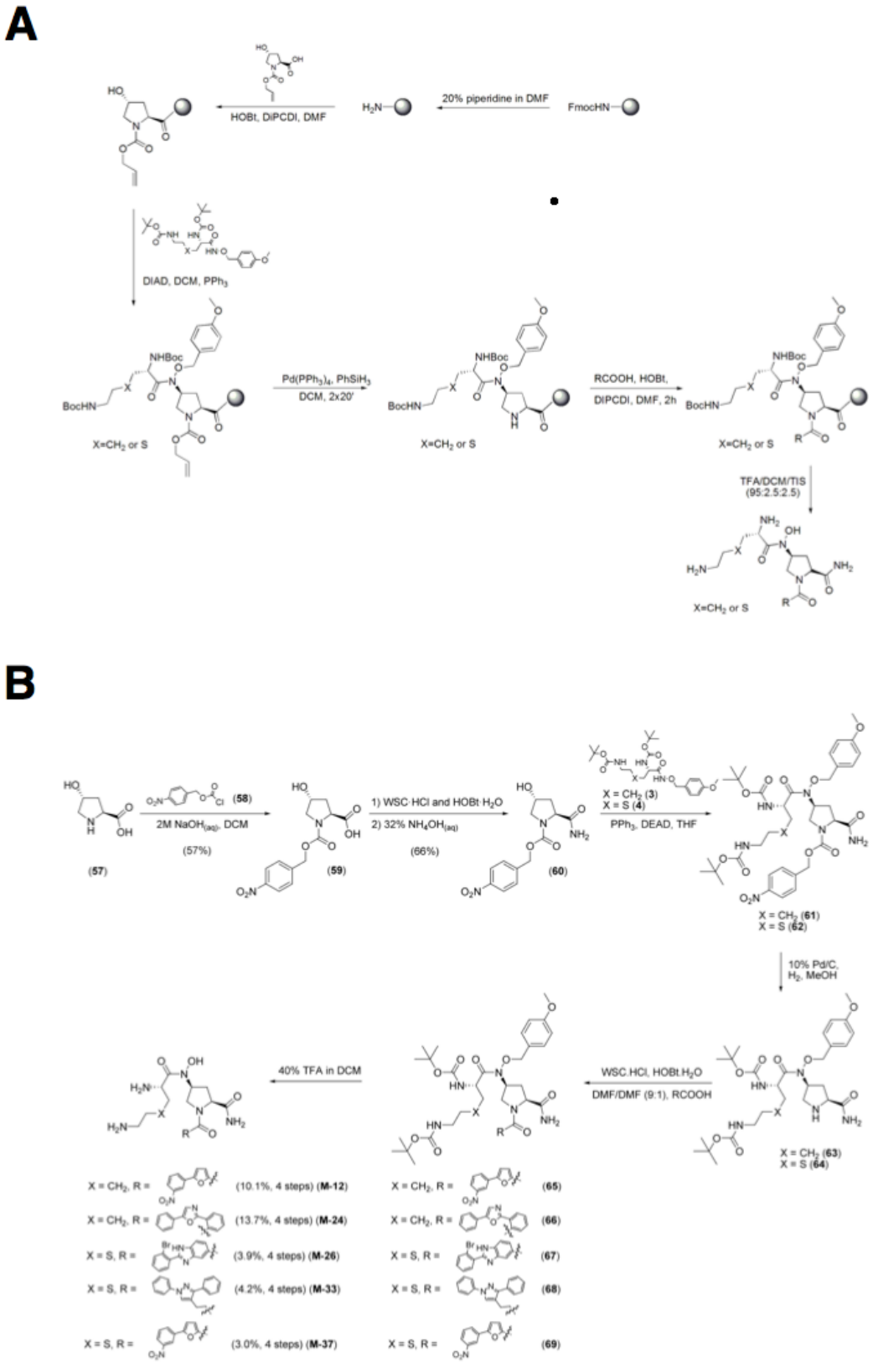


Figure 4. ***In vitro*** **inhibition of the aminoacylation reaction catalyzed by PfKRS-2, PfKRS-1 and HsKRS.** A) For the two most promising inhibitors, **M-26** and **M-37**, the IC_{50} was computed both at 48h (gray) and 96h (black). Both inhibitors show a clear delayed inhibition effect, which is typical of apicoplastic inhibitors. To verify the target of these inhibitors, we performed aminoacylation reactions with the PfKRS-2 enzyme, but also with PfKRS-1 and HsKRS enzymes (see Methods). As can be seen from the figure, both compounds inhibit PfKRS-2, whereas the aminoacylation activity of HsKRS and PfKRS-1 remains practically unaffected.

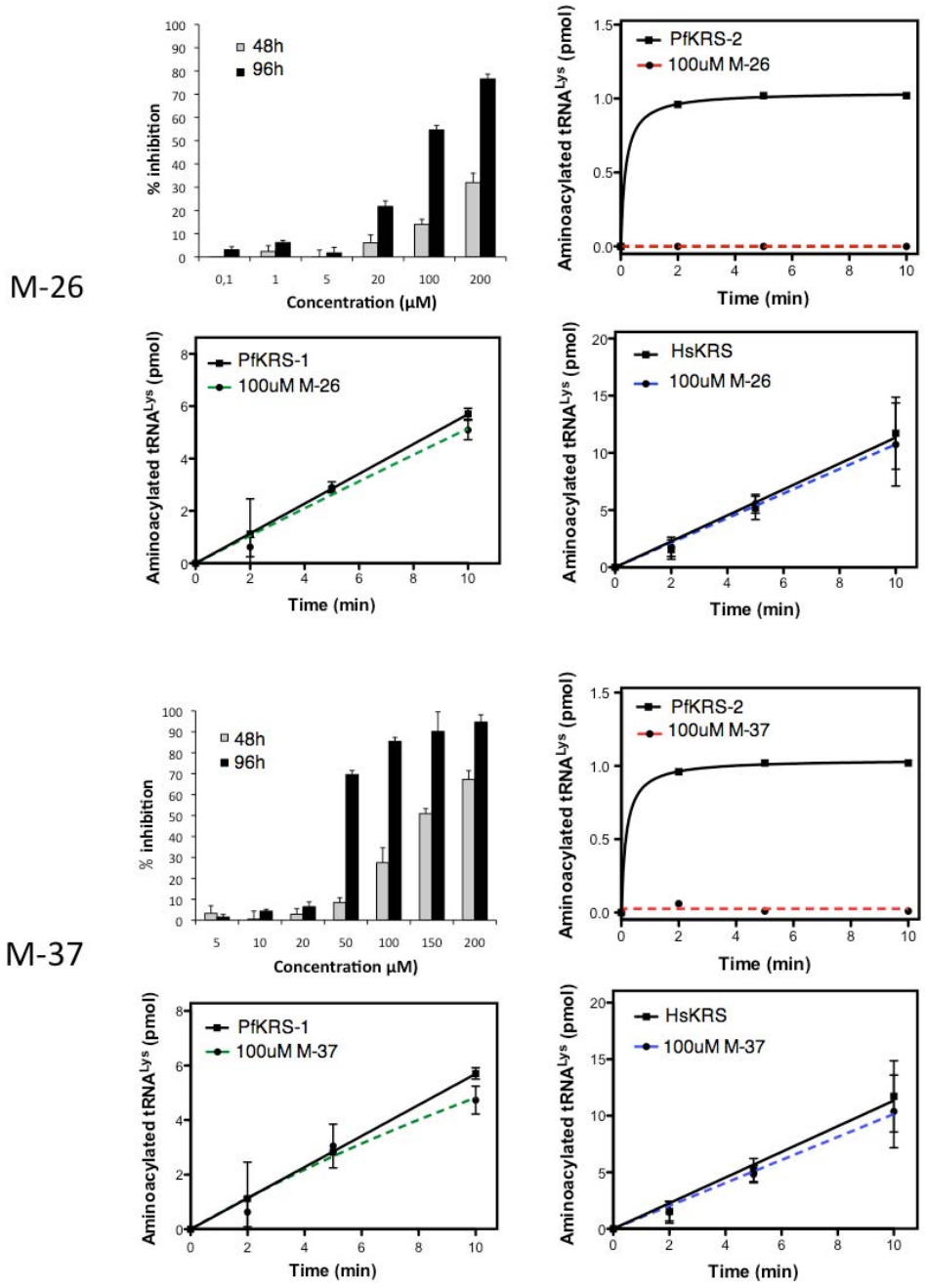
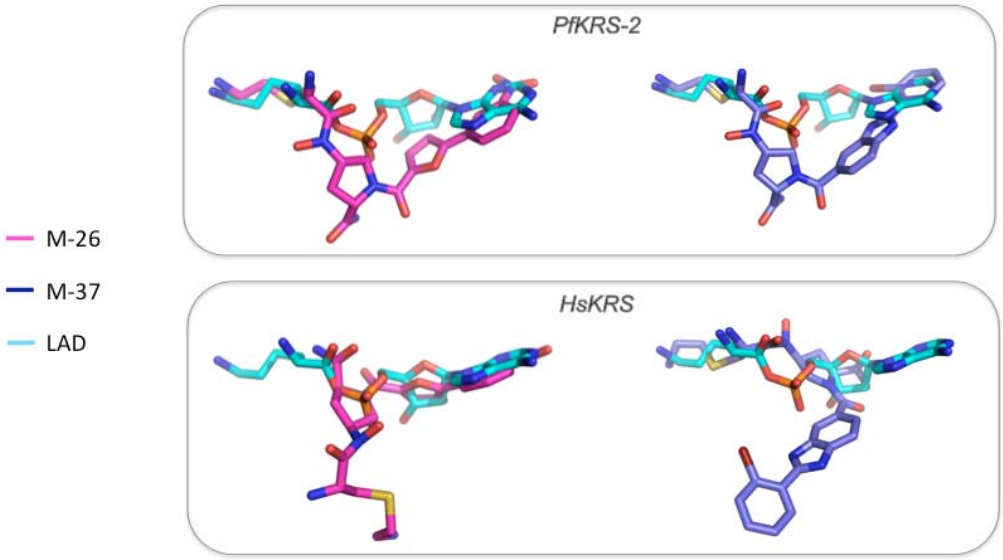


Figure 5. Structural analysis of M-26 and M-37 binding modes. On the top panel, the binding modes of M-26 (magenta) and M-37 (blue) docked into PfKRS-2 are shown. The natural ligand (lysyl-adenylate; LAD) is colored in cyan. Both compounds shown an lysyl-adenylate-like binding mode in the *P. falciparum* binding site (GlideScores of -11.99 and -10.54, for M-26 and M-37, respectively). On the lower panel, the binding modes of M-26 (magenta) and M-37 (blue) docked into HsKRS are shown. The natural ligand is shown in cyan. M-26 shows an adenine-like binding mode in *H. sapiens* (GlideScore = -10.85), whereas M-37 shows a lysine-like binding mode (GlideScore = -8.23).



SUPPLEMENTAL INFORMATION

Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from *Plasmodium falciparum*

Rob Hoen[†,#], Eva Maria Novoa[§,#], Alba López[†], Noelia Camacho[§], Laia Cubells[§], Pedro Vieira[&], Manuel Santos[&], Patricia Marin-Garcia[⊥], Jose Maria Bautista[⊥], Alfred Cortés[§], Lluís Ribas de Pouplana[*,§,±] and Miriam Royo[*,†]

[†] Combinatorial Chemistry Unit, Barcelona Science Park, University of Barcelona, C/ Baldiri Reixac 10, 08028 Barcelona, Catalonia, Spain.

[§] Institute for Research in Biomedicine, C/ Baldiri Reixac 10, Barcelona 08028, Catalonia, Spain.

[±]ICREA. Passeig Lluís Companys 1, Barcelona 08010, Catalonia, Spain.

[&]RNA Biology Laboratory, Department of Biology and Centre for Environmental and Marine Studies (CESAM), University of Aveiro, Aveiro, Portugal.

[⊥] Department of Biochemistry and Molecular Biology IV, Complutense University of Madrid, Madrid, Spain.

[#] These authors have contributed equally to this paper.

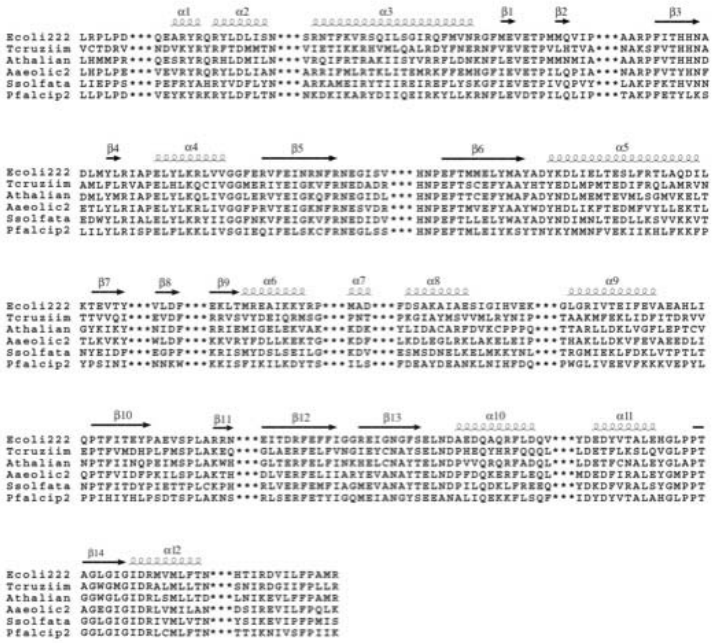
*Corresponding authors

1. SUPPLEMENTAL DATA

1.1. SUPPLEMENTAL FIGURES AND LEGENDS

Figure S1, related to Figure 1. Structural characterization of *P. falciparum* apicoplastic lysyl-tRNA synthetase. A) Structural alignment of the catalytic domain of lysyl-tRNA synthetases. Only a subset of sequences from the alignment is shown, including archaeal (*S.solfataricus*), bacterial (*E. coli, A. aeolicus*) and eukaryal (mitochondrial: *T. cruzii*, cytosolic: *A. thaliana*, apicoplastic: *P. falciparum*) organisms. The *P. falciparum* KRS sequence corresponds to PF14_0166 (PfKRS-2). B) Homology model of the apicoplastic PfKRS-2. One monomer is shown in red, whereas the other domain is colored according to its domains. The characteristic class II catalytic domain is shown in cyan, whereas the anticodon binding domain is shown in green. The natural ligand analogue has been docked to the structure (yellow). The bipartite apicoplastic targeting sequence has not been included in the homology model.

A


α1 α2 α3 β1 β2 β3
Ecoli222 LRPLPD***QEARYRQRYLDLISN***SRNTPKVRSQILSGIRQFMVNRGFMEVETPMMQVIP***AARPFITHHNA
Tcruziim VCTDRV***NDVKYRYRFTDMMTN***VIETIKKHVMLQALRDYFWERNFVEVETPVLNTVA***NAKSFVTHHNA
Athalian LHMMPR***QRSKYRQRHLDMILN***VRQIFRTRAKIISIVRRFLDNKNFLEVETPMMNMIA***AARPFVTHHND
Aaeolic2 LHPLPE***VEVRYRQRYLDLIAN***ARRIFMLRTKLITEMRKFFEMNGFIEVETPILQPIA***NARPFVTYHNF
Ssolfata LIEPPS***PEFRYAHRYVDFLYN***ARKAMEIRYTIIREIREFLYSKGFIEVETPIVQPVY***LAKPFKTHVNN
Pfalcip2 LLPLPD***VEYKYRKRYLDFLTN***NKDKIKARYDIIQEIRKYLLKRNFLEVDTPILQLIP***TAKPFETYLKS
β4 α4 β5 β6 α5
Ecoli222 DLMYLRIAPELYLKRLVVGGFERVFEINRNFRNEGISV***HNPEFTMMELYMAYADYKDLIELTESLFRTLAQDIL
Tcruziim AMLFLRVAPELHLKQCIVGGMERIYEIGKVFRNEDADR***HNPEFTSCEFYAAYETYEDLMPMTEDIFRQLAMRVN
Athalian DMLYMRIAPELYLKQLIVGGLERVYEIGKQFRNEGIDL***HNPEFTTCEFYMAFADYNDLMEMTEVMLSGMVKELT
Aaeolic2 ETLYLRIAPELYLKRLIVGGFPRVYEIGKNFRNESVDR***HNPEFTMVEFYAAYWDYHDLIKFTEDMFVYLLEKTL
Ssolfata EDWYLRIALELYLKRYIIGGFNKVFEIGKVFRNEDIDV***HNPEFTLLELYWAYADYNDIMNLTEDLLKSVVKKVT
Pfalcip2 LILYLRISPELFLKKLIVSGIEQIFELSKCFRNEGLSS***HNPEFTMLEIYKSYTNYKYMMNFVEKIIKHLFKKFP
β7 β8 β9 α6 α7 α8 α9
Ecoli222 KTEVTY***VLDFPKEKLTMREAIKKYRP***MAD***FDSAKAIAESIGIHVEK***GLGRIVTEIFEVAEAHLI
Tcruziim TTVVQI***EVDFPKERRVSVYDEIQRMSG***PNT***PKGIAYMSVVMLRYNIP***TAAKMFEKLIDFITDRVV
Athalian GYKIKY***NIDFPKRIEMIGELEKVAK***KDK***YLIDACARFDVKCPPPQ***TTARLLDKLVGFLEPTCV
Aaeolic2 TLKVKY***WLDFPKKVEYFDLLKEKTG***KDF***LKDLEGLRKLAKELEIP***THAKLLDKVFEVAEHDLI
Ssolfata NYEIDF***EGPFPKKRISMYDSLSEILG***KDV***ESMSDNELKELMKKYNL***TRGMIEKLFDKLVTPTLT
Pfalcip2 YPSINI***NNKWPKKISFIKILKDYTS***ILS***FDEAYDEAMKLNIHFDQ***PWGLIVEEVFKKKVEPYL
β10 β11 β12 β13 α10 α11
Ecoli222 QPTFITEYPAEVSPLARRN***EITDRFEFFIGGREIGNGFSELNDAEDQAQRFLDQV***YDEDYVTALEHGLPPT
Tcruziim EPTFVMDHPLPMSPLAKEQ***GLAERFELFVNGIEYCNAYSELNDPHEQYHRFQQQL***LDETFLKSLQVGLPPT
Athalian NPTFIINQPEIMSPLAKWH***GLTERFELFINKHELCNAYTELNDPVVQRQRFADQL***LDETFCNALEYGLAPT
Aaeolic2 QPTFVIDFPKILSPLAKTH***DLVERFELIIARYEVANAYTELNDPFDQKERFLEQL***MDEDFIRALEYGMPPT
Ssolfata NPTFITDYPIETTPLCKPH***RLVERFEMFIAGMEVANAYTELNDPILQDKLFREEQ***YDKDFVRALSYGMPPT
Pfalcip2 PPIHIYHLPSDTSPLAKNS***RLSERFETYIGQMEIANGYSEEANALIQEKKFLSQF***IDYDYVTALAHGLPPT
β14 α12
Ecoli222 AGLGIGIDRMVMLFTN***HTIRDVILFPAMR
Tcruziim AGWGMGIDRALMLLTN***SNIRDGIIFPLLR
Athalian GGWGLGIDRLSMLLTD***LNIKEVLFFPAMR
Aaeolic2 AGEGIGIDRLVMILAN***DSIREVILFPQLK
Ssolfata GGLGIGIDRIVMLVTN***YSIKEVIFFPMIS
Pfalcip2 GGLGIGIDRLCMLFTN***TTIKNIVSFPIIK

B

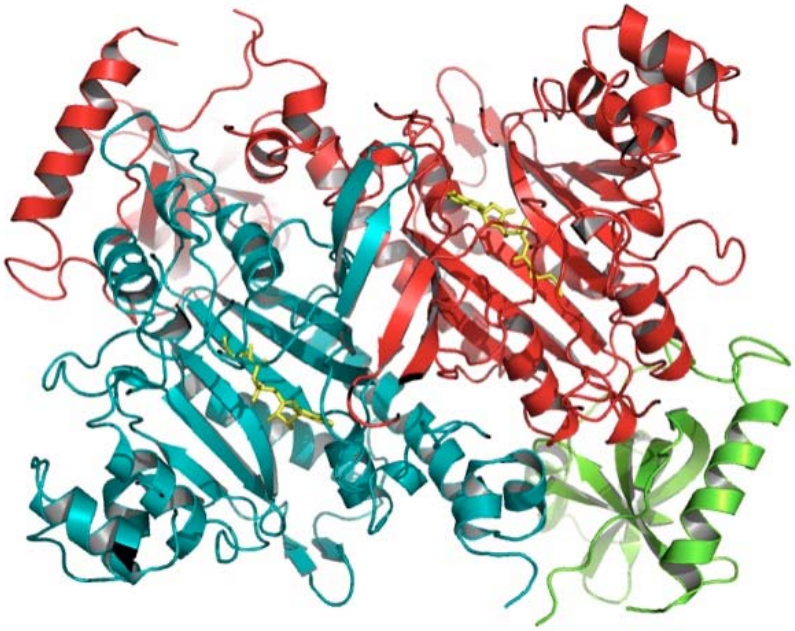
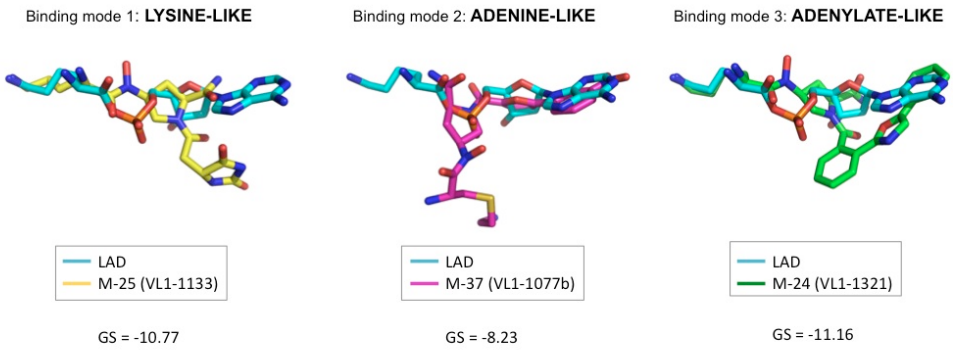


Figure S2, related to Table 1. Analysis of the docking predictions. A) Structural analysis of the predicted binding modes of the virtual library. The predicted binding modes of the docked compounds can be grouped into three different classes, depending on the part of the lysyl-adenylate (LAD) moiety that is being occupied by the analog: (1) Lysine-like binding mode (left), where only the lysine/thialysine is being recognized is a similar mode as the natural ligand, (2) Adenine-like binding mode (middle), where only the heterocycle mimetic is being recognized by the adenine pocket, and (3) Adenylate-like binding mode, where both lysine and heterocycle are being recognized similarly to the natural ligand binding mode. The GlideScores (GS) of each of the analogues on HsKRS are shown. Compounds showing adenylate-like binding mode tend to present highest GlideScores. B) Evaluation of the docking performance. of the virtual library.

A



B

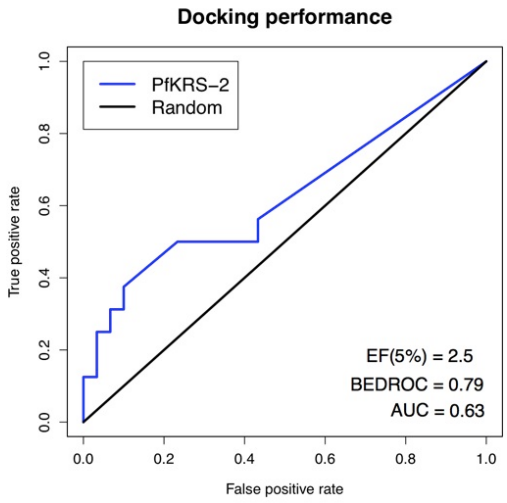


Figure S3. ***In vivo*** **toxicology and antimalarial activities of M-26 and M-37.** A) *In vivo* toxicology of **M-26** and **M-37** tested on zebrafish embryos, measured at 24, 48 and 72h. The development of the embryos is unaffected by the presence of the compounds (**M-26** and **M-37**) compared to the untreated embryos (C). B) *In vivo* antimalarial activity of **M-26** and **M-37** on *P. yoelii* infected mice, shown as percentage of mice survival.

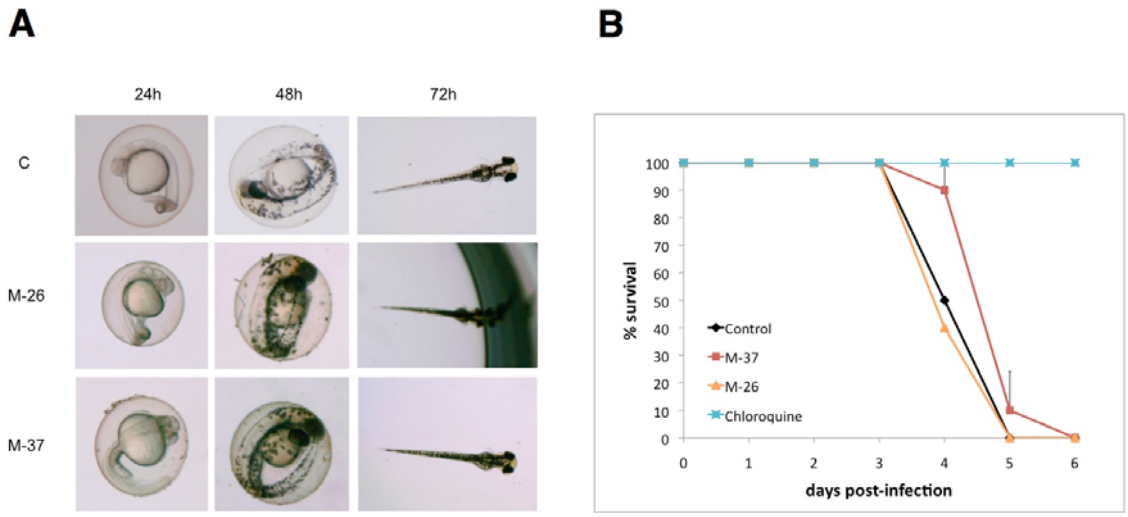


Figure S4, related to Figure 3. Synthesis of: (**A**) O-(4-methoxybenzyl)hydroxylamine (**73**), (**B**) di-*tert*-butyl 6-(4-methoxybenzyloxyamino)-6-oxohexane-1,5-diyldicarbamate (3) and (**C**) 2-((tert-butoxycarbonyl)amino)-3-((2-((tert-butoxycarbonyl)amino)ethyl)thio)propanoic acid (BocThiolysine(Boc)OH) (4)

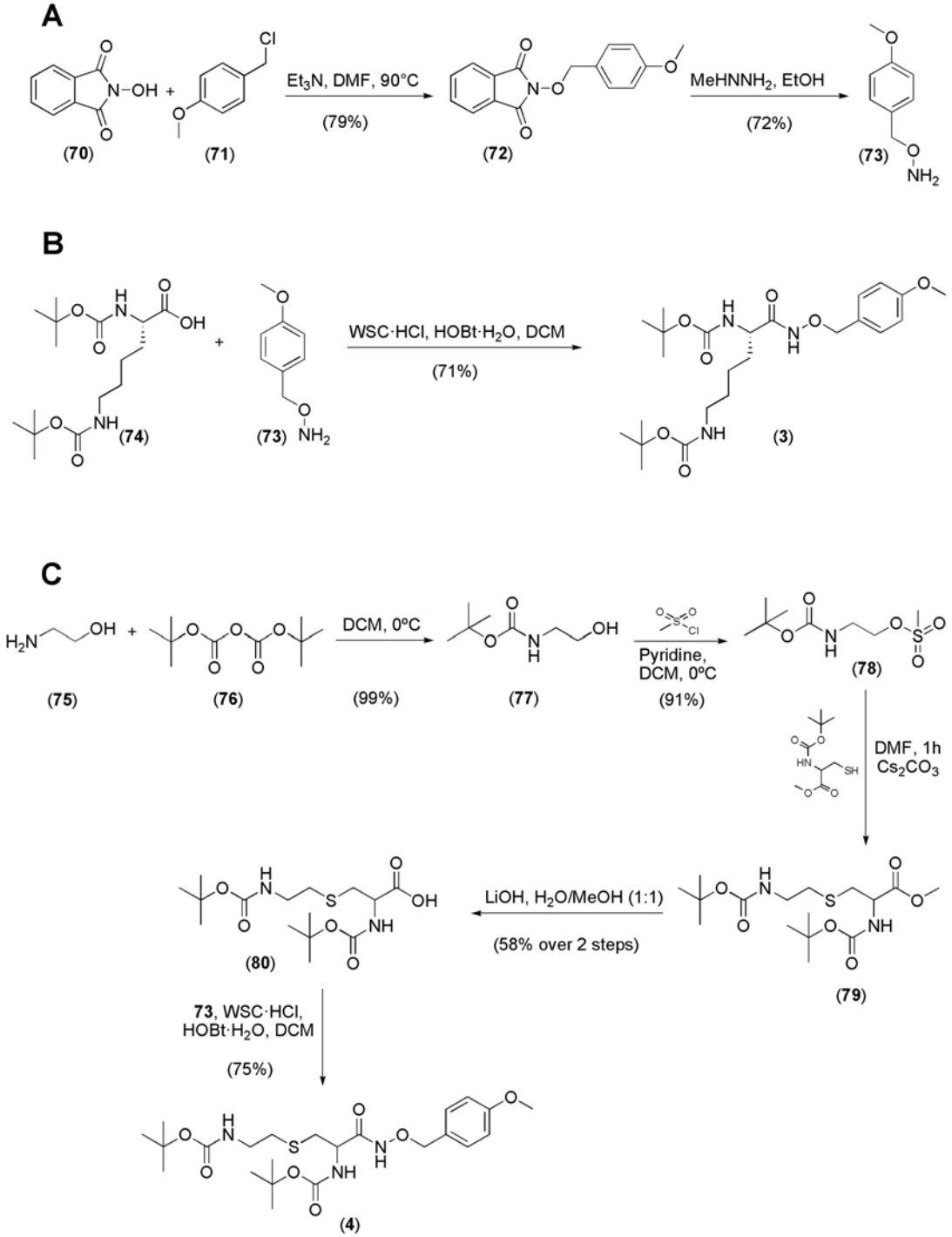
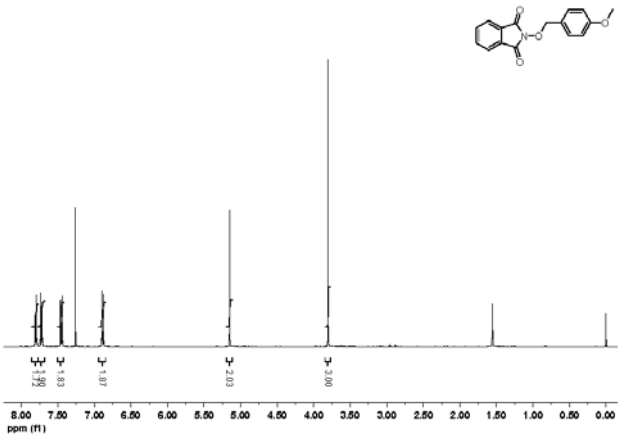


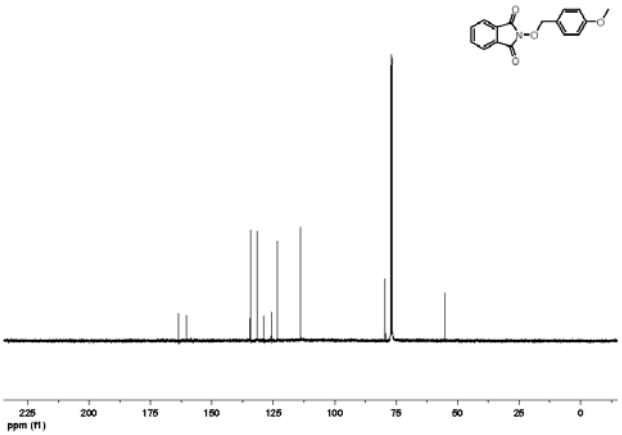
Figure S5, related to Table 2. NMR analysis of the synthesized compounds. (A) ^{1}H and ^{13}C NMR spectra of compound **72**, (B) ^{1}H and ^{13}C NMR spectra of compound **73**, (C) ^{1}H and ^{13}C NMR spectra of compound **3**, (D) ^{1}H and ^{13}C NMR spectra of compound **77**, (E) ^{1}H and ^{13}C NMR spectra of compound **78**, (F) ^{1}H and ^{13}C NMR spectra of compound **80,** (G) ^{1}H and ^{13}C NMR spectra of compound **4**, (H) ^{1}H and ^{13}C NMR spectra of compound **59**, (I) ^{1}H and ^{13}C NMR spectra of compound **60**, (J) ^{1}H, ^{13}C, Tocsy NMR spectra and LCMS chromatogram of compound **M-12**, (K) ^{1}H, ^{13}C, Tocsy NMR spectra and LCMS chromatogram of compound **M-24**, (L) ^{1}H, ^{13}C, Tocsy NMR spectra and LCMS chromatogram of compound **M-26**, (M) ^{1}H, ^{13}C, Tocsy NMR spectra and LCMS chromatogram of compound **M-33**, (N) ^{1}H, ^{13}C, Tocsy NMR spectra and LCMS chromatogram of compound **M-37**.

(A)

2-(4-methoxybenzyloxy)isoindoline-1,3-dione (72)



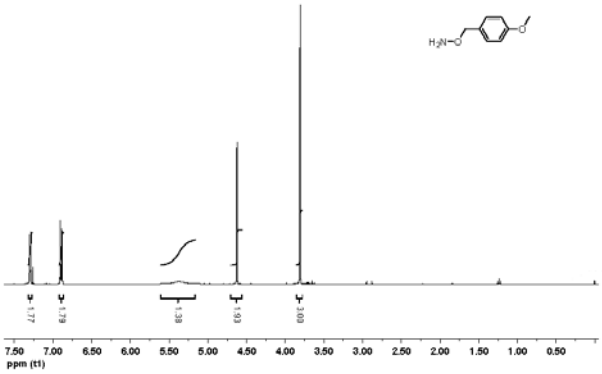
^{1}H-NMR (400 MHz, $CDCl_3$) δ = 7.83-7.77 (m, 2H), 7.75-7.71 (m, 2H), 7.45 (d, *J* = 8.70 Hz, 2H), 6.89 (d, *J* = 8.71 Hz, 2H), 5.15 (s, 2H), 3.80 (s, 3H)



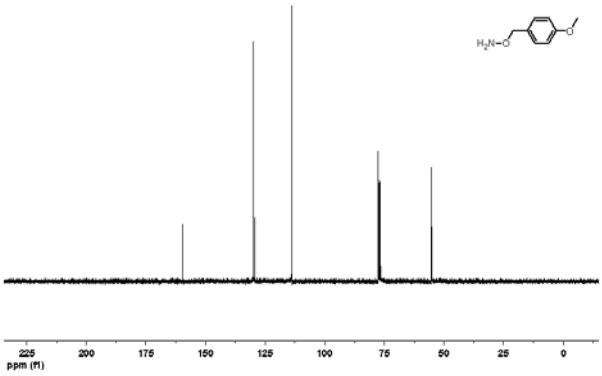
^{13}C-NMR (101 MHz, $CDCl_3$) δ = 163.5; 160.4; 134.3; 131.6; 128.8; 125.8; 123.4; 113.9; 79.4; 55.2;

(B)

O-(4-methoxybenzyl)hydroxylamine (73)



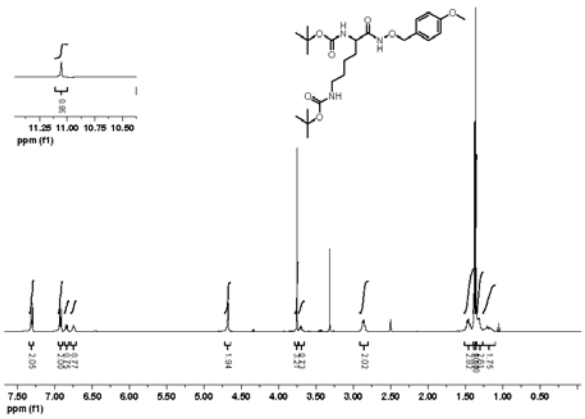
^{1}H-NMR (400 MHz, $CDCl_3$) δ = 7.30 (d, J = 8.65 Hz, 2H), 6.90 (d, J = 8.67 Hz, 2H), 5.37 (bs, 2H), 4.62 (s, 2H), 3.81 (s, 3H)



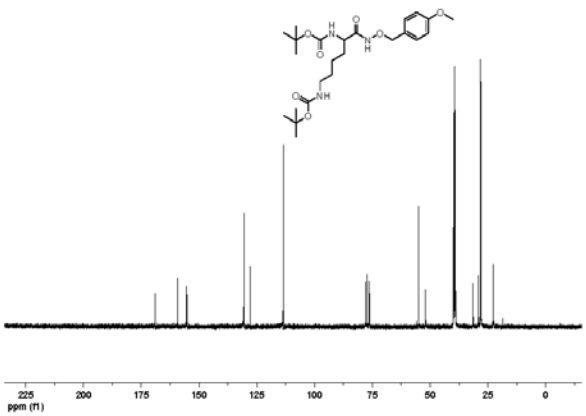
^{13}C-NMR (101 MHz, $CDCl_3$) δ = 159.4; 130.0; 129.3; 113.8; 77.5; 55.2

(C)

di-*tert*-butyl 6-(4-methoxybenzyloxyamino)-6-oxohexane-1,5-diyldicarbamate (3)



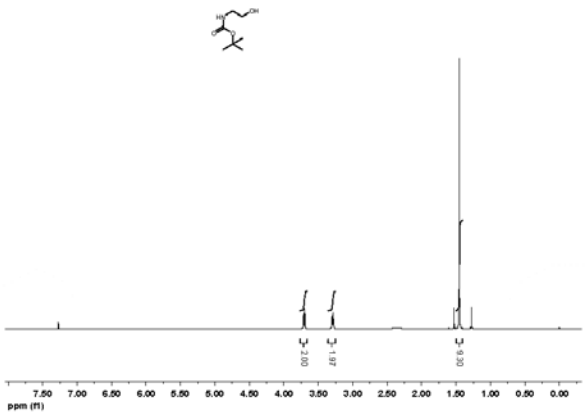
^{1}H-NMR (400 MHz, DMSO-$_{d6}$) δ = 11.05 (bs, 1H), 7.31 (d, *J* = 8.59 Hz, 2H), 6.92 (d, *J* = 8.57 Hz, 2H), 6.84 (bd, *J* = 7.82 Hz, 1H), 6.74 (bt, *J* = 4.94, 4.94 Hz, 1H), 4.68 (s, 2H), 3.75 (s, 3H), 3.73-3.67 (m, 1H), 2.90-2.82 (m, 2H), 1.50-1.41 (m, 2H), 1.37 (s, 9H), 1.36 (s, 9H), 1.34-1.26 (m, 2H), 1.26-1.10 (m, 2H)



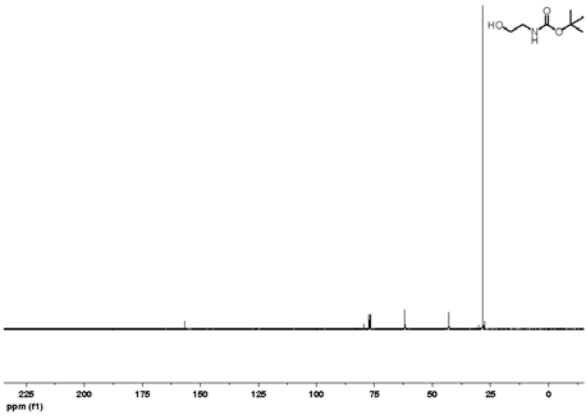
^{13}C-NMR (101 MHz, DMSO-$_{d6}$) δ = 168.9; 159.2; 155.4; 155.1; 130.5; 127.7; 113.5; 77.8; 77.2; 76.2; 54.9; 51.9; 31.4; 29.0; 28.1; 28.0; 22.6

(D)

tert-butyl (2-hydroxyethyl)carbamate (77)



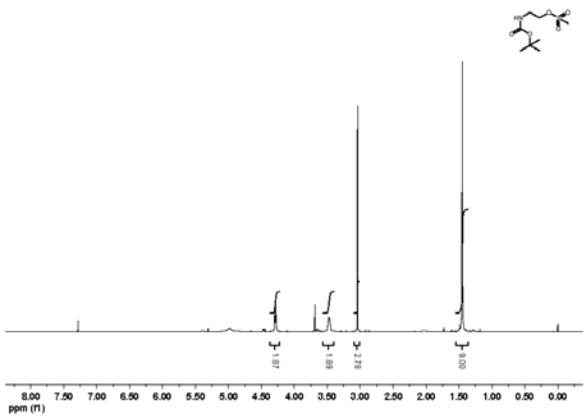
^{1}H NMR (400 MHz, $CDCl_3$) δ = 3.69 – 3.64 (m, 2H), 3.26 (t, J = 5.1 Hz, 2H), 1.43 (s, 9H)



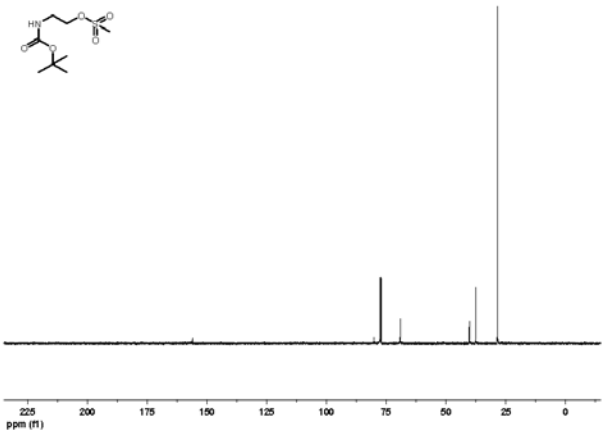
^{13}C-NMR (101 MHz, $CDCl_3$) δ = 156.6; 79.4; 61.9; 42.9; 28.2

(E)

2-((tert-butoxycarbonyl)amino)ethyl methanesulfonate (78)



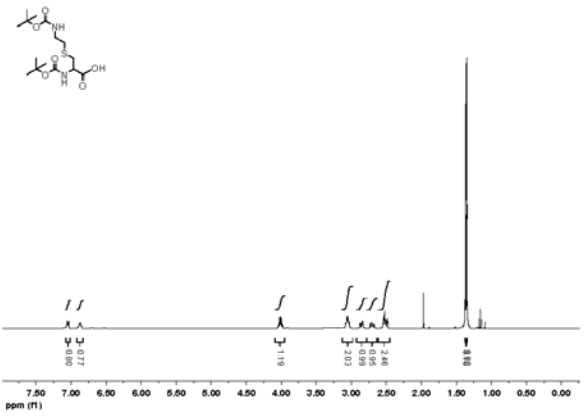
^{1}H NMR (400 MHz, $CDCl_3$) δ = 4.27 (t, *J* = 5.1 Hz, 2H), 3.45 (dd, *J* = 4.0 Hz, 2H), 3.02 (s, 3H), 1.43 (s, 9H)



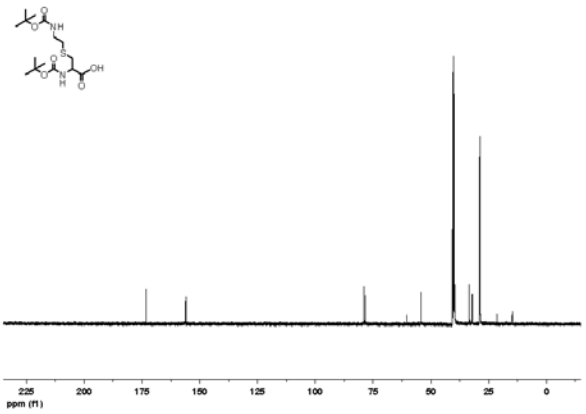
^{13}C NMR (101 MHz, $CDCl_3$)δ = 155.9, 80.1, 69.0, 40.1, 37.5, 28.4

(F)

2-((tert-butoxycarbonyl)amino)-3-((2-((tert-butoxycarbonyl)amino)ethyl)thio)propanoic acid (BocThialysine(Boc)OH) (80)



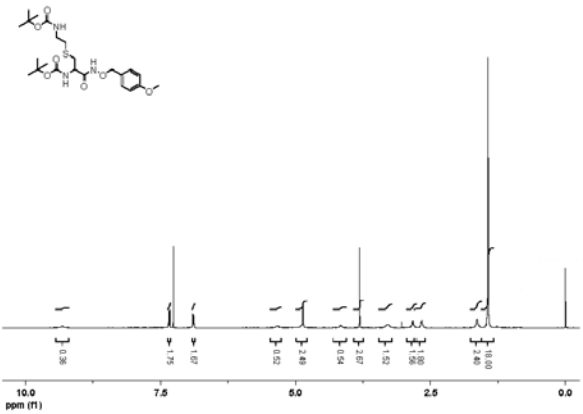
^{1}H NMR (400 MHz, DMSO-d_6) δ = 7.06 (d, J = 8.3 Hz, 1H), 6.88 (t, J = 5.2 Hz, 1H), 4.03 (td, J = 7.2, 4.6 Hz, 1H), 3.07 (dd, J = 13.3, 6.5 Hz, 2H), 2.87 (dd, J = 13.5, 4.6 Hz, 1H), 2.72 (dd, J = 13.4, 9.4 Hz, 1H), 2.57 – 2.51 (m, 2H), 1.38 (s, 9H), 1.37 (s, 9H)



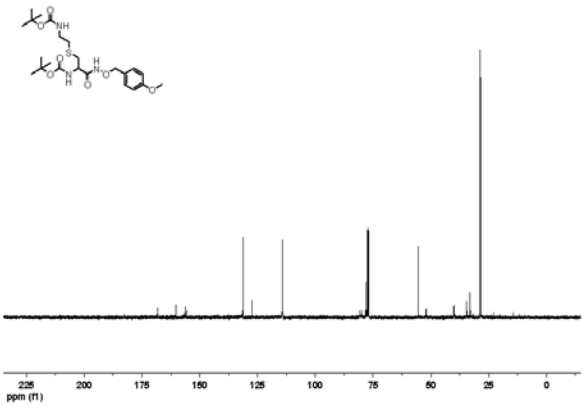
^{13}C NMR (101 MHz, DMSO-d_6) δ = 172.5, 155.4, 78.2, 77.7, 53.7, 32.7, 31.4, 28.2

(G)

BocThialysine(Boc)-NH-OPMB (4)



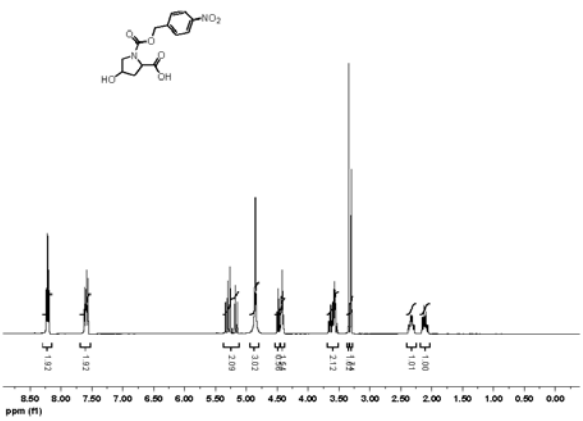
^{1}H NMR (400 MHz, $CDCl_3$) δ = 9.37 (s, 1H), 7.33 (d, *J* = 8.6 Hz, 2H), 6.88 (d, *J* = 8.5 Hz, 2H), 5.35 (d, *J* = 3.7 Hz, 1H), 4.91 (d, *J* = 6.2 Hz, 1H), 4.86 (s, 2H), 4.17 (s, 1H), 3.80 (s, 3H), 3.39 – 3.19 (m, 2H), 2.90 – 2.75 (m, 2H), 2.65 (t, *J* = 5.2 Hz, 2H), 1.73 (s, 1H), 1.43 (s, 18H)



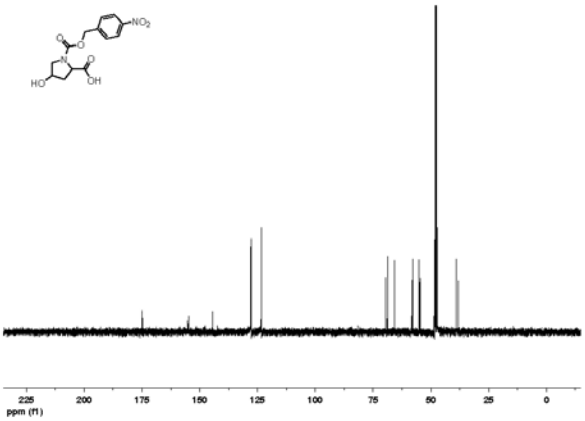
^{13}C NMR (101 MHz, $CDCl_3$) δ = 168.2, 160.1, 156.2, 155.5, 131.1, 127.3, 114.0, 78.1, 55.4, 51.9, 39.9, 34.4, 32.9, 28.5

(H)

(2S,4R)-4-hydroxy-1-((4-nitrobenzyloxy)carbonyl)pyrrolidine-2-carboxylic acid (59)



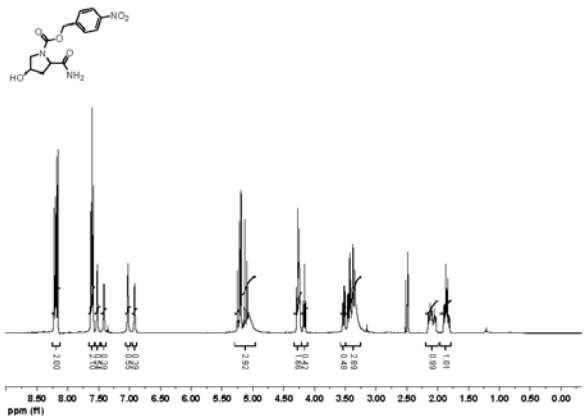
^{1}H NMR (400 MHz, CD_3OD) δ = 8.21 (t, *J* = 8.3 Hz, 2H), 7.58 (dd, *J* = 13.7, 8.8 Hz, 2H), 5.36 – 5.09 (m, 2H), 4.44 (dt, *J* = 25.9, 8.1 Hz, 2H), 3.69 – 3.48 (m, 2H), 2.39 – 2.25 (m, 1H), 2.16 – 2.03 (m, 1H)



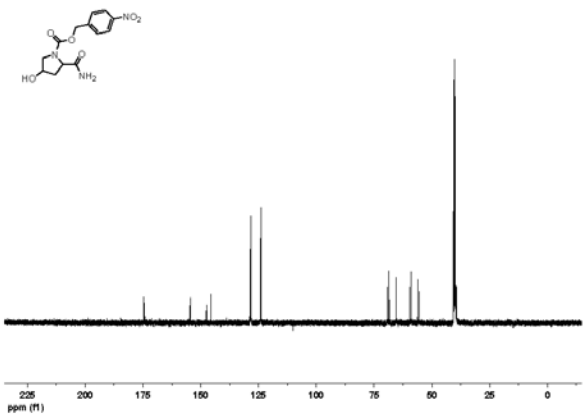
^{13}C NMR (101 MHz, CD_3OD) δ = 166.7, 166.4, 146.6, 136.2, 136.1, 119.7, 119.6, 115.1, 115.0, 61.2, 60.5, 57.5, 57.4, 49.9, 49.6, 46.7, 46.3, 39.5, 30.8, 29.9

(I)

(2S,4R)-4-nitrobenzyl 2-carbamoyl-4-hydroxypyrrolidine-1-carboxylate (60)



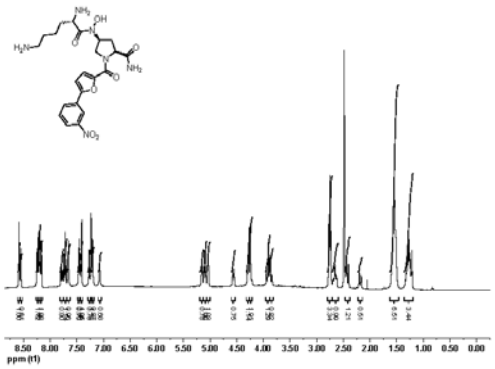
^{1}H NMR (400 MHz, DMSO-d_6) δ = 8.23 (d, J = 8.7 Hz, 1H), 8.19 (d, J = 8.8 Hz, 1H), 7.63 (t, J = 8.2 Hz, 4H), 7.53 (s, 1H), 7.43 (s, 1H), 7.04 (s, 1H), 6.93 (s, 1H), 5.25 (d, J = 14.5 Hz, 1H), 5.22 – 5.19 (m, 2H), 5.13 (d, J = 14.5 Hz, 1H), 5.06 (t, J = 3.8 Hz, 1H), 4.32 – 4.23 (m, 3H), 4.18 (t, J = 7.8 Hz, 1H), 3.54 (dd, J = 10.9, 4.4 Hz, 1H), 3.49 – 3.35 (m, 3H), 2.20 – 2.01 (m, 2H), 1.96 – 1.80 (m, 2H)



^{13}C NMR (101 MHz, DMSO-d_6) δ = 174.1, 173.6, 153.9, 153.6, 146.8, 145.0, 145.0, 128.0, 127.7, 123.5, 123.3, 68.6, 67.9, 64.8, 64.7, 58.8, 58.4, 55.5, 54.8, 39.5, 38.7.

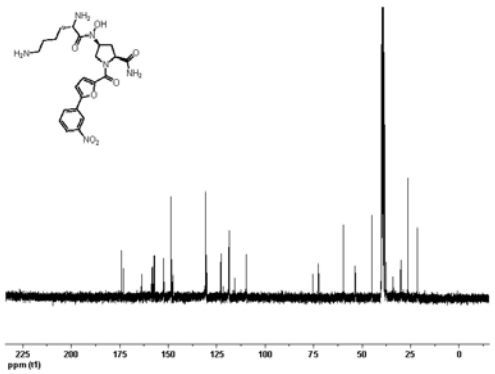
(J)

(2S,4S)-4-((S)-2,6-diamino-N-hydroxyhexanamido)-1-(2-(5-phenyloxazol-2-yl)benzoyl)pyrrolidine-2-carboxamide (M-12)

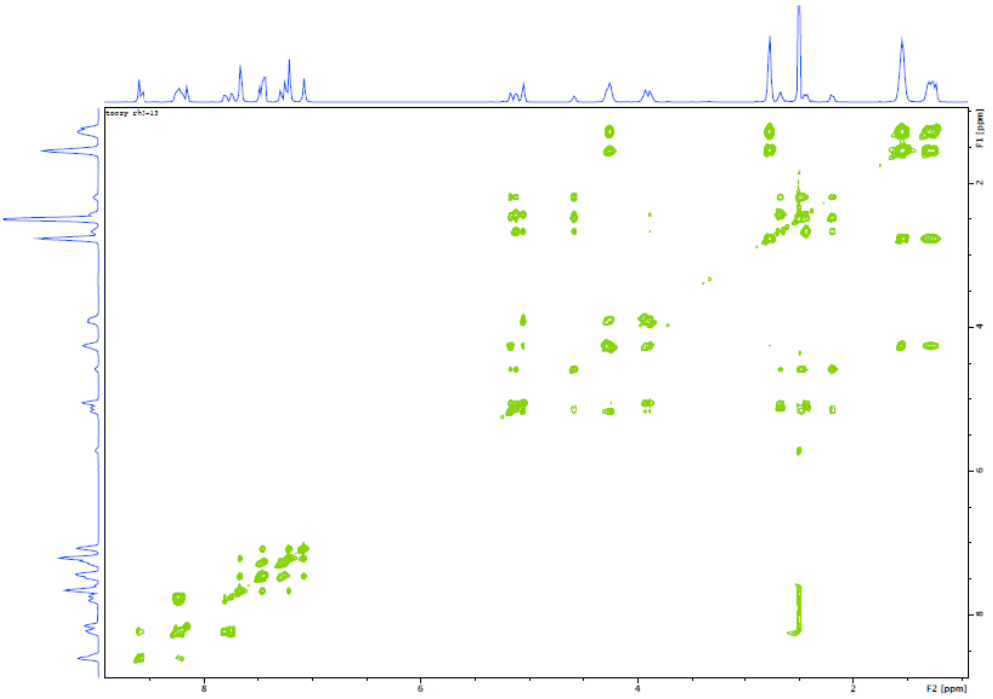


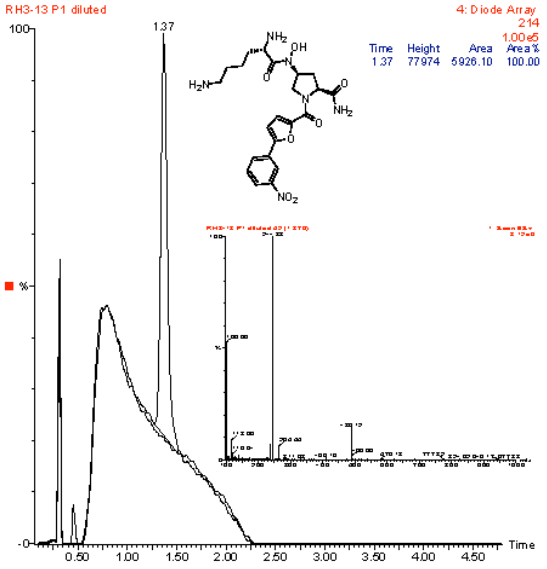
^{1}H NMR (400 MHz, DMSO-d_6, conformer 1) δ = 8.59 (s, 1H), 8.24 (d, *J* = 8.03 Hz, 1H), 8.20 (d, *J* = 1.53 Hz, 1H), 7.66 (s, 1H), 7.72 (t, *J* = 8.04 Hz, 1H), 7.41 (d, *J* = 3.62 Hz, 1H), 7.23 (d, *J* = 3.64 Hz, 1H), 7.21 (s, 1H) 5.10 (d, *J* = 9.67 Hz, 1H), 4.56 (d, *J* = 8.64 Hz, 1H), 4.29 (d, *J* = 7.75 Hz, 1H), 3.86 (dd, *J* = 13.71, 4.57 Hz, 2H), 2.76 (t, *J* = 7.50, 7.50 Hz, 2H), 2.73-2.59 (m, 1H), 2.47-2.46 (m, 1H), 2.45-2.38 (m, 1H), 2.18 (d, *J* = 14.04 Hz, 1H), 1.62-1.47 (m, 4H), 1.36-1.19 (m, 2H);

^{1}H NMR (400 MHz, DMSO-d_6, conformer 2) δ = 8.55 (s, 1H), 8.22 (d, *J* = 1.22 Hz, 1H), 8.19-8.15 (m, 1H), 7.78 (t, *J* = 8.04 Hz, 1H), 7.46 (d, *J* = 3.62 Hz, 1H), 7.45 (s, 1H), 7.27 (d, *J* = 3.58 Hz, 1H), 7.07 (s, 1H) 5.15 (s, 1H), 5.06-5.02 (m, 1H), 4.29 (d, *J* = 7.75 Hz, 1H), 4.28-4.22 (m, 1H), 3.92 (d, *J* = 13.68 Hz, 1H), 3.86 (dd, *J* = 13.71, 4.57 Hz, 2H), 2.76 (t, *J* = 7.50, 7.50 Hz, 2H), 2.71-2.60 (m, 1H), 2.42 (d, *J* = 14.35 Hz, 1H), 1.62-1.47 (m, 4H), 1.36-1.19 (m, 2H);



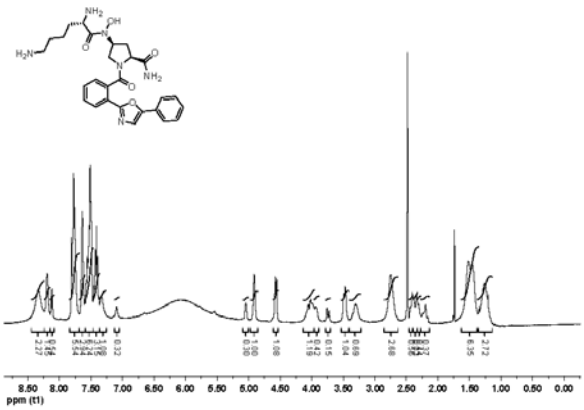
^{13}C NMR (101 MHz, DMSO-d_6) δ = 174.2; 173.0; 163.7; 158.2; 157.9; 157.6; 157.1; 156.8; 152.3; 152.2; 148.4; 148.3; 147.6; 130.8; 130.7; 130.5; 130.2; 130.0; 122.9; 122.8; 118.7; 118.6; 118.4; 118.4; 115.6; 109.8; 109.7; 75.3; 72.4; 59.6; 53.5; 53.3; 44.9; 38.4; 37.5; 33.9; 30.0; 30.0; 29.9; 26.4; 21.5.


F2 [ppm]
F1 [ppm]


RH3-13 P1 diluted
4: Diode Array
214
1.00e5
Time Height Area Area %
1.37 77974 5926.10 100.00
1.37
100
%
0
0.50 1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50
Time

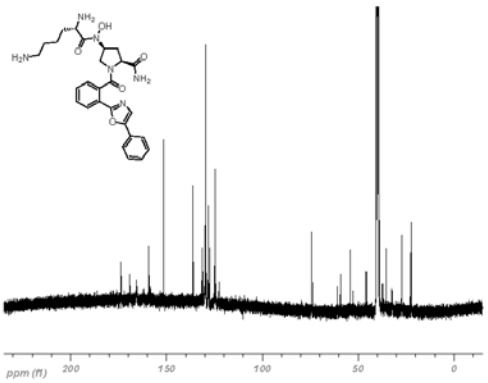
(K)

(2S,4S)-4-((S)-2,6-diamino-N-hydroxyhexanamido)-1-(5-(3-nitrophenyl)furan-2-carbonyl)pyrrolidine-2-carboxamide (M-24)

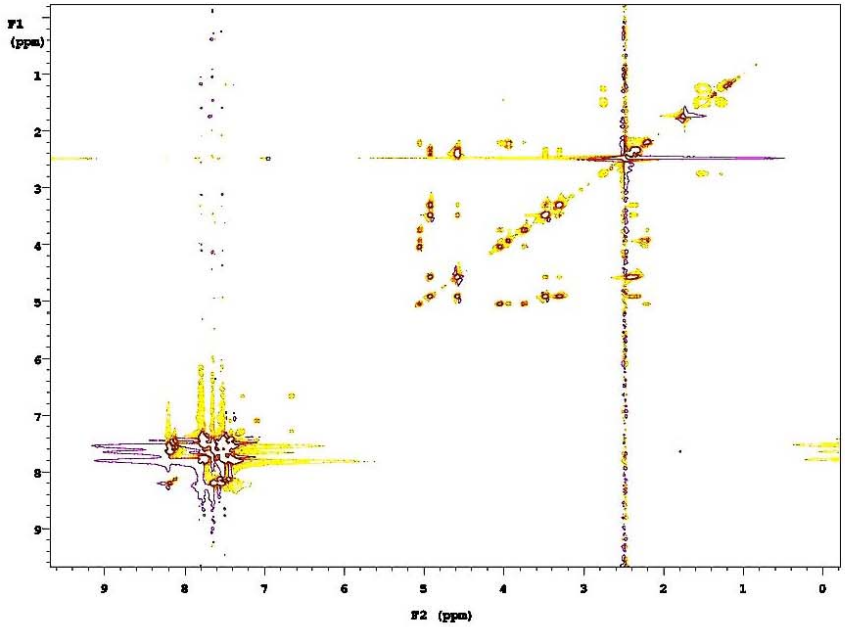


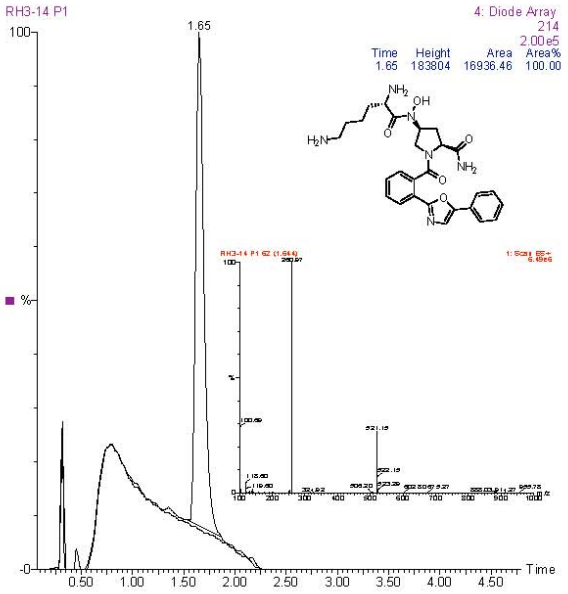
^{1}H NMR (400 MHz, DMSO-d_6, conformer 1) δ = 8.45-8.24 (m, 2H), 8.24-8.16 (m, 1H), 7.84-7.71 (m, 3H), 7.67-7.60 (m, 1H), 7.60-7.47 (m, 3H), 7.47-7.36 (m, 1H), 7.36-7.28 (m, 1H), 4.96-4.86 (m, 1H), 4.68-4.46 (m, 1H), 4.05-3.87 (m, 1H), 3.54-3.40 (m, 1H), 2.86-2.62 (m, 2H), 2.45-2.35 (m, 1H), 2.37-2.28 (m, 1H), 1.62-1.36 (m, 4H), 1.35-1.15 (m, 2H)

^{1}H NMR (400 MHz, DMSO-d_6, conformer 2) δ = 8.44-8.24 (m, 2H), 8.15-8.09 (m, 1H), 7.84-7.70 (m, 3H), 7.67-7.60 (m, 1H), 7.59-7.47 (m, 3H), 7.47-7.37 (m, 1H), 7.13-7.06 (m, 1H), 5.08-5.01 (m, 1H), 4.14-3.99 (m, 1H), 3.79-3.68 (m, 1H), 3.39-3.23 (m, 1H), 2.86-2.65 (m, 2H), 2.39-2.27 (m, 1H), 2.27-2.14 (m, 1H), 1.61-1.37 (m, 2H), 1.36-1.16 (m, 2H)



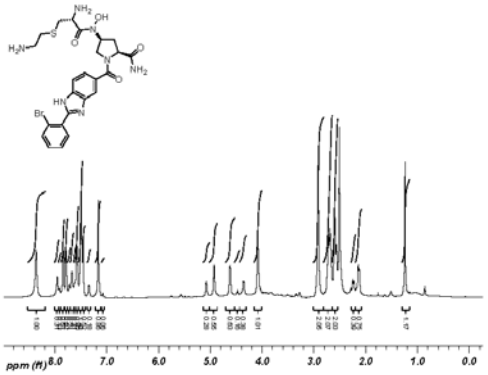
^{13}C NMR (101 MHz, DMSO-d_6) δ = 174.2; 173.0; 163.7; 158.2; 157.9; 157.6; 157.1; 156.8; 152.3; 152.2; 148.4; 148.3; 147.6; 130.8; 130.7; 130.5; 130.2; 130.0; 122.9; 122.8; 118.7; 118.6; 118.4; 118.4; 115.6; 109.8; 109.7; 75.3; 72.4; 59.6; 53.5; 53.3; 44.9; 38.4; 37.5; 33.9; 30.0; 30.0; 29.9; 26.4; 21.5


F1
(ppm)
F2 (ppm)

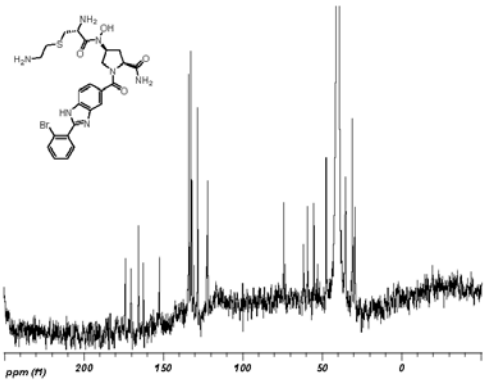

RH3-14 P1
4: Diode Array
214
2.00e5
Time Height Area Area%
1.65 183804 16936.46 100.00
1.65
%
Time

(L)

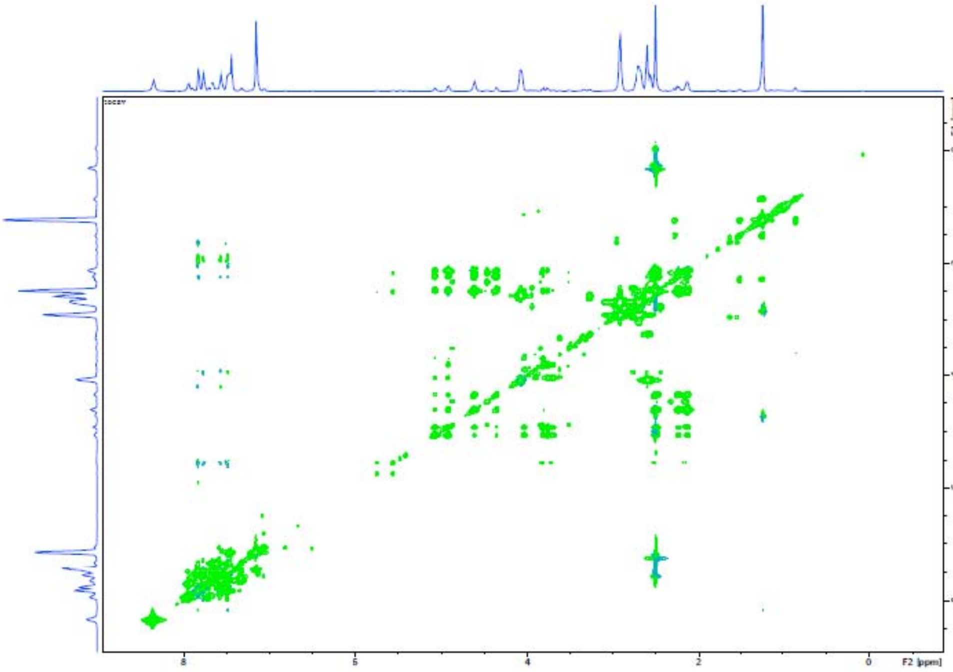
(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(2-(2-bromophenyl)-1H-benzo[d]imidazole-5-carbonyl)pyrrolidine-2-carboxamide (M-26)

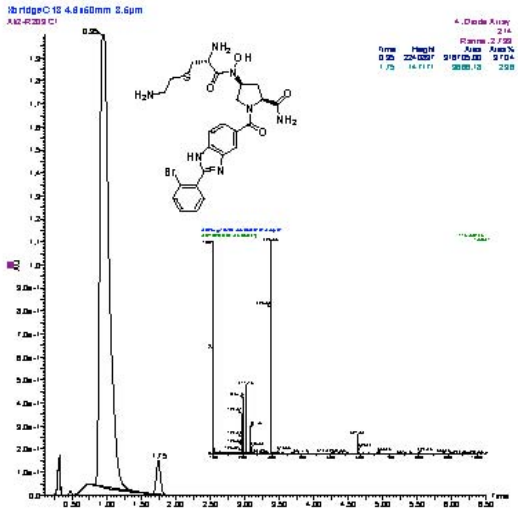


^{1}H NMR (600 MHz, DMSO) (Mixture of 2 Conformers) δ = 8.41 – 8.30 (s, 1H), 7.99 – 7.87 (m, 2H),7.87 – 7.80 (d, *J* = 7.1 Hz, 1H), 7.80 – 7.74 (d, *J* = 6.6 Hz, 1H), 7.71 – 7.53 (m, 2H), 7.53 – 7.42 (m, 3H), 7.40 – 7.29 (m, 1H), 7.20 – 7.10 (s, 1H), 7.10 – 7.02 (s, 1H), 5.13 – 5.00 (s, 1H), 4.99 – 4.85 (s, 1H), 4.71 – 4.54 (m, 1H), 4.40 – 4.26 (m, 1H), 4.16 – 3.96 (s, 1H), 3.00 – 2.82 (s, 2H), 2.76 – 2.63 (m, 2H), 2.62 – 2.57 (m, 2H), 2.57 – 2.53 (s, 1H), 2.33 – 2.17 (d, *J* = 13.2 Hz, 1H), 2.20 – 2.05 (d, *J* = 12.8 Hz, 1H), 1.31 – 1.14 (s, 1H);



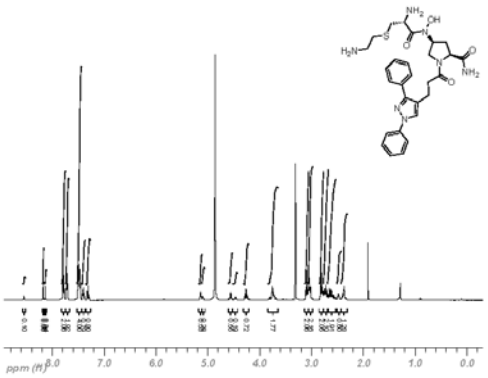
^{13}C NMR (151 MHz, DMSO) δ = 173.69, 173.42, 170.01, 169.74, 165.13, 162.10, 161.85, 152.07, 151.89, 133.45, 132.29, 132.13, 131.62, 130.23, 127.87, 122.13, 121.54, 73.72, 72.81, 61.21, 58.65, 54.72, 52.47, 47.01, 39.52, 34.93, 34.64, 30.46, 28.99.


F2 [ppm]
F1 [ppm]


XbridgeC18 4.6x50mm 3.5µm
Diode Array
NH2
OH
H2N
S
N
O
NH2
HN
N
Br
Time
Height
Area
Area%
0.95
1.75

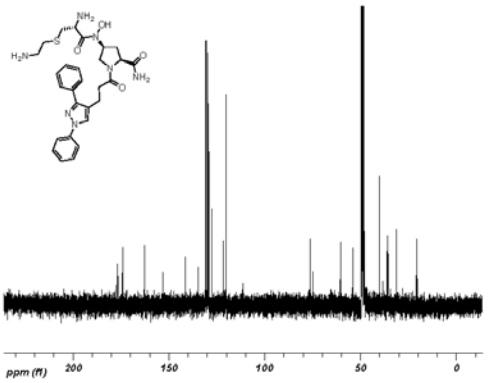
(M)

(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(3-(1,3-diphenyl-1H-pyrazol-4-yl)propanoyl)pyrrolidine-2-carboxamide (M-33)

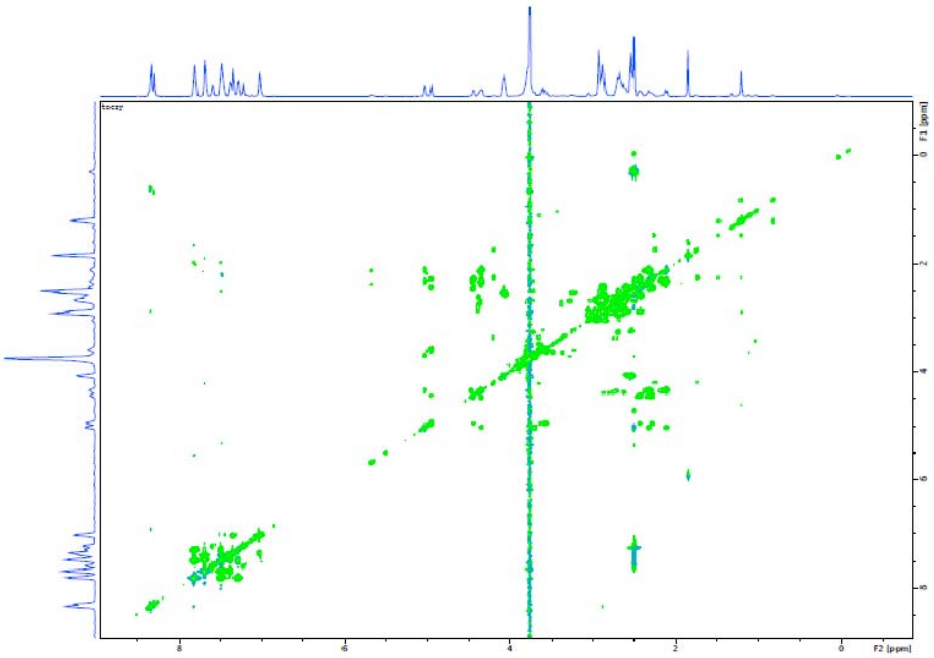


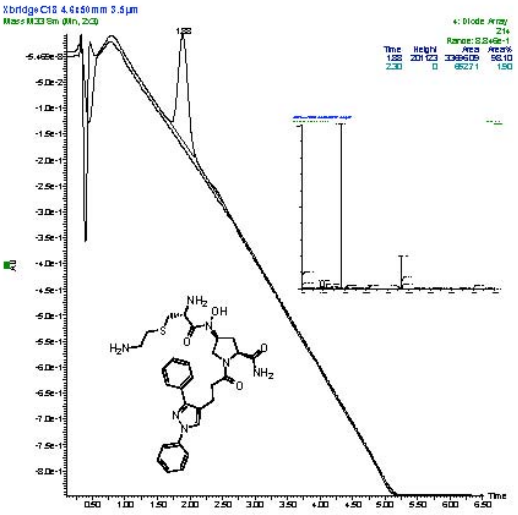
^{1}H NMR (400 MHz, CD_3OD; conformer 1) δ 8.25 – 8.14 (s, 1H), 7.83 – 7.75 (m, 2H), 7.75 – 7.68 (m, 2H), 7.56 – 7.43 (m, 4H), 7.44 – 7.36 (m, 1H), 7.34 – 7.24 (m, 1H), 5.22 – 5.12 (m, 1H), 5.16 – 4.99 (m, 1H), 4.66 – 4.52 (dd, J = 8.6, 3.1 Hz, 1H), , 4.34 – 4.14 (m, 1H), 3.88 – 3.60 (dq, J = 9.4, 4.4 Hz, 3H), 3.15 – 2.95 (m, 6H), 2.89 – 2.51 (m, 4H), 2.42 – 2.31 (m, 1H), 1.99 – 1.80 (s, 1H), 1.35 – 1.17 (s, 1H).

^{1}H NMR (400 MHz, CD_3OD conformer 2) 8.15 - 8.10 (s, 1H), 7.83 - 7.75 (m, 2H), 7.75 - 7.68 (m, 2H), 7.56 - 7.43 (m, 4H), 7.44 - 7.36 (m, 1H), 7.34 - 7.24 (m, 1H), 5.22 - 5.12 (m, 1H), 5.16 - 4.99 (m, 1H), 4.66 - 4.52 (dd, J = 8.6, 3.1 Hz, 1H), 4.53 - 4.41 (m, J = 5.4 Hz, 1H), 3.88 - 3.60 (dq, J = 9.4, 4.4 Hz, 3H), 3.15 - 2.95 (m, 6H), 2.89 - 2.51 (m, 4H), 2.52 - 2.44 (m, 1H), 1.99 - 1.80 (s, 1H), 1.35 - 1.17 (s, 1H).



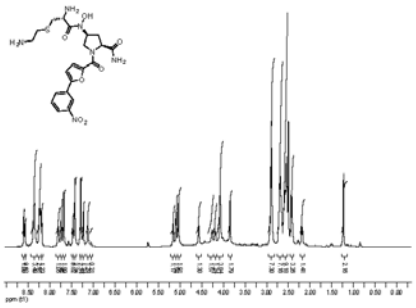
^{13}C NMR (101 MHz, CD_3OD) δ 176.74, 176.56, 174.20, 174.08, 162.50, 153.04, 141.34, 134.73, 130.58, 130.57, 129.66, 129.19, 129.14, 129.12, 129.06, 128.91, 127.51, 121.62, 120.02, 119.99, 76.22, 74.77, 60.76, 60.21, 54.05, 53.82, 47.74, 39.96, 38.38, 36.24, 35.93, 35.67, 35.51, 31.17, 31.12, 20.81, 20.46.


F1 [ppm]
F2 [ppm]


XBridge C18 4.6x50mm 3.5µm
4: Diode Array
Time
Height
Area
Area%
1.88
2.30
Time

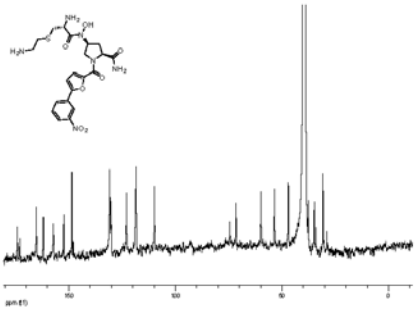
(N)

(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(5-(3-nitrophenyl)furan-2-carbonyl)pyrrolidine-2-carboxamide (M-37)

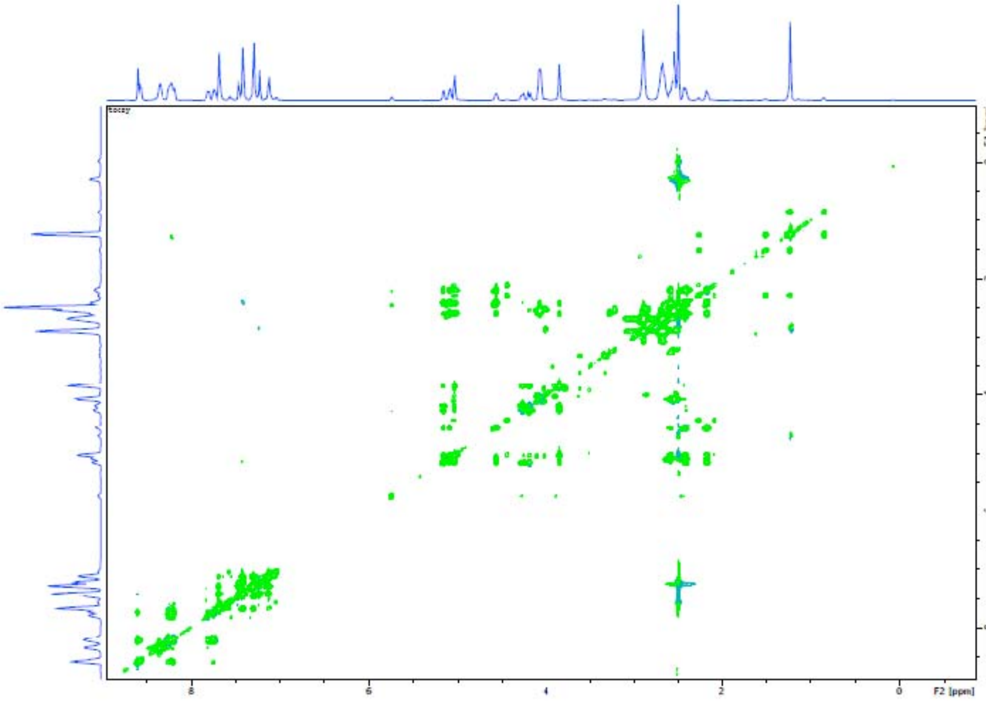


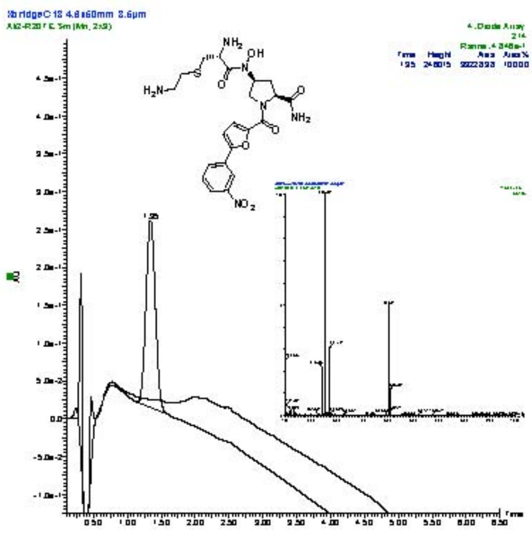
^{1}H NMR (600 MHz, DMSO-d_6, conformer 1) δ = 8.60 (s, 1H), 8.36 (s, 3H), 8.23 (dd, J = 15.6, 7.7 Hz, 2H), 7.80 (t, J = 7.7 Hz, 1H), 7.71 – 7.35 (m, 1H), 7.35 – 7.01 (m, 1H), 5.74 (s, 1H), 5.09 (d, J = 9.3 Hz, 1H), 5.04 (s, 1H), 4.07 (d, J = 5.6 Hz, 1H), 3.84 (s, 2H), 2.93 – 2.84 (m, 2H), 2.74 – 2.64 (m, 2H), 2.63 – 2.52 (m, 2H), 2.43 (d, J = 13.5 Hz, 1H), 2.19 (d, J = 13.3 Hz, 1H)

^{1}H NMR (600 MHz, DMSO-d_6, conformer 2) δ = 8.57 (s, 1H), 8.36 (s, 3H), 8.19 (d, J = 7.9 Hz, 2H), 7.74 (t, J = 7.8 Hz, 1H), 7.71 – 7.35 (m, 1H), 7.35 – 7.01 (m, 1H), 5.74 (s, 1H), 5.16 (s, 1H), 4.57 (d, J = 9.0 Hz, 1H), 4.26 (d, J = 7.2 Hz, 1H), 4.19 (d, J = 11.4 Hz, 1H), 4.07 (d, J = 5.6 Hz, 1H), 2.93 – 2.84 (m, 2H), 2.74 – 2.64 (m, 2H), 2.63 – 2.52 (m, 2H), 2.43 (d, J = 13.5 Hz, 1H), 2.19 (d, J = 13.3 Hz, 1H).



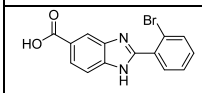
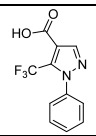
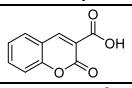
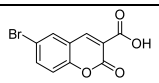
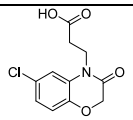
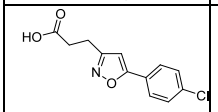
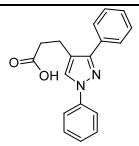
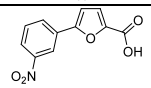
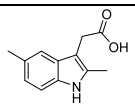
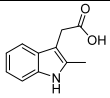
^{13}C NMR (151 MHz, DMSO) δ = 174.07, 172.96, 165.20, 161.92, 157.28, 157.08, 152.32, 148.56, 147.92, 130.87, 130.27, 122.89, 118.41, 109.81, 74.48, 71.57, 59.91, 53.57, 47.0, 37.67, 34.81, 34.11, 30.79, 29.00.

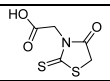
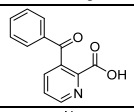
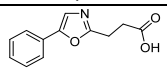
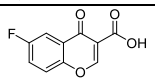
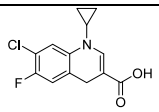
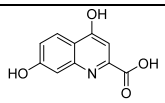
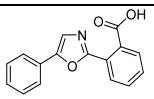
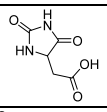
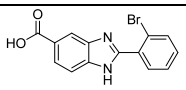
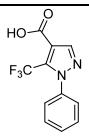
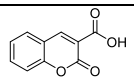
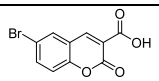
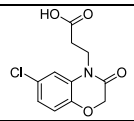
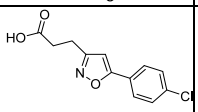
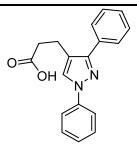
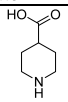
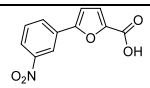


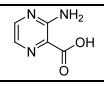
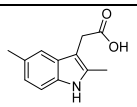
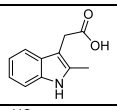
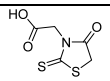
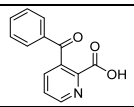
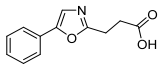
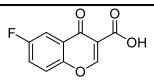
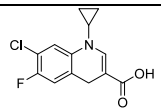
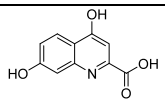
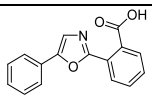
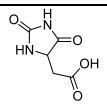


1.2. SUPPLEMENTAL TABLES

Table S1, related to Table 1: Yield, purity, purification methods and predicted binding modes of obtained products from the first library.

R-group	Code	Yield (mg)	Purity [1] (%)	Purification Method[2] (% ACN (time))	Calculated Mass	Found Mass (M+1)	Binding Mode[3]	
							pF	*HS*
	M-01	3.5	84.4	0-17 (18')	571.15	574.16	3	1 or 2
	M-02	9.3	95.0	0-17 (18')	511.22	512.17	3	1 or 2
	M-03	1.9	53.0	0-17 (18')	445.20	446.28	3	1 or 2
	M-04	1.6	90.2	0-17 (18')	523.11	524.23	3	2
	M-06	3.9	98.7	0-17 (18')	510.20	511.23	1 or 3	2
	M-07	4.0	93.9	0-17 (18')	444.21	445.21	3	1
	M-08	1.1	90.2	0-17 (18')	547.29	548.22	3	2
	M-09	3.4	91.4	Isocratic 0 (14')	383.16	384.09	1	2
	M-11	12.0	85.8	Isocratic 0 (14')	384.25	385.04	3	4
	M-12	6.3	89.4	0-17 (18')	488.20	489.18	3	2
	M-13	10.4	92.7	Isocratic 0 (14')	394.21	395.21	1	3
	M-14	4.2	86.5	Isocratic 0 (14')	367.19	368.01	1 or 3	2
	M-15	1.4	99.0	0-17 (18')	474.26	475.24	4	2
	M-16	0.9	93.9	Isocratic 0 (14')	444.25	445.21	3	3

	M-17	1.0	95.3	Isocratic 0 (14')	446.14	447.13	3	1 or 2
	M-18	4.2	86.7	0-17 (18')	482.23	483.15	1	3
	M-19	2.0	86.5	0-17 (18')	472.24	473.23	4	4
	M-20	1.3	95.3	Isocratic 0 (14')	445.20	446.22	3	2
	M-21	1.7	76.5	Isocratic 0 (14')	383.16	384.03	3	1
	M-22	2.0	89.5	0-17 (18')	536.20	537.17	1	1 or 2
	M-23	3.0	95.9	Isocratic 0 (14')	460.21	461.23	3	3
	M-24	0.5	99.0	0-17 (18')	520.24	521.31	4	3
	M-25	4.7	92.0	Isocratic 0 (14')	413.20	414.12	1	1
	M-26	3.2	96.0	10-20 (10')	589.11	592.18	3	1
	M-27	8.1	90.9	20-23 (10')	529.17	530.26	3	2
	M-28	-	-	0-20 (10')	463.15	-	1	2
	M-29	-	-	20-30 (10')	541.06	-	1 or 3	1 or 2
	M-31	-	-	15-20 (10')	528.15	-	1	2
	M-32	11.2	98.5	17-20 (10')	462.16	463.26	3	2
	M-33	6.5	92.0	35-45 (10')	565.24	566.38	3	2
	M-34	5.1	>99.9	10-30 (10')	401.11	402.23	1	3
	M-36	10.1	96.9	0-2 (10')	402.20	403.29	1	4
	M-37	8.3	88.2	15-25 (10')	506.15	507.24	3	2

	M-38	20.1	86.8	15-30 (10')	412.16	413.20	3	2
	M-39	12.6	84.9	0-10 (10')	385.14	386.23	1 or 3	2
	M-40	12.7	90.0	15-30 (10')	492.21	493.29	3	2
	M-41	11.4	88.1	15-30 (10')	462.20	463.26	3	2
	M-42	4.9	92.8	5-12 (10')	464.09	465.18	1 or 3	3
	M-43	8.6	92.6	15-30 (10')	500.18	491.24	1	2 or 3
	M-44	3.9	96.5	10-25 (10')	490.19	491.24	1 or 3	2
	M-45	-	-	5-30 (10')	463.15	-	3	2
	M-46	6.5	91.3	6-8.5 (10')	401.11	402.16	1	2
	M-47	9.0	92.7	15-25 (10')	554.15	555.20	1	2
	M-48	3.5	93.1	0-5 (6')	478.16	479.27	3	1 or 2
	M-49	9.0	95.8	20-30 (10')	538.19	539.26	4	3
	M-50	-	-	0-20 (10')	431.15	-	1	1

[1] Purities were determined by RP-HPLC-MS (220nm): Column C_{18} X-bridge (4.6 x 50 mm x 3.6 μm); gradient; 5 to 100% B in 4.5 minutes; A: H_2O-HCOOH (99.9:0.1), B: ACN-HCOOH (99.3:0.7)

[2] Final products were purified by preparative RP-HPLCMS using different linear gradients of H_2O (containing 0.1% HCOOH) and ACN (containing 0.07% HCOOH) at a flow rate of 16 mL/min. Column: Sunfire™ OBD C_{18}, (19 x 100 mm x 5 μm)

[3] 1 = Lysine-like; 2 = Adenine-like; 3 = LAD-like; 4 = none

2. SUPPLEMENTAL EXPERIMENTAL PROCEDURES

2.1 SOLID-PHASE SYNTHESIS OF THE LYSYL-ADENYLATE ANALOGUE LIBRARY

2.1.1 Materials and Equipment.

All chemical reagents were obtained from commercial suppliers and used without further purification. ^{1}H and ^{13}C NMR spectra were recorded at 298 K on a Varian Mercury-400 Fourier Transform spectrometer. Chemical shifts are reported in δ units (ppm) relative to the residual deuterated solvent signals of $CHCl_3$ (^{1}H NMR: δ 7.26; ^{13}C NMR: δ 77.0); DMSO (^{1}H NMR: δ 2.50; ^{13}C NMR: δ 39.5); MeOH (^{1}H NMR: δ 3.31 & 4.84; ^{13}C NMR: δ 49.1). The splitting patterns are designated as follows: s (singlet), d (doublet), t (triplet), q (quartet), b (broad). The RP-HPLC analyses were performed on a Waters Alliance instrument and RP-HPLC-MS on a Waters Alliance instrument coupled to a Micromass ZQ spectrometer with an electrospray (ES) probe. The purifications by preparative RP-HPLC were performed on a Waters HPLC Autopurification FractionLynx UV/MS system with an electrospray (ES) probe. Analytical thin-layer chromatography (TLC) was performed on precoated plates (Merck silica gel 60ACC, F254). Visualization of the developed chromatogram was achieved with UV light. Manual flash column chromatography was performed using silica (Merck, 70-230 mesh). Automated flash chromatography was performed on a Teledyne Isco module Companion® with photodiode array detector using Silica-RediSep® columns.

2.1.2. Reactions and couplings.

Fmoc removal. The Fmoc group was removed using the following reaction conditions: (i) DMF (5 × 1 min); (ii) piperidine/DMF (2:8) (1 × 1 and 2 × 15 min); (iii) DMF (5 × 1 min).

Coupling to resin. The resin was washed with anhydrous DCM (5 × 1 min) and DMF (5 × 1 min). After it was washed, it was treated with Alloc-Hyp-OH (5 eq), DIPCDI (5 eq), and HOBt (5 eq) in

DMF for 2 h. After the products were washed with DMF (5 × 1 min) and DCM (5 × 1 min), the extension of the coupling was monitored by the Kaiser test.

Mitsunobu Coupling. To the peptidyl resins preswollen in anhydrous DCM (5 mL) was added PPh_3 (7 eq) and Boc-L-Lys(Boc)-OPMB (**3**) (7 eq). The mixtures were shaken until the reagents were completely dissolved. Then, DIAD (7 eq) was added dropwise at 0 °C, and the mixture was shaken overnight at room temperature. After that, the solvent was removed by suction; the resins were washed with DCM (5 × 1 min) and then were washed twice with DMF, DCM, methanol, and finally, with DCM.

Alloc Group Removal. Removal of the Alloc group was achieved with $Pd(PPh_3)_4$ (0.1 eq) in the presence of $PhSiH_3$ (10 eq) in DCM under Ar (2 × 20 min, 25°C).

N-Acylation. After the NR-Alloc group had been removed, acylations of the R-amino groups were carried out using RCOOH (5 eq), DIPCDI (5 eq), and HOBt (5eq) in DMF for 2 h at 25 °C. Resins were washed with DMF (5 × 1 min) and DCM (5 × 1 min). Acylations were monitored by the chloranil test.

Cleavage from resin. Rink amide resins were cleaved with TFA/H2O (95:5) for 4 h at room temperature. TFA was evaporated; the compounds were dissolved in H2O-MeCN and then lyophilized.

2.1.3 Characterization of the compounds.

2-(4-methoxybenzyloxy)isoindoline-1,3-dione (72) - A mixture of 13.1 g (80.0 mmol) of N-hydroxyphthalimide (**70**), 26.6 ml (192 mmol) of Et_3N and 10.9 ml (80.0 mmol) of *p*-methoxybenzyl chloride (**71**) in 200 ml of DMF was stirred for 1 h at 90°C. The reaction mixture was poured into 250 ml of ice-water. A precipitate was formed which was collected by filtration. The precipitate was dried in vacuum and used in the next step as obtained. The product was obtained as a white solid (17.9 g; 63.3 mmol; 79%). ^{1}H NMR (400 MHz, $CDCl_3$) d = 7.83-7.77 (m, 2H), 7.75-7.71 (m, 2H), 7.45 (d, *J* = 8.70 Hz, 2H), 6.89 (d, *J* = 8.71 Hz, 2H), 5.15 (s, 2H), 3.80 (s,

3H); ^{13}C NMR (101 MHz, $CDCl_3$) d = 163.5; 160.4; 134.3; 131.6; 128.8; 125.8; 123.4; 113.9; 79.4; 55.2.

O-(4-methoxybenzyl)hydroxylamine (73) - To a suspension of 17.0 g (60.0 mmol) of 2-(4-methoxybenzyloxy)isoindoline-1,3-dione (**72**) in 300 ml of EtOH, 3.45 ml (65.9 mmol) of monomethylhydrazine were added. The reaction mixture was refluxed for 1 h. The reaction mixture was concentrated *in vacuo* and the resulting white solid was suspended in ether. The suspension was allowed to stand at RT for 30 min The solid was removed by filtration and the organic layer was concentrated to obtain the desired product as an yellow oil (6.63g; 43.3 mmol; 72.2%). ^{1}H NMR (400 MHz, $CDCl_3$) d = 7.30 (d, *J* = 8.65 Hz, 2H), 6.90 (d, *J* = 8.67 Hz, 2H), 5.37 (bs, 2H), 4.62 (s, 2H), 3.81 (s, 3H); ^{13}C NMR (101 MHz, $CDCl_3$) d = 159.4; 130.0; 129.3; 113.8; 77.5; 55.2.

di-*tert*-butyl 6-(4-methoxybenzyloxyamino)-6-oxohexane-1,5-diyldicarbamate (3) - A solution of 5.50 g (15.9 mmol) of BocLys(Boc)OH (**74**), 3.14 g (16.4 mmol) of WSC·HCl and 2.43 g (15.9 mmol) of HOBt·H_2O in 75 ml of DCM was stirred for 20 min at RT. To the reaction mixture, a solution of 2.44 g (15.9 mmol) of O-(4-methoxybenzyl)hydroxylamine was added (**73**) in 10 ml of DCM. The resulting reaction mixture was stirred overnight at RT. The organic phase was washed with 0.1 M $HCl_{(aq)}$ (2 x 75ml). During the first washing a white precipitate was formed which was removed by filtration. Then, the organic phase was washed twice with saturated $NaHCO_{3(aq)}$, brine, dried over $MgSO_4$, filtered and concentrated. The crude was purified by automated flash column chromatography. (ISCO; SiO_2, 120 g, Hexane/EtOH, 230 nm) (TLC: SiO_2; Hexane/EtOH (4:1); R_f = 0.45). The desired product was obtained as a white solid (5.43 g; 11.3 mmol; 71%). The purity was >98.5% (determined by C_{18} RP-HPLC-MS, t_R 3.08 min, λ = 220nm) MS (ES^+) calcd. for $C_{24}H_{39}N_3O_7$: (m/z) 481.28; found: [M+H]= 482.22. ^{1}H NMR (400 MHz, DMSO-d_6) d = 11.05 (bs, 1H), 7.31 (d, *J* = 8.59 Hz, 2H), 6.92 (d, *J* = 8.57 Hz, 2H), 6.84 (bd, *J* = 7.82 Hz, 1H), 6.74 (bt, *J* = 4.94 Hz, 1H), 4.68 (s, 2H), 3.75 (s, 3H), 3.73-3.67 (m, 1H), 2.90-2.82 (m, 2H), 1.50-1.41 (m, 2H), 1.37 (s, 9H), 1.36 (s, 9H), 1.34-1.26 (m, 2H), 1.26-1.10 (m, 2H); ^{13}C NMR (101 MHz, DMSO-d_6) d

= 168.9; 159.2; 155.4; 155.1; 130.5; 127.7; 113.5; 77.8; 77.2; 76.2; 54.9; 51.9; 31.4; 29.0; 28.1; 28.0; 22.6.

tert-butyl (2-hydroxyethyl)carbamate (77) - To an ice-cooled solution of 30.0 g (138mmol) of *tert*-butyl-dicarbonate (**76**) in 60 ml of anhydrous DCM under nitrogen was added 8.30 ml (138mmol) of 2-aminoethanol (**75**). The resultant reaction mixture was stirred for 1.5 h at 0°C. The organic phase was washed successively with saturated $NaHCO_{3(aq)}$, brine, dried on $MgSO_4$, filtered and concentrated *in vacuo* to give the title compound as a pale yellow oil which was used in the next step without any additional purification (22.0 g; 136mmol; 99%). MS (ES^+) calcd. for $C_7H_{15}NO_3$: (m/z) 161.11; found: [M+H]= 161.95. ^1H NMR (400 MHz, $CDCl_3$) d = 3.69 – 3.64 (m, 2H), 3.26 (t, *J* = 5.1 Hz, 2H), 1.43 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) d = 156.8, 79.6, 62.0, 43.1, 28.4.

2-((tert-butoxycarbonyl)amino)ethyl methanesulfonate (78) - To an ice-cooled stirred solution of 22.0 g (136 mmol) of tert-butyl 2-hydroxyethylcarbamate (**77**) and 22 mL of anhydrous pyridine (273 mmol) in 60 ml of anhydrous DCM under nitrogen was added slowly 12.8 mL (165 mmol) of methanesulfonyl chloride. The resulting mixture was stirred 50 min while warming up to RT. The reaction mixture was washed successively with 5% $HCl_{(aq)}$ and saturated $NaHCO_{3(aq)}$, dried on $MgSO_4$, filtered and concentrated *in vacuo* to give the compound as a yellow oil which was used in the next step without any additional purification (29.8 g; 125mmol; 91%). MS (ES^+) calcd. for $C_8H_{17}NO_5S$: (m/z) 239.08; found [M+H]= 239.92. ^1H NMR (400 MHz, $CDCl_3$) d = 4.27 (t, *J* = 5.1 Hz, 2H), 3.45 (dd, *J* = 4.0 Hz, 2H), 3.02 (s, 3H), 1.43 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) d = 155.9, 80.1, 69.0, 40.1, 37.5, 28.4.

Methyl 2-((tert-butoxycarbonyl)amino)-3-((2-((tert-butoxycarbonyl)amino)ethyl)thio)propanoate (79) - To an ice-cooled stirred solution of 25.6 mL (124 mmol) of N-(tert-Butoxycarbonyl)-L-cysteine methyl ester and 48.7 g (149mmol) of Cs_2CO_3 in 30 ml of DMF was added a solution of 29.8 g (125 mmol) of 2-((tert-

butoxycarbonyl)amino)ethyl methanesulfonate (**78**) in 10 ml of DMF. The resulting mixture was stirred at 0°C for 30 min. Stirring was continued for 4 h while the reaction mixture warmed up to RT. The reaction mixture was filtered to remove the excess of Cs_2CO_3, and the organic layer was concentrated *in vacuo*. The crude was dissolved in DCM, washed successively with brine, dried on $MgSO_4$, filtered and concentrated *in vacuo* to give the compound as yellow oil which was used in the next step without any additional purification. MS (ES^+) calcd. for $C_{16}H_{30}N_2O_6S$: 378,18 (m/z); found [M+H]=379.24.

2-((tert-butoxycarbonyl)amino)-3-((2-((tert-butoxycarbonyl)amino)ethyl)thio)propanoic acid (BocThialysine(Boc)OH) (80) – To a solution of crude methyl 2-((tert-butoxycarbonyl)amino)-3-((2-((tert-butoxycarbonyl)amino)ethyl)thio)propanoate (**79**) in 50 ml of a mixture of H_2O/MeOH (1:1) was added 14.9 g (623 mmol) of LiOH. The reaction mixture was stirred overnight at RT. The solution was filtered to remove the excess of LiOH and then acidified with 2 M $HCl_{(aq)}$ to pH = 3. The product was extracted with DCM, washed with brine, dried on $MgSO_4$, filtered and concentrated *in vacuo* to give the desired compound as a yellow oil (26.2 g, 71.8 mmol, 58% over 2 steps). The purity was >86.6% (determined by C_{18} RP-HPLC-MS, t_R 2.70 min, λ = 214nm) MS (ES^+) calcd. for $C_{15}H_{28}N_2O_6S$: 364.17 (m/z); found [M+H]= 365.06. ^{1}H NMR (400 MHz, DMSO-d_6) δ = 7.06 (d, *J* = 8.3 Hz, 1H), 6.88 (t, *J* = 5.2 Hz, 1H), 4.03 (td, *J* = 7.2, 4.6 Hz, 1H), 3.07 (dd, *J* = 13.3, 6.5 Hz, 2H), 2.87 (dd, *J* = 13.5, 4.6 Hz, 1H), 2.72 (dd, *J* = 13.4, 9.4 Hz, 1H), 2.57 – 2.51 (m, 2H), 1.38 (s, 9H), 1.37 (s, 9H). ^{13}C NMR (101 MHz, DMSO-d_6) δ = 172.5, 155.4, 78.2, 77.7, 53.7, 32.7, 31.4, 28.2.

BocThialysine(Boc)-NH-OPMB (4) - A solution of 26.2 g (71.8mmol) of BocThialysine(Boc)OH (**80**), 13.8 g (71.8 mmol) of WSC·HCl and 11.0 g (71.8 mmol) of HOBt·H_2O in 120ml of DCM was stirred for 20 min at RT. To the reaction mixture was added a solution of 11.0 g (71.8 mmol) of O-(4-methoxybenzyl)hydroxylamine (**73**) in 10 ml of DCM. The resulting reaction mixture was stirred overnight at RT. The organic phase was washed with 0.1 M $HCl_{(aq)}$ (2 x 75ml). During the first washing a white precipitate was formed which was removed by filtration. Then, the organic phase

was washed twice with saturated $NaHCO_{3(aq)}$, brine, dried on $MgSO_4$, filtered and concentrated. The product was precipitated from DCM/Hexane (20:80) at -4ºC. The desired product was obtained as a white solid (26.7 g; 53.5 mmol; 75%). The purity was >85.5% (determined by C_{18} RP-HPLC-MS, t_R 3.15 min, λ = 214nm) MS (ES^+) calcd. for $C_{23}H_{37}N_3O_7S$: (m/z) 499.24; found: [M+H]= 500.17. ^1H NMR (400 MHz, $CDCl_3$) δ = 9.37 (s, 1H), 7.33 (d, *J* = 8.6 Hz, 2H), 6.88 (d, *J* = 8.5 Hz, 2H), 5.35 (d, *J* = 3.7 Hz, 1H), 4.91 (d, *J* = 6.2 Hz, 1H), 4.86 (s, 2H), 4.17 (s, 1H), 3.80 (s, 3H), 3.39 – 3.19 (m, 2H), 2.90 – 2.75 (m, 2H), 2.65 (t, *J* = 5.2 Hz, 2H), 1.73 (s, 1H), 1.43 (s, 18H). ^{13}C NMR (101 MHz, $CDCl_3$) δ = 168.2, 160.1, 156.2, 155.5, 131.1, 127.3, 114.0, 78.1, 55.4, 51.9, 39.9, 34.4, 32.9, 28.5.

2.2 SYNTHESIS IN SOLUTION OF COMPOUNDS M-12, M-24, M-26, M-33 and M-37

2.2.1 Characterization of the intermediate compounds

(2S,4R)-4-hydroxy-1-((4-nitrobenzyloxy)carbonyl)pyrrolidine-2-carboxylic acid (59) - To an ice-cooled solution of 10.0 g (76.2 mmol) of *L*-hydroxyproline (**57**) in 80 ml of 2 M $NaOH_{(aq)}$ was added dropwise a solution of 16.4 g (76.2 mmol) of 4-nitrobenzyl carbonochloridate (**58**) in 30 ml of DCM. The resulting reaction mixture was stirred for 1 h. The 2 phases were separated and the aqueous phase was washed once with DCM. The aqueous phase was acidified with concentrated H_2SO_4 (~5ml) to pH = 2-3. The desired product precipitated and was collected by filtration. The aqueous solution was cooled in the refrigerator and a second batch of product was collected. The product was obtained as a pale yellow solid, which was used in the next step without any additional purification (13.4 g; 43.2 mmol; 56.7%). The product was obtained as a mixture of two rotamers. MS(ES^+) calculated for $C_{13}H_{14}N_2O_7$: (m/z) 310.08; found: [M+H]= 310.97. ^1H NMR (400 MHz, CD_3OD) δ = 8.21 (t, *J* = 8.3 Hz, 2H), 7.58 (dd, *J* = 13.7, 8.8 Hz, 2H), 5.36 – 5.09 (m, 2H), 4.44 (dt, *J* = 25.9, 8.1 Hz, 2H), 3.69 – 3.48 (m, 2H), 2.39 – 2.25 (m, 1H), 2.16 – 2.03 (m, 1H). ^{13}C NMR (101

MHz, CD_3OD) δ = 166.7, 166.4, 146.6, 136.2, 136.1, 119.7, 119.6, 115.1, 115.0, 61.2, 60.5, 57.5, 57.4, 49.9, 49.6, 46.7, 46.3, 39.5, 30.8, 29.9.

(2S,4R)-4-nitrobenzyl 2-carbamoyl-4-hydroxypyrrolidine-1-carboxylate (60) – To an ice-cooled solution of 11.4 g (36.7 mmol) of (2S,4R)-4-hydroxy-1-((4-nitrobenzyloxy)carbonyl)pyrrolidine-2-carboxylic acid (**59**) in 30ml of anhydrous acetonitrile was added 8.40 g (43.8 mmol) of WSC·HCl and 6.72 g (43.9 mmol) of HOBt·H_2O. The resulting reaction mixture was stirred overnight while warming up to RT. The reaction mixture was cooled in an ice-bath and 10 ml of a 32% $NH_4OH_{(aq)}$ solution were added. The reaction mixture was stirred for 30 min at 0ºC and then 1 h at RT. The insolubles were removed by filtration and the filtrate was concentrated. The obtained crude was purified by automated flash column chromatography. (ISCO-R_f; SiO_2, 120g, DCM/MeOH, 272nm) (TLC: DCM/MeOH (4:1); UV=254nm; R_f=0.65) The desired product was obtained as a white solid (7.5g; 24.3 mmol; 66.3%). The purity was >99.0% (determined by C_{18} RP-HPLC-MS, t_R 1.53 min, λ = 220nm) The product was obtained as a mixture of two rotamers. MS (ES^+) calcd. for $C_{13}H_{15}N_3O_6$: (m/z) 309.10; found: [M+H]= 309.97. ^{1}H NMR (400 MHz, DMSO-d_6) δ = 8.23 (d, *J* = 8.7 Hz, 1H), 8.19 (d, *J* = 8.8 Hz, 1H), 7.63 (t, *J* = 8.2 Hz, 4H), 7.53 (s, 1H), 7.43 (s, 1H), 7.04 (s, 1H), 6.93 (s, 1H), 5.25 (d, *J* = 14.5 Hz, 1H), 5.22 – 5.19 (m, 2H), 5.13 (d, *J* = 14.5 Hz, 1H), 5.06 (t, *J* = 3.8 Hz, 1H), 4.32 – 4.23 (m, 3H), 4.18 (t, *J* = 7.8 Hz, 1H), 3.54 (dd, *J* = 10.9, 4.4 Hz, 1H), 3.49 – 3.35 (m, 3H), 2.20 – 2.01 (m, 2H), 1.96 – 1.80 (m, 2H). ^{13}C NMR (101 MHz, DMSO-d_6) δ = 174.1, 173.6, 153.9, 153.6, 146.8, 145.0, 145.0, 128.0, 127.7, 123.5, 123.3, 68.6, 67.9, 64.8, 64.7, 58.8, 58.4, 55.5, 54.8, 39.5, 38.7.

2.2.2 Characterization of the final compounds M-12, M-24, M-26, M-33, M-37

A solution of 1.00 g (3.23 mmol) of (2S,4R)-4-nitrobenzyl 2-carbamoyl-4-hydroxypyrrolidine-1-carboxylate (**60**), 3.89 mmol of BocThialysine(Boc)-NH-OPMB (**4**) or BocLys(Boc)-NH-OPMB (**3**), and 2.55 g (9.72 mmol) of triphenylphosphine in 15 ml of THF was stirred for 15 min at RT under an Ar-atmosphere. Then 1.53 ml (9.72 mmol) of diethyl azodicarboxylate (DEAD) were added slowly. The resulting reaction mixture was stirred overnight at RT under an Ar-atmosphere. LCMS

showed the presence of the desired product in addition to some side-products. No starting material was detected. The reaction mixture was concentrated and the resulting residue was dissolved in 50 ml of Et_2O. The organic phase was left overnight in the fridge. A white precipitate was formed which was removed by filtration. The crude was purified by two automated flash column chromatographies (ISCO-R_f; SiO_2; DCM: DCM/MeOH (4:1); 270 and 210 nm). The desired product was obtained as a 1:1 mixture of the product with triphenylphosphine oxide. The mixture was used in the next steps without any further purification.

To a solution of the aforementioned crude in 10 ml of MeOH a spatula point of 10% Pd/C was added. The suspension was stirred overnight under an H_2-atmosphere. The reaction mixture was filtered over Celite© and the filtrate was concentrated. The crude yellow oil was divided and used in the next step without any further purification. To a solution of 1 equivalent of the appropriate carboxylic acid, 1 equivalent of HOBt·H_2O and 1 equivalent of WSC·HCl in 15 ml of DCM/DMF (9:1) was added 1 batch of the aforementioned crude. The resulting reaction mixture was stirred overnight. The reaction mixture was diluted with 20 ml of DCM. The organic layer was washed with 5% $NaHCO_{3(aq)}$ (2 x 15ml), 0.5% citric $acid_{(aq)}$ (15 ml) and brine (15 ml), filtered over a phase separator and concentrated. The crude was dissolved in 10 ml of 40% of TFA in DCM. The solution was stirred for 18 h at RT. The reaction mixture was concentrated and the obtained crude was purified immediately by semi-preparative HPLC. (X-bridge; ACN/ H_2O (20mM $NH_4HCO_{3(aq)}$)); detection by corresponding mass and wavelength. The products were lyophilized and re-purified by semi-preparative HPLC. (X-bridge; ACN/H_2O (0.1% HCO_2H).

(2S,4S)-4-((S)-2,6-diamino-N-hydroxyhexanamido)-1-(2-(5-phenyloxazol-2-yl)benzoyl)pyrrolidine-2-carboxamide (M-12) – Purification: detection mass 521 m/z and wavelength 311 nm. The product was obtained as a white solid after lyophilizing. (78.9 mg; 0.15 mmol; 10.1% over 4 steps) The purity was >99.5% (determined by C_{18} RP-HPLC-MS, t_R 1.65 min, λ = 254nm). The product was obtained as a mixture of 2 conformers in a ratio of 10:7. HRMS (ES^+) calcd. for $C_{27}H_{32}N_6O_5$: (m/z) 520.2434; found: [M+H]= 521.2508. ^{1}H NMR (400 MHz, DMSO-d_6, conformer 1) d = 8.59 (s, 1H), 8.24 (d, *J* = 8.03 Hz, 1H), 8.20 (d, *J* = 1.53 Hz, 1H), 7.66 (s, 1H),

7.72 (t, *J* = 8.04 Hz, 1H), 7.41 (d, *J* = 3.62 Hz, 1H), 7.23 (d, *J* = 3.64 Hz, 1H), 7.21 (s, 1H) 5.10 (d, *J* = 9.67 Hz, 1H), 4.56 (d, *J* = 8.64 Hz, 1H), 4.29 (d, *J* = 7.75 Hz, 1H), 3.86 (dd, *J* = 13.71, 4.57 Hz, 2H), 2.76 (t, *J* = 7.50, 7.50 Hz, 2H), 2.73-2.59 (m, 1H), 2.47-2.46 (m, 1H), 2.45-2.38 (m, 1H), 2.18 (d, *J* = 14.04 Hz, 1H), 1.62-1.47 (m, 4H), 1.36-1.19 (m, 2H); ^{1}H NMR (400 MHz, DMSO-d_6, conformer 2) d = 8.55 (s, 1H), 8.22 (d, *J* = 1.22 Hz, 1H), 8.19-8.15 (m, 1H), 7.78 (t, *J* = 8.04 Hz, 1H), 7.46 (d, *J* = 3.62 Hz, 1H), 7.45 (s, 1H), 7.27 (d, *J* = 3.58 Hz, 1H), 7.07 (s, 1H) 5.15 (s, 1H), 5.06-5.02 (m, 1H), 4.29 (d, *J* = 7.75 Hz, 1H), 4.28-4.22 (m, 1H), 3.92 (d, *J* = 13.68 Hz, 1H), 3.86 (dd, *J* = 13.71, 4.57 Hz, 2H), 2.76 (t, *J* = 7.50, 7.50 Hz, 2H), 2.71-2.60 (m, 1H), 2.42 (d, *J* = 14.35 Hz, 1H), 1.62-1.47 (m, 4H), 1.36-1.19 (m, 2H); ^{13}C NMR (101 MHz, DMSO-d_6) δ = 174.2; 173.0; 163.7; 158.2; 157.9; 157.6; 157.1; 156.8; 152.3; 152.2; 148.4; 148.3; 147.6; 130.8; 130.7; 130.5; 130.2; 130.0; 122.9; 122.8; 118.7; 118.6; 118.4; 118.4; 115.6; 109.8; 109.7; 75.3; 72.4; 59.6; 53.5; 53.3; 44.9; 38.4; 37.5; 33.9; 30.0; 30.0; 29.9; 26.4; 21.5.

(2S,4S)-4-((S)-2,6-diamino-N-hydroxyhexanamido)-1-(5-(3-nitrophenyl)furan-2-carbonyl)pyrrolidine-2-carboxamide (M-24) – Purification: detection mass 489 m/z and wavelength 304 nm. The product was obtained as a white solid after lyophilizing. (100 mg; 0.20 mmol; 13.7% over 4 steps) The purity was >99.5% (determined by C_{18} RP-HPLC-MS, t_R 1.27 min, λ = 254nm). HRMS (ES$^+$) calcd. for $C_{22}H_{28}N_6O_7$: (m/z) 488.2019; found: [M+H]= 489.2091. ^{1}H NMR (400 MHz, DMSO-d_6, conformer 1) d = 8.45-8.24 (m, 2H), 8.24-8.16 (m, 1H), 7.84-7.71 (m, 3H), 7.67-7.60 (m, 1H), 7.60-7.47 (m, 3H), 7.47-7.36 (m, 1H), 7.36-7.28 (m, 1H), 4.96-4.86 (m, 1H), 4.68-4.46 (m, 1H), 4.05-3.87 (m, 1H), 3.54-3.40 (m, 1H), 2.86-2.62 (m, 2H), 2.45-2.35 (m, 1H), 2.37-2.28 (m, 1H), 1.62-1.36 (m, 4H), 1.35-1.15 (m, 2H) ^{1}H NMR (400 MHz, DMSO-d_6, conformer 2) d = 8.44-8.24 (m, 2H), 8.15-8.09 (m, 1H), 7.84-7.70 (m, 3H), 7.67-7.60 (m, 1H), 7.59-7.47 (m, 3H), 7.47-7.37 (m, 1H), 7.13-7.06 (m, 1H), 5.08-5.01 (m, 1H), 4.14-3.99 (m, 1H), 3.79-3.68 (m, 1H), 3.39-3.23 (m, 1H), 2.86-2.65 (m, 2H), 2.39-2.27 (m, 1H), 2.27-2.14 (m, 1H), 1.61-1.37 (m, 2H), 1.36-1.16 (m, 2H); ^{13}C NMR (101 MHz, DMSO-d_6) d = 174.2; 173.0; 163.7; 158.2; 157.9; 157.6; 157.1; 156.8; 152.3; 152.2; 148.4; 148.3; 147.6; 130.8; 130.7; 130.5; 130.2; 130.0; 122.9; 122.8;

118.7; 118.6; 118.4; 118.4; 115.6; 109.8; 109.7; 75.3; 72.4; 59.6; 53.5; 53.3; 44.9; 38.4; 37.5; 33.9; 30.0; 30.0; 29.9; 26.4; 21.5

(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(2-(2-bromophenyl)-1H-benzo[d]imidazole-5-carbonyl)pyrrolidine-2-carboxamide (M-26) - Purification: detection mass 592 m/z and wavelength 214 and 290 nm.The product was obtained as a white solid after lyophilizing. (37.3 mg; 0.063 mmol; 3.9% over 4 steps). The purity was >99.5% (determined by C_{18} RP-HPLC-MS, t_R 1.00 min, λ = 214nm). HRMS (ES^+) calcd. for $C_{24}H_{28}BrN_7O_4S$: (m/z) 589.1107; found: [M+H]= 592.1155. ^{1}H NMR (600 MHz, DMSO) (Mixture of 2 Conformers) δ = 8.41 – 8.30 (s, 1H), 7.99 – 7.87 (m, 2H),7.87 – 7.80 (d, *J* = 7.1 Hz, 1H), 7.80 – 7.74 (d, *J* = 6.6 Hz, 1H), 7.71 – 7.53 (m, 2H), 7.53 – 7.42 (m, 3H), 7.40 – 7.29 (m, 1H), 7.20 – 7.10 (s, 1H), 7.10 – 7.02 (s, 1H), 5.13 – 5.00 (s, 1H), 4.99 – 4.85 (s, 1H), 4.71 – 4.54 (m, 1H), 4.40 – 4.26 (m, 1H), 4.16 – 3.96 (s, 1H), 3.00 – 2.82 (s, 2H), 2.76 – 2.63 (m, 2H), 2.62 – 2.57 (m, 2H), 2.57 – 2.53 (s, 1H), 2.33 – 2.17 (d, *J* = 13.2 Hz, 1H), 2.20 – 2.05 (d, *J* = 12.8 Hz, 1H), 1.31 – 1.14 (s, 1H); ^{13}C NMR (151 MHz, DMSO) δ = 173.69, 173.42, 170.01, 169.74, 165.13, 162.10, 161.85, 152.07, 151.89, 133.45, 132.29, 132.13, 131.62, 130.23, 127.87, 122.13, 121.54, 73.72, 72.81, 61.21, 58.65, 54.72, 52.47, 47.01, 39.52, 34.93, 34.64, 30.46, 28.99.

(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(3-(1,3-diphenyl-1H-pyrazol-4-yl)propanoyl)pyrrolidine-2-carboxamide (M-33) - Purification: detection mass 566 and 567 m/z and wavelength 214 and 254 nm. The product was lyophilized and repurified by semi-preparative HPLC. (X-bridge; ACN/H_2O (0.1% HCO_2H). The product was obtained as a white solid after lyophilizing. (51.3 mg; 0.091 mmol; 4.2% over 4 steps). The purity was >99.5% (determined by C_{18} RP-HPLC-MS, t_R 1.88 min, λ = 214nm). HRMS (ES^+) calcd. for $C_{28}H_{35}N_7O_4S$: (m/z) 565.2471; found: [M+H]= 566.2541. ^{1}H NMR (400 MHz, CD_3OD; conformer 1) d 8.25 – 8.14 (s, 1H), 7.83 – 7.75 (m, 2H), 7.75 – 7.68 (m, 2H), 7.56 – 7.43 (m, 4H), 7.44 – 7.36 (m, 1H), 7.34 – 7.24 (m, 1H), 5.22 – 5.12 (m, 1H), 5.16 – 4.99 (m, 1H), 4.66 – 4.52 (dd, J = 8.6, 3.1 Hz, 1H), , 4.34 – 4.14 (m, 1H),

3.88 – 3.60 (dq, J = 9.4, 4.4 Hz, 3H), 3.15 – 2.95 (m, 6H), 2.89 – 2.51 (m, 4H), 2.42 – 2.31 (m, 1H), 1.99 – 1.80 (s, 1H), 1.35 – 1.17 (s, 1H). ^{1}H NMR (400 MHz, CD_3OD conformer 2) 8.15 - 8.10 (s, 1H), 7.83 - 7.75 (m, 2H), 7.75 - 7.68 (m, 2H), 7.56 - 7.43 (m, 4H), 7.44 - 7.36 (m, 1H), 7.34 - 7.24 (m, 1H), 5.22 - 5.12 (m, 1H), 5.16 - 4.99 (m, 1H), 4.66 - 4.52 (dd, J = 8.6, 3.1 Hz, 1H), 4.53 - 4.41 (m, J = 5.4 Hz, 1H), 3.88 - 3.60 (dq, J = 9.4, 4.4 Hz, 3H), 3.15 - 2.95 (m, 6H), 2.89 - 2.51 (m, 4H), 2.52 - 2.44 (m, 1H), 1.99 - 1.80 (s, 1H), 1.35 - 1.17 (s, 1H). ^{13}C NMR (101 MHz, CD_3OD) δ 176.74, 176.56, 174.20, 174.08, 162.50, 153.04, 141.34, 134.73, 130.58, 130.57, 129.66, 129.19, 129.14, 129.12, 129.06, 128.91, 127.51, 121.62, 120.02, 119.99, 76.22, 74.77, 60.76, 60.21, 54.05, 53.82, 47.74, 39.96, 38.38, 36.24, 35.93, 35.67, 35.51, 31.17, 31.12, 20.81, 20.46.

(2S,4S)-4-(2-amino-3-((2-aminoethyl)thio)-N-hydroxypropanamido)-1-(5-(3-nitrophenyl)furan-2-carbonyl)pyrrolidine-2-carboxamide (M-37) Purification: detection mass 507 m/z and wavelength 214 and 254 nm. The product was obtained as a white solid after lyophilizing. (24.4 mg; 0.048 mmol; 3% over 4 steps) The purity was >99.5% (determined by C_{18} RP-HPLC-MS, t_R 1.33 min, λ = 214 nm). HRMS (ES^+) calcd. for $C_{21}H_{26}N_6O_7S$: (m/z) 506.1584; found: [M+H]= 507.1654. ^{1}H NMR (600 MHz, DMSO-d_6, conformer 1) d = 8.60 (s, 1H), 8.36 (s, 3H), 8.23 (dd, J = 15.6, 7.7 Hz, 2H), 7.80 (t, J = 7.7 Hz, 1H), 7.71 – 7.35 (m, 1H), 7.35 – 7.01 (m, 1H), 5.74 (s, 1H), 5.09 (d, J = 9.3 Hz, 1H), 5.04 (s, 1H), 4.07 (d, J = 5.6 Hz, 1H), 3.84 (s, 2H), 2.93 – 2.84 (m, 2H), 2.74 – 2.64 (m, 2H), 2.63 – 2.52 (m, 2H), 2.43 (d, J = 13.5 Hz, 1H), 2.19 (d, J = 13.3 Hz, 1H); ^{1}H NMR (600 MHz, DMSO-d_6, conformer 2) d = 8.57 (s, 1H), 8.36 (s, 3H), 8.19 (d, J = 7.9 Hz, 2H), 7.74 (t, J = 7.8 Hz, 1H), 7.71 – 7.35 (m, 1H), 7.35 – 7.01 (m, 1H), 5.74 (s, 1H), 5.16 (s, 1H), 4.57 (d, J = 9.0 Hz, 1H), 4.26 (d, J = 7.2 Hz, 1H), 4.19 (d, J = 11.4 Hz, 1H), 4.07 (d, J = 5.6 Hz, 1H), 2.93 – 2.84 (m, 2H), 2.74 – 2.64 (m, 2H), 2.63 – 2.52 (m, 2H), 2.43 (d, J = 13.5 Hz, 1H), 2.19 (d, J = 13.3 Hz, 1H); ^{13}C NMR (151 MHz, DMSO) δ = 174.07, 172.96, 165.20, 161.92, 157.28, 157.08, 152.32, 148.56, 147.92, 130.87, 130.27, 122.89, 118.41, 109.81, 74.48, 71.57, 59.91, 53.57, 47.0, 37.67, 34.81, 34.11, 30.79, 29.00.

PUBLICATION 4:

Systematic study on Plasmodium falciparum aminoacyl-tRNA synthetases as antimalarial drug targets.

Camacho C, **Novoa EM**, Cubells L, Wilkinson B, Martin P, Bautista JM, Cortés A and Ribas de Pouplana L.

To be submitted

Systematic study on *Plasmodium falciparum* aminoacyl-tRNA synthetases as antimalarial drug targets

Noelia Camacho[1], Eva Maria Novoa[1], Laia Cubells[1], Barrie Wilkinson[2], Patricia Marín[3], Jose Maria Bautista[3], Alfred Cortés[1] and Lluis Ribas de Pouplana[1,4]*

[1] *Institute for Research in Biomedicine (IRB), c/ Baldiri Reixac 15-21 08028, Barcelona, Spain.*

[2] *Biotica Technology Ltd. 3 Riverside, Suite 5, Granta Park, Great Abington, Cambridge, CB21 6AD, United Kingdom.*

[3] *Department of Biochemistry and Molecular Biology IV, Complutense University of Madrid, Madrid, Spain.*

[4] *Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.*

* Corresponding author: lluis.ribas@irbbarcelona.org

ABSTRACT

Malaria remains a major global health problem, and emerging resistance to existing drugs results in an increasing urgency for new antimalarials. Protein translation is the target of several antimalarial drugs currently in use. In order to explore the potential of the aminoacyl-tRNA synthetase (ARS) family as source of antimalarial drug targets, we have treated *Plasmodium falciparum* cultures with a battery of both known and novel ARS inhibitors, and compared their activities. Amongst the compounds tested, borrelidin, a natural inhibitor of threonyl-tRNA synthetase (ThrRS), stands out for its potent antimalarial effect. Despite its promising antimalarial activity, borrelidin also inhibits human ThrRS, and is highly toxic to human cells. To circumvent this problem we have explored the antimalarial activities of a library of borrelidin derivatives, and evaluated their cytotoxicity in human cells. We find that some of these compounds present higher selectivity towards the *P. falciparum* enzyme, whilst maintaining their antiparasitic activity both *in vitro* and *in vivo*. We propose that borrelidin is a promising antimalarial scaffold that should be further explored for the search of novel antimalarial drugs.

INTRODUCTION

With approximately 250 million clinical cases, and over 1 million attributed deaths per year (WHO report 2010), malaria is one of the most severe infectious diseases that lacks an effective vaccine. Numerous antibacterials are known to kill malaria parasites (Jomaa et al., 1999; Ralph et al.,2001; Seeber, 2003; Surolia and Surolia, 2001), and are commonly used in malaria prophylaxis or as components of multiple drug therapies (Borrmann et al., 2004; Miller, 1974). However, the emergence of multi-resistant parasites compromise the efficacy of many existing chemotherapies (Andriantsoanirina et al., 2009; Bonnet et al., 2009; Carrara et al., 2009), leading to a growing urgency for the search of new antimalarials.

Protein translation has been a major focus for drug development. *Plasmodium* possesses three compartments active in protein synthesis: the cytosol, the mitochondrion and a relict plastid termed apicoplast. The prokaryotic origin of the apicoplast and mitochondria make many of their enzymes attractive targets for drug design. Doxycycline, a tetracycline known to specifically inhibit the replication of the apicoplast, is widely used for the prevention and treatment of malaria (Dahl et al., 2006). The inhibition of the apicoplast's metabolism causes a phenotype known as "delayed death" (Dahl and Rosenthal, 2007), in which the parasites do not show growth inhibition during the first asexual cycle during drug treatment (48h), but die in the second asexual cycle (98h), even if the drug is removed after the first cycle (Goodman et al., 2007).

Among the less exploited targets of the translation machinery is the family of aminoacyl-tRNA synthetases (ARS), which are essential enzymes (reviewed in Ochsner et al., 2007) that catalyze the correct attachment of amino acids to their cognate tRNAs. These enzymes are proven antibacterial drug targets of both experimental (Hurdle et al., 2005; Kim et al., 2003**;** Bennett et al.,1999; Schimmel et al.,1998) and commercially available drugs (Bactroban, GlaxoSmithKline). ARS have also been proposed as potential antimalarial drug targets (Jackson et al., 2011; Istvan et al 2010). Sequencing of the *P. falciparum* genome has allowed the identification of the complete set of plasmodial ARS genes, including several duplicated genes that likely correspond to enzymes that are active in the organelles (Bhatt et al, BMC Genomics). Although the three compartments are translationally active, only the cytosol and apicoplast are expected to contain ARS, while mitochondria are thought to import charged tRNAs (Pino et al., 2010).

The large evolutionary distance between *Plasmodium* ARS of apicoplastic origin and their human homologues supports their potential as antimalarial drug targets. Particularly interesting are those ARS that are unique to *Plasmodium*, and not present in human cells. For instance, lysyl-tRNA synthetase (LysRS) is present in two forms in *P. falciparum*, whereas mammalian cells contain

only one protein that is alternatively spliced, and is dually targeted both to the cytosol and mitochondria (Hoen et al., 2012).

Here we have analyzed the effect of known inhibitors of ARS and their potential use as antimalarials. We have tested and compared a battery of known inhibitors that target 10 different ARS with predicted cytosolic, apicoplastic or dual localization. Amongst the compounds tested, we find that borrelidin shows excellent antimalarial activity both *in vitro* and *in vivo*. Borrelidin has been already described to possess antiangiogenic (Moss et al. 2006; Kawamura et al. 2003), antimalarial (Otoguro et al. 2003; Ishiyama et al. 2011) and antimicrobial (Nass et al 1969) properties. However, its high cytotoxicity in humans limit its use (Wilkinson et al., 2006). Thus, we tested whether other borrelidin derivatives could be more target specific. We have investigated the antimalarial activity of a series of 30 borrelidin derivatives, and have tested their cytotoxicity in human cells. Our results show that several derivatives present higher selectivity than borrelidin, indicating that modifications of this compound can lead to increased selectivity without losing its antimalarial activity. The most potent compounds have been tested in *P. yoelii*-infected mice, where they clear the infection efficiently. These promising results suggest that borrelidin should be further explored as an antimalarial scaffold, and confirm that aminoacyl-tRNA synthetases are promising antimalarial drug targets that can be selectively inhibited.

RESULTS

1. Screening of Known ARS Inhibitors on *P. falciparum*

In order to identify the antimalarial activity of a battery of ARS inhibitors, a phenotypic screen was run against *in vitro* cultures of *P. falciparum* (Table 1). Our collection of inhibitors included: i) analogues of the natural ligands or reaction intermediates (mechanism-based inhibitors), ii) natural inhibitors and their derivatives, and iii) novel scaffolds targeting ARS (Figure 1). To address the issue of variable drug effects at different life stages, inhibitory compounds were applied to tightly synchronized cells, and analyzed by comparing the multiplication of parasites between treated and control parasites (Giemmsa-smears counting), both at 48 (first asexual cycle), and 96 hours (second asexual cycle). A drug was considered to cause delayed-death if it met the two following conditions: i) there was no measurable growth inhibition after 48h of treatment (first asexual cycle) at drug concentrations 10-fold higher than those needed to inhibit 50% of parasite growth at 96h (second asexual cycle), and ii) parasite growth inhibition was insensitive to the presence or absence of drug during the second asexual cycle (Goodman et al. 2007).

Mechanism-based inhibitors presented inhibition of plasmodial ARS in the nanomolar range (Table 1, see also Figure S1), with the exception of thialysine. Our results suggest that these compounds inhibit cytosolic ARS. Small decreases in IC_{50}(96h) values cannot be considered as delayed-death phenotypes, and may only suggest that the apicoplastic enzyme is more sensitive than its respective cytosolic homologue.

Natural ARS inhibitors were also tested on *P. falciparum* cultures (Table 1). Pseudomonic acid (PA) is produced by *Pseudomonas fluorescens* and is a potent and specific inhibitor of bacterial isoleucyl-tRNA synthetases, with an 8000-fold selectivity with respect to to mammalian enzymes (Hugues et al., 1980; Farmer et al., 1992). In our assays PA was relatively inactive at 48h (IC_{50}(48h)=257μM), but was active in the nanomolar range in the second asexual cycle (IC_{50}(96h)=93nM). This delayed-death phenotype is consistent with its high selectivity towards bacterial-type enzymes (Ward et al., 1986; Parenti et al., 1987), such as the apicoplast-targeted isoleucyl-tRNA synthetase (IleRS-2). This increased potency after a second cycle was seen even if PA was removed from the culture after the first cycle of incubation.

Borrelidin, a potent macrolide antibiotic produced by *Streptomyces rochei,* inhibits mammalian, bacterial and protozoan threonyl-tRNA synthetases (ThrRS) (Nass et al., 1969; Gerken and Arfin, 1984; Otoguro et al., 2003; Ishiyama et al. ,2011). As previously described (Otoguro et al., 2003), we observed a strong inhibition of parasite growth by borrelidin within the first 48h (Table 1), thus

showing an immediate death phenotype. Cispentacin, a proline analog that inhibits prolyl-tRNA synthetase, is isolated from *Bacillus cereus* and *Streptomyces setonii,* and has been shown to effectively protect against systemic *Candida albicans* and *Cryptococcus neoformans* infections (Oki et al., 1989). In our tests, however, cispentacin is a weak inhibitor of *Plasmodium* cultures. This could be due to the fact that in fungi, cispentacin accumulates at high intracellular levels through an active transport mechanism (Capobianco et al., 1993) that might be missing in *Plasmodium*.

AN2729 (Anacor Pharmaceuticals Inc) is a member of a new class of broad-spectrum antifungals that inhibit LeuRS (Rock et al., 2007; Barak and Loo, 2007). We investigated the effect of AN2729 on the parasite growth, and observed that this compound is capable of inhibiting *P. falciparum* cultures at low micromolar concentrations at 48h, indicating that boron-based drugs might also be useful for antimalarial drug design.

2. **Dual Targeting Approaches**

It is generally accepted that antimalarial treatments based on drug combinations help to prevent the appearance of drug resistance. Usually these combinations are based on compounds that target different enzymes and metabolic routes. However, we hypothesized that the combination of drugs that target the same metabolic process, e.g. tRNA aminoacylation, could produce a synergistic effect. Thus, we decided to test a combination of pseudomonic acid and borrelidin, given that amongst the battery of tested ARS inhibitors, these two compounds were shown to inhibit most efficiently plasmodial cultures *in vitro*.

To test whether the simultaneous use of both drugs modified their respective kinetics, we analyzed the effect of the combination of both compounds *in vitro*, at different concentrations around their respective IC_{50} concentrations as previously described (Cokol et al., 2011; Pereira et al., 2011), and analyzed the parasitemia after a second cycle (96h). As can be seen in Figure 2, neither synergy nor antagonism was observed, given that the effect of the combination of the two compounds corresponds exactly to the sum of their individual effects, indicating that the two metabolic pathways targeted are not affected by the inhibition of the other. This may be due to the fact that pseudomonic acid is primarily affecting the apicoplastic translational machinery -specifically IleRS-2- whereas borrelidin is inhibiting the cytosolic ThrRS. The fact that protein synthesis in cytosol and apicoplast occur at different times of the life cycle (Bozdech et al., 2003) could explain why a synergistic effect cannot be observed when using PA:borrelidin drug combinations.

We then investigated the antimalarial activity of both PA and borrelidin using *in vivo P. yoelii*-infected mice (Figure2C). PA did not reduce the parasitemia of infected mice. This result was expected as it has been described that, following intravenous or oral administration, this compound is rapidly metabolized to its inactive metabolite monic acid by the effect of esterases (Boyce JM., 2001; Hurdle et al., 2005). On the other hand, treatment with borrelidin reduced 95% of the parasitemia of infected mice, which is similar to the behavior observed with the positive control chloroquine. Complementarily, we performed survival time experiments with infected mice that consisted in drug treatment on the 3rd day post-infection followed by removal of the treatment, and a final measure of the mice survival. *P. yoelii*-infected mice treated with chloroquine survived in average of 6 days post-infection as compared to untreated infected mice which survived until day 4. *P. yoelii*-infected mice treated with borrelidin also survived in average 6 days, thus performing similar to chloroquine (Table 2). These results are in agreement with previous reports (Otoguro et al., 2003; Ishiyama et al., 2011)

3. Antimalarial Activity of Borrelidin Derivatives

Borrelidin as an antimalarial scaffold

With an IC_{50} of 0.97nM, borrelidin is a more potent antimalarial drug than artemether, artesunate and chloroquine (Otoguro et al., 2003; Ishiyama et al., 2011). Its activity is thought to arise from the inhibition of threonyl tRNA synthetase (ThrRS) (Vong et al., 2004). Previous reports have shown that borrelidin is a noncompetitive inhibitor with respect to threonine, and inhibits the amino acid activation step, as shown by ATP-PPi exchange and transient kinetic assay (Ruan et al., 2005). Genetic selection of *E. coli* borrelidin-resistant mutants has shown that borrelidin binds to a hydrophobic region proximal to the zinc at the active site of the *E. coli* ThrRS (Ruan et al., 2005). Indeed, the fact that other ARS do not have such a hydrophobic core in this part of their active site may explain why borrelidin only inhibits ThrRS but not any other ARS. To further confirm that the inhibition of *P. falciparum* cultures is due to the inhibition of ThrRS, we confirmed that the cluster of amino acids binding borrelidin in *E. coli* ThrRS is conserved in *P. falciparum* ThrRS (Figure S2).

In vitro antimalarial activity of borrelidin derivatives.

The results presented in this study point to borrelidin as a potential antimalarial agent (IC_{50}= 0,97nM). However, borrelidin also causes human cell toxicity in the nanomolar range (IC_{50}=345nM, tested on HEK 293T cells), presumably as a consequence of the inhibition of human ThrRS. Therefore, we decided to test other borrelidin derivatives to search for more potent and selective antimalarial compounds.

For this, a library of 30 borrelidin derivatives was tested against *P. falciparum* cultures at 100nM. We find that 13 of these derivatives are active *in vitro* (Figure 4A). We calculated the IC_{50} values for these 13 compounds both in *P. falciparum* and human cell cultures, finding that all borrelidin derivatives show lower inhibitory activities -higher IC_{50}- against *P. falciparum* cultures than borrelidin (Table 3). However, most of these molecules also appeared to be more selective than borrelidin towards the plasmodial enzyme (Table 3). Indeed, the selectivity of borrelidin for *P. falciparum* ThrRS versus its human homologue is 355-fold, whereas some of the derivatives tested present up to 16.000-fold selectivity towards the *P. falciparum* enzyme (e.g. BC-195), which implies a 50-fold increase in selectivity compared to borrelidin (Figure 4B). These encouraging results indicate that borrelidin derivatives are potent antimalarials in the low nanomolar range that show increased selectivity towards the plasmodial enzyme, showing low cross-reactivity towards its human homologue.

In vivo treatment of P.yoelii-infected mice with borrelidin derivatives

Amongst the 13 borrelidin derivatives that presented antimalarial activity at 100nM, we selected 5 compounds that presented highest selectivity (B-194, B-195, B-196, B-220 and B-240) for *in vivo* studies with *P. yoelii*-infected mice. Each of the 5 borrelidin analogues was tested at two different concentrations: 0,25 mg/kg/day -which is the effective dose of borrelidin-, and 6mg/kg/day –which is the effective dose of chloroquine- (Table 4). For each compound and each dose, the average parasitemia was measured after treatment with each of the compounds (Figure 4C), and the mice survival was monitored during over 20 days after drug treatment (Figure S3).

Our results show that two out the five tested compounds (BC-196 and BC-220) yield 100% survival at 6mg/kg/day (Table 4). The most promising compound is BC-220, which at 6mg/kg/day completely clears the parasitemia -even better than borrelidin or chloroquine- using the 4-day-test. Indeed, this compound is already showing 80% mice survival and a very good suppression at 0,25mg/kg/day. Interestingly, this compound was not the most potent analogue *in vitro*, suggesting that the compound bioavailability or other ADME properties are important in determining which hits are best *in vivo*.

DISCUSSION

Aminoacyl-tRNA synthetases have been proposed for many years to be druggable targets that can be used for drug discovery (Kim et al. 2003; Schimmel et al. 1998). Although for many years plasmodial aminoacyl-tRNA synthetases have remained unexplored as drug targets, recent works have shown not only that these enzymes are druggable but also that selective inhibition of these enzymes versus its human homologues is feasible (Istvan et al. 2011; Hoepfner et al. 2012; Hoen et al. 2012)

In this work we first evaluated and tested a series of known ARS inhibitors on *P. falciparum* cell cultures, to explore which compounds showed stronger antimalarial activities. Amongst the tested compounds, we find that borrelidin is the strongest antimalarial inhibitor (IC_{50}=2nM), which efficiently clears the plasmodial infection. However, borrelidin is not selective enough for clinical applications (Otoguro et al., 2003).

In spite of these results, we tested library of borrelidin analogues to find active compounds that were also selective. From our library of analogues, we find five compounds that present at least a 10-fold increase in selectivity compared to borrelidin, which were selected for further *in vivo* assays. Importantly, one of these compounds, BC-220, shows strong *in vivo* activity with parasite clearance comparable to chloroquine and 100% mice survival.

These encouraging results suggest that borrelidin is a good scaffold for antimalarial drug design, validate threonyl-tRNA synthetase as a druggable antimalarial drug target, and present a series of compounds that should be further characterized for future clinical testing.

MATERIALS AND METHODS

Reagents

Natural aminoacyl-tRNA synthetase inhibitors were purchased from the following companies: pseudomonic acid (GlaxoSmithKline), borrelidin (Fluorochem), cispentacin (Acros organics), thialysine (Sigma). AN2729 was a gift from ANACOR. The battery of borrelidin derivatives was obtained from Biotica Technology Ltd (Cambridge, UK). Sulfamoyl adenosine analogues were a kind gift from Magali Frugier (CNRS, France).

IC_{50} determinations

IC_{50} determinations were performed with synchronous 3D7A parasite cultures. Parasites were cultured in human erythrocytes in RPMI 1640 medium supplemented with glutamine. FACS was used to measure calculate the IC_{50} of the most active compounds, by using Syto-11 to discriminate parasitized from non-parasitized RBCs. Each sample was diluted at 1:100 in PBS and 0.5mM Syto-11 in DMSO was added to a final concentration of 0.5μM. Samples were excited at 488nm and analyzed using an FC500 flow cytometer. Data analysis was performed with the software package Prism.

Cell-based drug inhibition assays

Initial screens to test the activity were performed using the lactate dehydrogenase (LDH) activity assay. To perform the assay, 20 μl of sorbitol-synchronized infected erythrocytes (3% hematocrit) in each well of a 96-well plate was mixed with 100μl of Malstat reagent, 10μl of 2mg/ml of nitroblue tetrazolium and 10μl of 0,2μg/ml phenylethyl sulfate. After 30min incubation in the dark, the reaction was stopped by adding 100μl of 5% acetic acid to each well. Absorbance at 590nm (A_{590}) was measured on a plate reader to quantify the LDH activity proportional to the parasitemia. Smears were also prepared for each drug assay to visually confirm the absorbance results. For each tested compound, parasite LDH activity was measured both at 48 and 96h in order to check for a delayed death phenotype.

Cytotoxicity assays on human cells

Cytoxicity was measured using the Cell Profileration assay WST-1 (Roche) on HEK293T cells. Cells were cultured in a 96-well microplate, and incubated during 2-4h with WST-1. During this incubation period, viable cells convert WST-1 to a soluble formazan salt, which is quantified at 450nm with an ELISA plate reader.

***In vivo P. yoelii* infected mice treatments**

The *in vivo* study was carried out according to standard protocol following the "4 day Test" (Moneriz et al., Malar J 2011). *P. yoelii*-infected mice were divided into non-treated control, chloroquine-treated, borrelidin-treated, and borrelidin analogue-treated groups. Each group consisted of 5 to 10 mice. On day 0, each mouse was injected with 200μl of $2x10^6$ infected red blood cells intravenously. Two hours after inoculation, each mouse was orally treated with 200 μl of borrelidin, borrelidin analogue, pseudomonic acid or chloroquine. Control mice were given 200μl of distilled water orally. The treatment was repeated for the next 3 days for all the groups of animals. Every day, thin blood smears were obtained and stained with Giemsa to compute the parasitemia. The number of dead mice was recorded daily from all the study groups to determine the average of survival time of the infected mice after treatment.

ACKNOWLEDGMENTS

This work has been supported by grant BIO2009-09776 from the Spanish Ministry of Education and Science, and by grant MEPHITIS-223024 from the European Union. E.M.N is supported by a La Caixa/IRB International Ph.D. Programme Fellowship.

REFERENCES

Andriantsoanirina, V., Ratsimbasoa, A., Bouchier, C., Jahevitra, M., Rabearimanana, S., Radrianjafy, R., Andrianaranjaka, V., Randriantsoa, T., Rason, M.A., Tichit, M., et al. (2009). Plasmodium falciparum drug resistance in Madagascar: facing the spread of unusual pfdhfr and pfmdr-1 haplotypes and the decrease of dihydroartemisinin susceptibility. Antimicrob Agents Chemother 53, 4588-4597.

Barak, O., and Loo, D.S. (2007). AN-2690, a novel antifungal for the topical treatment of onychomycosis. Curr Opin Investig Drugs 8, 662-668.

Bennett, I., Broom, N.J., Cassels, R., Elder, J.S., Masson, N.D., and O'Hanlon, P.J. (1999). Synthesis and antibacterial properties of beta-diketone acrylate bioisosteres of pseudomonic acid A. Bioorg Med Chem Lett 9, 1847-1852.

Bhatt, T.K., Kapil, C., Khan, S., Jairajpuri, M.A., Sharma, V., Santoni, D., Silvestrini, F., Pizzi, E., and Sharma, A. (2009). A genomic glimpse of aminoacyl-tRNA synthetases in malaria parasite Plasmodium falciparum. BMC Genomics 10, 644.

Bonnet, M., Broek, I., van Herp, M., Urrutia, P.P., van Overmeir, C., Kyomuhendo, J., Ndosimao, C.N., Ashley, E., and Guthmann, J.P. (2009). Varying efficacy of artesunate+amodiaquine and artesunate+sulphadoxine-pyrimethamine for the treatment of uncomplicated falciparum malaria in the Democratic Republic of Congo: a report of two in-vivo studies. Malar J 8, 192.

Borrmann, S., Issifou, S., Esser, G., Adegnika, A.A., Ramharter, M., Matsiegui, P.B., Oyakhirome, S., Mawili-Mboumba, D.P., Missinou, M.A., Kun, J.F., et al. (2004). Fosmidomycin-clindamycin for the treatment of Plasmodium falciparum malaria. J Infect Dis 190, 1534-1540.

Boyce, J.M. (2001). MRSA patients: proven methods to treat colonization and infection. J Hosp Infect 48 Suppl A, S9-14.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. (2003). The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1, E5.

Capobianco, J.O., Zakula, D., Coen, M.L., and Goldman, R.C. (1993). Anti-Candida activity of cispentacin: the active transport by amino acid permeases and possible mechanisms of action. Biochem Biophys Res Commun 190, 1037-1044.

Carrara, V.I., Zwang, J., Ashley, E.A., Price, R.N., Stepniewska, K., Barends, M., Brockman, A., Anderson, T., McGready, R., Phaiphun, L., et al. (2009). Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. PLoS One 4, e4551.

Cokol, M., Chua, H.N., Tasan, M., Mutlu, B., Weinstein, Z.B., Suzuki, Y., Nergiz, M.E., Costanzo, M., Baryshnikova, A., Giaever, G., et al. (2011). Systematic exploration of synergistic drug pairs. Mol Syst Biol 7, 544.

Dahl, E.L., and Rosenthal, P.J. (2007). Multiple antibiotics exert delayed effects against the Plasmodium falciparum apicoplast. Antimicrob Agents Chemother 51, 3485-3490.

Dahl, E.L., Shock, J.L., Shenai, B.R., Gut, J., DeRisi, J.L., and Rosenthal, P.J. (2006). Tetracyclines specifically target the apicoplast of the malaria parasite Plasmodium falciparum. Antimicrob Agents Chemother 50, 3124-3131.

Farmer, T.H., Gilbart, J., and Elson, S.W. (1992). Biochemical basis of mupirocin resistance in strains of Staphylococcus aureus. J Antimicrob Chemother 30, 587-596.

Gerken, S.C., and Arfin, S.M. (1984). Chinese hamster ovary cells resistant to borrelidin overproduce threonyl-tRNA synthetase. J Biol Chem 259, 9202-9206.

Goodman, C.D., Su, V., and McFadden, G.I. (2007). The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. Mol Biochem Parasitol 152, 181-191.

Hoen, R., Novoa, E. M., Cubells, L., López, A., Marín-Garcia, P., Bautista, J. M., Vieira, P., Santos, M., Cortés, A., Ribas de Pouplana, L., and Royo, M. (2012). Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum. J Med Chem, under review.

Hoepfner, D.; McNamara, C. W.; Lim, C. S.; Studer, C.; Riedl, R.; Aust, T.; McCormack, S. L.; Plouffe, D. M.; Meister, S.; Schuierer, S.; Plikat, U.; Hartmann, N.; Staedtler, F.; Cotesta, S.; Schmitt, E. K.; Petersen, F.; Supek, F.; Glynne, R. J.; Tallarico, J. A.; Porter, J. A.; Fishman, M.

C.; Bodenreider, C.; Diagana, T. T.; Movva, N. R.; Winzeler, E. A., Selective and specific inhibition of the plasmodium falciparum lysyl-tRNA synthetase by the fungal secondary metabolite cladosporin. Cell Host Microbe 2012, 11 (6), 654-63.

Hughes, J., and Mellows, G. (1980). Interaction of pseudomonic acid A with Escherichia coli B isoleucyl-tRNA synthetase. Biochem J 191, 209-219.

Hurdle, J.G., O'Neill, A.J., and Chopra, I. (2005). Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents. Antimicrob Agents Chemother 49, 4821-4833.

Ishiyama, A., Iwatsuki, M., Namatame, M., Nishihara-Tsukashima, A., Sunazuka, T., Takahashi, Y., Omura, S., and Otoguro, K. (2011). Borrelidin, a potent antimalarial: stage-specific inhibition profile of synchronized cultures of Plasmodium falciparum. J Antibiot (Tokyo) 64, 381-384.

Istvan, E.S., Dharia, N.V., Bopp, S.E., Gluzman, I., Winzeler, E.A., and Goldberg, D.E. (2011). Validation of isoleucine utilization targets in Plasmodium falciparum. Proc Natl Acad Sci U S A 108, 1627-1632.

Jackson, K.E., Pham, J.S., Kwek, M., De Silva, N.S., Allen, S.M., Goodman, C.D., McFadden, G.I., de Pouplana, L.R., and Ralph, S.A. (2012). Dual targeting of aminoacyl-tRNA synthetases to the apicoplast and cytosol in Plasmodium falciparum. Int J Parasitol 42, 177-186.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., et al. (1999). Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285, 1573-1576.

Kawamura, T., Liu, D., Towle, M.J., Kageyama, R., Tsukahara, N., Wakabayashi, T., and Littlefield, B.A. (2003). Anti-angiogenesis effects of borrelidin are mediated through distinct pathways: threonyl-tRNA synthetase and caspases are independently involved in suppression of proliferation and induction of apoptosis in endothelial cells. J Antibiot (Tokyo) 56, 709-715.

Kim, S., Lee, S.W., Choi, E.C., and Choi, S.Y. (2003). Aminoacyl-tRNA synthetases and their inhibitors as a novel family of antibiotics. Appl Microbiol Biotechnol 61, 278-288.

Miller, L.H., Glew, R.H., Wyler, D.J., Howard, W.A., Collins, W.E., Contacos, P.G., and Neva, F.A. (1974). Evaluation of clindamycin in combination with quinine against multidrug-resistant strains of Plasmodium falciparum. Am J Trop Med Hyg 23, 565-569.

Moneriz, C., Marin-Garcia, P., Garcia-Granados, A., Bautista, J.M., Diez, A., and Puyet, A. (2011). Parasitostatic effect of maslinic acid. I. Growth arrest of Plasmodium falciparum intraerythrocytic stages. Malar J 10, 82.

Moss, S.J., Carletti, I., Olano, C., Sheridan, R.M., Ward, M., Math, V., Nur, E.A.M., Brana, A.F., Zhang, M.Q., Leadlay, P.F., et al. (2006). Biosynthesis of the angiogenesis inhibitor borrelidin: directed biosynthesis of novel analogues. Chem Commun (Camb), 2341-2343.

Nass, G., Poralla, K., and Zahner, H. (1969). Effect of the antibiotic Borrelidin on the regulation of threonine biosynthetic enzymes in E. coli. Biochem Biophys Res Commun 34, 84-91.

Nass, G., Poralla, K., and Zahner, H. (1969). Effect of the antibiotic Borrelidin on the regulation of threonine biosynthetic enzymes in E. coli. Biochem Biophys Res Commun 34, 84-91.

Ochsner, U.A., Sun, X., Jarvis, T., Critchley, I., and Janjic, N. (2007). Aminoacyl-tRNA synthetases: essential and still promising targets for new anti-infective agents. Expert Opin Investig Drugs 16, 573-593.

Oki, T., Hirano, M., Tomatsu, K., Numata, K., and Kamei, H. (1989). Cispentacin, a new antifungal antibiotic. II. In vitro and in vivo antifungal activities. J Antibiot (Tokyo) 42, 1756-1762.

Olano, C., Moss, S.J., Brana, A.F., Sheridan, R.M., Math, V., Weston, A.J., Mendez, C., Leadlay, P.F., Wilkinson, B., and Salas, J.A. (2004). Biosynthesis of the angiogenesis inhibitor borrelidin by Streptomyces parvulus Tu4055: insights into nitrile formation. Mol Microbiol 52, 1745-1756.

Otoguro, K., Ui, H., Ishiyama, A., Kobayashi, M., Togashi, H., Takahashi, Y., Masuma, R., Tanaka, H., Tomoda, H., Yamada, H., et al. (2003). In vitro and in vivo antimalarial activities of a non-glycosidic 18-membered macrolide antibiotic, borrelidin, against drug-resistant strains of Plasmodia. J Antibiot (Tokyo) 56, 727-729.

Parenti, M.A., Hatfield, S.M., and Leyden, J.J. (1987). Mupirocin: a topical antibiotic with a unique structure and mechanism of action. Clin Pharm 6, 761-770.

Pereira, M.R., Henrich, P.P., Sidhu, A.B., Johnson, D., Hardink, J., Van Deusen, J., Lin, J., Gore, K., O'Brien, C., Wele, M., et al. (2011). In vivo and in vitro antimalarial properties of azithromycin-

chloroquine combinations that include the resistance reversal agent amlodipine. Antimicrob Agents Chemother 55, 3115-3124.

Pino, P., Aeby, E., Foth, B.J., Sheiner, L., Soldati, T., Schneider, A., and Soldati-Favre, D. (2010). Mitochondrial translation in absence of local tRNA aminoacylation and methionyl tRNA Met formylation in Apicomplexa. Mol Microbiol 76, 706-718.

Ralph, S.A., D'Ombrain, M.C., and McFadden, G.I. (2001). The apicoplast as an antimalarial drug target. Drug Resist Updat 4, 145-151.

Rock, F.L., Mao, W., Yaremchuk, A., Tukalo, M., Crepin, T., Zhou, H., Zhang, Y.K., Hernandez, V., Akama, T., Baker, S.J., et al. (2007). An antifungal agent inhibits an aminoacyl-tRNA synthetase by trapping tRNA in the editing site. Science 316, 1759-1761.

Ruan, B., Bovee, M.L., Sacher, M., Stathopoulos, C., Poralla, K., Francklyn, C.S., and Soll, D. (2005). A unique hydrophobic cluster near the active site contributes to differences in borrelidin inhibition among threonyl-tRNA synthetases. J Biol Chem 280, 571-577.

Schimmel, P., Tao, J., and Hill, J. (1998). Aminoacyl tRNA synthetases as targets for new anti-infectives. FASEB J 12, 1599-1609.

Seeber, F. (2003). Biosynthetic pathways of plastid-derived organelles as potential drug targets against parasitic apicomplexa. Curr Drug Targets Immune Endocr Metabol Disord 3, 99-109.

Surolia, N., and Surolia, A. (2001). Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum. Nat Med 7, 167-173.

Vong, B.G., Kim, S.H., Abraham, S., and Theodorakis, E.A. (2004). Stereoselective total synthesis of (-)-borrelidin. Angew Chem Int Ed Engl 43, 3947-3951.

Ward, A., and Campoli-Richards, D.M. (1986). Mupirocin. A review of its antibacterial activity, pharmacokinetic properties and therapeutic use. Drugs 32, 425-444.

Wilkinson, B., Gregory, M.A., Moss, S.J., Carletti, I., Sheridan, R.M., Kaja, A., Ward, M., Olano, C., Mendez, C., Salas, J.A., et al. (2006). Separation of anti-angiogenic and cytotoxic activities of borrelidin by modification at the C17 side chain. Bioorg Med Chem Lett 16, 5814-5817.

Figure 1. Structures of known ARS inhibitors tested on *P. falciparum* cultures

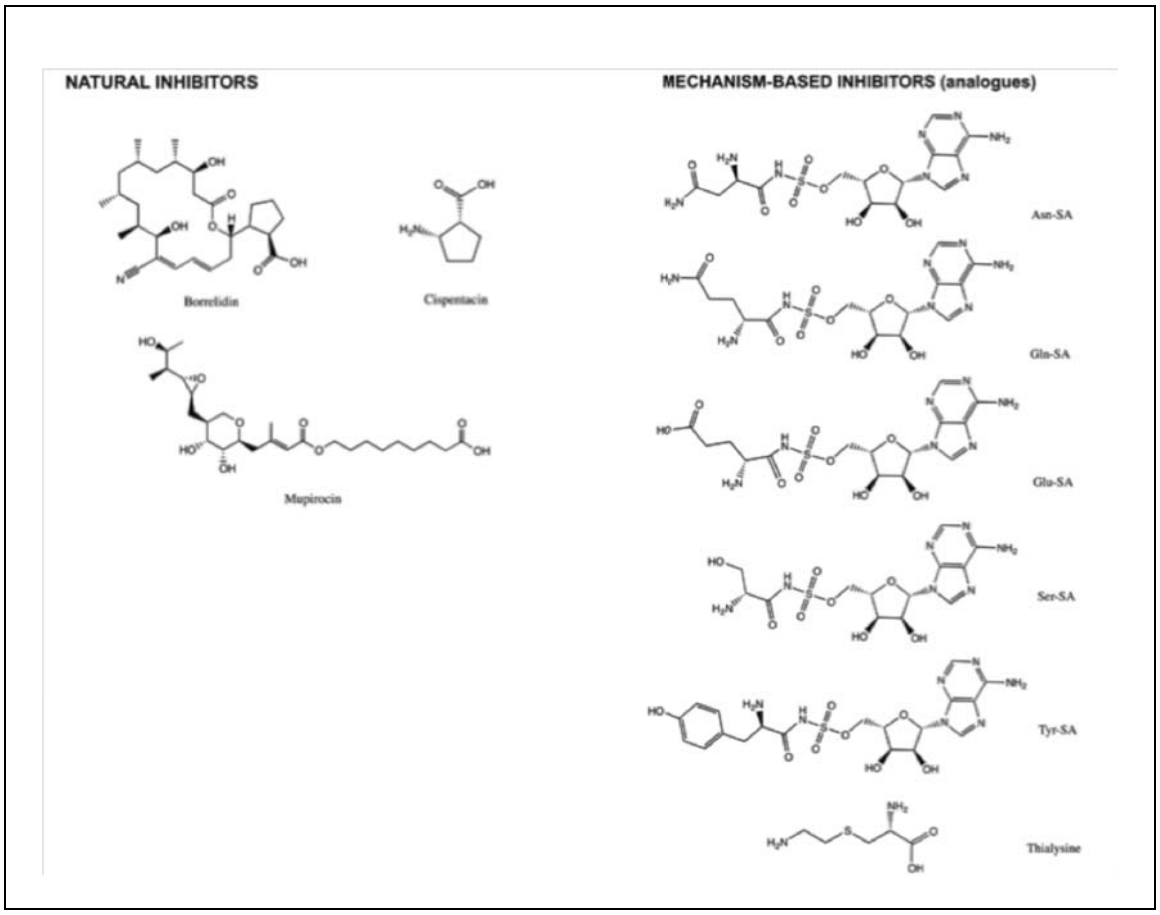
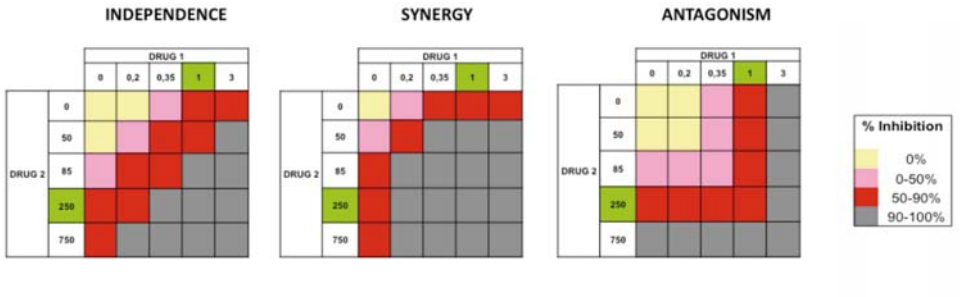


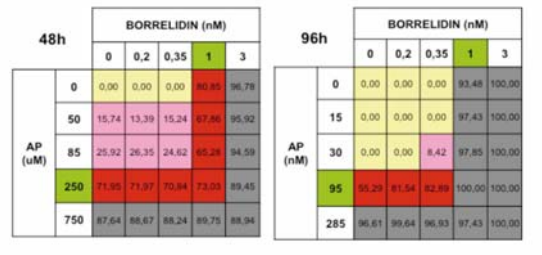
Figure 2. Dual targeting approaches.

A) Experimental set-up for classification of drug interactions. Inhibition rates were measured for all pairwise combinations of five drug concentrations, linearly increasing from 0 to 3 times the IC_{50} inhibitory concentration (green). Isophenotypic curves parallel to the diagonal for independent drug pairs, concave for synergistic drug pairs, and convex for antagonistic drug pairs, according to Loewe additivity. B) *In vitro* testing for multiple targeting effect produced by the combination of pseudomonic acid and borrelidin on parasite cultures. C) *In vivo* inhibitory activities in *P.yoelii*-infected mice of known ARS inhibitors that presented most potent *in vitro* inhibition: borrelidin and pseudomonic acid. Chloroquine has been used as positive control. Survival times with the removal of the drug on the 3rd day post infection with the different drug chemotherapies tested.

A



B



C

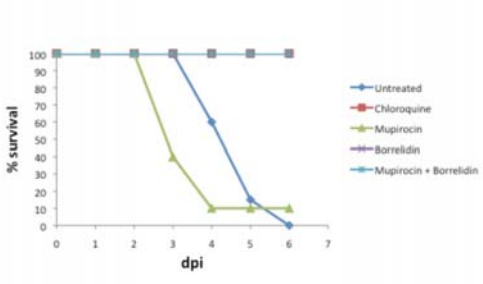


Figure 3. Library of borrelidin analogues. Library of borrelidin derivatives tested on cell-based assays of *P. falciparum* cultures.

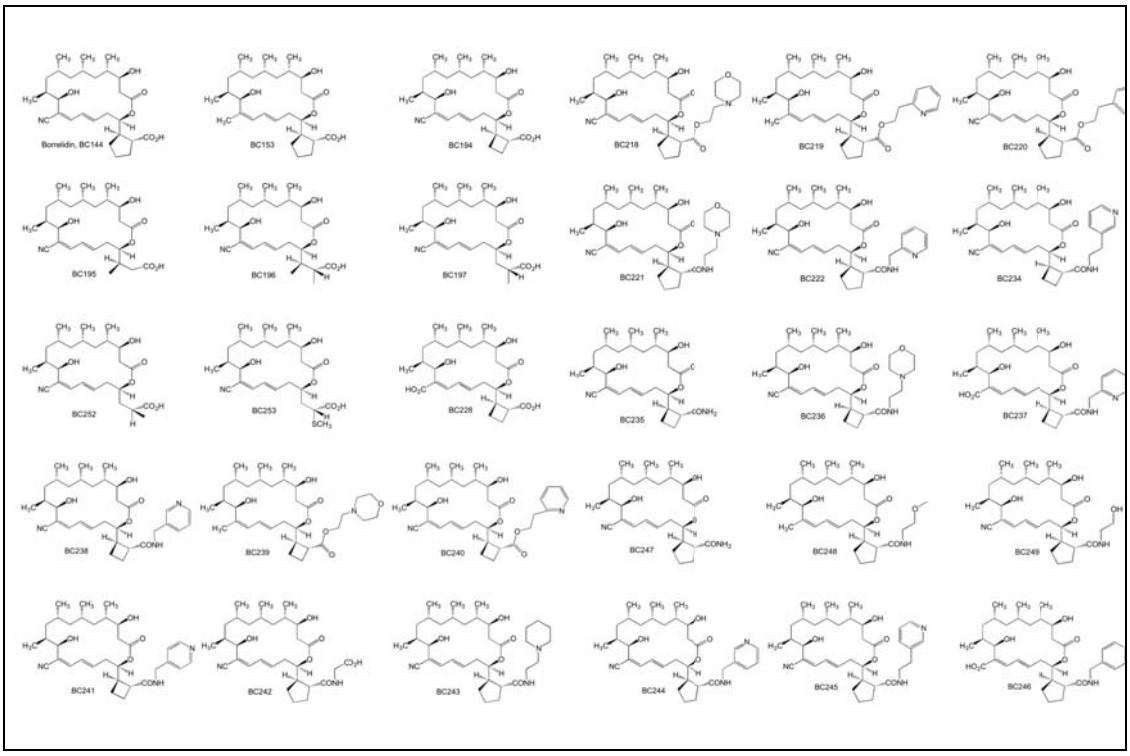
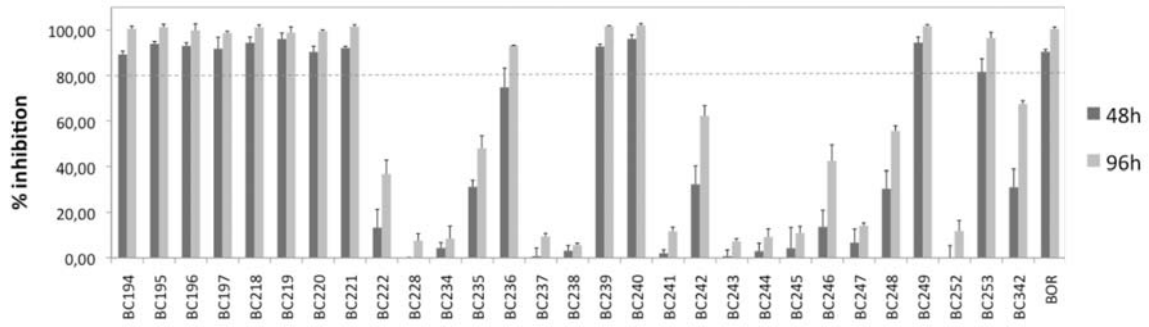
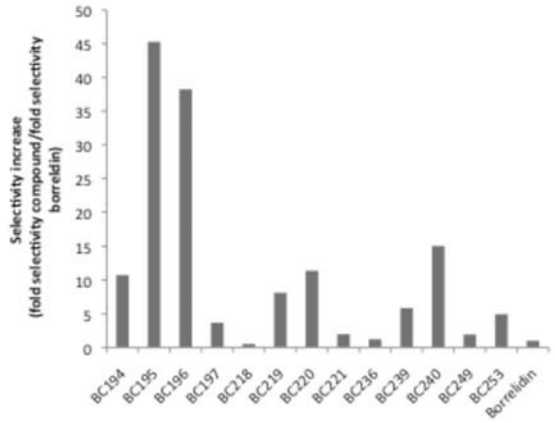


Figure 4. Antimalarial activity of the library of borrelidin derivatives. A) *In vitro* antimalarial activities of borrelidin derivatives tested at 100nM. Borrelidin (BOR) has also been included as positive control. Compounds inhibiting over 80% at 100nM (13 amongst the 30 compounds tested) were considered to be active, and were selected for IC_{50} determination in both *P. falciparum* and Hek293T cultures. B) Fold selectivity comparison of the 13 selected borrelidin derivatives compared to borrelidin. Borrelidin fold selectivity has been normalized to 1. C) Average parasitemias in *P. yoelii*-infected mice treated with the different compounds, determined by microscopic examination of Wright's stained flood films taken on day 4 after drug treatment. Compound BC-240 was not measured at 0.25 mg/kg.

A



B



C

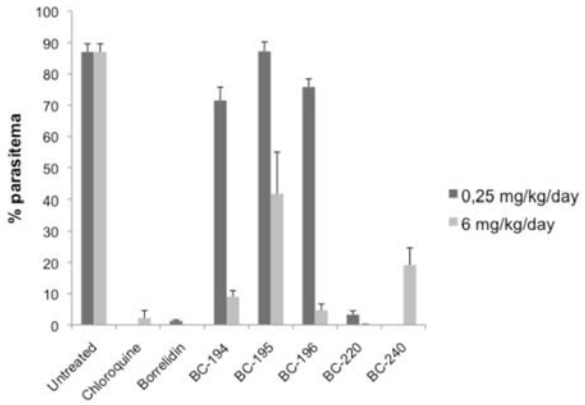


Table 1. List of known ARS inhibitors tested on *P. falciparum* cultures

Inhibitor	Target	IC50 (48h)	IC50 (96h)	Fold 96h/48h
Mechanism-Based inhibitors (analogues)				
Glu-SA	Glutamyl-tRNA synthetase PF13_0257 / MAL13P1.281	372,2 nM	463,5 nM	1
Gln-SA	Glutaminyl-tRNA synthetase PF13_0170	172,4 nM	150,3 nM	1
Asn-SA	Asparaginyl-tRNA synthetase PFB0525w / PFE0475w	88,3 nM	73,35 nM	1
Tyr- SA	Tyrosyl-tRNA synthetase MAL8P1.125 / PF11_0181	98,5 nM	84,6 nM	1
Ser-SA	Seryl-tRNA synthetase PF07_0073 / PFL0770w	39,5 nM	16,9 nM	2
Thialysine	Lysyl-tRNA synthetase PF13_0262 / PF14_0166	484,8 μM	154,3 μM	3
Natural inhibitors				
Mupirocin	Isoleucyl-tRNA synthetase PF13_0179 / PFL1210w	257 μM	93 nM	2763
Borrelidin	Threonine-tRNA synthetase PF11_0270	1,24 nM	0,97 nM	1
Cispentacin	Prolyl-tRNA synthetase PFL0670c / PFI1240c	573,4 μM	462,8 μM	1
Novel scaffolds				
AN2729	Leucyl-tRNA synthetase PFF1095w / PF08_0011	1,03 μM	0,68 μM	2

Table 2. *In vivo* intraperitoneal antimalarial activities and curative activity of known ARS inhibitors as antimalarial treatments against *P. yoelii yoelii* 17XL during 4-day test.

Compound	Dose (mg/kg/day)	Average % parasitemia	Average % suppression	% Survival	Survival time (days)
Negative control		94.54 ± 1.1	0	0	4.15 ± 0.1
Chloroquine	30	0 ± 0	100 ± 0	100	--
Mupirocin	2.5	92.73 ± 1.7	10.05 ± 7.6	40	4.6 ± 0.3
Borrelidin	0.25	1.3 ± 0.5	98.86 ± 0.5	100	--
Mupir + borr	2.5 & 0.25	0.66 ± 0.4	99.31 ± 0.4	100	--

Table 3. *In vitro* inhibitory activities of borrelidin derivatives both in *P. falciparum* cultures and human HEK293 cells.

	Plasmodium falciparum		**Human cells - Hek293T**	**Fold selectivity**
	IC_{50} 48h	**IC_{50} 96h**	**IC_{50} 72h**	**IC_{50}72h (Hs) / IC_{50}96h (Pf)**
BC194	3,88 nM	3,49 nM	13,32 μM	3816
BC195	7,15 nM	4,48 nM	72,06 μM	16084
BC196	6,73 nM	4,4 nM	59,75 μM	13579
BC197	93,68 nM	76,71 nM	> 100 μM	>1303
BC218	17,4 nM	25 nM	4,61 μM	184
BC219	25,35 nM	18,59 nM	53,63 μM	2884
BC220	44,36 nM	23,71 nM	95,99 μM	4048
BC221	36,08 nM	24,19 nM	16,87 μM	697
BC236	57,3 nM	55,21 nM	24,02 μM	435
BC239	8,137 nM	8,05 nM	16,76 μM	2081
BC240	9,625 nM	17,25 nM	92,26 μM	5348
BC249	11,06 nM	25 nM	17,14 μM	685
BC253	100,6 nM	56,95 nM	> 100 μM	> 1755
Borrelidin	1,24 nM	0,97 nM	345 nM	355

Table 4. *In vivo* intraperitoneal antimalarial activities and curative activity of the 5 most selective borrelidin derivatives (B-194, B-195, B-196, B-220 and B-240) against *P. yoelii yoelii* 17XL-infected mice (n=4-10).

Compound	Dose (mg/kg/day)	Average % parasitemia	Average % suppresion	% survival	Survival time in fatal cases (days)
Negative control	-	87.0 ± 2.6	0	0	4.8 ± 0.3
Chloroquine	6	2.2 ± 2.4	97.5 ± 2.75	100	-
Borrelidin	0.25	1.3 ± 0.3	98.9 ± 0.5	100	-
BC-194	0.25	71.5 ± 4.25	17.8 ± 4.9	0	5.0 ± 0.1
	6	9.0 ± 1.95	89.7 ± 2.2	40	8.33 ± 0.3
BC-195	0.25	87.2 ± 3.0	1.7 ± 2.1	0	4.7 ± 0.7
	6	41.7 ± 13.3	52.0 ± 15.3	25	6.3 ± 0.3
BC-196	0.25	75.8 ± 2.6	12.8 ± 3.0	0	5.2 ± 0.2
	6	4.7 ± 2.0	94.6 ± 2.3	100	-
BC-220	0.25	3.3 ± 1.2	96.2 ± 1.3	80	9.0 ± 0.1
	6	0.1 ± 0.1	99.9 ± 0.2	100	-
BC-240	6	19.1 ± 5.5	78.1 ± 6.3	60	7.5 ± 0.3

Figure S1. Peripheral blood smears of infected mice treated with different drugs. Blood smears were taken from each treated mouse on the 2nd, 4th, and 6th day post-infection (dpi) stained with Wright's eosin methylene blue solution and evaluated under microscope. Representative blood smears taken from i) infected mice treated with vehicle, ii) chloroquine, iii) mupirocin, iv) borrelidin, and v) mupirocin & borrelidin.

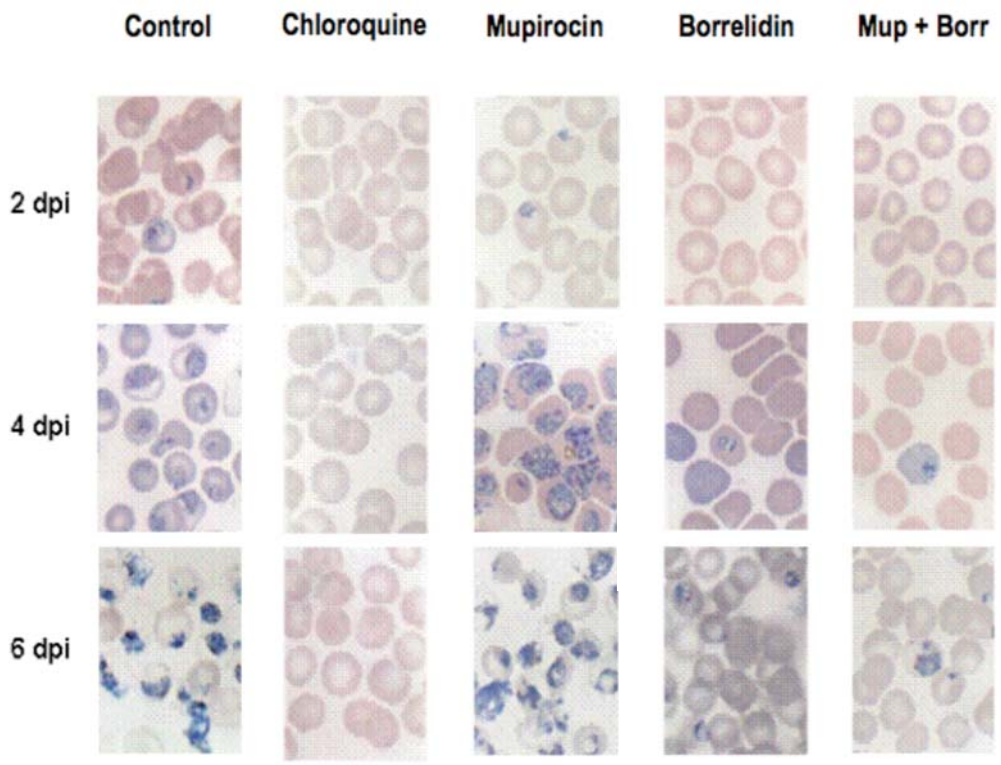


Figure S2. Multiple sequence alignment of threonyl-tRNA synthetases. Residues involved in borrelidin recognition are labeled with an asterisk. The three first sequences (*Escherichia coli, Homo sapiens* and *Plasmodium falciparum*) correspond to the bacterial-type ThrRS, and present conserved residues at the hydrophobic cluster, and are therefore expected to be sensitive to borrelidin. On the other hand, *Pyrococcus abyssi* ThrRS corresponds to an archaeal-type ThrRS, which does not conserve the hydrophobic cluster residues identified to interact with borrelidin, and has been shown to be non-sensitive to borrelidin.

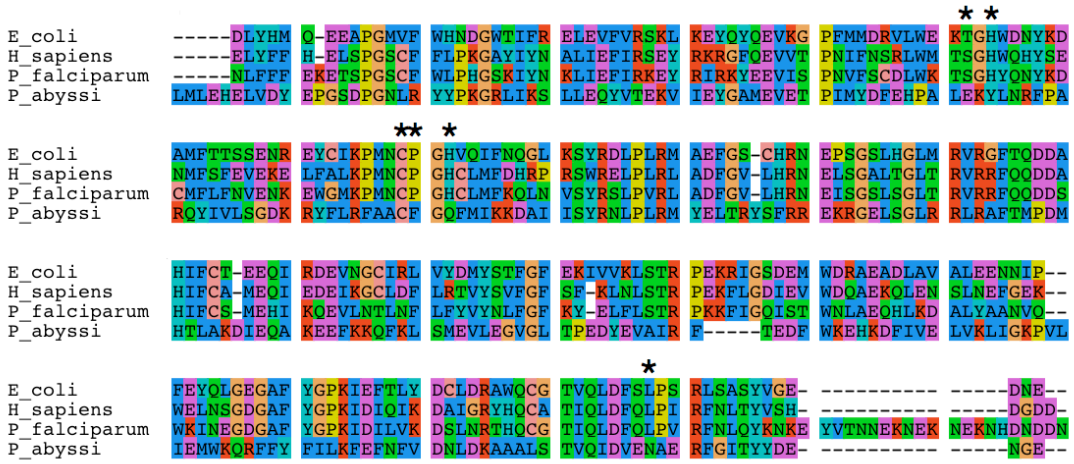
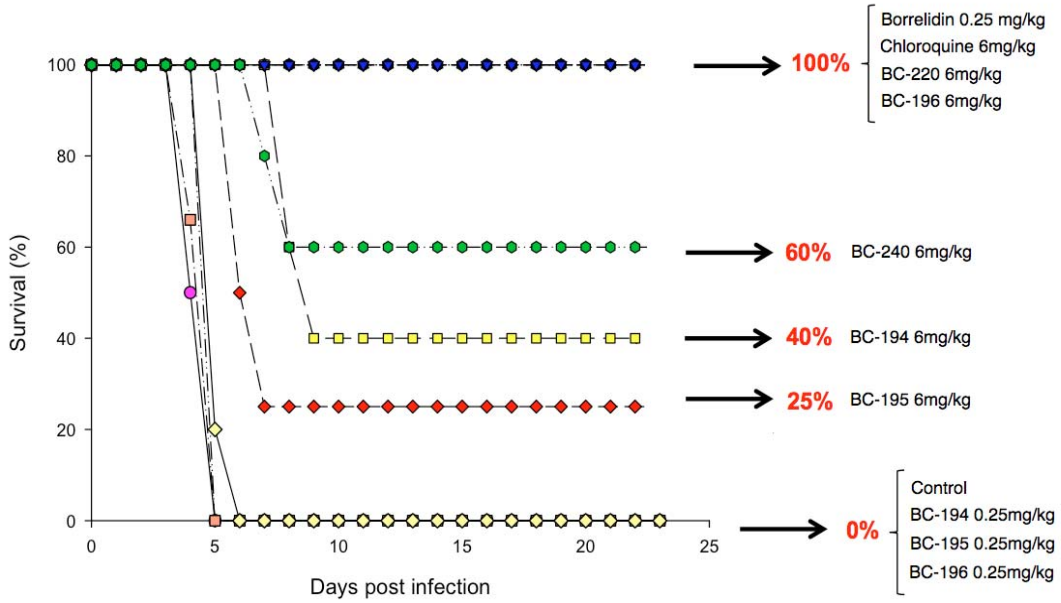


Figure S3. ***In vivo*** **mice survival of activities in *P. yoelii*-infected mice treated with the selected subset of borrelidin derivatives**. Percentage of mice survival of *P. yoelii*-infected mice, measured over 20 days after drug treatment. Chloroquine (Cq) has been used as positive control.



4.3. Chapter 3: Method development

4.3.1. Introduction

4.3.1.1. A sequence-based prediction method to identify pathogenicity-related proteins.

4.3.1.1.1. Use of sequence-based homology searches to predict protein function

Genome sequencing efforts are providing us with large amounts of sequence data. Without including metagenomic data, there are over 3.000 bacterial complete genomes published (www.genomesonline.org). The genomes of prokaryotes possess specific and relatively well-understood promoter sequences, such as transcription factor binding sites, that are relatively easy to identify, allowing us to annotate its open reading frames (ORFs).

However, identifying a gene and understanding its function are altogether different matters. At least one-fourth of genes that are identified in bacterial genomes are "hypothetical" or of unknown function. For many genome sequences, the only annotation that will be available will be based on computational predictions and comparisons with related microorganisms.

The dominant method of function "prediction" uses sequence homology software, because most proteins generally fall into a relatively small number of homologous protein families of related structure and usually of at least somewhat related function. Indeed, two proteins that diverged through evolution from a common ancestral sequence tend to have structural and functional characteristics in common. In this regard, computer programs for sequence-database homology search –e.g. BLAST, HMMER and FASTA- can be used to discern whether a given protein is homologous to an already known sequence or sequence family.

In this work we have developed a method to predict if a protein is related with pathogenicity, based exclusively on its sequence. For this aim, developed an algorithm that builds a protein profile for each input sequence, and is launched against a curated dataset of complete sequenced genomes –which includes a set of known human pathogens- to find its homologues. Finally, we consider a protein as pathogenicity-related if it is over-represented in

a set of proteomes from human pathogens compared to what should be expected by chance (**Publication 5**).

However, it is important to remember that homology offers only a "low-resolution" prediction of function. Thus, sequence homology analysis can often determine what a protein is likely to do, but not whether if it will be pathogenic or not, because a small number of changes may have a profound biological effect (Yoshida et al., 2001). Therefore, from our sequence-based analysis we cannot state whether a given protein sequence will be pathogenic or not, but the fact that its homologues–although with unknown function- are over-represented in human pathogens can give us a hint about its potential role in pathogenicity. Thus, this tool has the potential to shorten the list of proteins with unknown function that should deserve further characterization as potential drug targets.

4.3.1.1.2. Application of the method to functionally-related proteins

During their extended evolution genes coding for aaRS have experienced numerous instances of duplication, insertion and deletion of domains. The aaRS-related proteins that have resulted from these genetic events are generally known as aminoacyl-tRNA synthetase-like proteins (aaRS-like). This heterogeneous group of polypeptides carries out an equally varied number of functions that need not be related to gene translation. Several of these proteins remain uncharacterized. At least sixteen different aaRS-like proteins have been identified to date, but their functions remain incompletely understood.

Importantly, several of these aaRS-like proteins have been related to pathogenicity in several species. For this reason, we decided to investigate whether specific aaRS-like enzymes are found to be over-represented in pathogenic species compared to non-pathogenic species. To this end, we combined the analysis of the phylogenetic distribution of bacterial aaRS-like proteins with a simple and rapid algorithm for the identification of proteins that are over-represented in human pathogenic organisms. Our method positively identifies AsnA as over-represented in pathogenic species. Interestingly, AsnA has already been described as important in bacterial pathogens of plants and animals, and we suggest that its importance in infection may be extended to human microbial infections, and thus its role in pathogenicity should be further investigated (**Publication 5**).

4.3.1.1.3. Application of the method to whole genomes

Amongst the set of fully sequenced genomes, the *P. falciparum* genome contains by far the largest percentage –around 50%- of unknown or "hypothetical" proteins (Gardner et al., 2002). This lack of knowledge clearly limits the development of novel antimalarials and impedes a better understanding of the biology of the parasite. Thus, choosing useful targets for antimalarial drug design can be a difficult task.

In order to short-list which proteins could be priorized as antimalarial drug targets, we applied our sequence-based method to predict pathogenicity-related proteins to the full set of *P. falciparum* proteins (**Figure 4.23**). We considered those proteins with ER-species>2 and ER-proteins>2 as pathogenicity-related proteins, obtaining a list of 1209 ORFs. From these, if we remove those that present human homologs, we obtain a final list of 798 ORFs (~15% of the proteome).

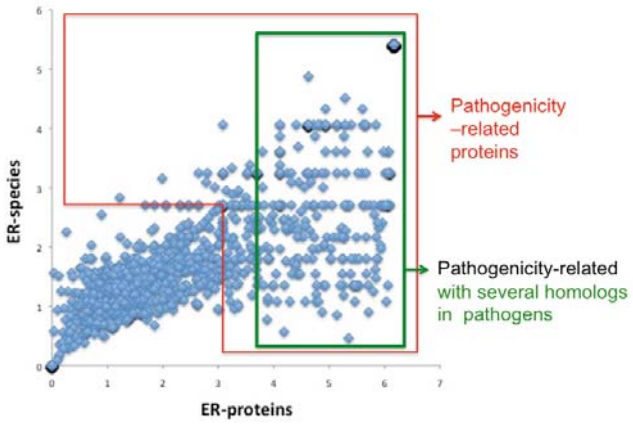


Figure 4.23. *P. falciparum* whole proteome analysis.

Previous works have already suggested to short-list the number of potential uncharacterized proteins that deserve further attention for antimalarial drug design. It was suggested to focus on apicoplast-targeted enzymes, because the prokaryotic origin of these nuclear-encoded apicoplast-targeted sequences makes them excellent drug targets (McFadden and Roos, 1999), and several inhibitors targeting these enzymes have already been shown to kill the parasite (Fichera and Roos 1997; Jomaa et al. 1999). Therefore, DeRisi and colleagues cross-referenced the list of predicted apicoplast-targeted sequences (plasmoDB.org) with those genes found to be maximally expressed between 33 and 36 hours post-infection (hpi) –which correspond to the expression times of the plastid genome-, resulting in a list of 124 in-phase

apicoplast-targeted genes (Bozdech et al., 2003). Importantly, from these, 76 ORFs (62%) were of unknown function, and are likely to include excellent candidate drug targets.

In a similar fashion, they also checked which genes presented expression profiles with similar characteristics to those involved in merozoite invasion–amongst them are seven of the best-known malaria vaccine candidates, including AMA1, MSP1, MSP3, MSP5, EBA175, RAP1 and RESA1-, obtaining a list of 262 ORFs (Bozdech et al., 2003). Another work also provided a list of 425 ORFs that were shown to be associated to *P. falciparum* heterochromatin protein 1 (PfHP1) (Flueck et al. 2009). This protein is a major structural component of virulence gene island throughout the genome, and is highly associated with the majority of known exported proteins involved in host-parasite interactions -e.g. *var*, *rif*, *stevor*, *surfin*, *pfmc-2tm*- (Flueck et al. 2009).

Taking into account these different approaches, we decided to provide a more solid list of candidate proteins, related with pathogenicity or to be used as drug/vaccine targets. For this aim, we cross-referenced our set of pathogenicity-related proteins (798 ORFs) with: i) the 124 set of in-phase apicoplast-targeted genes (Bozdech et al. 2003); ii) the 262 ORFs presenting similar expression profiles as merozoite invasion proteins (Bozdech et al. 2003); and iii) the 425 PfHP1-associated ORFs (Flueck et al. 2009). Two different intersections between the lists have been built, depending on whether the aim is to find candidate drug targets (**Figure 4.24a**) or candidate vaccine targets (**Figure 4.24b**). The specific ORFs found in the center of the Venn diagrams are shown in **Table 5**.

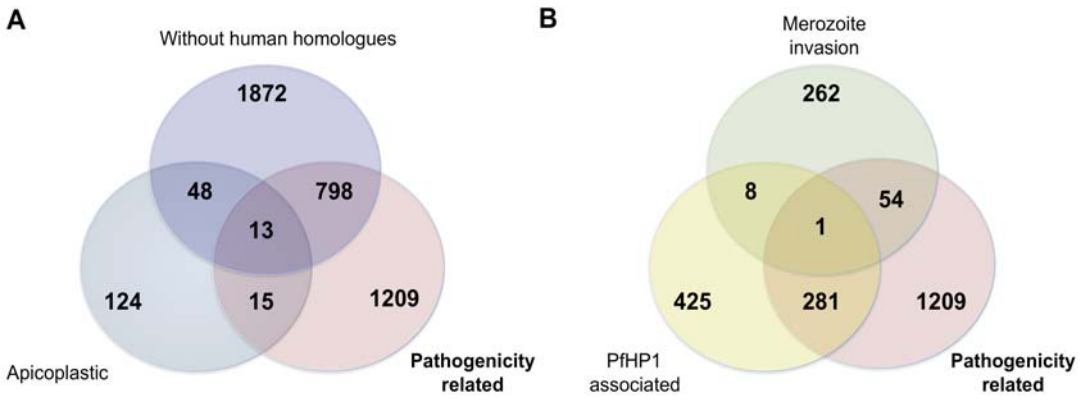


Figure 4.24. Venn diagrams showing intersections between several lists of plasmodial ORFs to decipher potential vaccine and drug candidates. A) Strategy to find candidate drug targets, by intesecting in-phase apicoplast-targeted genes, pathogenicity-related genes and genes without human homologues. **B**) Strategy to find candidate vaccine targets, by intersecting in-phase merozoite invasion genes, PfHP1-associated genes and pathogenicity-related genes.

Table 5. List of potential *P .falciparum* drug and vaccine candidates

Gene ID	PlasmoDB Annotation
Drug candidates	
MAL8P1.61	conserved Plasmodium protein, unknown function
PF08_0101	conserved Plasmodium protein, unknown function
PF10_0030	conserved Plasmodium protein, unknown function
PF10_0207	conserved Plasmodium membrane protein, unknown function
PF11_0324	conserved Plasmodium protein, unknown function
PF13_0025	apical membrane antigen 1 (AMA1)
PF14_0249	conserved Plasmodium protein, unknown function
PF14_0566	conserved Plasmodium protein, unknown function
PFC0435w	parasite-infected erythrocyte surface protein (PIESP1)
PFC0670c	conserved Plasmodium protein, unknown function
PFD0760c	conserved Plasmodium protein, unknown function
PFE0710w	conserved Plasmodium protein, unknown function
PFL0875w	conserved Plasmodium protein, unknown function
Vaccine candidates	
PF10_0355	merozoite surface protein (MSP3.8)

4.3.1.2. Ensemble docking from homology models

The genome sequencing efforts are providing us with a lot of information for hundreds of organisms, including humans. We are now faced with describing, controlling and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by protein three-dimensional structures, which are best determined by experimental methods such as X-ray crystallography and NMR spectroscopy. Despite significant advances in these techniques, the gap between the number of known sequences and structures continues to grow (Baker and Sali, 2001). Protein structure prediction methods attempt to bridge this gap, being comparative modeling the most reliable of the available methods to predict the 3D structure of a protein (Marti-Renom et al., 2000).

It is widely accepted that docking to comparative models is more challenging and less successful than docking to crystallographic structures (**Figure 4.25**), although a series of papers demonstrate the success in the use of comparative models in computational drug design studies (Schafferhans and Klebe, 2001; Evers and Klebe, 2004). However, little work has been done to quantify the accuracy of docking to comparative models (McGovern and Shoichet, J Med Chem 2003).

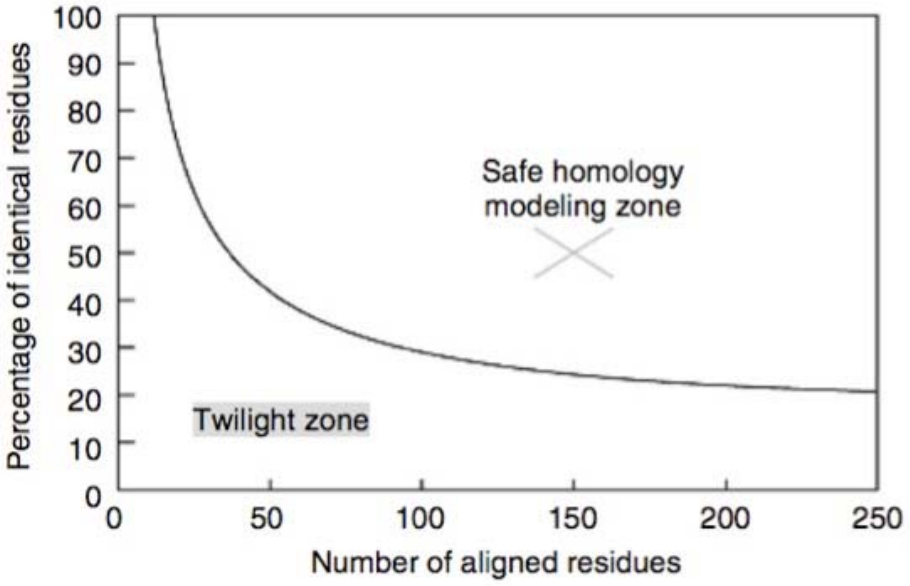


Figure 4.25. The reliabiity of homology models depends on the target-template sequence identity. Homology models over the 30% sequence identity are considered to be in the safe homology modeling zone. Below 30% sequence identity, serious errors might occur, and result in the basic fold being mis-predicted (Sander and Schneider, 1991).

We have attempted to determine the minimum sequence identity required to obtain docking results sufficiently similar to those obtained with crystallographic structures. We explored with great detail the quality of proteins structures derived from homology modeling for high throughput docking using *state-of-the-art* computational methods. We find that, contrary to common believe, structures derived from homology modeling are often of similar quality for docking purposes than the real crystal structure, even in cases where the template used to create the structural model shows a moderate sequence identity with the protein of interest. Indeed, we designed an "ensemble docking" approach (Craig et al., 2010) based on homology models that outperforms in most cases the docking performance using single experimental structures (**Figure 4.26**). Using this approach we estimate that the number of human proteins ameanable to high throughput docking for the design of increases five times, raising the possibility to perform proteome-scale docking experiments (**Publication 6** and **Publication 7**).

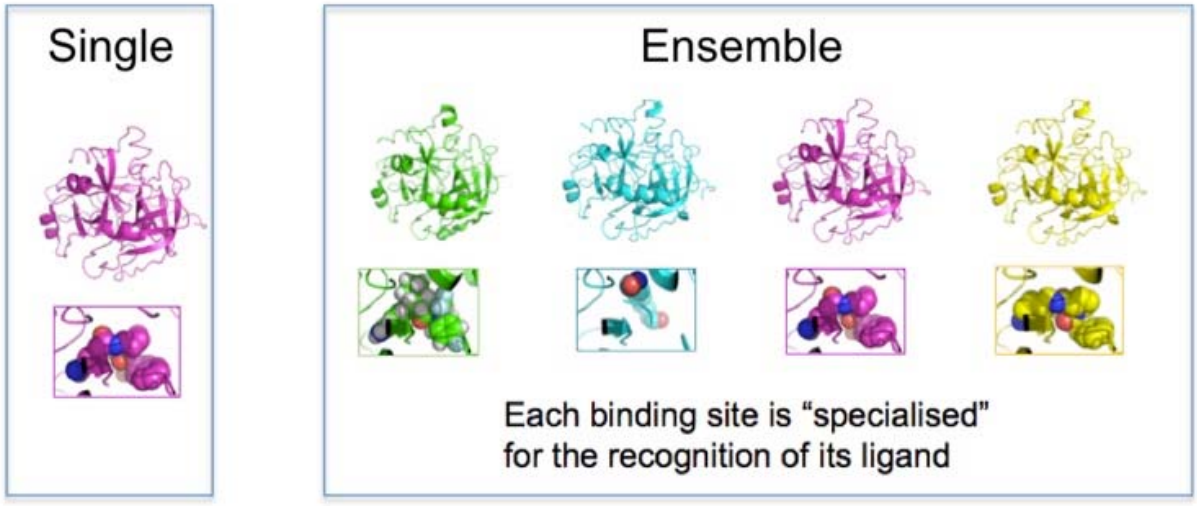


Figure 4.26. Rationale behind the ensemble docking approach. The classical docking approach consists in single docking one protein structure –generally an holo structure- against a set of ligands. The ensemble docking approach consists in docking a set of protein structures –holo structures with different ligands bound in its active site- against the set of ligands. Since each binding site is specialized for the recognition of its ligand –and similar scaffolds-, each structure of the ensemble will recognize a different subset of active ligands that otherwise would not be found using a single structure.

4.3.2. Publications

PUBLICATION 5:

A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins.

Novoa EM, Castro de Moura M, Orozco M and Ribas de Pouplana L.

FEBS Lett 2010, 584 (2): 460-466.

FEBS Letters 584 (2010) 460–466

journal homepage: www.FEBSLetters.org

A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins

Eva Maria Novoa [a], Manuel Castro de Moura [a], Modesto Orozco [a], Lluís Ribas de Pouplana [a,b,*]

[a] Institute for Research in Biomedicine (IRB), c/ Baldiri Reixac 15-21, 08028 Barcelona, Spain
[b] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

ARTICLE INFO

Article history:
Received 1 October 2009
Revised 3 November 2009
Accepted 8 November 2009
Available online 12 November 2009

Edited by Manuel Santos

Keywords:
Aminoacyl-tRNA synthetase-like protein
Pathogenicity
Proteome
Bacteria

ABSTRACT

During their extended evolution genes coding for aminoacyl-tRNA synthetases (ARS) have experienced numerous instances of duplication, insertion and deletion of domains. The ARS-related proteins that have resulted from these genetic events are generally known as aminoacyl-tRNA synthetase-like proteins (ARS-like). This heterogeneous group of polypeptides carries out an equally varied number of functions that need not be related to gene translation. Several of these proteins remain uncharacterized. At least 16 different ARS-like proteins have been identified to date, but their functions remain incompletely understood. Here we review the individual phylogenetic distribution of these proteins in bacteria, and apply a new genomics method to determine their potential implication in pathogenicity.

1. Introduction

Aminoacyl-tRNA synthetases represent an extraordinary example of functional and structural conservation [1]. Across all living species most of these enzymes display an almost identical structure, providing one of the few cases where phylogenetic and structural analyses can be expected to yield information about the first evolutionary steps of cellular life on earth [2–4]. As would be expected from a large group of enzymes, with complicated modular structures and extremely long evolutionary lives, a large group of related proteins has formed as a result of total or partial duplications of ARS genes [5,6]. In addition, some ARS-like proteins may exist that are coded by ancestral genes that were lately fused to a pre-existing ARS. Differentiating between these two possibilities can be difficult.

Functionally speaking ARS-like proteins are not a homogeneous class. However, a global analysis of their distribution is interesting because it provides information on the evolutionary history of ARS, and it might help to identify tendencies in the functional roles that ARS-related domains adopt when they diverge from their ancestral enzymes. Moreover, the species distribution of each ARS-like protein is likely to provide information on its biological role. More specifically, the search for correlations between gene distribution and complex biological phenotypes can be a powerful tool for the identification of biological function.

Here we combine the analysis of the phylogenetic distribution of bacterial ARS-like proteins with a simple and rapid algorithm for the identification of proteins that are over-represented in human pathogenic organisms. First, we have applied our method to re-examine the different ARS-like proteins found in bacteria, clustering them according to a sequence-similarity profile. Secondly, we have analyzed whether each of the 11 bacterial ARS-like proteins that we obtain is functionally linked to bacterial virulence (Fig. 1). Our method positively identifies AsnA as over-represented in pathogenic species. AsnA has already been described as important in bacterial pathogens of plants and animals [7,8]. We suggest that its importance in infection may be extended to human microbial infections.

2. Methods

2.1. Protein profile generation and determination of phylogenetic distributions

We selected 16 well-documented ARS-like proteins for our study (Table 1). For each of them, a multiple alignment was built with ClustalW [29] using the Gonnet protein matrix, followed by a Hidden Markov profile building using the HMMER package [30]. Each protein profile was used as query to find all existing homologues in the Uniprot database (www.uniprot.org). In order to apply a consistent criterion to the determination of each protein's distribution we applied a cutoff value to the search for

* Corresponding author. Address: Institute for Research in Biomedicine (IRB), c/ Baldiri Reixac 15-21, 08028 Barcelona, Spain.
E-mail address: lluis.ribas@irbbarcelona.org (L. Ribas de Pouplana).

0014-5793/$36.00
doi:10.1016/j.febslet.2009.11.016

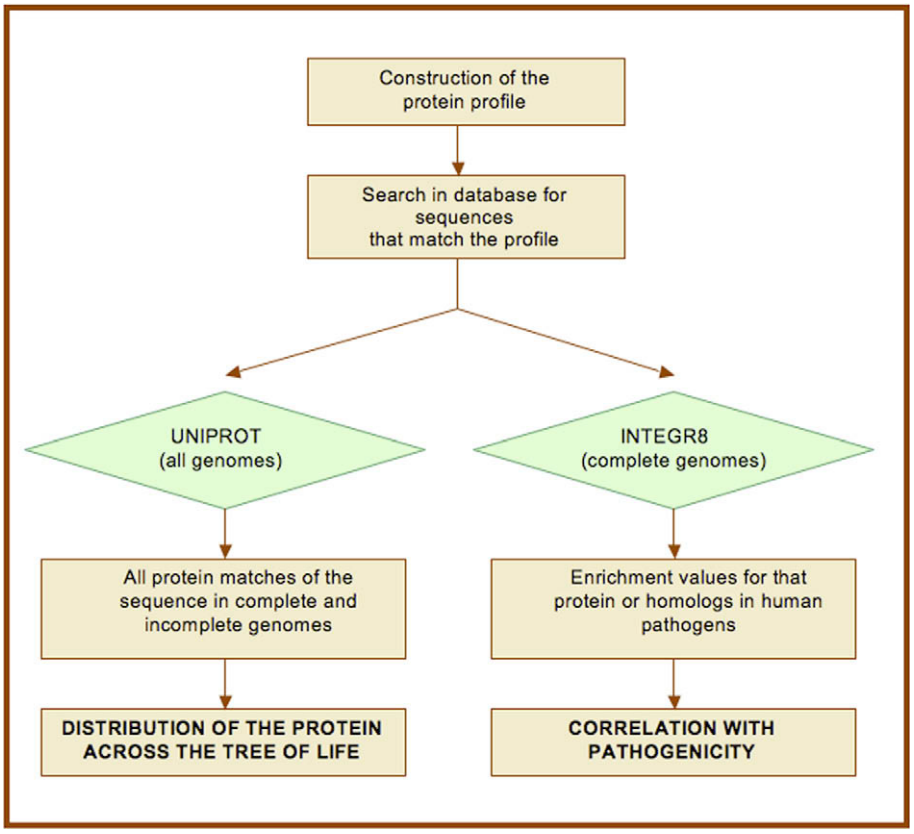


Fig. 1. Schematic representation of the over-representation analysis performed in this work.

Table 1
List of the 16 ARS-like proteins considered in this study.

Synthetase-like	aaRS paralog	Reference
Ybak	ProRS	[9,10]
HisZ	HisRS	[11,12]
AlaX	AlaRS	[5,13]
PrdX (ProX)	ProRS	[13,14]
GluX (YadB)	GluRS	[15]
CTP	Class I ARS	[16]
ATPS	Class I ARS	[16]
EMAP-II	MetRS, TyrRS	[17–19]
Arc1p	MetRS, TyrRS	[20]
Trbp111	MetRS, TyrRS	[21]
BirA	SerRS	[22,23]
AsnA	AspRS, AsnRS	[24,25]
ThrRS-ed	ThrRS	[26]
Gcn2	HisRS	[11]
Pol gamma B	GlyRS	[27]
PoxA/GenX	LysRS	[28]

homologues (per-sequence *E*-value cutoff of 10.0). This procedure identified clusters of proteins that were considered as evolutionarily related and treated as a single family. Those families present in bacteria were selected for further analysis. The distribution found for each bacterial ARS-like family was graphically displayed through the quantification of all its homologous sequences in the main bacterial phyla and the representation of these frequencies on a model phylogenetic tree of bacteria [31] (Fig. 2).

To correct for the fact that not all bacterial phyla are equally represented among the Uniprot database, a standardization of the values of the ARS-like proteins was done in order to obtain final values comparable among the different bacterial phyla. The relative abundance of each protein in a phylum was computed by dividing the number of protein hits found in that phylum by the total number of proteins found for the phylum in the Uniprot database:

$$\text{Relative abundance} = \frac{\text{No. “X” in phylum}}{\text{No. proteins of phylum in Uniprot}} \quad (1)$$

Since a protein of a given species may be represented more than once in the Uniprot database – e.g. same protein from different strains –, only semi-quantitative values can be obtained from this analysis. Nevertheless, the calculation is accurate enough to provide an estimation of the distribution of each ARS-like protein among bacterial phyla.

2.2. Correlation analysis of protein distributions and pathogenicity

2.2.1. Database preparation and construction of the set of human pathogens

In order to identify proteins over-represented in pathogenic species the curated set of complete proteomes from the Integr8 database (2069 complete proteomes) was used (www.ebi.ac.uk/integr8). This collection was further modified to obtain our final proteome dataset (viral proteomes were removed and only one proteome per species was used) of 910 complete proteomes.

In over-representation studies a carefully curated dataset is essential to avoid artificial over-representation of data (e.g. fragments of the proteins, point mutations, more that one strain per species) that leads to non-reliable values of enrichment. From the final dataset of 910 complete proteomes, 168 were identified as belonging to human pathogens. This was done with the help of different curated databases: HAMAP database (http://www.expasy.ch/sprot/hamap/), pathogenic bacteria database (bac.hs.med.kyoto-u.ac.jp), national microbiology pathogen data resource (www.nmpdr.org), pathogenic fungi database (www.pfdb.net),

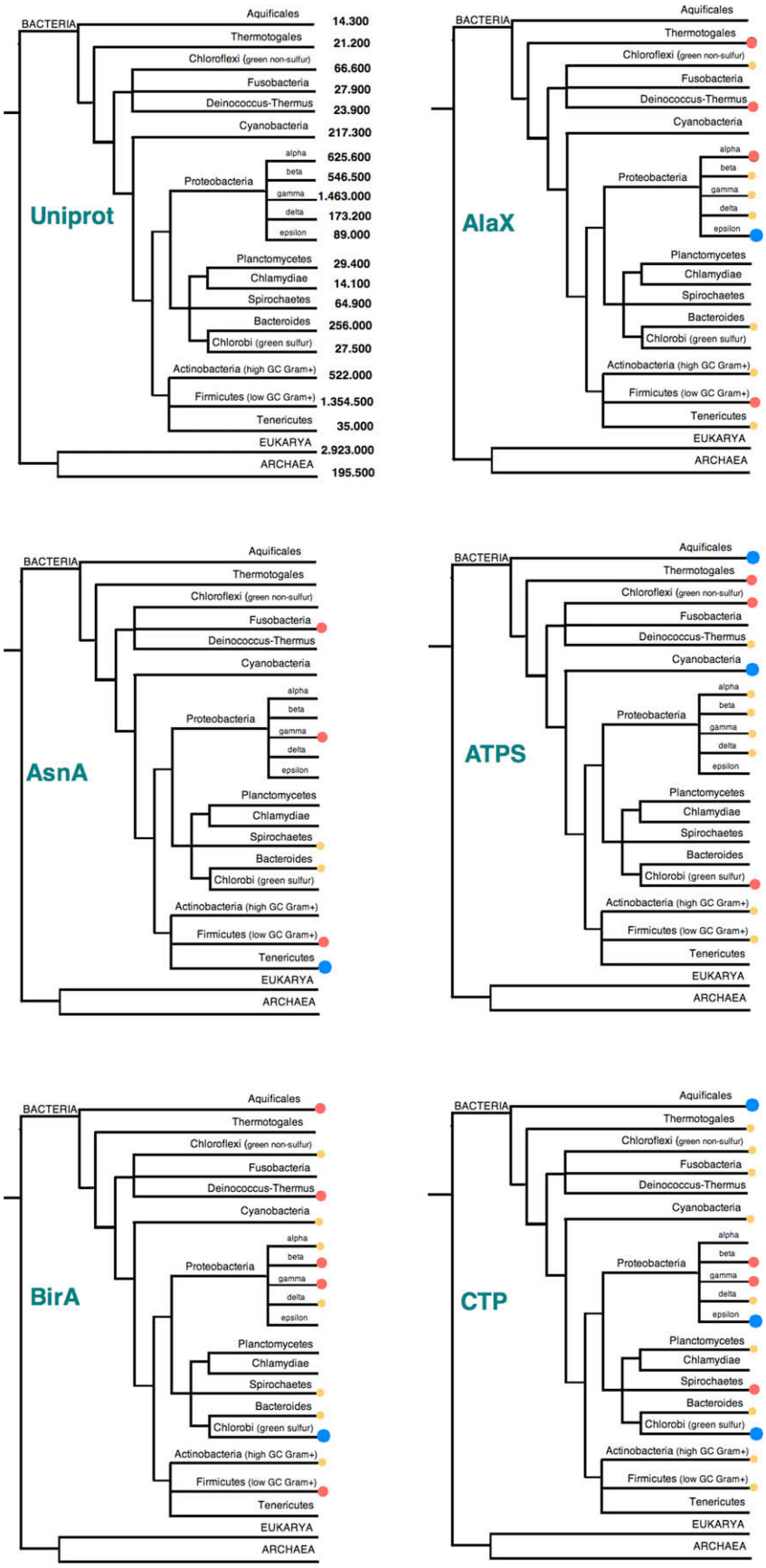


Fig. 2. Phylogenetic distribution and relative abundance of the 11 bacterial ARS-like proteins considered in this work. Each tree is labeled according to the protein whose distribution is being analyzed. The tree labeled UNIPROT shows the number of proteins per phylum that are included in the database used. The relative abundance (r.a.) of each protein in each phylum is represented by a colored circle at the end of the phylum's branch: blue (r.a. $\geqslant$ 25), salmon (r.a. $\geqslant$ 10) and yellow (0 > r.a. > 10). Only bacterial relative abundances are shown.

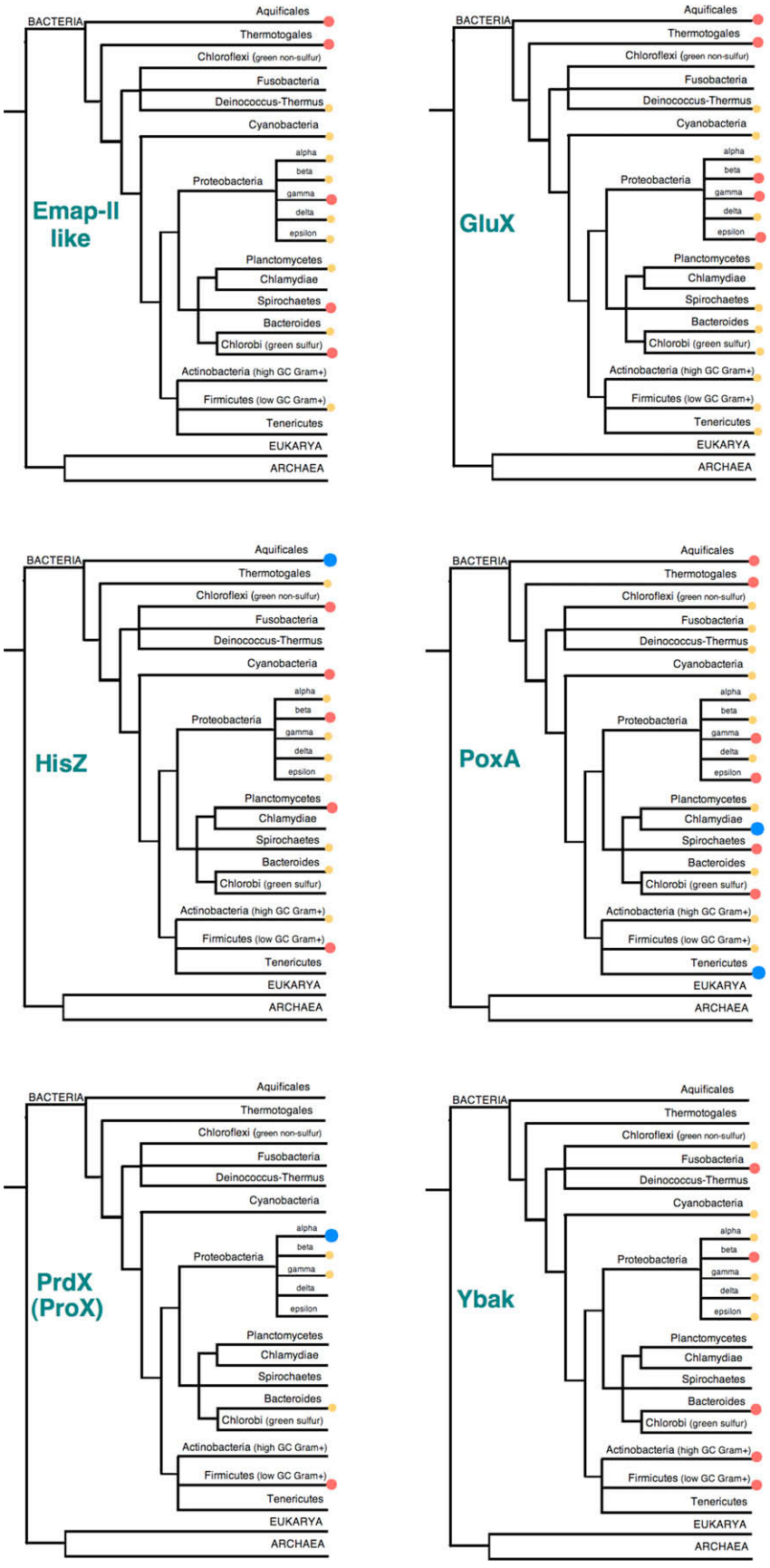


Fig. 2 (*continued*)

eukaryotic pathogens database (eupathdb.org), and pathogen portal (www.pathogenportal.org). The final list of human pathogens includes 146 bacteria, 11 fungi and 12 protozoa.

2.2.2. Construction of a control dataset

Both positive and negative controls were included in the study for external validation of the method. Negative controls used are proteins not expected to be over-represented in human pathogens (tubulin, enolase, alanyl-tRNA synthetase, lactate dehydrogenase, and pyruvate dehydrogenase). Positive controls were built with proteins known to be linked to pathogenicity – e.g. virulence factors – (haemolysin, gamma-glutamyl transpeptidase, CapC, fim2 fimbrial subunit precursor, lipopolysaccharide transferase, sycE secretion chaperone, heme exporter protein CcmC, long polar fimbrial chaperone, adhesin, cholera enterotoxin, streptococcal exotoxin I and lipoteichoic acid synthase).

2.2.3. Calculation of over-representation indices

Protein profiles were built for both controls and test cases (ARS-like proteins) following the same procedure explained above. Each protein profile was compared to our curated set of Integr8 proteomes, to obtain the complete list of homologues for each of the proteins among the 910 proteomes. We considered a protein as pathogenicity-related if it was found over-represented in the set of human pathogens compared to what is expected by chance. Over-representation was measured using two different indices: enrichment rate of the number of proteins (ER-proteins) and enrichment rate of the number of species (ER-species), which are computed as follows:

$$\text{ER-species} = \frac{\text{Pathogenic species with “X”/Species with “X”}}{\text{Pathogenic species/Species in database}} \quad (2)$$

$$\text{ER-proteins} = \frac{\text{No. “X” in path. spp/No. “X” in database}}{\text{Proteins in path. spp/Proteins in database}} \quad (3)$$

where “X” is the queried protein of interest.

Although both ratios quantify the over-representation of a given protein among pathogen species they may produce different enrichment ratios because a species can have one or more homologues of the queried protein. Thus, enrichment must be quantified both in terms of number of proteins and number of species.

Significance testing on protein distribution results was performed using a one-tailed test, and threshold values were computed both for 1% and 0.1% false positive rates (FP) [32,33]. In one-tailed tests, we can compute the threshold or cutoff value depending on the false positive rates (FP) that we accept:

$$\text{Threshold } (5\%\ \text{FP}) = \overline{X} \pm 1.64\,\widehat{\sigma} \quad (4)$$

$$\text{Threshold } (1\%\ \text{FP}) = \overline{X} \pm 2.32\,\widehat{\sigma} \quad (5)$$

$$\text{Threshold } (0.1\%\ \text{FP}) = \overline{X} \pm 3.09\,\widehat{\sigma} \quad (6)$$

where $\bar{X}$ is the population mean and $\widehat{\sigma}$ is the estimator of the standard deviation of the population. By plotting ER-species as a function of ER-proteins, control proteins that are not linked to pathogenicity should be clustered around the (1, 1) coordinates. A protein that is not over-represented is expected to fall into the normal distribution of the negative controls, with cutoff values that depend on the rate of false positives that we accept.

3. Results

3.1. Distribution of bacterial ARS-like proteins

Analysis of the phylogenetic distributions among the different bacterial phyla was performed for the complete set of ARS-like proteins (Fig. 2). From the 16 ARS-like proteins initially analyzed (Table 1) Arc1p, Gcn2, ThrRS-ed, Polγβ and AlaX2 were excluded because their distribution was found to be limited to eukarya (Gcn2 and Arc1p), archaea (ThrRS-ed), or eukarya and archaea (AlaX2, Polγβ). Emap-II and Trpb111 sequences were merged into one unique class because 90% of the sequences identified as Trbp111 are also present in the Emap-II profile. The distributions of the resulting 11 ARS-like proteins present in bacterial phyla are shown in Fig. 2. Minority phyla have not been represented in order to simplify the presentation of the results.

3.2. Identification of pathogenicity-related ARS-like proteins

We have constructed a simple and fast algorithm to determine whether a given protein is significantly over-represented in pathogenic organisms, and we have applied the method to bacterial ARS-like proteins. We consider a protein as pathogenicity-related if it is over-represented in a set of proteomes from human pathogens compared to what it should be expected by chance.

We computed the enrichment values (ER-proteins and ER-species, see Section 2), both for the set of controls and for the ARS-like proteins (Table 2). By plotting the enrichment rates (Fig. 3), we can clearly distinguish two differently distributed populations, corresponding to the negative and positive controls. The negative control distribution is centered around ER-proteins = 1 and ER-species = 1, whereas the positive control distribution (pathogenicity-related) has a higher variance and goes from non-enrichment values to high enrichment values. ARS-like proteins are mainly distributed among the negative control distribution, with the exception of AsnA, which clusters with pathogenicity-related proteins.

Table 2
Overrepresentation values for the different ARS-like proteins, including the negative and positive controls used in this study.

	ER-proteins	ER-species
Negative controls		
Tubulin	1.1	1.55
Enolase	1.01	0.96
Alanyl-tRNA synthetase	1.14	1.02
Lactate deshydrogenase	0.89	0.86
Pyruvate deshydrogenase	0.93	0.92
Positive controls		
Lipoteichoic acid synthase	1.24	1.17
Adhesin yadA	6.17	5.42
Haemolysin	2.06	1.81
Glutamyl transpeptidase	0.95	0.91
CapC	1.68	1.55
Fimbrial subunit precursor	2.26	2.71
LPS transferase	2.92	2.56
SycE secretion chaperone	3.08	2.71
Heme exporter protein	1.32	1.17
Fimbrial chaperone	2.46	1.87
Cholera enterotoxin	6.17	5.42
Streptococcal exotoxin	6.17	5.42
Coagulase	5.11	4.42
HifA – pilin	3.08	2.71
ARS-like proteins		
AlaX	0.5	0.51
ThrX	1.11	1
AsnA	2.53	1.79
ATPS	0.76	0.69
BirA	0.97	0.87
CTP	1	0.91
GluX	1.21	1.43
HisZ	0.44	0.4
PoxA	1.29	0.97
PrdX (ProX)	0.86	0.77
Ybak	1.2	1.06
EMAP-II	1.13	0.98

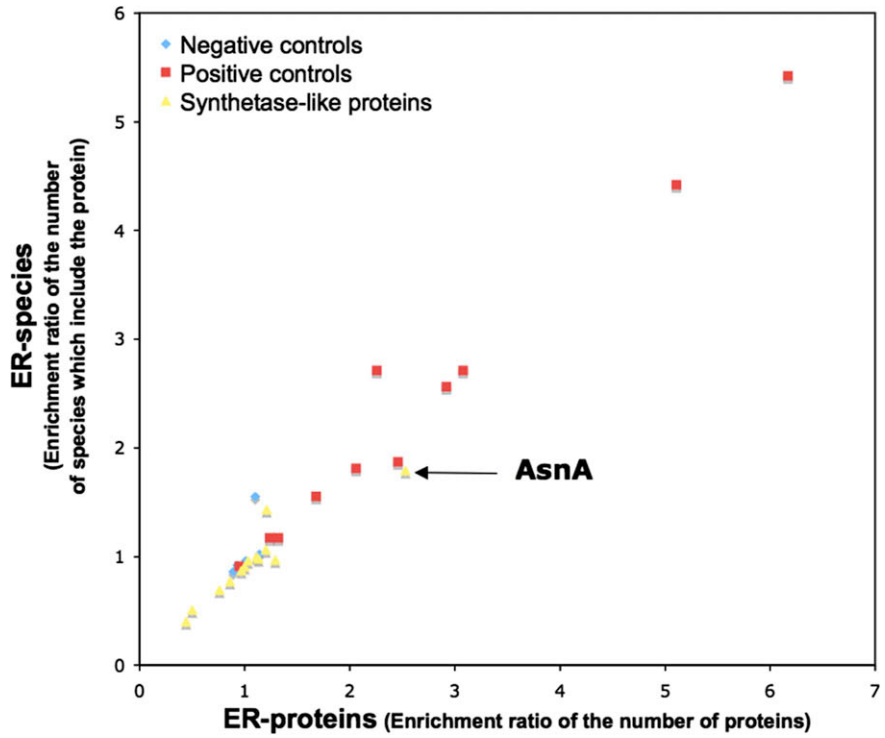


Fig. 3. Distribution of over-representation values for all ARS-like proteins (yellow boxes), and positive or negative controls for pathogenicity (red squares and blue diamonds, respectively). The position of AsnA is marked by an arrow and labeled accordingly.

Significance testing on the distribution results for AsnA was performed using a one-tailed test as described above. Since the ER-proteins mean for the negative controls is 1.014 ± 0.107, the thresholds corresponding for 5% FP, 1% FP and 0.1% FP are 1.19, 1.26 and 1.34, respectively. Taking this into account, AsnA is not a member of the negative control distribution with a *P*-value that approaches zero even at 0.1% FP. Thus, our results suggest that AsnA might be correlated with pathogenicity. GluX slightly deviates from the negative control set, however ER-proteins and -species values for GluX are below its respective cutoffs for a 1% false positive rate. Thus we can conclude that this deviation is not statistically significant and that GluX is not over-represented in human pathogens.

4. Discussion

The evolutionary relationships between ARS and ARS-like proteins have been analyzed previously through the use of phylogenetic methods [3,34,35]. This approach represents the best available strategy for the identification of cladistic relationships, but it is easily confounded by the extremely long evolutionary times experienced by aminoacyl-tRNA synthetases and their related proteins. Irrespectively of clade relationships, the species distribution of genes represents important information that can be linked to function and, indirectly, to evolutionary origin. Here we have analyzed the distribution of an ARS-like proteins families in bacteria and built a simple algorithm to analyze correlations between the distribution of a given protein and the pathogenicity of the species where it is present. The 11 ARS-like protein families that we have analyzed display very different distribution patterns among bacterial phyla. *A grosso modo*, we can distinguish between proteins that are universally or almost universally present, those that are present in the majority of phyla, and those that are present only in a minority of the main bacterial groups.

A wide distribution of a protein possibly reflects an ancient origin of the gene but lateral gene transfer, which is particularly widespread among bacteria, should always be considered an alternative explanation. This is the case for the proteins CTP, EMAP II, YadB, HisZ, and PoxA. Among this group are enzymes whose function is completely unrelated to gene translation (CTP, HisZ, and PoxA) and others that remain linked to tRNA biology (EMAP II and YadB). Interestingly, PoxA is a well-known pathogenicity factor in *Salmonella* [28]. However, its wide distribution suggests that its biological function is not exclusively linked to the establishment of infection, and the protein does not appear to be over-represented in pathogenic species (Figs. 2 and 3). Obviously negative values for enrichment in pathogens do not eliminate the possibility that a protein is a virulence factor. However, significant positive enrichment rates should be indicative of proteins whose function is pathogenicity-related.

Abundant but not universally distributed bacterial ARS-like families represent an important fraction of the set analyzed here (AlaX, ATPS, BirA, YbaK). Interestingly, two trans editing domains are present in this group, indicating that the need for misacylation correction may not be universal among bacteria. The scattered distribution of these enzymes may suggest that lateral gene transfer occurred among those species where the fidelity of the genetic code is particularly compromised and benefits from the function of in-trans editing domains [26].

It should be stressed that this situation needs not to be related to the specific kinetic behavior of the concerned ARS but can be caused by environmental conditions that, for instance, change the relative availability of similar amino acids. This situation would clearly favor the lateral transfer of these genes among species under similar environmental stresses.

Finally a small set of proteins (AsnA and PrdX) present a very limited distribution among bacteria. PrdX was originally described as the *trans*-editing enzyme ProX from *Clostridium sticklandii*, and shown to specifically deacylate alanyl-tRNA^{Pro} [13,36]. PrdX and YbaK are two different *trans*-editing enzymes that hydrolyze different forms of mischarged tRNA^{Pro} [13]. Consistent with previous reports, YbaK and PrdX groups do not overlap in our analysis. How-

ever, they do display overlapping distributions at the phylum level, as would be expected from two independent editing domains that recognize different substrates. Despite its more limited distribution PrdX is not over-represented in pathogenic bacteria (Fig. 3).

Asparagine synthetase (AsnA) is a paralog of asparagine- and aspartyl-tRNA synthetases that displays a limited distribution among bacterial phyla. AsnA is unique among the ARS-like proteins analyzed here because it is significantly over-represented in human pathogenic bacteria. AsnA has been shown to act as a virulence factor in fish and plant pathogens, although the molecular bases for this role in virulence remain unknown [7,8]. From our data it is reasonable to predict that AsnA may also be a virulence factor among human pathogens that, as such, deserves further analysis and consideration as a potential therapeutic target.

Acknowledgements

This work has been supported by Grants BIO2006-01551 from the Spanish Ministry of Science and Education, and HEALTH-F3-2009-223024 (MEPHITIS) from the European Union.

References

[1] Ibba, M. and Soll, D. (2000) Aminoacyl-tRNA synthesis. Annu. Rev. Biochem. 69, 617–650.
[2] Ribas de Pouplana, L. and Schimmel, P. (2001) Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends Biochem. Sci. 26, 591–596.
[3] Wolf, Y.I., Aravind, L., Grishin, N.V. and Koonin, E.V. (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. 9, 689–710.
[4] O'Donoghue, P. and Luthey-Schulten, Z. (2003) On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol. Mol. Biol. Rev. 67, 550–573.
[5] Schimmel, P. and Ribas De Pouplana, L. (2000) Footprints of aminoacyl-tRNA synthetases are everywhere. Trends Biochem. Sci. 25, 207–209.
[6] Francklyn, C. (2003) TRNA synthetase paralogs: evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to de novo biosynthesis. Proc. Natl. Acad. Sci. USA 100, 9650–9652.
[7] Menendez, A., Fernandez, L., Reimundo, P. and Guijarro, J.A. (2007) Genes required for *Lactococcus garvieae* survival in a fish host. Microbiology 153, 3286–3294.
[8] Olea, F. et al. (2004) Up-regulation and localization of asparagine synthetase in tomato leaves infected by the bacterial pathogen *Pseudomonas syringae*. Plant Cell Physiol. 45, 770–780.
[9] Wong, F.C., Beuning, P.J., Silvers, C. and Musier-Forsyth, K. (2003) An isolated class II aminoacyl-tRNA synthetase insertion domain is functional in amino acid editing. J. Biol. Chem. 278, 52857–52864.
[10] Wong, F.C., Beuning, P.J., Nagan, M., Shiba, K. and Musier-Forsyth, K. (2002) Functional role of the prokaryotic proline-tRNA synthetase insertion domain in amino acid editing. Biochemistry 41, 7108–7115.
[11] Sissler, M., Delorme, C., Bond, J., Ehrlich, S.D., Renault, P. and Francklyn, C. (1999) An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. Proc. Natl. Acad. Sci. USA 96, 8985–8990.
[12] Bond, J.P. and Francklyn, C. (2000) Proteobacterial histidine-biosynthetic pathways are paraphyletic. J. Mol. Evol. 50, 339–347.
[13] Ahel, I., Korencic, D., Ibba, M. and Soll, D. (2003) Trans-editing of mischarged tRNAs. Proc. Natl. Acad. Sci. USA 100, 15422–15427.
[14] Murayama, K. et al. (2005) Structure of a putative trans-editing enzyme for prolyl-tRNA synthetase from *Aeropyrum pernix* K1 at 1.7 Å resolution. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 61, 26–29.
[15] Dubois, D.Y. et al. (2004) An aminoacyl-tRNA synthetase-like protein encoded by the *Escherichia coli* yadB gene glutamylates specifically tRNAAsp. Proc. Natl. Acad. Sci. USA 101, 7530–7535.
[16] Ibba, M., Francklyn, C. and Cusack, S. (2005) The Aminoacyl-tRNA Synthetases, Landes Bioscience: Eurekah.com, Georgetown, TX, USA.
[17] Kao, J. et al. (1994) Characterization of a novel tumor-derived cytokine. Endothelial-monocyte activating polypeptide II. J. Biol. Chem. 269, 25106–25119.
[18] Kim, Y., Shin, J., Li, R., Cheong, C., Kim, K. and Kim, S. (2000) A novel anti-tumor cytokine contains an RNA binding motif present in aminoacyl-tRNA synthetases. J. Biol. Chem. 275, 27062–27068.
[19] Wakasugi, K. and Schimmel, P. (1999) Highly differentiated motifs responsible for two cytokine activities of a split human tRNA synthetase. J. Biol. Chem. 274, 23155–23159.
[20] Simos, G., Segref, A., Fasiolo, F., Hellmuth, K., Shevchenko, A., Mann, M. and Hurt, E.C. (1996) The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases. EMBO J. 15, 5437–5448.
[21] Nomanbhoy, T., Morales, A.J., Abraham, A.T., Vortler, C.S., Giege, R. and Schimmel, P. (2001) Simultaneous binding of two proteins to opposite sides of a single transfer RNA. Nat. Struct. Biol. 8, 344–348.
[22] Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2002) Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res. 12, 1507–1516.
[23] Artymiuk, P.J., Rice, D.W., Poirrette, A.R. and Willet, P. (1994) A tale of two synthetases. Nat. Struct. Biol. 1, 758–760.
[24] Nakatsu, T., Kato, H. and Oda, J. (1998) Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. Nat. Struct. Biol. 5, 15–19.
[25] Roy, H., Becker, H.D., Reinbolt, J. and Kern, D. (2003) When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. Proc. Natl. Acad. Sci. USA 100, 9837–9842.
[26] Korencic, D. et al. (2004) A freestanding proofreading domain is required for protein synthesis quality control in Archaea. Proc. Natl. Acad. Sci. USA 101, 10260–10265.
[27] Fan, L., Sanschagrin, P.C., Kaguni, L.S. and Kuhn, L.A. (1999) The accessory subunit of mtDNA polymerase shares structural homology with aminoacyl-tRNA synthetases: implications for a dual role as a primer recognition factor and processivity clamp. Proc. Natl. Acad. Sci. USA 96, 9527–9532.
[28] Kaniga, K., Compton, M.S., Curtiss 3rd, R. and Sundaram, P. (1998) Molecular and functional characterization of *Salmonella enterica* serovar *typhimurium* poxA gene: effect on attenuation of virulence and protection. Infect. Immun. 66, 5599–5606.
[29] Larkin, M.A. et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.
[30] Wistrand, M. and Sonnhammer, E.L. (2005) Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics 6, 99.
[31] Dworkin, M. and Falkow, S. (2006) The Prokaryotes: A Handbook on the Biology of Bacteria, Springer, New York, London.
[32] Fisher, R.A. (1925) Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh.
[33] Freund, J.E. and Perles, B.M. (2006) Modern Elementary Statistics, Prentice Hall, Paramus, NJ, USA.
[34] Ibba, M. and Soll, D. (2004) Aminoacyl-tRNAs: setting the limits of the genetic code. Genes Dev. 18, 731–738.
[35] Beebe, K., Ribas De Pouplana, L. and Schimmel, P. (2003) Elucidation of tRNA-dependent editing by a class II tRNA synthetase and significance for cell viability. EMBO J. 22, 668–675.
[36] An, S. and Musier-Forsyth, K. (2004) Trans-editing of Cys-tRNAPro by *Haemophilus influenzae* YbaK protein. J. Biol. Chem. 279, 42359–42362.

PUBLICATION 6:

***Ensemble docking in homology models*.**

Novoa EM, Ribas de Pouplana L, Barril X and Orozco M.

J Chem Theory Comput 2010, 6 (8): 2547-2557.

JCTC Journal of Chemical Theory and Computation

Ensemble Docking from Homology Models

Eva Maria Novoa,[†] Lluis Ribas de Pouplana,[‡,§] Xavier Barril,[§,||] and Modesto Orozco*[,†,⊥]

Joint IRB-BSC Research Program in Computational Biology, Institute for Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Spain, Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain, Departament de Fisicoquímica, Facultat de Farmàcia, Avgda Diagonal sn, Barcelona 08028, Spain, and Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Spain

Received May 10, 2010

Abstract: We present here a systematic exploration of the quality of protein structures derived from homology modeling when used as templates for high-throughput docking. It is found that structures derived from homology modeling are often similar in quality for docking purposes than real crystal structures, even in cases where the template used to create the structural model shows only a moderate sequence identity with the protein of interest. We designed an "ensemble docking" approach based on the use of multiple homology models. The method provides results which are usually of better quality than those expected from single experimental X-ray structures. The use of this approach allows us to increase around five times the universe of use of high-throughput docking approaches for human proteins, by covering over 75% of known human therapeutic targets.

Introduction

New algorithms and computers are making possible the use of atomistic docking approaches in a high-throughput (HTD) regime, being possible to screen in silico libraries containing 10^5-10^6 compounds against a limited number of protein targets.[1−3] However, we cannot ignore that the requirement of computational efficiency implies the introduction of severe simplifications in both the description of molecular interactions and the coverage of the conformational space of ligands and proteins.[4−6] As a result, docking methods have problems in representing ligand-induced conformational changes in the protein, and in general the quality of docking algorithms decreases as the docked drug differs from that bound in the crystal structure.[7,8] However, despite all these limitations, the power of current docking algorithms is beyond all doubt, and many authors have demonstrated that their use largely enriches the possibility to find a good binder from a large library of decoys and that the proposed optimal poses are good starting points for lead-optimization processes.[9−12] It is not surprising, then, that virtual screening based on docking algorithms is a routine task in medicinal chemistry laboratories.[11,13,14]

The inputs of docking algorithms are ligand and protein structures, and the outputs are a series of "poses", i.e., possible configurations of the protein−ligand complex, which are then scored using an empirically refined function yielding to a small subset of preferred binding modes with the associated binding affinity.[15−17] Given the number of approximations done in a docking algorithm, the practical purpose of HTD is not the accurate ranking of potential

* Corresponding author phone: 0034-93-4037155; e-mail:modesto@mmb.pcb.ub.es.

[†] Joint IRB-BSC Research Program in Computational Biology, Institute for Research in Biomedicine.

[‡] Cell and Developmental Biology, Institute for Research in Biomedicine.

[§] Institució Catalana per la Recerca i Estudis Avançats.

[||] Facultat de Farmàcia.

[⊥] Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine.

10.1021/ct100246y
Published on Web 07/16/2010

binders, but the enrichment of true binders among the top-ranked compounds and the recovery of good leads for refinement.

The need to have a three-dimensional structure of the target protein strongly limits the use of docking algorithms, and despite the impressive advance of structural genomics, the number of proteins for which experimental structure is known represents only a small fraction of the total proteome. Thus, the 2010 version of the Protein Data Bank (PDB) contains around 60000 entries, but only 42.5% (25560) of them correspond to unique proteins from which only 15% (3935) are human.[18,19] In comparison, sequence analysis suggests that the total number of human proteins ranges between 20332—Swissprot[20]—and 93110—RefSeq[21]—probably twice or more if spliced forms are considered,[22] which means that the PDB covers only between 2 and 19% of the human proteins. The gap between structure and sequence becomes even larger if we consider proteins from virus, bacteria, or other pathogens for which little structural information exists.

Protein structure can be predicted by a variety of computational methods,[23] homology modeling (also named comparative modeling) being the most accurate one in cases where there is a clear sequence identity between the target protein and at least one template with known three-dimensional structure.[24,25] The quality of the structure derived from homology modeling roughly correlates with the sequence identity between the target protein and template proteins.[26] Thus, it is accepted that for sequence identities below 30% less than half of the residues have their C_α correctly placed.[27,28] The percentage of correctly placed residues increases to 85% for identities ranging from 30 to 50%, and most of the C_αs are well-positioned for sequence identities above 50%. Inside the high-quality range no direct correlation exists between the accuracy of the model and the sequence identity with the template, and evaluation of the expected quality of a model is still an unsolved problem.[29] In fact, the concept of "goodness" is not unique, since it depends on its planned use.[30] For example a model with an accuracy around 3.5 Å in backbone positioning may be good enough for understanding protein function or designing mutations but is expected to be of small utility for prediction of ligand binding.[26−31]

Different authors have tried to evaluate the quality of homology models for docking experiments. Thus, McGovern and Shoichet performed high-throughput docking on 10 target enzymes for which apo, holo, and homology model structures were available, finding that they were useful for enriching the screening, but not as powerful as the holo-crystal structure.[32] Diller and Li reported good enrichments (in some cases similar to those obtained with the crystal structure) when model structures of six kinases obtained for identities in the range of 30−50% were used to screen a large library.[33] Similar results were obtained by Oshiro et al.[34] in the study of two targets (CDK2 and factor VIIa), by Gilson's group with a set of five targets,[35] and by Ferrara and Jacoby in the analysis of insulin growth factor I receptor.[36] In a very recent paper Fan et al.[37] found good results when ensembles of homology models of several proteins were used to screen for ligands in the DUD database[38] using the DOCK computer program.[39] All these studies illustrate the power of homology models to guide docking experiments but also underline their limitations related to the lack of "a priori" evaluations on the quality of the model for docking purposes and on the problems of selecting a priori a structural model from the battery of solutions given by homology modeling routines (for discussion see ref 36).

The introduction of protein flexibility is the next step in docking, and there is a significant amount of work focused in this direction.[40,41] Among the different approaches suggested, "ensemble docking" (also known as multiple docking) is one of the most popular ones. It assumes that the effect of target flexibility in docking can be represented by using a Boltzmann ensemble of conformations for the protein instead of just a single rigid structure. Different methods for generating ensembles have been proposed, including molecular dynamics[42,43] (from a known experimental structure of the target), crystallographic (X-ray),[44−47] and spectroscopic (NMR).[48,49] All these approaches require experimental knowledge of protein structure and are then able to cover just a small fraction of proteome. In this contribution, following the pioneering work by Fan et al.,[37] we explore the possibility of using ensembles derived from homology (comparative) modeling. This approach is simple and fast and, if successful, would allow us to dramatically expand the range of applicability of ensemble docking approaches. We explored, with a wide range of metrics and for a large number of proteins, not only the ability of the approach to enrich in active ligands drug libraries but also the structural quality of the docking predictions, a crucial element in lead optimization procedures. We designed and tested a procedure to perform ensemble docking based on the combination of Modeller[50] and Glide,[51] finding that the results are in general of better quality than those expected when a single-crystal structure is used as a template in docking experiments.

Methods

Protein Data Sets. We defined two sets of proteins of our study: one for training and another for testing. The training set was defined considering proteins for which at least 30 crystal ligand-bound structures were available in PDB (with the same sequence or at most one single mutation). PDBs with point mutations were only used to build the set of active ligands but were not included in the set of docked proteins. This set of proteins includes thrombin (2cn0, 1ay6, 1bmm, 1tom, 1xm1), renin (2g24, 1bil, 1hrn, 1rne, 2g1r), cyclin-dependent kinase 2—CDK2—(1aq1, 1e1v, 1gz8, 1jsv, 3ddq), and protein tyrosine phosphatase 1B—PTP-1B—(2f71, 1c83, 1g7g, 1ony, 2h4g). The test set was created using less restrictive conditions in terms of the number of crystallized structures available—at least eight—and contained α-momorcharin (1mrg, 1aha, 1mom, 1f8q, 1ahb), trypsin (1tng, 1tnl, 1f0t, 1lqe, 2by5), p38 kinase (3hp2, 1w7h, 2baj, 3c5u, 3cg2), HIV retrotranscriptase (3jyt, 1dtq, 1rt1, 1s6p, 3dol), factor Xa (2vvc, 1ezq, 1fax, 1lpk, 1nfu), and heat shock protein 90—HSP90—(1yet, 1osf, 1uy6, 1yc4, 2ccs).

Homology Modeling. The derivation of model structures was performed using scripts designed for HTD production

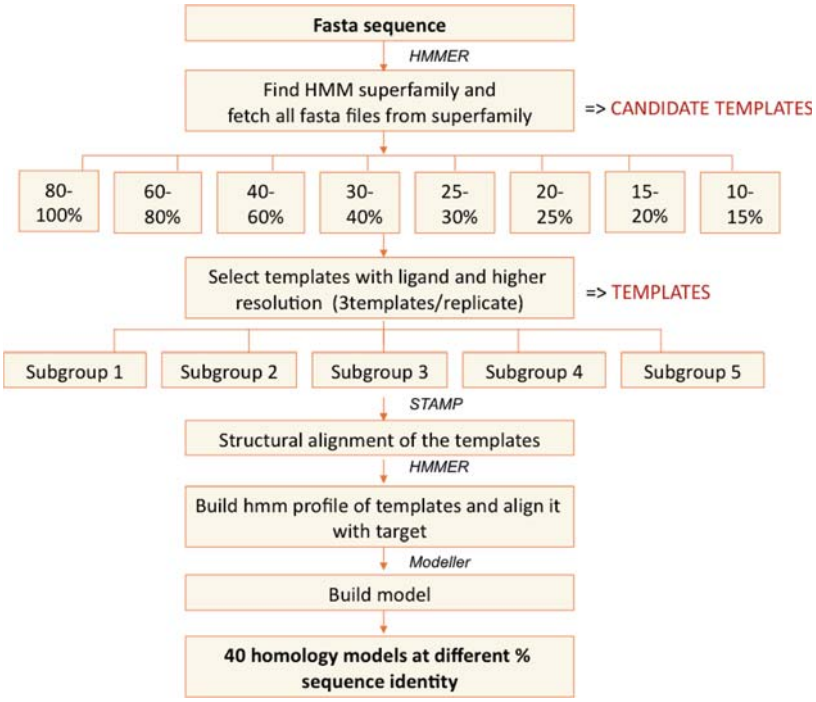


Figure 1. Comparative model building work flow. The process is automated such that a FASTA sequence is given as input, and a total of 40 homology models that range from 10 to 100% sequence identity are obtained. The software used at each step is detailed in Methods.

trying to mimic the standard expert procedure for homology modeling (see Figure 1). We are aware that by using automatic protocols homology modeling might be prone to errors, related mostly to misalignment problems, which can be easily corrected by manual refinement. However, to evaluate a pure HTD scenario, no human refinement was done here, which means that results presented here can be considered a lower limit of accuracy for the technique. Accordingly, the sequence of each target protein was extracted from the PDB, transformed to FASTA-format, and launched against the Pfam-A database[52] using HMMER[53] to assign the sequence to a superfamily. All the FASTA files for which there is a PDB corresponding to this same superfamily—all the candidate templates—were retrieved and aligned to the target sequence using ClustalW.[54] After this procedure each template was placed into different categories depending on its percentage sequence identity: 80−100, 60−80, 40−60, 30−40, 25−30, 20−25, 15−20, and 10−15%. For each sequence-identity category we selected 15 templates considering only proteins bound to ligand and solved at the highest resolution possible. Each set of 15 templates was divided into five subsets in order to build five different models per sequence-identity category. It is important to remark that the five models per sequence-identity category were built on the basis of different templates. Such templates were structurally aligned by STAMP,[55] creating then a profile using HMMER, which was introduced as a meta-template for alignment of the target sequence (see Figure 1). Finally, the 9v5 version of MODELLER[50] was used to create structural models using default options.

Ligand Selection. The active ligands to dock were downloaded from the PDB database (www.rcsb.org), by selecting all available X-ray ligands from PDB complexes for each of the proteins of the study. All the available ligands were subjected to similarity analysis using MOE[56] implementation of MACCS structural fingerprints[57] and distributed in 80% identity clusters. Only one compound per cluster was selected, which guarantees the diversity of the ligands, avoiding bias derived from the overrepresentation of the same scaffold. The set of known ligands was mixed with 1000 diverse "decoys" (molecules not described as binders for these proteins) which were selected from the most populated clusters obtained using Reynolds' algorithm at a similarity cutoff level of 60%[58] on a local database containing 1.7 million commercially available compounds—already filtered by drug-likeness criteria: Lipinski rules, Veber rules, and lack of reactive groups.[59−61] The percentage of active ligands ranged from 0.5 to 10%, depending on the protein.

Docking Procedure. Ligand screening and docking was performed using the Glide 5.0 program.[51] The extraprecision Glide docking (Glide XP) protocol was used for the training set, while the standard-precision (Glide SP) protocol was used for the test set, trying then to mimic a normal HTD procedure (in practice, we found very small differences between both scoring functions). Starting from the PDB structures, ligands were prepared using the LigPrep[62] facility in Schrödinger utility MAESTRO,[63] by generating low-energy ionization and tautomeric states within the range of pH 7.0 ± 2.0. All ligands were energy-minimized using the OPLS_2005 force field implemented in MAESTRO.[63] The setup of proteins was done with the Protein Preparation Wizard facility, which included hydrogen optimization, protonation, and geometry optimization using again the OPLS_2005 force field. The receptor grid defining the docking universe was built centered on the crystallographic ligand, which was then removed as any other nonprotein molecule.

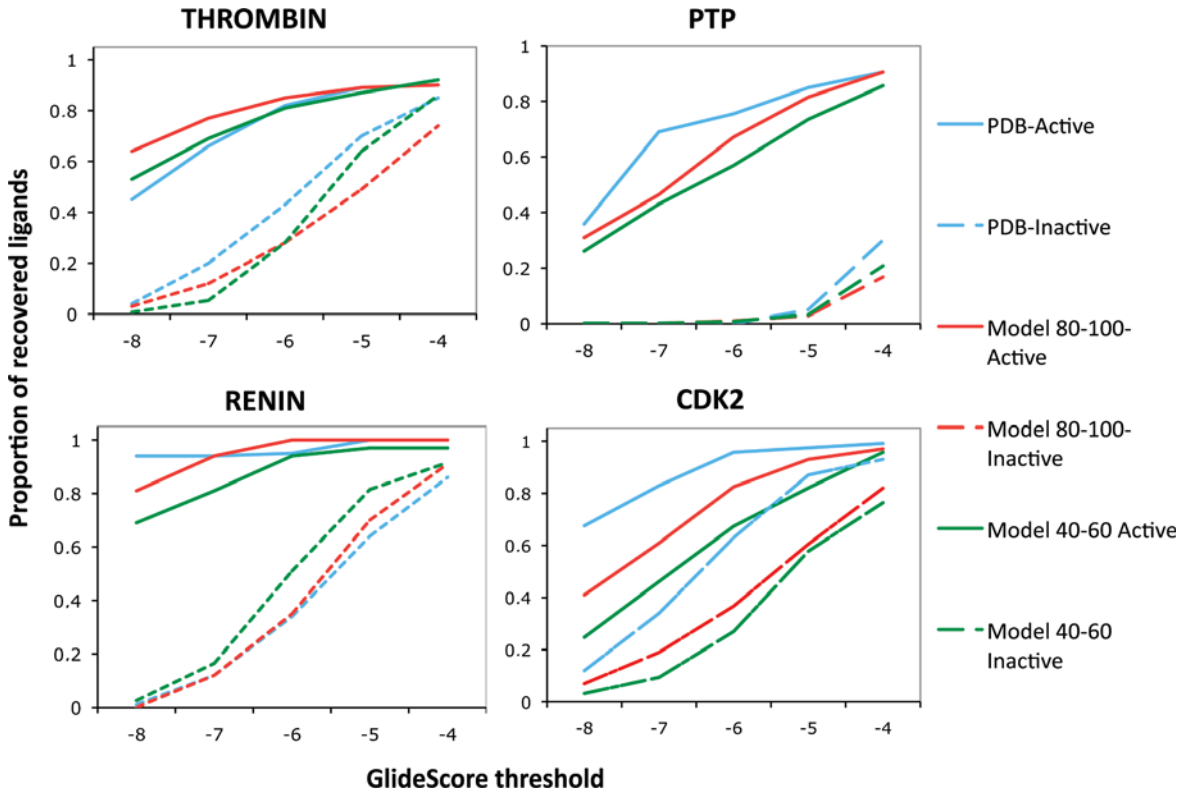


Figure 2. Recovery of active (nondashed lines) and inactive ligands (dashed lines) for each of the proteins of the training set. As can be seen in the plots, both active and inactive ligand recovery increases as the GS threshold decreases. The different colors correspond to different sequence identity ranges: blue (PDB), red (model 80−100), and green (model 40−60). The selected PDBs are all high-resolution holo conformations: 2cn0, 1aq1, 2g24, and 2f71, which correspond to thrombin, cdk2, renin, and PTP-1B, respectively.

Metrics for Preevaluation of Model Quality. The structural quality of the model was evaluated using both global and local parameters. The global quality indexes included global root-mean-square deviation (rmsd; model-reference PDB), global sequence identity, number of gaps in the alignment, and sequence coverage of the model. The local parameters were always referred to the binding site (defined as the set of residues with at least one atom at less than 5 Å from the crystal ligand) and included binding site rmsd, binding site sequence identity, and atom conservation in binding site structure. All rmsd measures were computed using the MMTSB tool set.[64]

Metrics for Evaluation of Success in Docking. The success of docking was measured by analyzing the following: (1) the ability of the models to predict the structure of the ligand−protein complex and (2) the applicability of the models for virtual screening purposes. The ability of models to predict the structure of the complex was assessed by (i) measuring the proportion of docked poses with rmsd below 2 Å from crystal structure using an SVL script in MOE, (ii) measuring the rmsd obtained when comparing the best-docked pose (rmsd-based selection) and the best-ranked pose (GlideScore-based selection) with the crystallographic ligand, and (iii) measuring the similarity between ligand−protein contact maps in models and crystal structures, which are determined by comparing the number of atoms that are conserved from those found at less than 5 Å from the docked ligand—compared to the original PDB where the docked ligand is found. Thus, for each docked ligand, a different ligand−protein map is built and compared to its corresponding PDB.

The utility of the models for virtual screening purposes was evaluated by assessing the performance of the homology models to discriminate between active compounds and decoys (inactive). A virtual screening run selects a list of molecules (n) from a given database of N entries, which includes both actives (true positive compounds, TP) and decoys (false positive compounds, FP). Actives (A) that have not been found by the screening method are false negatives (FN), and decoys that have not been selected are true negatives (TN). The optimum screening is that able to recover all true positives, without recovering any false positive.

Many different enrichment descriptors described in the literature have been considered in this work.[65,66] First we computed the *sensitivity* (true positive rate; TPR; see eq 1) and the *specificity* (true negative rate; TNR see eq 2) indexes. The first indicates the ability of the method to recover the real ligands, while the second informs on its ability to avoid decoys.

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (1)$$

$$\text{specificity} = \text{TNR} = \frac{\text{TN}}{(\text{FP} + \text{TN})} = 1 - \text{FPR} \quad (2)$$

where FPR stands for false positive rate.

The *accuracy* (Acc; eq 3) index was used to describe the percentage of molecules which have been correctly classified by the screening protocol, while *precision* (positive predictive value; PPV) was used to describe the proportion of true positives among the list of selected compounds given by the docking (eq 4).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{N} = \frac{A}{N}\mathrm{TPR} + \left(1 - \frac{A}{N}\right)\mathrm{TNR} \qquad (3)$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \qquad (4)$$

To assess the ability of the models to obtain true actives among the first ranked compounds (an extra requirement in HTD studies[67]), the *enrichment factor* (EF, eq 5) was used.

$$\mathrm{EF} = \frac{\mathrm{TP}/n}{A/N} \qquad (5)$$

ROC (receiver operating characteristic; true positive versus false positive rates) curves and the associated AUC curves (area under the ROC curve) have also been used to determine the discriminatory power of the virtual screening procedure. These metrics are especially powerful since they are not dependent on the ratio of actives to decoys of the database.[68−70]

Results and Discussion

Structural Quality of the Models. Modeller[50] provides good global models when using structural templates with sequence identities above 25% (Supporting Information Figures S1 and S2). The use of templates with sequence identities below such a threshold can yield wrong structures due mainly to alignment errors or to the presence of large unfolded regions. The atom conservation—i.e., the similarity between ligand−protein contact maps—at binding sites grows faster than global sequence identity, and for identities as small as 25−30% around 60−70% of the atoms at the experimental binding site are conserved in the model (Supporting Information Figure S3). The heavy-atoms rmsd between model and real binding sites are typically below 2 Å for sequence identities above 25% (Supporting Information Figure S4). Clearly, then, structural models created using homology modeling not only reproduce well global protein structure but also provide quite important details of the binding site. Whether or not the quality of these details is enough for drug docking studies will be the main subject of discussion in the remaining of our communication.

Docking Enrichment Using Single-Structure Homology Models. The second point to analyze was the quality of single homology models when used to recover specifically active ligands from a mixture of ligands and decoys. Within the Glide framework the number of hits recovered in a docking depends on the scoring (GS) threshold. For very restrictive GS values very few decoys (false positives) are recovered, but many real ligands might be lost. On the contrary, when very permissive GS values are used, all real ligands are recovered, but at the expense of increasing dramatically the number of incorrectly selected decoys. Results shown in Figure 2 demonstrate that using a single PDB structure as

Table 1. Training Set Enrichment Descriptors[a]

		PDB		model 80−100		model 60−80		model 40−60		model 30−40		model 25−30		model 20−25		model 15−20		model 10−15	
	GS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS
sensitivity	−8	0.56	0.81	0.49	0.77	0.43	0.74	0.38	0.67	0.32	0.60	0.26	0.55	0.22	0.37	0.14	0.26	0.11	0.20
	−7	0.76	0.92	0.66	0.86	0.62	0.88	0.57	0.83	0.47	0.75	0.45	0.79	0.39	0.66	0.27	0.50	0.20	0.38
	−6	0.87	0.97	0.82	0.94	0.78	0.92	0.74	0.91	0.63	0.88	0.62	0.91	0.54	0.80	0.41	0.70	0.30	0.51
specificity	−8	0.95	0.85	0.94	0.88	0.96	0.88	0.96	0.88	0.97	0.91	0.98	0.93	0.98	0.96	0.99	0.97	0.99	0.97
	−7	0.88	0.71	0.88	0.76	0.90	0.76	0.90	0.75	0.94	0.80	0.94	0.77	0.94	0.81	0.91	0.86	0.94	0.90
	−6	0.69	0.44	0.73	0.51	0.74	0.48	0.71	0.45	0.80	0.51	0.81	0.49	0.82	0.56	0.80	0.66	0.85	0.75
EF (1%)	–	24.40	26.49	21.72	26.09	22.3	24.58	19.01	21.49	17.95	23.08	15.19	19.66	17.67	21.29	11.35	13.62	6.84	9.74
accuracy	−8	0.93	0.85	0.91	0.87	0.92	0.87	0.92	0.87	0.93	0.90	0.93	0.91	0.93	0.92	0.93	0.92	0.93	0.92
	−7	0.88	0.72	0.87	0.77	0.88	0.77	0.88	0.76	0.91	0.80	0.91	0.78	0.90	0.80	0.87	0.83	0.88	0.86
	−6	0.70	0.48	0.74	0.53	0.74	0.51	0.71	0.48	0.80	0.54	0.80	0.52	0.80	0.57	0.77	0.66	0.81	0.73
PPV	−8	0.52	0.37	0.33	0.29	0.40	0.31	0.39	0.32	0.40	0.34	0.47	0.50	0.49	0.45	0.48	0.41	0.48	0.41
	−7	0.42	0.31	0.36	0.27	0.39	0.30	0.43	0.33	0.53	0.35	0.46	0.39	0.49	0.39	0.42	0.33	0.41	0.33
	−6	0.26	0.18	0.26	0.17	0.25	0.16	0.25	0.16	0.26	0.16	0.28	0.22	0.28	0.20	0.23	0.18	0.23	0.18
AUC	–	0.85	0.95	0.86	0.92	0.84	0.90	0.83	0.89	0.82	0.88	0.78	0.88	0.73	0.83	0.70	0.78	0.56	0.59

[a] Six different enrichment descriptors (sensitivity, specificity, EF for the top 1% ranked compounds, accuracy, PPV and AUC) have been computed for models and PDBs using both single docking (SG) and ensemble docking (ENS) approaches. For each of the cases, enrichment descriptors have been quantified taking different scoring (GS) thresholds: −8, −7, and −6, allowing us to see the difference between models and PDBs not only depending on the docking approach used−single or ensemble−but also depending on the chosen GS threshold. In the case of the EF (1%) and AUC enrichment descr iptors, no GS threshold has been used, given that these descriptors are GS-threshold independent.

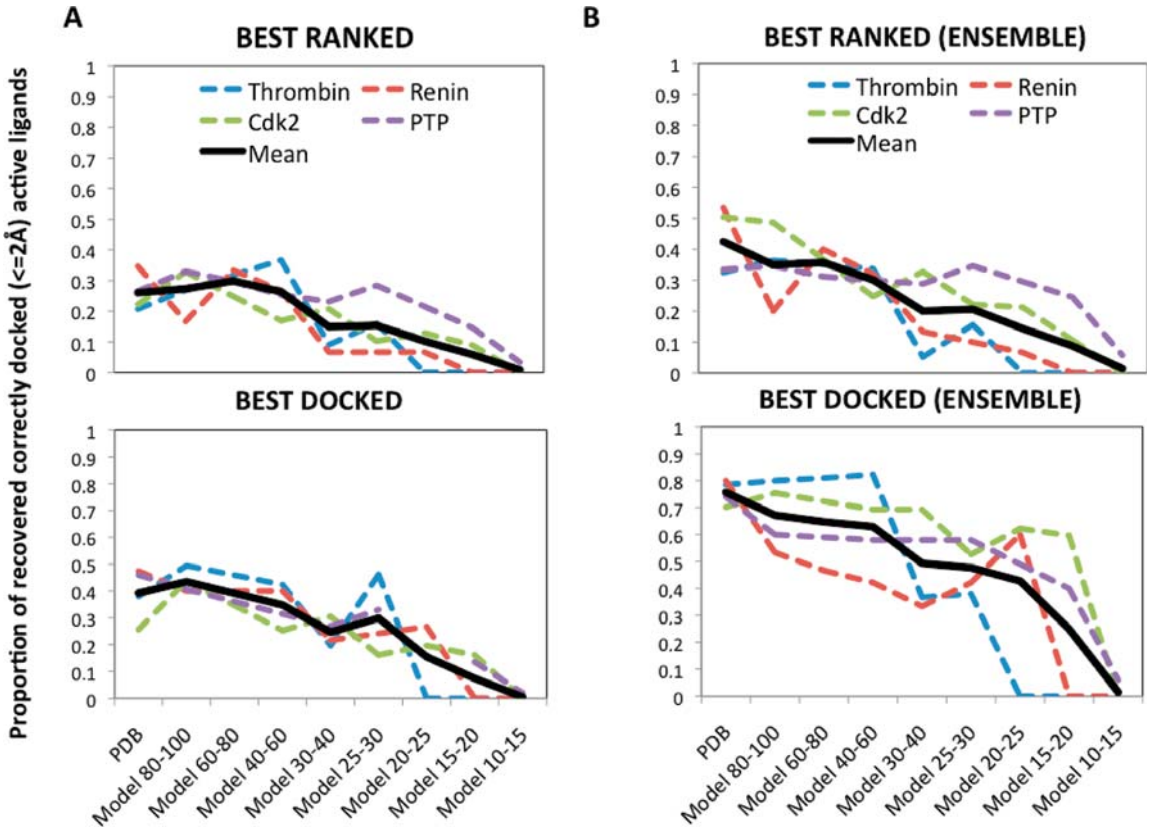


Figure 3. Recovery of correctly docked ligands versus sequence identity of the models. The recovery is defined as the fraction of correctly active docked ligands—less than 2 Å rmsd from the crystal structure—with respect to the total active docked and scored ligands. (A) In the upper plot, the *best-ranked* active ligand pose is chosen from all the proposed poses by using a score-based selection, whereas in the lower plot the *best-docked* ligand pose is chosen by using an rmsd-based selection. (B) Recovery of correctly docked ligands versus sequence identity when using an ensemble docking approach. Both score-based selection—i.e., best ranked—and rmsd-based selection—i.e., best docked—are shown. Each protein of the training set is labeled accordingly, and the mean value of the four training set proteins is shown in black.

template Glide is able to recover typically between 40 and 90% of the real ligands with a small number of false positives for a very strict scoring function threshold ($-GS = 8$). The ratio of true positive increases about 10 percentile points for $-GS = 7$ and 5−10 extra points for $-GS = 6$, keeping still an acceptable rate of recovery of false positives; for larger $-GS$ values the rate of false positives becomes unacceptable. In any case, the improvement with respect to random selection is very clear, demonstrating the performance of the Glide docking algorithm.

When homology models are used for docking, the performance of Glide is not lost (Figure 2 and Table 1), even in cases where the models are built using proteins with a modest level of homology as templates. It is especially encouraging that in some cases homology models outperform experimental structures for drug docking, a result already found by other authors[32,37] and which encourages the use of modeled protein structures for drug design experiments. The fact that homology models outperform X-ray structure for thrombin might appear surprising but is on the line of previous works with this protein[32] which demonstrated that probably the holo structure of thrombin is overspecialized for ligand binding, with problems arising in cross-docking experiments similar to those performed here. Homology models, less refined for a particular ligand binding mode, are then more successful.

Structural Quality of the Docking Poses Obtained Using Single-Structure Homology Models. The ability of the docking algorithm to capture specifically the maximum of active ligands is the major requirement for hit finding. However, to guide the optimization of the hit, there is an additional requirement: the drug needs to be correctly placed at the binding site. When using an experimental PDB structure as template, Glide is able to find poses that are very close (rmsd < 2 Å) to the bound conformation found in crystal in around 50% of cases, and in fact in more than 30% of cases the best scored poses (typically $-GS > 8$) match the experimental conformation (Figure 3A). Very interestingly, the global performance of the method does not change significantly when single homology models built from sequence identities above 40% are used, and even models built from templates with sequence identities around 25% can provide reasonable results. Again, it is remarkable that for some proteins homology models can provide more accurate binding mode predictions than the experimental structure—e.g., thrombin homology models recover on average 20% more correctly docked ligands compared to the crystallographic structures.

Ensemble Docking versus Single-Structure Docking. Proteins adapt their structure to the bound ligand, which explains the problems of docking methods to recognize active

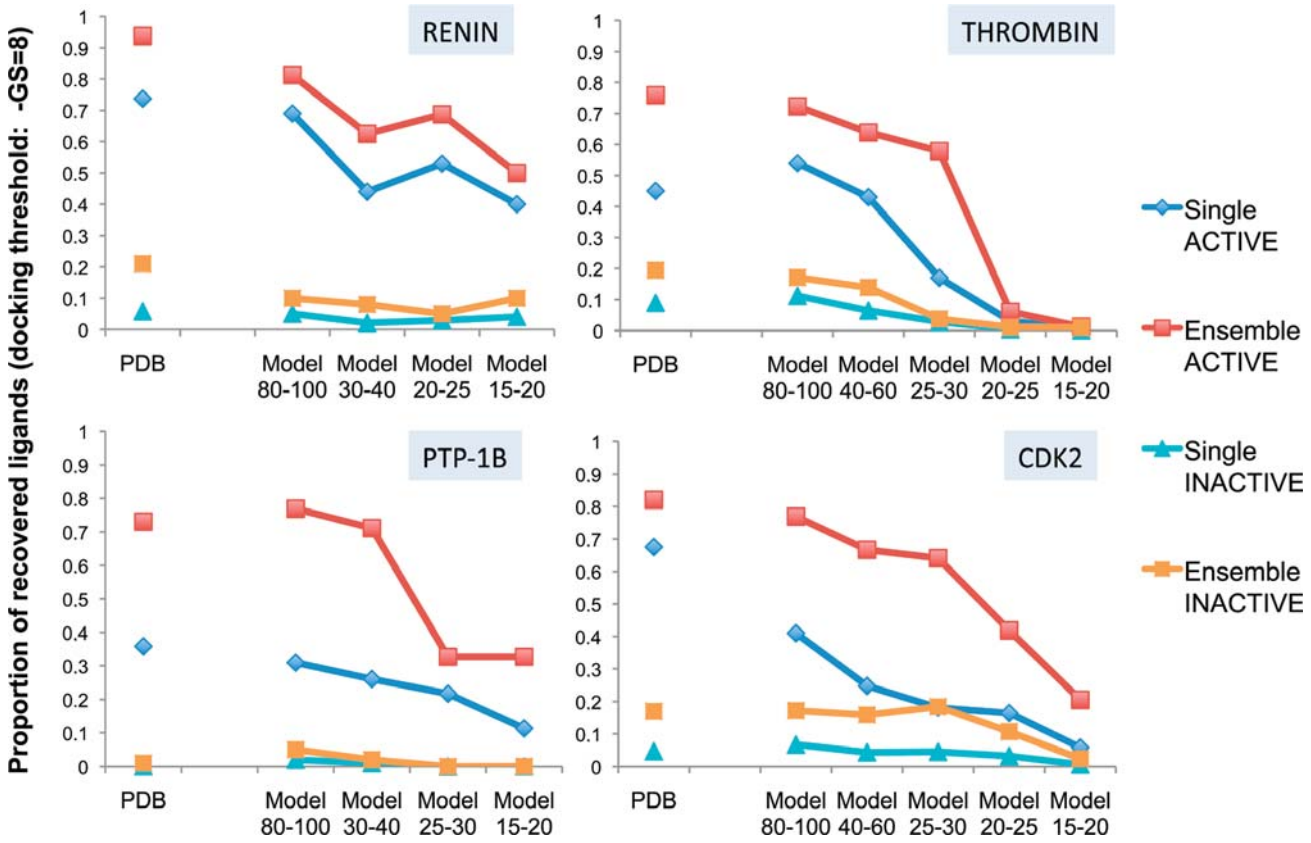


Figure 4. Ensemble docking versus single docking approach. The performance of both approaches is being compared in terms of recovered active ligands and decoys for the four proteins of the training set. The single docking approach performance is shown with blue and cyan lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the red and orange lines correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. In all cases, the difference between active and inactive recovery is higher when using ensemble docking. Results shown correspond to $-GS = 8$.

ligands when the protein structure has been solved in the presence of a very different compound. This problem is graphically illustrated in Figures 3B and Supporting Information Figure S5, which show the dispersion of results that can be obtained for a given protein when different high-resolution X-ray structures are used for docking. We can alleviate this problem by docking the drug against all the protein structures, selecting then as optimal docking mode that with the best scoring. This strategy is known as multiple docking or ensemble docking, which has been used and described in previous papers.[40,71−76] An ensemble of receptor conformations provides a structural degree of freedom that cannot be achieved with other flexible-receptor docking methods, such as induced-fit docking (IFD).[77] In our ensemble docking procedure, we have used five different structures, which is in accordance with the number of receptor structures used in previous papers.[71,73] This ensemble docking procedure (using at this point only experimental structures) leads to a clear improvement with respect to the average situation found when docking was done for single structures if a restrictive GS threshold is used (see Table 1). In fact, for strict threshold values the ensemble docking approach yields in most cases better results than those obtained by using the best "dockable" experimental structure, while the performance can decay for permissive thresholds due to the retrieval of false positives. It is also worth noting that the ensemble docking approach improves also the chances to recover good structural models for lead optimization procedures (compare Figure 3A with Figure 3B, and see Supporting Information Figure S6).

Ensemble Docking from Homology Models. The preceding analysis suggests that in general better docking results are obtained if all the experimental structural information of a protein is used as input for an ensemble docking procedure. The question is now, whether or not this situation is maintained for the less accurate ensembles generated by comparative modeling. Results in Table 1 demonstrate that the use of ensembles increases very significantly sensitivity (70−100%) with respect to single models, decreasing only slightly the specificity (around 6% for $-GS \geq 8$), leading to an overall improvement in the docking results. Thus, improvement made by the use of ensemble docking is more important in cases where the initial structures have lower AUCs, such as those in homology models.

Homology-modeling based ensemble docking coupled with good structural models and strict scoring thresholds outperforms in most cases single-structure docking performed using experimental structures (Table 1 and Figure 4). In fact, the quality of the ensemble docking results for accurate homology models (sequence identity above 80%) is indistinguishable from those obtained using experimental ensembles, and on average more than 80% of active ligands are recovered with a small percentage of recovered decoys

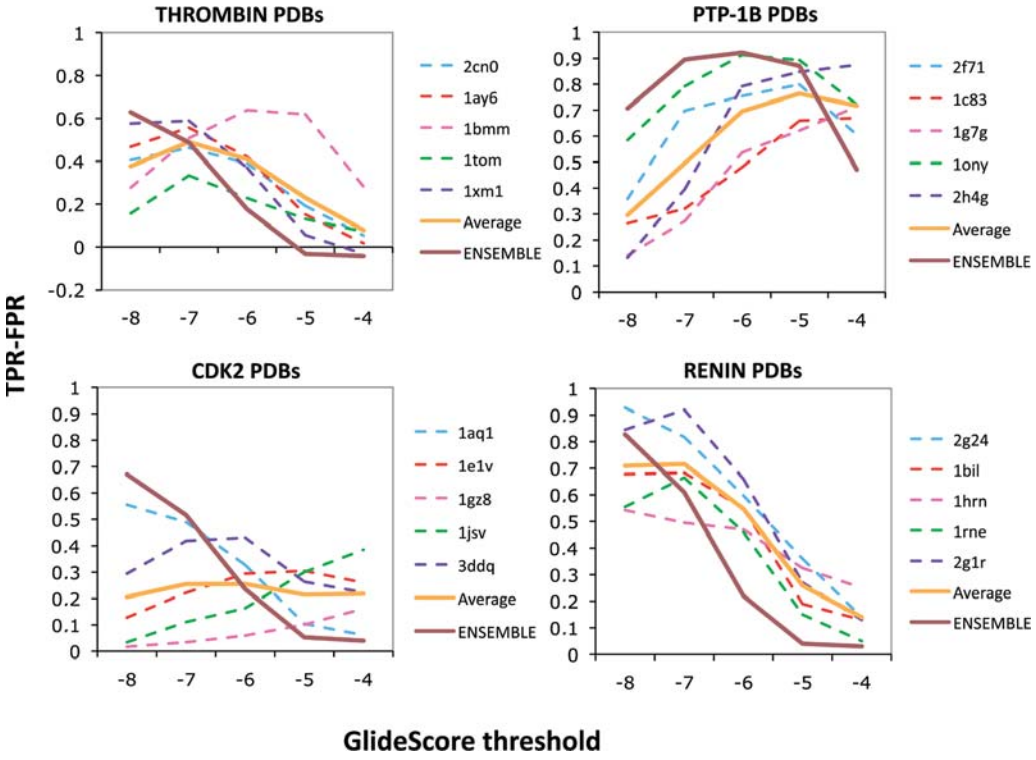


Figure 5. Performance of ensemble PDBs versus single PDBs. Performance (*y*-axis) is measured as the difference between the true positive rate (TPR) and the false positive rate (FPR). The dashed lines correspond to single PDBs, whereas the nondashed lines correspond to the average and ensemble of the PDBs, labeled with orange and brown, respectively. In all cases except for renin, the ensemble performs better than any single PDB at strict GS thresholds. However, in all proteins of the training set, the ensemble's performance decreases more rapidly—i.e., has higher slope—than any of the single PDBs.

(Figure 4) when homology ensembles are used. The ensemble docking protocol is very robust to the decrease in sequence identity, given that models with sequence identities in the range of 30−40% still provide good results. On the contrary, the protocol outlined here is very sensitive to the scoring threshold used, and less strict GS values increase excessively the recovery of false positives (Table 1 and Supporting Information Figures S7 and S8).

Finally, it is worth noting the large structural quality of the complexes obtained in homology-derived ensemble docking even when templates were not very homologous (Figure 5 and Supporting Information Figure S6). In fact, in most cases docking using ensembles of homology models outperform single experimental structure docking (Figure 4).

Validation of Results. Analysis on four proteins for which a large amount of structural data exist suggested (see above) that ensemble docking using homology models with sequence identity above 30−40% displayed a good ability to specifically recover active ligands when used as input for Glide calculations. Furthermore, the suggested complexes were in general reasonably close to the experimental binding modes, suggesting that the derived poses could be safely used in lead-optimization procedures. Analysis of the data suggests that the best balance between sensitivity and specificity is obtained when strict Glide scoring values were used to discriminate between active and decoy complexes. It is however unclear whether these results are general or specific for the proteins considered up to now. To analyze this point, we studied the ability of Glide on homology modeling ensembles of six unrelated proteins (see Methods).

Results summarized in Figure 6 demonstrate the good screening performance of the ensemble-docking approach performed with homology models also in the completely unrelated set of proteins used for validation. It is difficult to extend results of this small set of proteins to the entire proteome, but results suggest that docking performed using ensembles of homology models created using templates with sequence identity in the range of 30−40% leads to results which are of similar quality (according to most metrics) than those obtained using a single experimental structure. The screening performance of docking using ensembles of high-quality homology models is in general superior to that of docking using a single experimental structure and similar to docking procedures using an ensemble of experimental structures. Finally, Figure 7 confirms the geometrical quality of the complexes resulting from the homology model based docking procedure and accordingly its potential use in lead optimization processes. Our results indicate that the use of ensembles of homology models—built with Modeller—as input for Glide—using strict scoring thresholds—improves both the retrieval of active ligands from a chemical library and also the recovery of good structural complexes for lead optimization processes.

Gain in the Coverage of the Dockable Proteome. Results above suggest that an identity range of 30−40% is enough to build ensembles of homology models which can significantly enrich chemical libraries in active ligands. These results allow us to expand the applicability of structure-based drug design to a large universe of targets. Thus, while only 19% of (20332—Swissprot-annotated) human proteins can

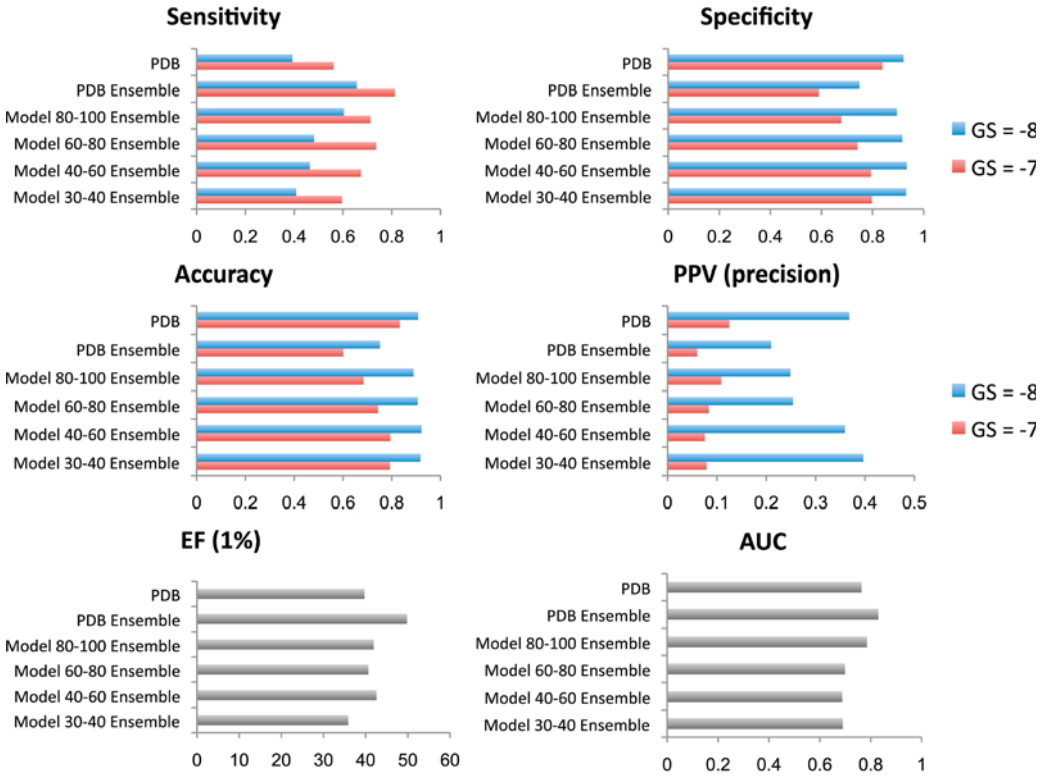


Figure 6. Enrichment descriptors for the test set. Only ensemble results for sequence identities>30% are shown for simplification. In the four top plots (sensitivity, specificity, accuracy, and PPV), enrichment descriptors are computed for −GS = 8 (blue) and −GS = 7 (red).

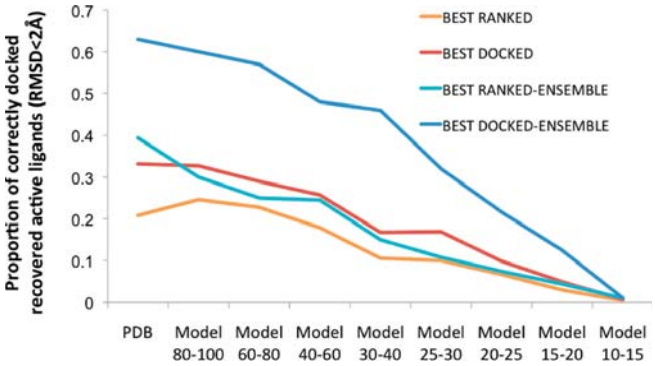


Figure 7. Recovery of correctly docked active ligands of the test set. A ligand is considered as correctly docked when its rmsd with the crystallographic ligand is below 2 Å. Both score-based selection—i.e., best ranked—and rmsd-based selection—i.e, best docked—are shown. Single docking averages are shown in red and orange, whereas ensemble docking averages are shown in blue and cyan.

be subjected to docking experiments using experimental structures, around 55% of (Swissprot) known human proteins can be studied by ensemble docking using homology models built from templates with 40% identity (Supporting Information Figure S9). Furthermore, less than 50% of human proteins of pharmacological interest have crystal structure available (DrugBank[78]). This coverage increases 41%—i.e., covering over 75% of the human drug targets—when using homology models up to 30% identity (see Supporting Information Figure S10).

With all the required cautions needed in the use of homology models for docking purposes (related mostly to the problems in finding good templates and in determining "a priori" the quality of the model), we suggest that the use of comparative models can enlarge dramatically the universe of applicability of small-molecule docking approaches, opening the possibility to analyze all potential cross-interactions of drug candidates, warning on potential adverse effects, opening new horizons both in the development of "dirty" drugs and in the determination of new indications for already annotated drugs.

ABBREVIATION. A, actives; Acc, accuracy; AUC, area under ROC curve; CDK2, cyclin-dependent kinase 2; EF, enrichment factor; ENS, ensemble; FN, false negatives; FP, false positives; FPR, false positive rate; GS, glide score; HIV, human immunodeficiency virus; HSP90, heat shock protein 90; HTD, high throughput docking; IFD, induced-fit docking; MACCS, molecular access system; PPV, positive predictive value; PTP-1B, protein tyrosine phosphatase 1B; rmsd, root-mean-square deviation; ROC, receiver operating characteristic; seq id, sequence identity; SVL, scientific vector language; TN, true negatives; TNR, true negative rate; TP, true positives; TPR, true positive rate.

Acknowledgment. This work has been supported by the Spanish Ministry of Science (Grants BIO2009-10964 and SAF2009-08811), the Instituto Nacional de Bioinformática, the Consolider E-science project, the ISCIII- COMBIOMED project, and the Fundación Marcelino Botin.

Supporting Information Available: Figures S1−S10 showing global rmsd between the homology models and the reference pdb, the correlation between the percentages of sequence identity and their sequence coverage, correlation

between the binding site sequence conservation and the percentage of sequence identity of the model, rmsd of the binding site, ROC curve plots for thrombin pdbs, similarity between ligand and protein contact maps, ensemble versus single docking approach, coverage of the human proteome, and structural coverage of human targets of pharmaceutical interest. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schneider, G.; Bohm, H. J. *Drug Discovery Today* **2002**, *7*, 64–70.

(2) Alvarez, J. C. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365–370.

(3) Lyne, P. D. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(4) Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. *Curr. Pharm. Des.* **2005**, *11*, 323–333.

(5) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. *J. Med. Chem.* **2008**, *51*, 6237–6255.

(6) Jacobson, M. P.; Sali, A. *Annual Reports in Medicinal Chemistry*; Academic Press: London, 2004; pp 259−276.

(7) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Proteins* **2006**, *65*, 15–26.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(9) Abagyan, R.; Totrov, M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375–382.

(10) Cavasotto, C. N.; Orry, A. J. *Curr. Top. Med. Chem.* **2007**, *7*, 1006–1014.

(11) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(12) Shoichet, B. K. *Nature* **2004**, *432*, 862–865.

(13) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851–5815.

(14) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(15) Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

(16) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409–443.

(17) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(19) O'Donovan, C.; Apweiler, R.; Bairoch, A. *Trends Biotechnol.* **2001**, *19*, 178–181.

(20) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45–48.

(21) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res.* **2005**, *33*, D501–D504.

(22) Clark, F.; Thanaraj, T. A. *Hum. Mol. Genet.* **2002**, *11*, 451–464.

(23) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342–348.

(24) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.

(25) Koehl, P.; Levitt, M. *Nat. Struct. Biol.* **1999**, *6*, 108–111.

(26) Marti-Renom, M. A.; Madhusudhan, M. S.; Fiser, A.; Rost, B.; Sali, A. *Structure* **2002**, *10*, 435–440.

(27) Eswar, N.; Sali, A. Comparative Modeling of Drug Target Proteins. In *Computer-Assisted Drug Design, Comprehensive Medicinal Chemistry II*; Taylor, J., Triggle, D., Mason, J. S., Eds.;Elsevier: Oxford, U.K., 2007; Vol. 4, pp 215−236.

(28) Sanchez, R.; Pieper, U.; Melo, F.; Eswar, N.; Marti-Renom, M. A.; Madhusudhan, M. S.; Mirkovic, N.; Sali, A. *Nat. Struct. Biol.* **2000**, *7*, 986–990.

(29) Eramian, D.; Eswar, N.; Shen, M. Y.; Sali, A. *Protein Sci.* **2008**, *17*, 1881–1893.

(30) Cavasotto, C. N.; Phatak, S. S. *Drug Discovery Today* **2009**, *14*, 676–683.

(31) Baker, D.; Sali, A. *Science* **2001**, *294*, 93–96.

(32) McGovern, S. L.; Shoichet, B. K. *J. Med. Chem.* **2003**, *46*, 2895–2907.

(33) Diller, D. J.; Li, R. *J. Med. Chem.* **2003**, *46*, 4638–4347.

(34) Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L.; Lanctot, J. K.; Putta, S.; Stanton, R.; Grootenhuis, P. D. *J. Med. Chem.* **2004**, *47*, 764–767.

(35) Kairys, V.; Fernandes, M. X.; Gilson, M. K. *J. Chem. Inf. Model.* **2006**, *46*, 365–379.

(36) Ferrara, P.; Jacoby, E. *J. Mol. Model.* **2007**, *13*, 897–905.

(37) Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. *J. Chem. Inf. Model.* **2009**, *49*, 2512–2527.

(38) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(39) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 380–397.

(40) Totrov, M.; Abagyan, R. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.

(41) B-Rao, C.; Subramanian, J.; Sharma, S. D. *Drug Discovery Today* **2009**, *14*, 394–400.

(42) Paulsen, L. P.; Anderson, A. C. *J. Chem. Inf. Model.* **2009**, *49*, 2813–2819.

(43) Armen, R. S.; Chen, J.; Brooks, C. L. *J. Chem. Theory Comput.* **2009**, *5*, 2909–2923.

(44) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.

(45) Huang, S. Y.; Zou, X. *Proteins* **2007**, *66*, 399–421.

(46) Rueda, M.; Bottegoni, G.; Abagyan, R. *J. Chem. Inf. Model.* **2009**, *50*, 186–193.

(47) Craig, I. R.; Essex, J. W.; Spiegel, K. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.

(48) Damm, K. L.; Carlson, H. A. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.

(49) Huang, S. Y.; Zou, X. *Protein Sci.* **2007**, *16*, 43–51.

(50) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

(51) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(52) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. *Nucleic Acids Res.* **2006**, *34*, D247–D251.

(53) Eddy, S. R. *Bioinformatics* **1998**, *14*, 755–763.

(54) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.

(55) Russell, R. B.; Barton, G. J. *Proteins* **1992**, *14*, 309–323.

(56) *Molecular Operating Environment (MOE)*, Version 2007 09; Chemical Computing Group: Montreal, Quebec, Canada, 2007.

(57) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.

(58) Reynolds, C. H.; Druker, R.; Pfahler, L. B. *J. Chem. Comput. Sci.* **1998**, *38*, 305–312.

(59) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug. Delivery Rev.* **2001**, *46*, 3–26.

(60) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(61) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.

(62) *LigPrep*, Version 2.2; Schrödinger: New York, NY, 2008.

(63) *Maestro*, Version 8.5; Schrödinger: New York, NY, 2008.

(64) Feig, M.; Karanicolas, J.; Brooks, C. L. 3rd. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.

(65) Langer, T.; Hoffmann, R. D., *Pharmacophores and Pharmacophore Searches*; Wiley-VCH: Weinheim, Germany, 2006.

(66) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

(67) Truchon, J. L.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(68) Nicholls, A. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

(69) Jain, A. N.; Nicholls, A. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(70) Witten, I. H.; Frank, E. Credibility: Evaluating what's been learned. In *Data mining—Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005; pp 161−176.

(71) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. *J. Mol. Biol.* **1997**, *266*, 424–440.

(72) Yoon, S.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.

(73) Cavasotto, C. N.; Abagyan, R. A. *J. Mol. Biol.* **2004**, *12*, 209–225.

(74) Duca, J. S.; Madison, V. S.; Voigt, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 659–668.

(75) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. *Eur. Biophys. J.*, in press.

(76) Barril, X.; Morley, S. D. *J. Med. Chem.* **2005**, *48*, 4432–4443.

(77) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.

(78) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

CT100246Y

ENSEMBLE DOCKING FROM HOMOLOGY MODELS.

Eva Maria Novoa[1], Lluis Ribas de Pouplana[2,3] Xavier Barril[3,4] and

*Modesto Orozco[1,5]**

SUPPORTING INFORMATION

Figure S1. Global RMSD between the homology models and the reference PDB. The proteins of the training set are labeled in cyan (thrombin), orange (cdk2), green (renin) and purple (PTP-1B).

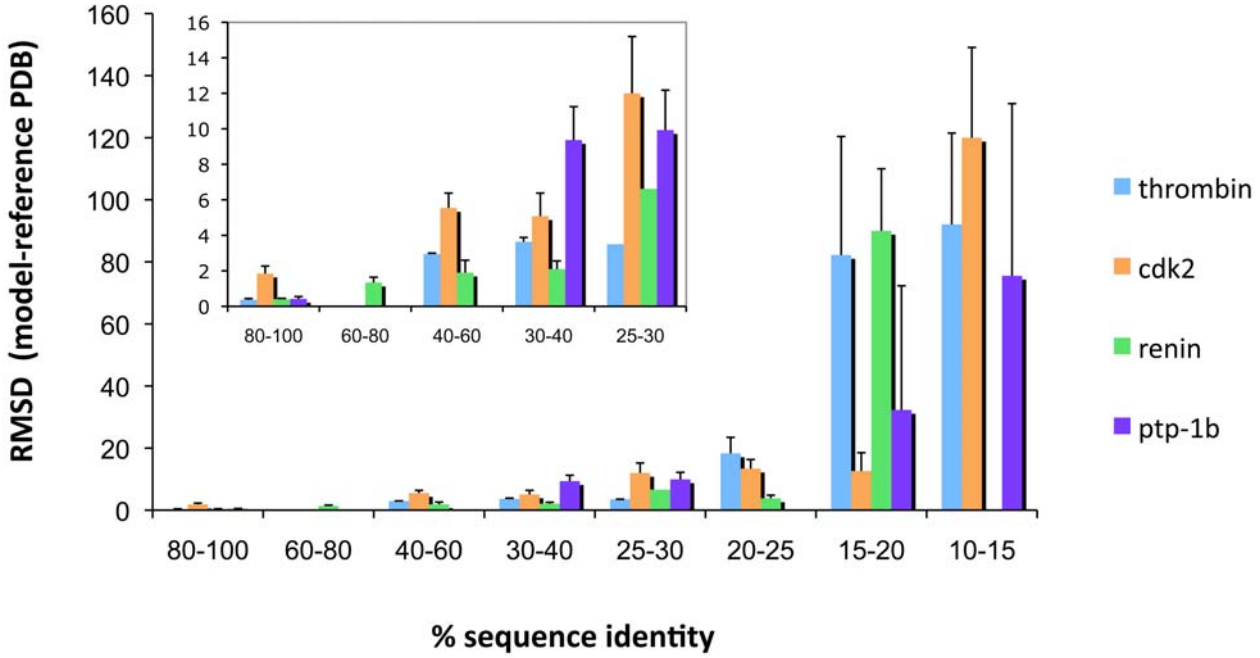


Figure S2. Correlation between percentage of sequence identity of the models and their percentage of sequence coverage. Homology models over 30% present very high sequence coverage values, in all cases over 80% sequence coverage. The proteins are labeled in cyan (cdk2), red (thrombin), green (renin) and purple (PTP-1B).

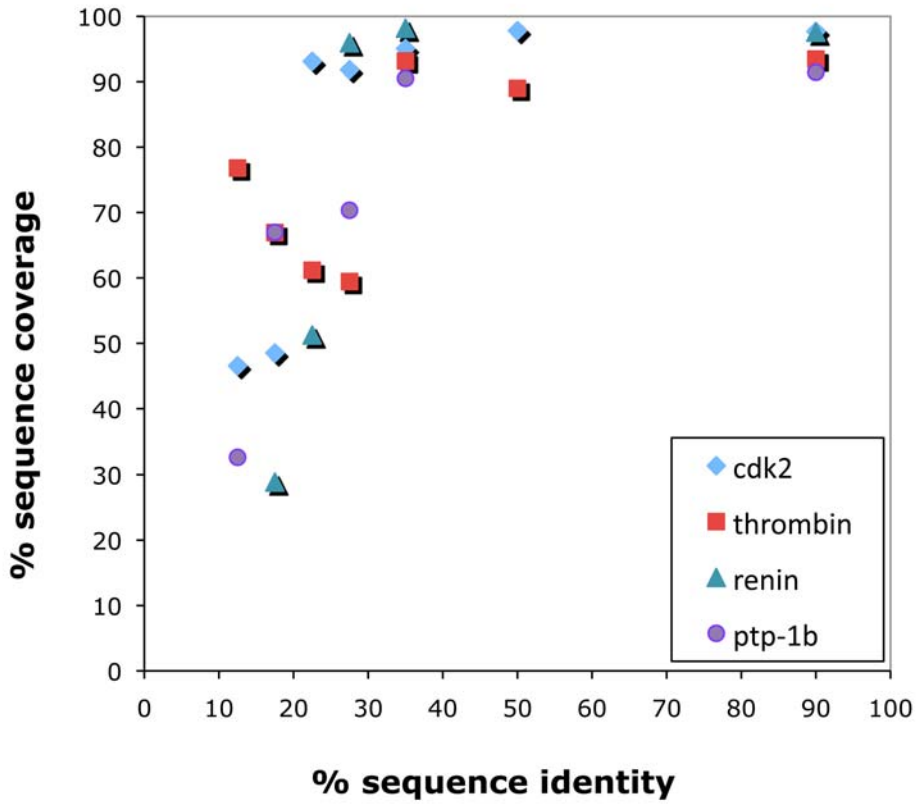


Figure S3. Correlation between the binding site sequence conservation and the percentage of sequence identity of the model. The percentage of conserved atoms in the binding site has been extracted from the structure of the model -not from its sequence alignment-. The binding site has been defined as the atoms of the protein that are found at least at less than 5Å from the reference ligand. To compute the percentage of conserved atoms in the binding site, each model was structurally superimposed to its reference PDB. Once superimposed, atoms of the binding site of the model were defined as those found at 5Å from the reference PDB ligand. The chosen reference ligand is originally found in an "holo" structure: 2cn0 (thrombin), 1aq1 (cdk2), 2g24 (renin), 2f71 (PTP-1B). Each protein is labeled accordingly.

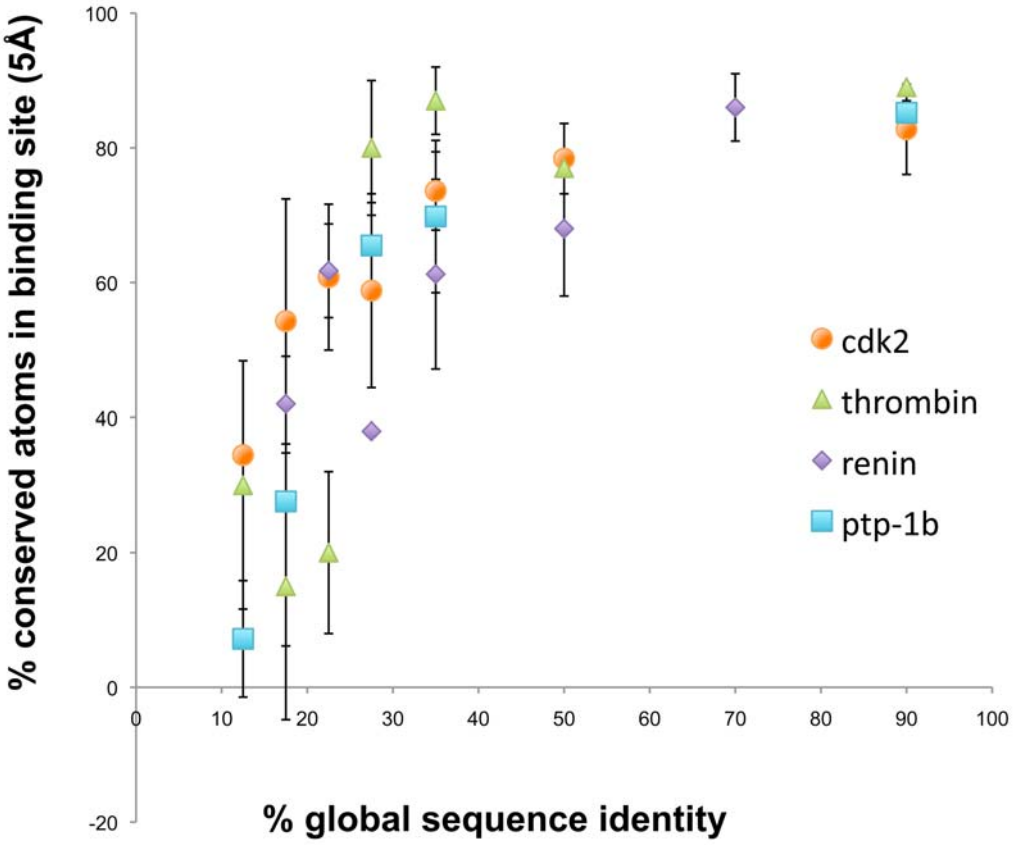


Figure S4. RMSD of the binding site for the comparative models of the 4 proteins of the training set. The binding site is defined as those residues that have at least one atom at less than 5Å from the ligand. Each protein is labeled accordingly.

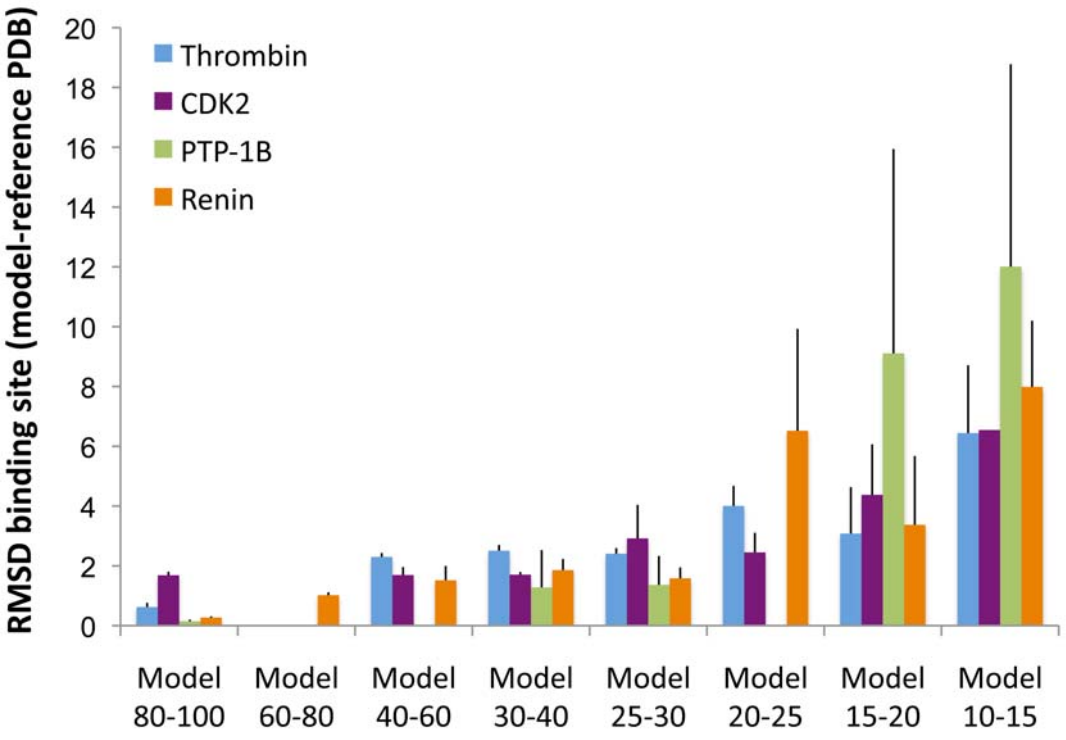


Figure S5. ROC curve plots for thrombin PDBs. Dashed lines correspond to ROC curves of single thrombin PDBs (1ay6, 1bmm, 1tom, 1xm1 and 2cn0), whereas non-dashed lines correspond to the average of all single PDBs (orange) and ensemble of these PDBs (brown). Clearly, the ensemble performs better than the best-performing PDB.

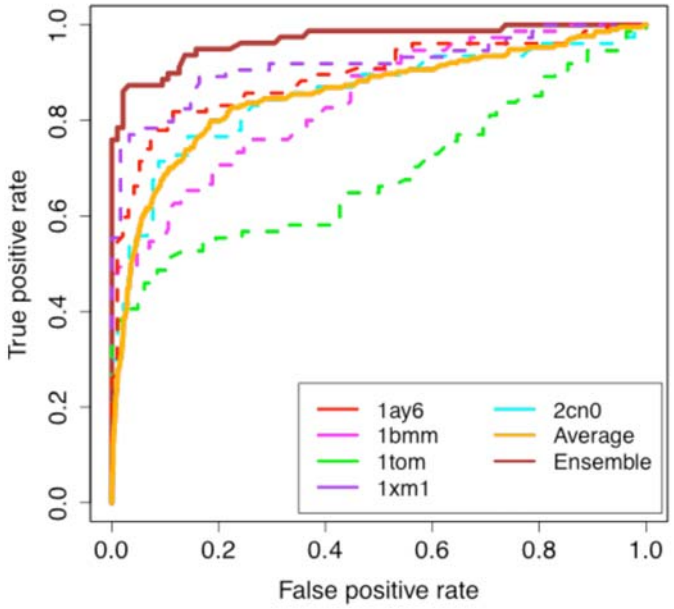


Figure S6. Similarity between ligand-protein contact maps. For each docked ligand, a contact map - i.e., a list of atoms or residues found at less than 5Å from the ligand- is compared to the contact map of the original PDB that included experimentally the ligand. Thus, these results indicate the maximum percentage of atoms that can be doing right contacts. For the homology models, only ensemble docking results are shown. The upper plot shows the percentage of common atoms of the contact maps, whereas the lower plot shows the percentage of common residues. Each of the proteins of the training set is labeled accordingly, and the mean value is shown in black.

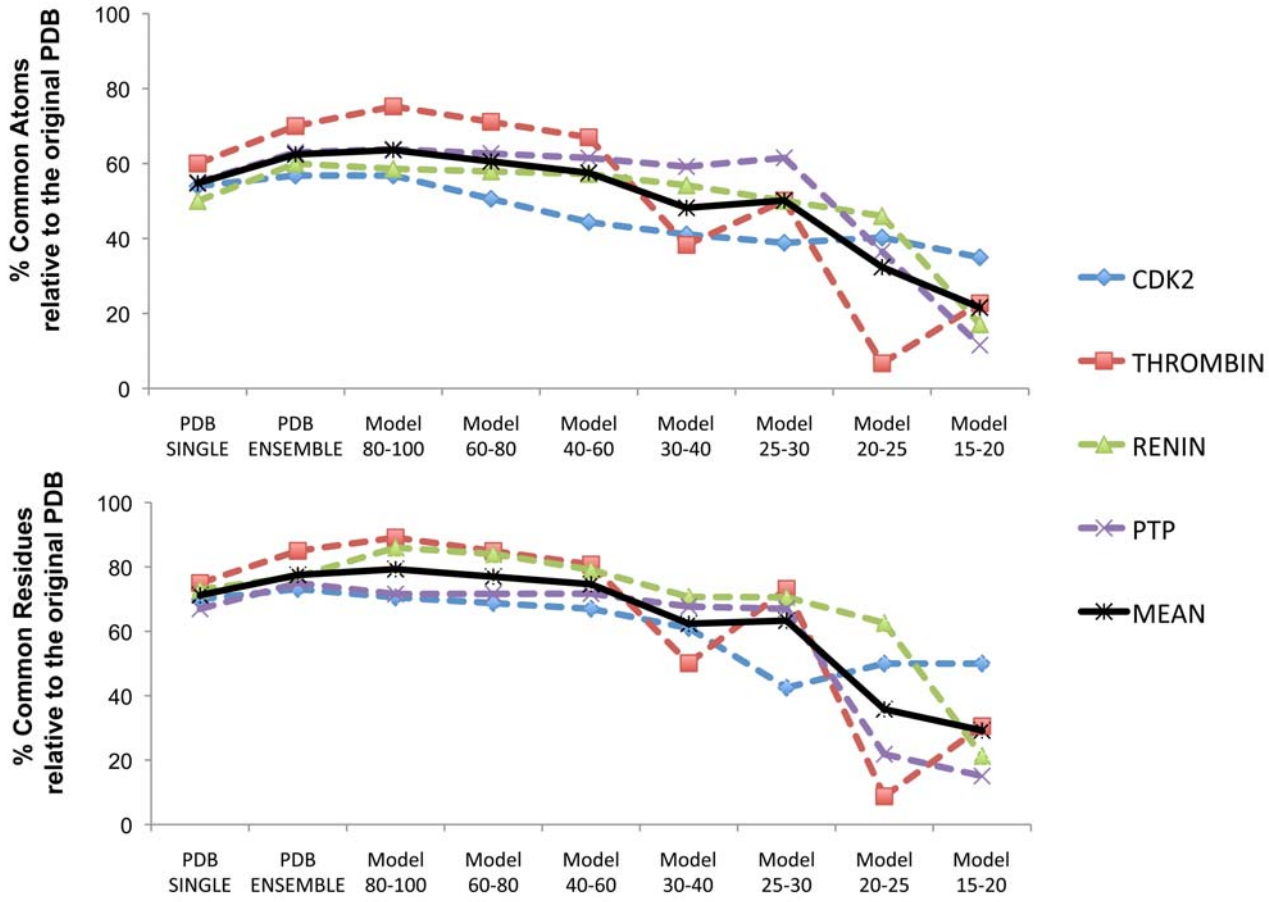


Figure S7. Ensemble docking approach versus single docking approach. The single docking approach performance is represented with blue and cyan lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the red and orange lines correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. Results shown correspond to -GS=7.

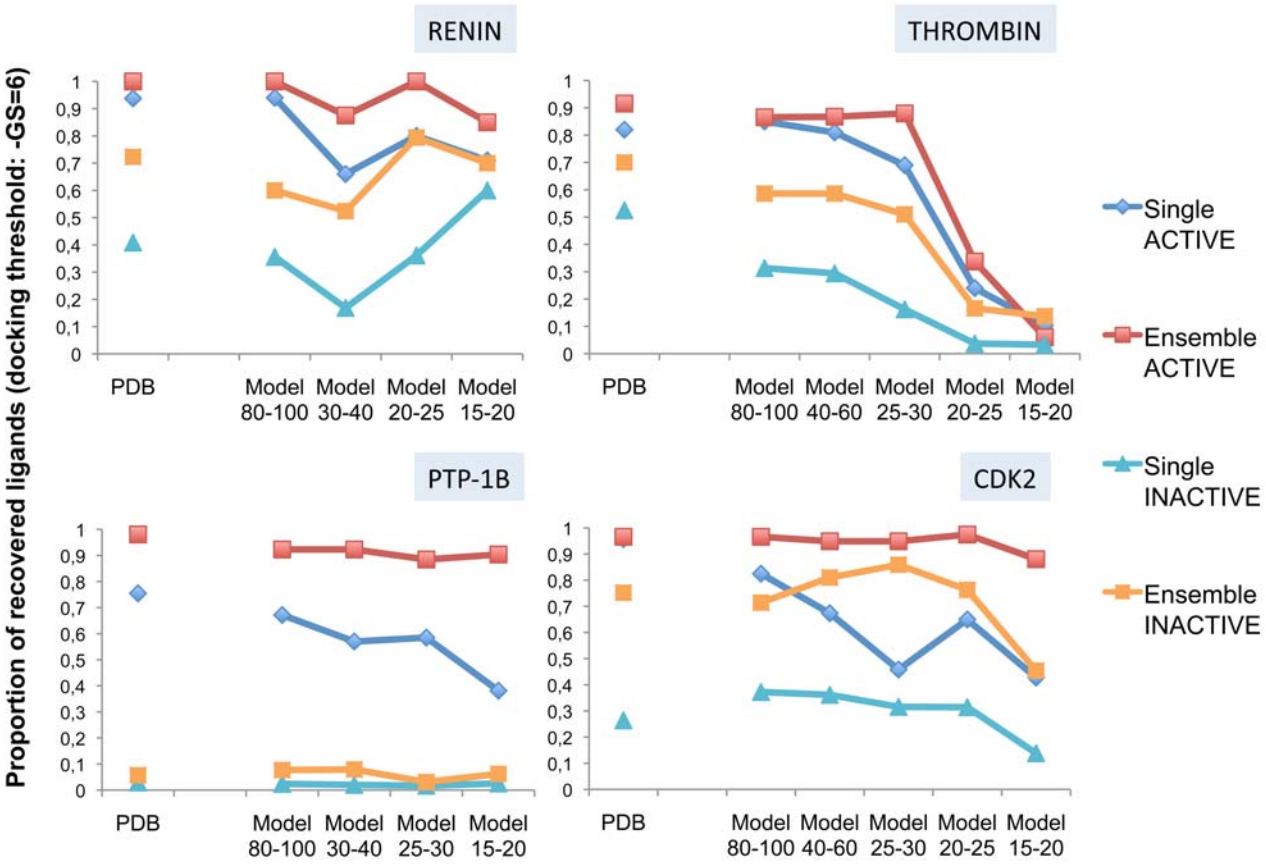


Figure S8. Ensemble docking approach versus single docking approach. The single docking approach performance is represented with blue and cyan lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the red and orange lines correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. Results shown correspond to -GS=6.

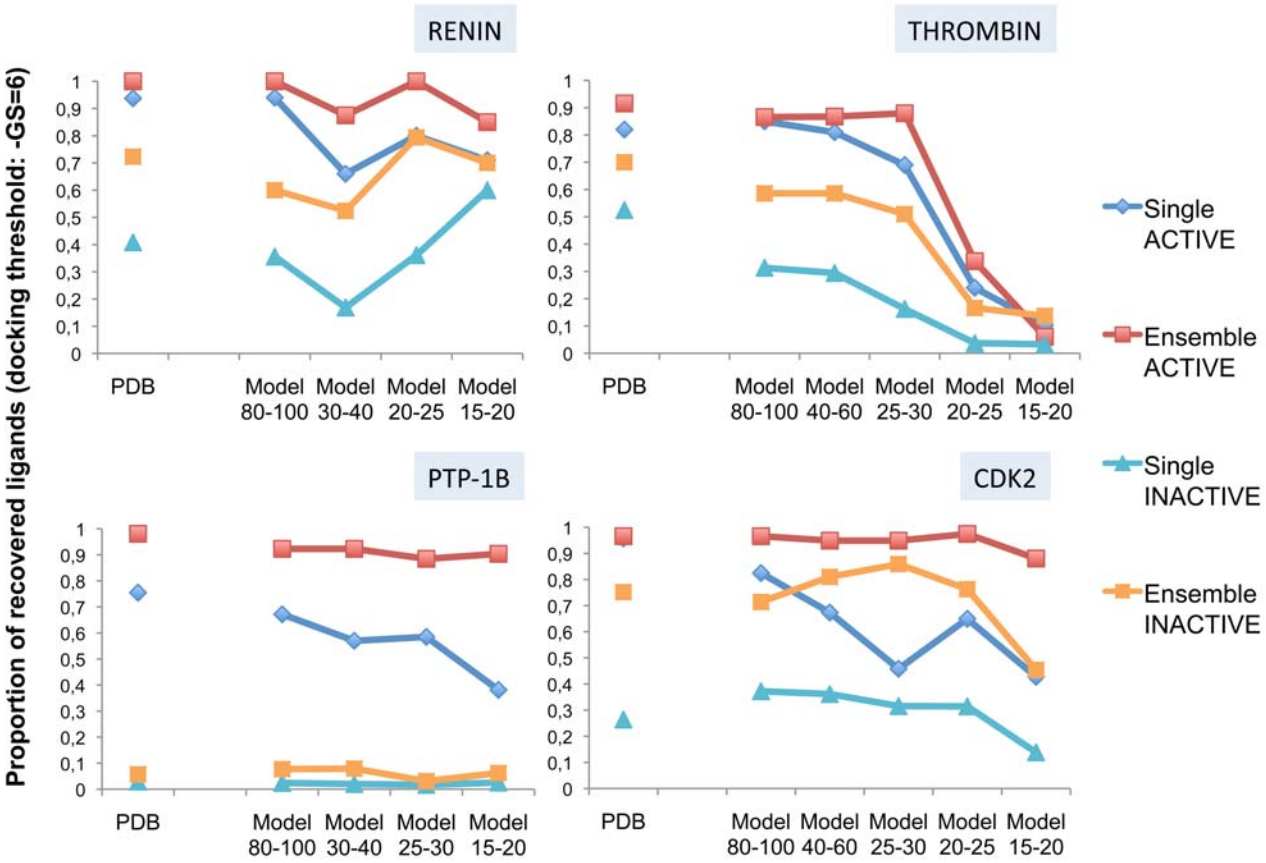


Figure S9. Coverage of the human proteome using comparative modeling. At a 50% sequence identity, less than half of the human proteome can be modeled. However, below 50%, the number of proteins that can be modeled rapidly increases, reaching a 78% of human proteins that can be modeled with a threshold of 30% sequence identity.

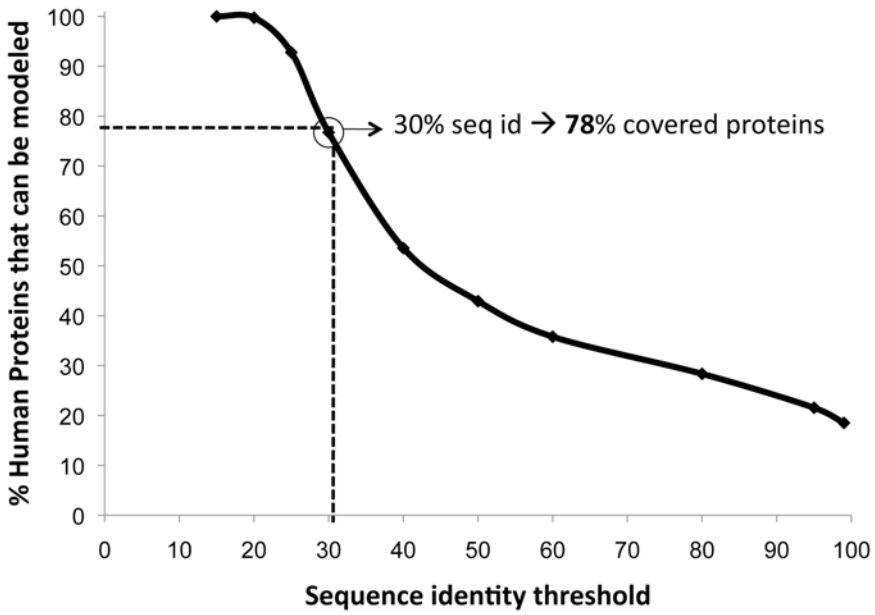
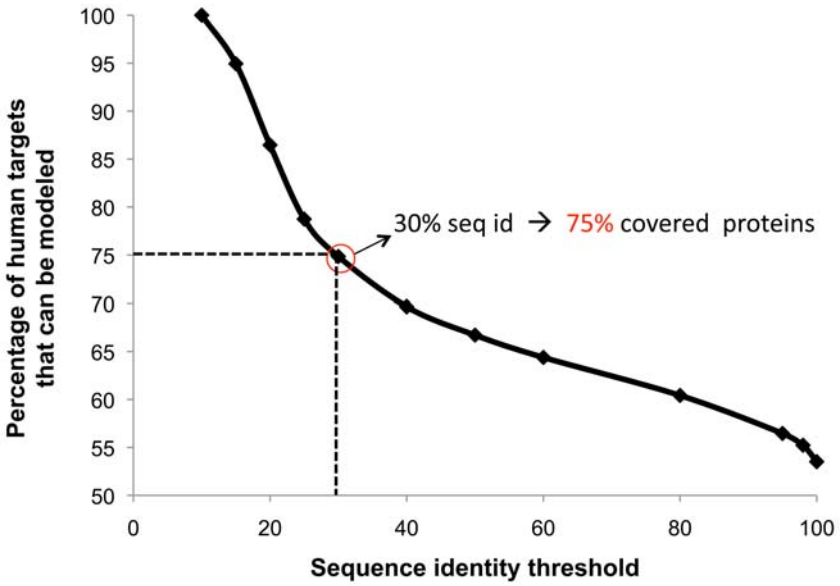


Figure S10. Structural coverage of human targets of pharmacological interest depending on the sequence identity threshold. A 30% sequence identity threshold - which still gives very good results when using the ensemble docking approach - allows us to cover 41% more human drug targets, obtaining a final coverage of 75% of the human drug targets. The list of targets has been obtained from the DrugBank database.



PUBLICATION 7:

Small molecule docking from theoretical structural models.

Novoa EM, Ribas de Pouplana L and Orozco M.

In: “Computational Modeling of Biological Systems: From Molecules to Pathways”. Ed Springer, New York (USA). Vol 4, pp 75-96.

Small Molecule Docking from Theoretical Structural Models

Eva Maria Novoa, Lluis Ribas de Pouplana, and Modesto Orozco

1 Docking as a Method for Drug Design

Structural approaches to rational drug design rely on the basic assumption that pharmacological activity requires, as necessary but not sufficient condition, the binding of a drug to one or several cellular targets, proteins in most cases. The traditional paradigm assumes that drugs that interact only with a single cellular target are specific and accordingly have little secondary effects, while promiscuous molecules are more likely to generate undesirable side effects. However, current examples indicate that often efficient drugs are able to interact with several biological targets [1]

E.M. Novoa
Joint IRB-BSC Research Program in Computational Biology, Barcelona Supercomputing Center and Institute for Research in Biomedicine, IRB, Josep Samitier 1–5, Barcelona 08028, Spain

Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain
e-mail: eva.novoa@irbbarcelona.org

L.R. de Pouplana
Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain

Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain
e-mail: lluis.ribas@irbbarcelona.org

M. Orozco (✉)
Joint IRB-BSC Research Program in Computational Biology, Barcelona Supercomputing Center and Institute for Research in Biomedicine, IRB, Josep Samitier 1–5, Barcelona 08028, Spain

Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain

Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine, Josep Samitier 1–5, Barcelona 08028, Spain
e-mail: modesto@mmb.pcb.ub.es

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_4,

and in fact some dirty drugs,[1] such as chlorpromazine, dextromethorphan, and ibogaine exhibit desired pharmacological properties [2]. These considerations highlight the tremendous difficulty of designing small molecules that both have satisfactory ADME properties and the ability of interacting with a limited set of target proteins with a high affinity, avoiding at the same time undesirable interactions with other proteins. In this complex and challenging scenario, computer simulations emerge as the basic tool to guide medicinal chemists during the drug discovery process.

Since early works in the 1980s, molecular docking has arised as a leading simulation technique to facilitate the drug design. The traditional paradigm of docking, known as rigid-body docking approach, assumes implicitly the Fisher's lock-and-key model [3], and considers that the ligand-induced structural changes of the protein are negligible [4]. However, drugs generally exhibit a certain degree of flexibility, and the bioactive conformation might not be the most stable conformation in solution [5, 6]. This fact leads to the need of considering drug flexibility for a successful docking simulation. Furthermore, analysis of the Protein Data Bank [7] reveals that ligand binding can introduce non-negligible changes in protein structure which often affect the binding site, raising tremendous difficulties for docking techniques, especially in cases where structural changes are not only binding-specific, but also drug-specific [8]. A second limitation in docking experiments arises from the evaluation of the ligand-binding free energy. Free-energy simulation techniques are expensive calculations that remain impractical for the evaluation of large numbers of ligands [9]. Current docking strategies are based on the combination of very fast functions, which intend to predict binding poses and rank them by means of a more complex equation (the "scoring function"), which has been parameterized to reproduce experimental binding data of protein–drug complexes [10]. However, scoring functions implemented in docking programs make various assumptions and simplifications, and do not fully account for all phenomena that determine molecular recognition.

Despite all the challenges, the major practical limitation for docking procedures does not emerge from technical uncertainties in the evaluation or scoring of docking poses, but comes from the lack of experimentally solved protein structures. Indeed, despite the massive effort focused in the experimental resolution of protein structures, 2010 version of the PDB contains less than 4,000 unique human proteins, while RefSeq [11] suggests the existence of nearly 100,000 human proteins, twice or more if splicing variants are considered. Therefore, the current version of PDB is covering only around 4% of the known human proteome [12]. This sequence-structure gap becomes even larger if we consider proteins from virus, bacteria, or other pathogens for which less amount of structural information exists.

The evaluation of the potential interactions of drugs with multiple targets is severely limited if the analysis relies exclusively on experimentally solved structures. Fortunately, this limitation can be partially solved with the use of predicted models of proteins as templates for docking (Fig. 1). In this chapter, we very briefly

[1] Drugs that bind to several molecular targets or receptors, and therefore tend to have a wide range of effects and possibly negative side effects.

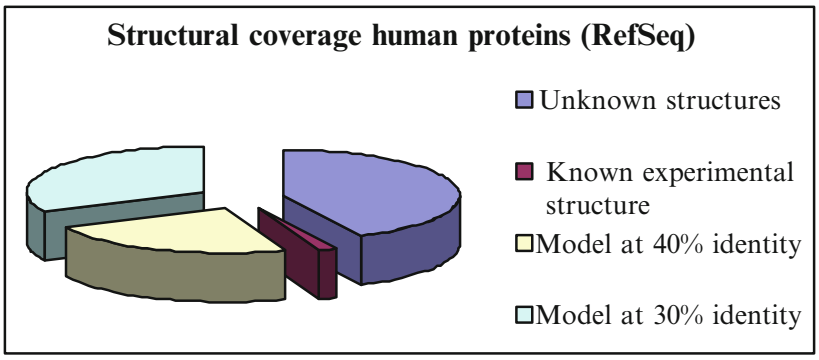


Fig. 1 Structural coverage of human proteins according to RefSeq without including splicing variants

review the state-of-the-art of docking procedures, making special emphasis on the potential use of ensembles of structural protein models derived from homology modeling in high-throughput docking experiments.

2 Docking Algorithms

There is a plethora of docking algorithms and strategies that have been implemented in a large variety of computer programs, some local and used by a restricted community, and others commercially available that have a wide user community. It is out of our scope to review all of them here, and we just outline the basic formalism behind the most popular ones. The reader is addressed to excellent reviews to gain a more complete view on current algorithms [10, 13–16].

In principle, all docking algorithms follow a stepwise procedure: (1) several estimates of the ligand–protein complex (binding poses) are proposed, and (2) these poses are then ranked using a scoring function and offered to the user, who typically focuses his/her attention to the best scored ones. Given that scoring functions are fitted against experimental binding data, scoring values have "free energy of binding" units. Therefore, they can be used to differentiate between good and bad drug candidates and even to have an estimate of the binding free energy of the drug.

The differences between the different docking programs rely on (1) the method used to explore the drug-binding landscape, (2) the method used to introduce flexibility, and (3) the nature and the parameterization of the scoring function. For example, DOCK [17], one of the first widely used docking programs, performs a geometrically based docking of the ligands based on isomorphic subgraph matching algorithms [18], which is later refined by considering the chemical nature of the ligand and the binding site. Different scoring functions—mostly in the AMBER [19] force-field—are used during the different stages of the fitting and ranking process, including complex physical functions calling to atomistic force-field calculations coupled to Generalized Born or Poisson–Boltzmann calculations. The popular AUTODOCK program [20] offers a variety of optimizers including Monte Carlo simulated annealing and different genetic algorithms using smoothed potential

energy terms precomputed in a regular grid.[2] Scoring is performed considering ligand-entropic terms and desolvation contributions in addition to ligand–protein interaction terms. GOLD [21], another very popular program, uses a sampling protocol similar to the genetic algorithm implemented in AUTODOCK and a very wide range of well-validated scoring functions, which include specific corrections such as those for metal ions and covalent interactions [22]. This program includes also specific scoring functions for kinases and offers the possibility to incorporate user-refined scoring functions. The program FLExX [23], which has also an excellent record of success, uses a geometry-fitting algorithm derived from computer vision engineering, where drugs grow in optimum orientations and conformations at the binding site from an original seed fragment. The program permits the introduction of knowledge-based pharmacological restraints and the incorporation of essential water molecules and crucial metal ions in the binding site. Scoring is based on a simple physical scoring function based on OPLS [24] force-field parameters. ICM [25], a powerful program to fit small ligands to proteins, uses a smoothed atomistic energy function coupled with a Monte Carlo algorithm in internal coordinates to sample the drug–protein binding space. Its scoring function contains the usual contributions plus two desolvation correction terms. GLIDE [26], a widely used docking program in the pharmaceutical industry, uses a "funnel strategy" where each pose passes a series of hierarchical filters that evaluate the ligand–receptor interactions, including spatial fit, complementarity of interactions using a grid-based method, and finally an evaluation and minimization using OPLS-AA nonbonded ligand–receptor interaction energy. GLIDE incorporates a variety of scoring functions with increasing computational complexity. MedusaDock [27], a recently developed software, is a docking method which models both ligand and receptor flexibility in a rapid manner by using sets of discrete rotamers, obtaining quite good results with targets which are known to be very flexible.

In addition to those implemented in standard programs, many other scoring functions have been developed (for a review see [28]), using experimentally calibrated master equations similar to that in (1).

$$\Delta G_{\text{binding}} = \alpha E_{\text{ele}} + \beta E_{\text{vW}} + \chi E_{\text{HBond}} + \delta G_{\text{desolv}} + \varepsilon S_{\text{lig}} + \phi E_{\text{dist}}^{\text{lig}} + \varphi G_{\text{others}}, \quad (1)$$

where E_{ele} and E_{vW} stand for usual electrostatic and van der Waals terms—typically smoothed to avoid nuclei discontinuities. Hydrogen bonds contribution is sometimes explicitly included in E_{Hbond}, while in others it is captured by E_{ele} and E_{vW}. The ligand and protein desolvation contribution (typically computed from occluded surface/volumes) are included in G_{desolv}, the loss of ligand entropy upon binding is introduced in S_{lig} (typically roughly approximated by counting the number of rotable bonds in the ligand), and the constrained energy is captured by E_{lig}. Other additional terms can be included, such as corrections for covalent interactions,

[2]Representation of the receptor energetic contributions (mainly electrostatic and van der Waals) to be read during the ligand scoring.

cation–π contacts, special metal–ligand interactions, presence of buried waters in the binding cavity, and many others. All these different terms are weighted using parameters that are fitted against empirical data. As discussed above, different programs offer the user the possibility of using family specific scoring functions and to incorporate his/her own scoring functions. However, the large number of available scoring functions has generated an obvious confusion in the users community and has driven to the popularization of strategies based on consensus or meta-scoring functions. Future work needs to be done by the community to order this explosion of different scoring strategies.

Flexibility is treated at different levels by various programs. Ligands with potential drug-like properties tend to be small and moderately flexible, which facilitates the determination of the optimum docking conformation by different methods such as energy minimization, Monte Carlo, genetic algorithms, molecular dynamics, and many others. The complexity here arises from the need to determine which is the optimum geometry in solution [6]. As noted above, the incorporation of the protein flexibility is much more difficult due to the large number of protein degrees of freedom, and none "final" algorithm has been yet developed. Many programs allow the user to refine a reduced number of residues in the protein—generally limited to side chains—by using rotamer libraries [29], Monte Carlo [30], or restrained molecular dynamics [31]. Nevertheless, one of the most popular strategies consists in the "ensemble" docking approach, which assumes that the effect of target flexibility in docking can be represented by using a Boltzmann ensemble of conformations for the protein instead of just a single rigid structure. Different methods for generating ensembles have been proposed, including molecular dynamics from a known experimental structure of the target [32, 33], crystallographic (X-ray) [34–37], and spectroscopic (NMR) [38, 39]-derived structures.

A common feature in most descriptions of new docking methods is the claim that it is more accurate than the competitors. In our experience, the performance of docking algorithms changes in each version and depends quite significantly on the nature of the problem and the skills of the modeler running the project, factors that hinder the validity of the conclusions derived from blind test experiments [40]. An estimate of the market share taken by the different docking algorithms is also difficult to determine, particularly in a scenario of site-licenses, cost-related decisions in the selection of docking engines and where publication is not often a priority. However, a simple analysis of the literature (ISI CITATION MANAGER) in 2009 reveals that the market is quite equally divided among different codes (see Fig. 2).

3 Scenario for Docking Use

The literature is full of examples of use of docking algorithms in drug design procedures, and the documentation accompanying the different computer programs illustrates many examples where docking has been crucial to derive significant results. Even though most docking studies are done inside pharmaceutical industries

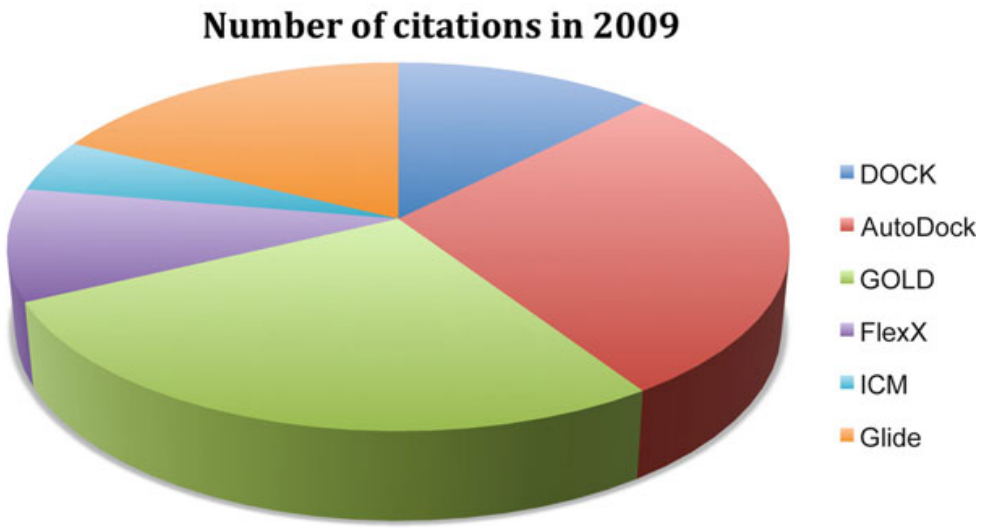


Fig. 2 Number of citations in scientific literature of commonly used docking algorithms in 2009

and are never published, analysis of the literature reveals that the word "docking" has been used in the title or abstract in 1,565 publications during 2009.

Docking can be done in quite different scenarios, where objectives and success criteria can be quite different:

1. Derivation of structural binding mode for a known binder
2. Determination of primary or secondary targets for a drug
3. Virtual high-throughput screening (vHTS)

The derivation of a structural binding mode for a small molecule is probably the most traditional use of docking algorithms. Within this paradigm, the process starts after high-throughput experimental studies (or alternative methods) that detect one or several small molecules which display activity against a given target. However, there are many factors that determine whether these "hits" can become "leads" or can be modified to improve their properties. Such a lead optimization process requires a quite detailed knowledge of the binding mode, something that only in silico docking can provide with the required velocity. In this context, the use of docking methods is defined by the limited number of drugs to consider and by the existence of a single target protein. The accuracy is, however, crucial since errors in the placement of the drug can completely misguide the lead optimization process. A basic metric commonly used for evaluating the accuracy of the predicted binding modes of docking programs is the root mean square deviation (RMSD) between the predicted conformation and the native pose of the ligand:

$$\mathrm{RMSD} = \left(\sum_{N} \frac{(R_i - R_j)^2}{N} \right)^{1/2}, \tag{2}$$

where R stands for the ligand coordinates in the predicted binding mode (i) and in the native pose (j), and N is the total number of atoms. In many practical

cases, the predicted binding mode can be useful even if there is a significant RMSD, provided that some key groups are properly located. Then, it is also convenient to use more case-specific descriptors for the validation of docking methods such as the generalized RMSD:

$$\mathrm{RMSD} = \left(\sum_{N} \frac{\xi_n (R_i - R_j)^2}{N} \right)^{1/2}, \tag{3}$$

where N is the total number of atoms in the drug and the weighting factor ξ_n reflects the importance of the residue n in defining the bioactive drug–protein complex. Many other qualitative measures of structural quality of the docking poses have been suggested [41].

Docking programs do not provide a single pose as an output, but a series of them ranked according to the scoring function. Thus, it is not an uncommon situation that the real binding mode is detected, but not top-ranked by the scoring function. Thus, an additional requirement for the derivation of a structural binding mode is the correct ranking of the good docking solution, which would guarantee that the final user does not disregard it in a further study. A quite common global estimate of the accuracy of the predicted binding mode is the "2 Å RMSD rule," which consists in computing the percentage of predicted binding modes of the ligands that are found at less than 2 Å from the native pose. In a recent study [12], we found that for a selected set of proteins, around 30% of the correctly predicted docked poses are disregarded due to a failure in the scoring of these poses. Thus, instead of correctly predicting the binding mode of 43% of the poses, only 30% of the poses are correctly predicted and scored (see Fig. 3).

The determination of primary or secondary targets for a drug is an increasing field of application for docking algorithms, especially due to the emergence of "drug repositioning" strategies [42], i.e., the identification of new indications for existing drugs. Both new indications and adverse drug reactions are caused by unexpected ligand–protein interactions on secondary targets, and can be explored through docking experiments. The objective here is not necessarily to predict the binding mode with extreme accuracy, but to detect possible targets for a drug.

During the last decades, the dominant philosophy in drug design has been the "one gene, one drug, one disease" paradigm. However, many effective drugs have shown to act via modulation of multiple proteins rather than single targets. Indeed, recent studies suggest that selective compounds compared to multitarget drugs may exhibit lower clinical efficacy [43,44]. In this regard, parallel large-scale multitarget virtual screening is a promising method to derive secondary targets.

The use of docking in vHTS is a common practice in pharmacological research due to its reduced cost compared to experimental HT techniques and to the existence of large virtual chemical libraries—containing over a million of potential ligands—available for screening [10]. The main objective of this type of projects is to mine the original library and derive a small subset of compounds, which has a larger percentage of promising ligand candidates, a process that is known as "enrichment."

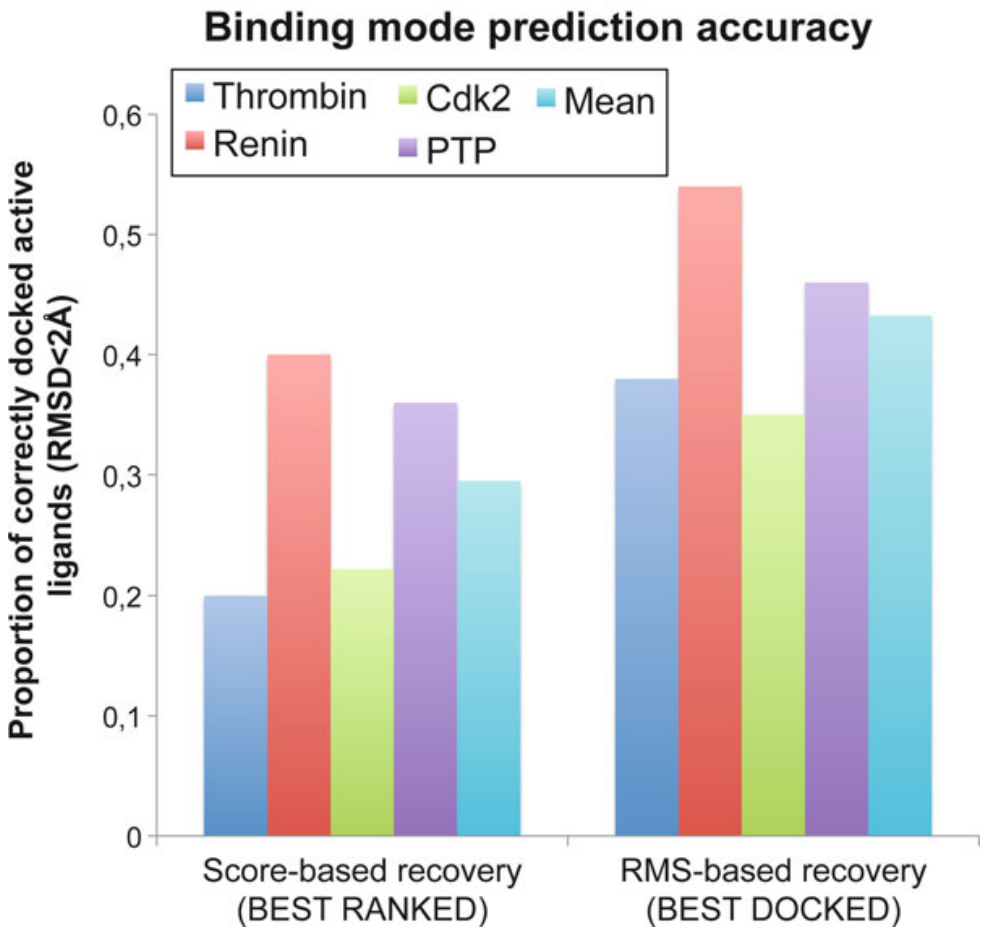


Fig. 3 Binding mode prediction accuracy of for five different human proteins: thrombin, rennin, cyclin-dependent kinase 2 (CDK2), and protein phosphatase 1B (PTP-1B)

Technically, vHTS requires very fast computer strategies, especially in cases where primary and secondary targets are screened simultaneously. Current protocols for vHTS are based on filtering strategies, where basic geometrical or pharmacological criteria are used to obtain a more focused chemical library.

The evaluation of the performance of docking methods is especially important considering the cost of the calculation. Here, the most important objective is to check the ability of the method to discriminate between active compounds and decoys (inactive). A virtual screening run selects a list of molecules (n) from a given database of N entries, which includes both actives (true positive compounds, TP) and decoys (false positive compounds, FP). Actives (A) that have not been found by the screening method are false negatives (FN) and decoys that have not been selected are true negatives (TN). The optimum screening is that able to recover all the true positives, without recovering any false positive. Although it is clear that virtual screening methods can be assessed by their ability to discriminate between active and inactive compounds, assessing the enrichment in a virtual screening procedure is a nontrivial task. Many different enrichment descriptors have been described in the literature [45,46], and they can all provide different information on

the performance of the screening. A combination of several enrichment descriptors is recommended if the aim is to evaluate the performance of a docking algorithm.

The most popular descriptors used to evaluate the quality of docking experiments in this scenario are the *sensitivity* [true positive rate; TPR; see (4)], which indicates the ability of the method to recover the true ligands, and the *specificity* [true negative rate; TNR see (5)], which informs on its ability to avoid decoys.

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} = 1 - \text{FPR}, \tag{5}$$

where FPR stands for false positive rate. Also, *accuracy* [*Acc*; (6)] describes the percentage of molecules which have been correctly classified by the screening protocol, and the *precision* (positive predictive value; *PPV*) gives accounts for the proportion of true positives among the list of selected compounds given by the docking (7).

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{N}} = \frac{\text{A}}{\text{N}} \times \text{TPR} + \left(1 - \frac{\text{A}}{\text{N}}\right) \times \text{TNR}. \tag{6}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

In order to o assess the ability of the models to obtain true actives among the first ranked compounds (an extra requirement in high-throughput docking) [47], the *enrichment factor* [EF, (8)] can be used:

$$\text{EF} = \frac{\text{TP}/n}{A/N}. \tag{8}$$

Recently, receiver operating characteristic (ROC; true positive vs. false positive rates) curves and the associated area under the ROC curves (AUC) have also become very popular to evaluate the discriminatory power of the virtual screening procedure [48–50]. The main advantage of these metrics is that they are independent on the ratio of actives to decoys of the database and accordingly they are good measures of the global performance of a docking algorithm in a vHTS procedure.

4 Protein Structure Prediction

One of the major practical limitations to the use of docking in pharmacological research lies in the need of high accurate structural data for the protein. Fortunately, protein structure can be predicted by a variety of computational methods, homology-modeling (also named comparative modeling) being the most accurate one in cases where there is a clear homolog with known structure [51, 52]. Building a protein structure from homology modeling requires a template—a protein with

similar amino acid sequence—and involves four major steps: fold assignment, sequence alignment, model building, and model refinement. Several computer packages are available to perform all this process automatically, such as the SWISS-MODEL software [53], the 3D-JIGSAW package [54], or the ModWeb tool [55]. Nevertheless, the general consensus [52] is that manually curated models derived from the use of programs, such as MODELLER [56], are more reliable than automatic procedures.

One of the most critical steps in homology modeling is the identification of the proper template. The simplest method that can be used for this purpose is a simple BLAST search [57] against the PDB database. However, methods based on multiple sequence alignments or profiles have demonstrated to be much more sensitive in identifying distantly related homologs [57, 58]. Choosing the best template among the candidates derived from multiple alignments is crucial for the final accuracy of the model and in addition to sequence identity we need to consider that "holo" structures are always better templates than "apo" ones [59]. In the case that several holo candidates are available, we should favor the structure containing a similar ligand to the one that we aim to dock [60, 61].

Another crucial step in the model generation is the alignment of the target with the template(s). This procedure can be done easily with standard alignment algorithms in cases of large identity between template(s) and target protein. However, in difficult cases (below 30% sequence identity), the alignment obtained by standard methods needs to be refined by:

1. Including structural information of the template, i.e., avoiding gaps in secondary structure elements, in buried regions, or between two residues that are far in space [62–65].
2. Building a multiple structure-based alignment of the templates and use them to align the target sequence to it.
3. Calculating the target and template sequence profiles by aligning them with sequences sufficiently similar to the target and template sequences respectively, so that they can be aligned without significant errors. The final target-template alignment is then obtained by aligning the two profiles [66, 67].

In general, the use of multiple structures and multiple sequences benefits from the evolutionary and structural information about the templates and target sequence, and often produces a better alignment for modeling than pairwise alignment methods [68, 69]. In any case, once the template is selected and the target protein is aligned, the structural model can be generated using different approaches. In this context, MODELLER [56], one of the most widely used homology modeling engines, typically builds models by enforcing spatial restraints derived from the template structure(s).

The quality of the structure derived from homology modeling roughly correlates with the sequence identity between the target and the template proteins [70]. Thus, it is accepted that for sequence identities below 30% less than half of the residues have their Cα correctly placed [71, 72]. The percentage of correctly placed residues increases to 85% for identities ranging from 30 to 50% and most of the Cαs are well

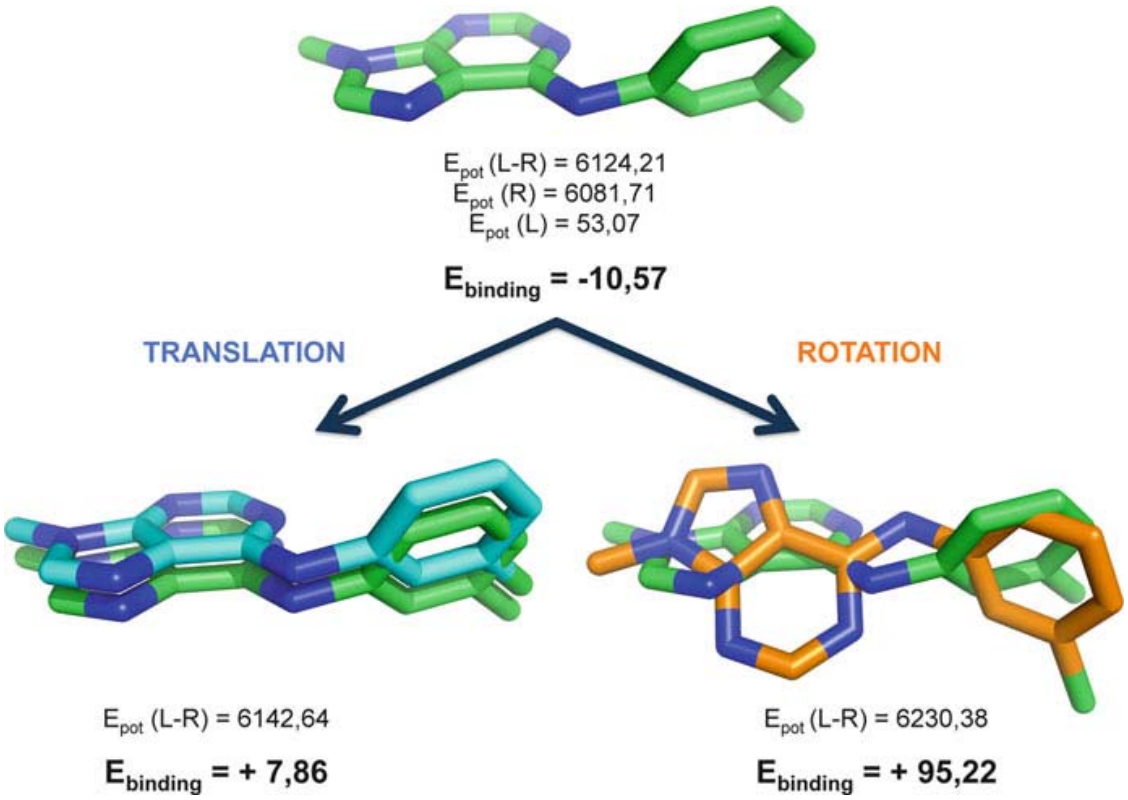


Fig. 4 The ligand–receptor interaction energy is strongly altered by slight translation and/or rotation movements of the ligand. The ligand–receptor binding energy (Ebinding) has been computed as the difference of the potential energy of the complex [Epot(L–R)] with respect to the individual potential energies of the ligand [Epot(L)] and the receptor [Epot(R)]. The ligand shown has been taken from the structure of a human CDK2 (PDB code 1ckp)

positioned for sequence identities above 50%. Inside the high-quality range no direct correlation exists between the accuracy of the model and the sequence identity with the template, and evaluation of the expected quality of a model is still an unsolved problem [73]. In fact, the concept of "accuracy of the model" can be arbitrary, since it depends on its planned use. For example, a model with accuracy around 3.5 Å in backbone positioning may be sufficient for understanding protein function or designing mutations, but is expected to be of small utility for predicting ligand binding [74, 75], since the strong dependence of the ligand–receptor interaction energy on fine geometrical details (Fig. 4) implies that small structural errors might cause a large bias in the binding calculation. A deep discussion on this point is presented in the next section of this chapter.

5 HT Docking from Homology Modeled Structures

The use of homology models in docking calculations has been recently explored by different groups, finding in general quite encouraging results. McGovern and Shoichet [59] performed a high-throughput docking on ten enzymes for which

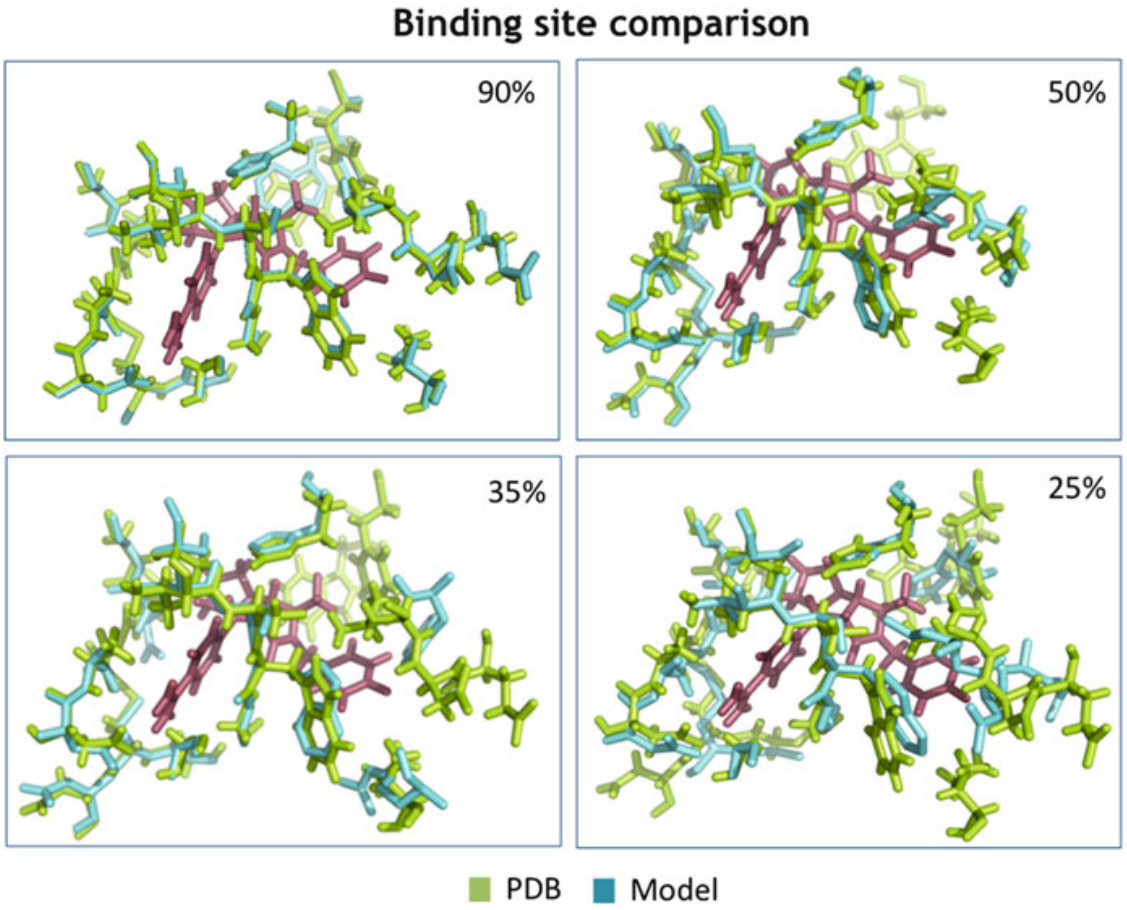


Fig. 5 Binding site comparison of thrombin PDB structure (2cn0, shown in *green*) with homology models of different sequence identity (*blue*). The ligand shown (*magenta*) corresponds to the crystallized ligand in the 2cn0 structure. As can be seen from the figure, even at low sequence identities the binding site structure is still reasonably conserved

apo, holo, and homology model structures were available, suggesting that they were useful for enriching the screening, but in general not as powerful as the holo-crystal structure. Diller and Li [76] reported significant enrichments of the homology models of six kinases with identities in the range of 30–50% when used to screen a large chemical library. Similar results were obtained by Oshiro et al. [77] in the study of two targets (cyclin-dependent kinase 2 (CDK2) and factor VIIa), by Gilson's group [78] with a set of five targets, and by Ferrara and Jacoby [79] in the analysis of insulin growth factor I receptor. All these results suggest that the conservation of the binding sites in modeled structures is sufficient, and does not affect docking accuracy significantly (Fig. 5).

Recently, various groups have suggested [12, 80] using ensembles of homology models as templates, developing automatic strategies valid within the HT-regime (Fig. 6). The use of the ensemble docking approach coupled to homology modeling has two main advantages: (1) there is no need to identify the "best" performing homology model and (2) protein flexibility is implicitly included in the docking run. When using the ensemble docking approach, each homology model is built on

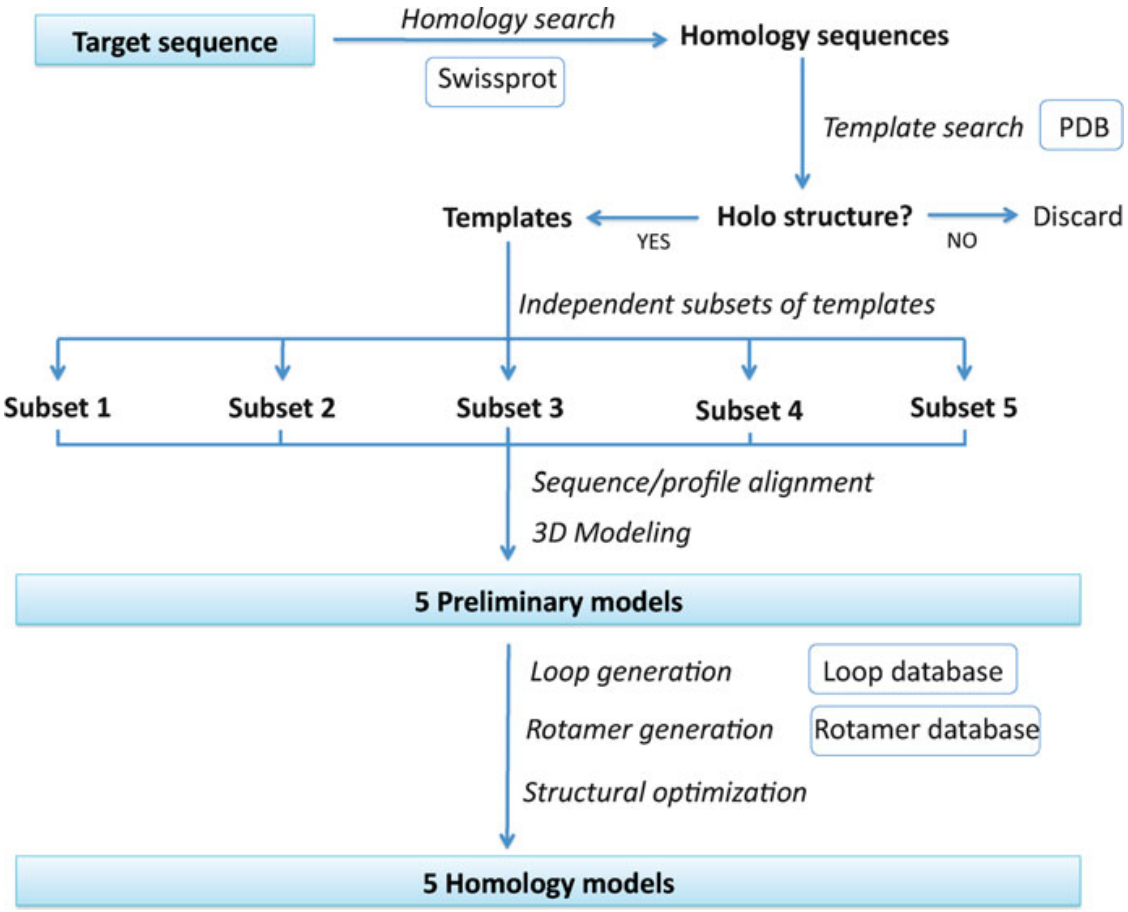


Fig. 6 Example of workflow [12] for building homology models to be used in the ensemble docking approach

a basis of a different template, and thus the binding site is specialized to recognize a different subset of active ligands. As a result, there is an improvement in the probabilities of detecting "true positives" (Fig. 6).

Different studies using ensemble docking with experimental structures have obtained controversial results. Some authors [34] state that ensemble docking clearly improves the performance of the docking process, while others [37, 81, 82] complain about the increase in "false negatives" and suggest that the enrichment of the results using ensembles is not so different when compared to a good-performing crystal structure (although the rules to select "a priori" which is a good-performing crystal structure are not evident). The situation when using homology models is more evident, since in this case the use of ensembles increases very significantly the sensitivity with respect to single models, decreasing only slightly the specificity and leading to an overall clear improvement of the docking results [12]. Figure 7 illustrates the increase in the proportion of correctly predicted binding modes when using ensemble docking compared to single model docking—only homology models are being considered in the figure. In this example, single models produce moderate binding mode predictions, being able to recover 30% of correctly docked

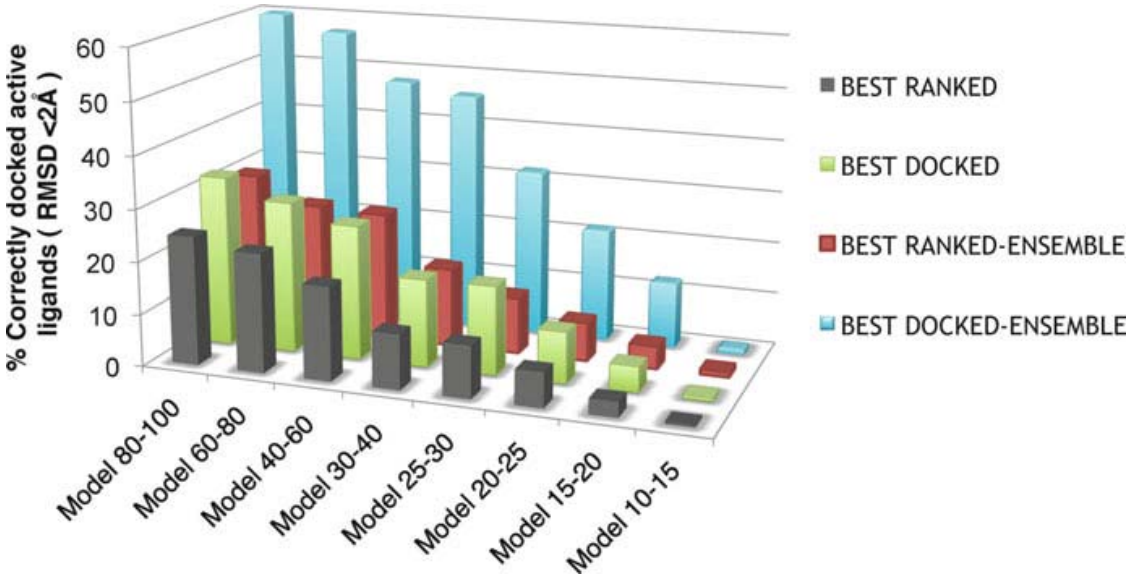


Fig. 7 Recovery of correctly docked active ligands for a selected set of proteins (α-momorcharin, trypsin, p38 kinase, HIV retrotranscriptase, factor Xa, and heat shock protein 90). As can be seen from the figure, the correctly docked ligand recovery is dependent both on the strategy of docking (ensemble docking versus single docking) and on the sequence identity of the template. A ligand is considered as correctly docked when its RMSD with the crystallographic ligand is below 2 Å. Both score-based selection—i.e., best ranked—and RMSD-based selection—i.e., best docked—are shown. Single docking averages are shown in *black* and *green*, whereas ensemble docking averages are shown in *red* and *cyan*. These results were obtained by docking a database containing both known actives and decoys, using Glide docking program in an SP—standard precision—mode (data from [12])

ligands (21% if we only take into account the best ranked solution), whereas the ensemble docking approach increases the correctly docked ligands to 57% (29% when considering the best ranked solution).

Homology modeling-based ensemble docking coupled with good structural models and strict scoring methods can outperform single PDB docking (Fig. 8). Furthermore, the ensemble docking protocol is very robust to the decrease in sequence identity, given that models with sequence identities in the range of 30–40% still provide good results for most proteins.

A better view on the overall quality of the homology-based ensemble docking approach is obtained by analyzing simultaneously its ability for vHTS (i.e., its capability to recover specifically active ligands from the dataset) and in the context of structural determination of binding modes (i.e., its capability to yield good structural solution as the top ranked ones). Results displayed in Figs. 8 and 9 provide evidence on the power of the ensemble docking approach in a wide range of working scenarios.

As a summary, the general accepted "rule" is that only models built with more than 50% sequence identity are accurate enough for docking and the accuracy in docking is higher with holo structures than homology models [75, 77, 83]. However, recent available studies using ensemble docking with homology models

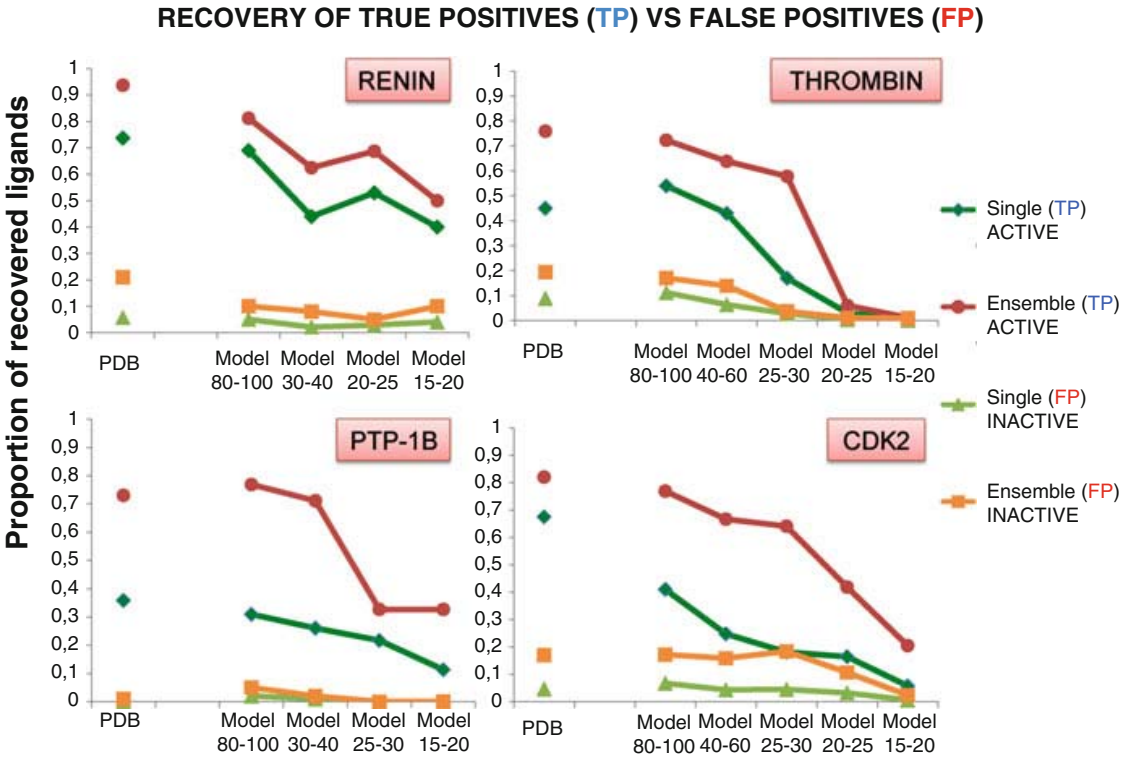


Fig. 8 Ensemble docking versus single docking approach. The performance of both approaches is being compared in terms of recovered active ligands and decoys for four human proteins: renin, thrombin, cyclin- CDK2, and PTP-1B. The single docking approach performance is represented with *green* and lime lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the *red* and *orange lines* correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. In all cases, the difference between active (true ligands) and inactive (decoys) recovery is higher when using ensemble docking. Results where obtained using Glide computer program in an extra-precision (XP) mode with a GlideScore (GS) threshold $= -8$ (data from [12])

[12] strongly suggest that models with sequence identity above 30–40% display a considerable ability to specifically recover active ligands, and can even outperform single crystal structures. Although it is difficult to extend results of the small set of proteins used in these studies to the entire proteome, the use of ensemble docking is extremely recommended over single docking, especially when using homology models. Moreover, the use of homology models is not limited to the retrieval of active ligands from a chemical library, but can also provide structural complexes with sufficient accuracy for lead optimization processes.

6 Increasing Coverage

As noted in the beginning of this chapter, despite the tremendous effort focused for many decades in the experimental determination of protein structures, the current version of PDB covers only a small fraction of human proteins. This coverage is

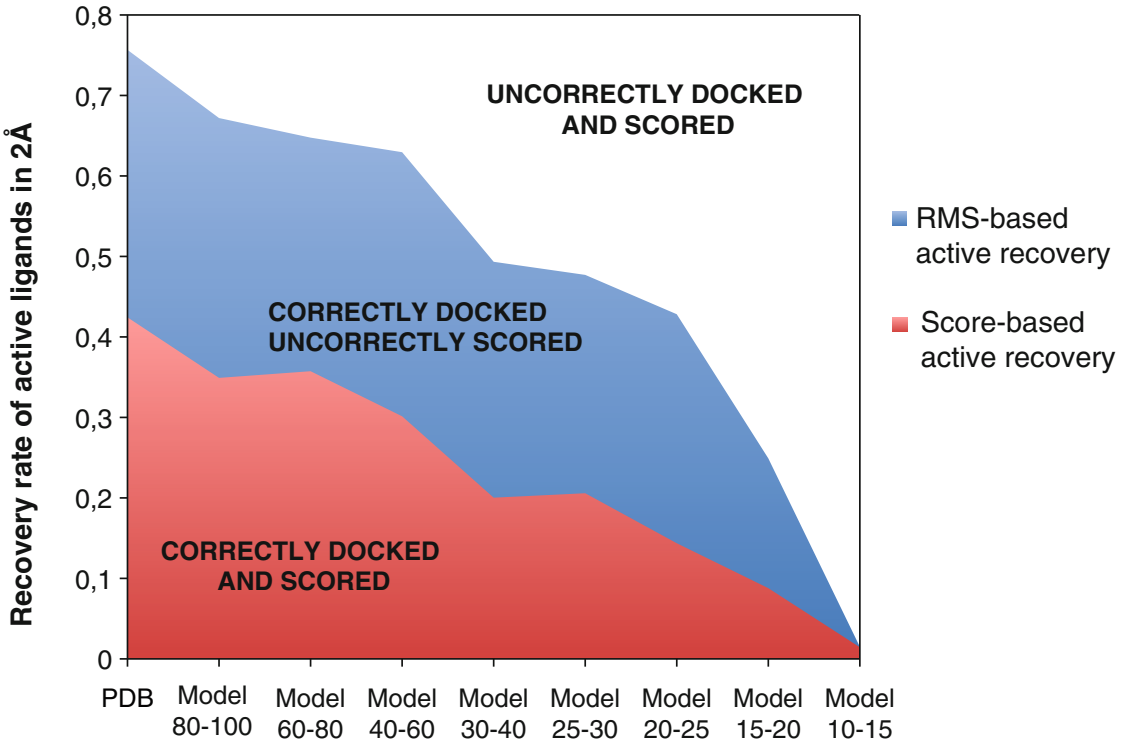


Fig. 9 Expected percentage of success in docking experiments performed using the ensemble docking approach (both for X-ray and homology models, the later obtained from templates with different degrees of sequence identities). Recovery rates have been computed as average recovery rates of four human proteins: renin, thrombin, CDK2, and PTP-1B

even smaller if we focus on protein structures coming from pathogens. Our group and others [12, 80] have suggested that homology models derived from templates with identity ranges of 30–40% can significantly enrich chemical libraries. These results allow us to expand dramatically the universe of use of docking techniques (Fig. 1), especially in the case of human proteins with pharmacological interest (taken from DrugBank database; [84]), which are covered over 75% when using homology models up to 30% identity (Fig. 10).

Thus, with all the required cautions needed in the use of homology models for docking purposes (related mostly to the problems in finding good templates and in determining "a priori" the quality of the model), the use of comparative models can enlarge dramatically the universe of applicability of small-molecule docking approaches. Ensemble docking performed on homology models provides results of similar, or even better quality than those obtained with single crystal structures, leading to a clear enrichment in the chemical libraries, and producing poses of good structural quality, even in cases where ligand binding implies non-negligible changes in protein structure. Altogether ensemble docking from homology modeling appears as a promising alternative to extend the use of docking strategies in drug-design pipelines.

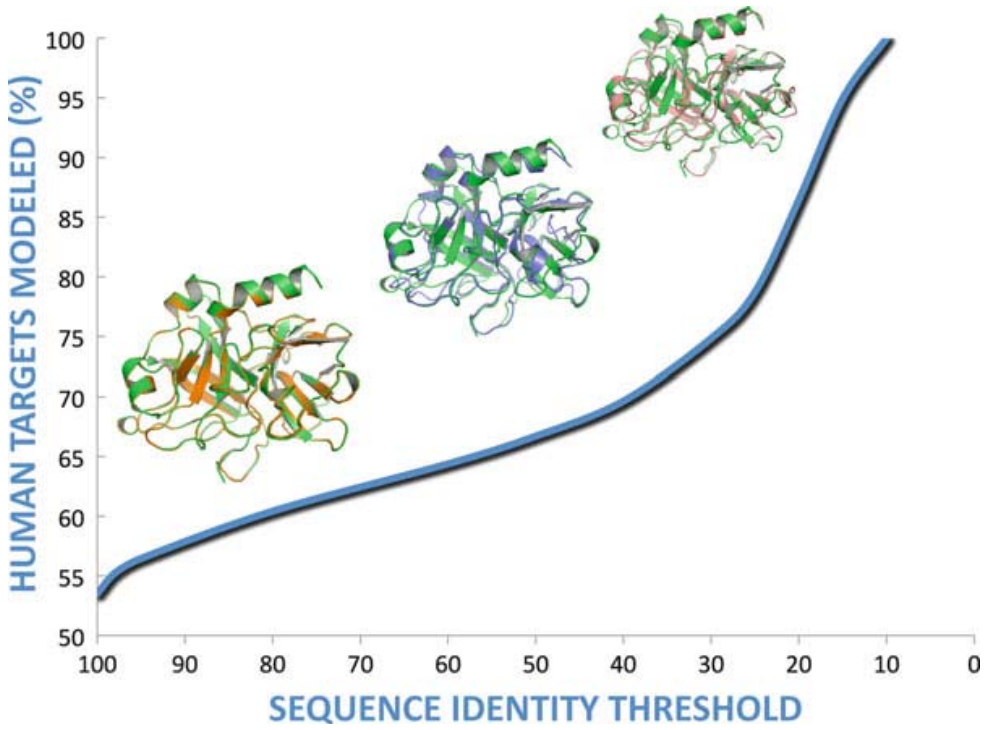


Fig. 10 Structural coverage of human targets of pharmacological interest depending on the sequence identity threshold used in homology modeling. A 30% sequence identity threshold—which still gives very good results when using the ensemble docking approach—allows us to cover 41% more human drug targets, obtaining a final coverage of 75% of the human drug targets. The superimposition of the crystal structure of thrombin (*green*, PDB code 2cn0) and homology models built with different sequence identity—90% (*orange*), 50% (*blue*), and 30% (salmon)—is also shown

References

1. Campbell, S.J., Gold, N.D., Jackson, R.M., Westhead, D.R.: Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. **13**(3), 389–395 (2003)
2. Keiser, M.J., et al.: Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. **25**(2), 197–206 (2007)
3. Fisher, E.: Einfluss der Konfiguration auf die Wirkung der Enzyme. Berichte der Deutschen Chemischen Gesellschaft. **27**, 2985–2993 (1894)
4. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E.: A geometric approach to macromolecule–ligand interactions. J. Mol. Biol. **161**(2), 269–288 (1982)
5. Butler, K.T., Luque, F.J., Barril, X.: Toward accurate relative energy predictions of the bioactive conformation of drugs. J. Comput. Chem. **30**(4), 601–610 (2009)
6. Perola, E., Charifson, P.S.: Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J. Med. Chem. **47**(10), 2499–2510 (2004)
7. Berman, H.M., et al.: The protein data bank. Nucleic. Acids. Res. **28**(1), 235–242 (2000)
8. Cozzini, P., et al.: Target flexibility: an emerging consideration in drug discovery and design. J. Med. Chem. **51**(20), 6237–6255 (2008)
9. Merz, K.M.: Limits of free energy computation for protein–ligand interactions. J Chem. Theory. Comput. **6**(4), 1018–1027 (2010)

10. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug. Discov. **3**(11), 935–949 (2004)
11. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic. Acids. Res. **33**(Database issue), D501–504 (2005)
12. Novoa, E.M., de Pouplana, L.R., Barril, X., Orozco, M.: Ensemble docking from homology models. J. Chem. Theory. Comput. **6**(8), 2547–2557 (2010)
13. Halperin, I., Ma, B., Wolfson, H., Nussinov, R.: Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins. **47**(4), 409–443 (2002)
14. Leach, A.R., Shoichet, B.K., Peishoff, C.E.: Prediction of protein–ligand interactions. Docking and scoring: successes and gaps. J. Med. Chem. **49**(20), 5851–5855 (2006)
15. Shoichet, B.K., McGovern, S.L., Wei, B., Irwin, J.J.: Lead discovery using molecular docking. Curr. Opin. Chem. Biol. **6**(4), 439–446 (2002)
16. Sousa, S.F., Fernandes, P.A., Ramos, M.J.: Protein-ligand docking: current status and future challenges. Proteins. **65**(1), 15–26 (2006)
17. Shoichet, B.K., Bodian, D.L., Kuntz, I.D.: Molecular docking using shape descriptors. J. Comput. Chem. **13**, 380–397 (1992)
18. Gardiner, E.J., Willett, P., Artymiuk, P.J.: Graph-theoretic techniques for macromolecular docking. J. Chem. Inf. Comput. Sci. **40**(2), 273–279 (2000)
19. Ponder, J.W., Case, D.A.: Force fields for protein simulations. Adv. Protein. Chem. **66**, 27–85 (2003)
20. Morris, G.M. et al.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function J. Comput. Chem. **19**, 1639–1662 (1998)
21. Jones, G., Willett, P., Glen, R.C.: Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. **245**(1), 43–53 (1995)
22. Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D.: Improved protein-ligand docking using GOLD. Proteins. **52**(4), 609–623 (2003)
23. Rarey, M., Kramer, B., Lengauer, T., Klebe, G.: A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. **261**(3), 470–489 (1996)
24. Jorgensen, W.L., Tirado-Rives, J.: The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and Crambin. J. Am. Chem. Soc. **110**, 1657–1666 (1988)
25. Abagyan, R., Totrov, M., Kuznetsov, D.: ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. **15**, 488–506 (1994)
26. Friesner, R.A., et al.: Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. **47**(7), 1739–1749 (2004)
27. Ding, F., Yin, S., Dokholyan, N.V.: Rapid flexible docking using a stochastic rotamer library of ligands. J. Chem. Inf. Model. **50**(9), 1623–1632 (2010)
28. Gohlke, H., Klebe, G.: Statistical potentials and scoring functions applied to protein-ligand binding. Curr. Opin. Struct. Biol. **11**(2), 231–235 (2001)
29. Dunbrack, R.L., Jr. Karplus, M.: Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. **230**(2), 543–574 (1993)
30. Holm, L., Sander, C.: Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins. **14**(2), 213–223 (1992)
31. Brunger, A.T., Kuriyan, J., Karplus, M.: Crystallographic R factor refinement by molecular dynamics. Science. **235**(4787), 458–460 (1987)
32. Armen, R.S., Chen, J., Brooks, C.L.: An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. J. Chem. Theory. Comput. **5**(10), 2909–2923 (2009)
33. Paulsen, J.L., Anderson, A.C.: Scoring ensembles of docked protein:ligand interactions for virtual lead optimization. J. Chem. Inf. Model. **49**(12), 2813–2819 (2009)
34. Craig, I.R., Essex, J.W., Spiegel, K.: Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. J. Chem. Inf. Model. **50**(4), 511–524 (2010)

35. Huang, S.Y., Zou, X.: Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins. **66**(2), 399–421 (2007a)
36. Rao, S., et al.: Improving database enrichment through ensemble docking. J. Comput. Aided. Mol. Des. **22**(9), 621–627 (2008)
37. Rueda, M., Bottegoni, G., Abagyan, R.: Recipes for the selection of experimental protein conformations for virtual screening. J. Chem. Inf. Model. **50**(1), 186–193 (2010)
38. Damm, K.L., Carlson, H.A.: Exploring experimental sources of multiple protein conformations in structure-based drug design. J. Am. Chem Soc. **129**(26), 8225–8235 (2007)
39. Huang, S.Y., Zou, X.: Efficient molecular docking of NMR structures: application to HIV-1 protease. Protein. Sci. **16**(1), 43–51 (2007b)
40. Hawkins, P.C., Warren, G.L., Skillman, A.G., Nicholls, A.: How to do an evaluation: pitfalls and traps. J. Comput. Aided. Mol. Des. **22**(3–4), 179–190 (2008)
41. Warren, G.L., et al.: A critical assessment of docking programs and scoring functions. J. Med. Chem. **49**(20), 5912–5931 (2006)
42. Yang, L., et al.: Identifying unexpected therapeutic targets via chemical-protein interactome. PLoS ONE. **5**(3), e9568 (2010)
43. Petrelli, A., Giordano, S.: From single- to multi-target drugs in cancer theraphy: when aspecificity becomes an advantage. Curr. Med. Chem. **15**, 422–432 (2008)
44. Wermuth, C.G.: Multitarget drugs: the end of the 'one-target-on-disease' phylosophy? Drug. Discov. Today. **9**, 826–827 (2004)
45. Kirchmair, J., Markt, P., Distinto, S., Wolber, G., Langer, T.: Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? J. Comput. Aided. Mol. Des. **22**(3–4), 213–228 (2008)
46. Langer T., Hoffmann RD.: Pharmacophores and Pharmacophore Searches. Wiley-VCH, Weinheim, Germany, pp. 338–343 (2006)
47. Truchon, J.F., Bayly, C.I.: Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J. Chem. Inf. Model. **47**(2), 488–508 (2007)
48. Jain, A.N., Nicholls, A.: Recommendations for evaluation of computational methods. J. Comput. Aided. Mol. Des. **22**(3–4), 133–139 (2008)
49. Nicholls, A.: What do we know and when do we know it? J. Comput. Aided. Mol. Des. **22**(3–4), 239–255 (2008)
50. Witten, I.H., Frank, E.: Credibility: Evaluating what's been learned. In: Data minings: Practical machine learning tools and techniques, 2nd ed; Morgan Kaufmann: San Francisco, CA, pp. 161–176 (2005)
51. Koehl, P., Levitt, M.: A brighter future for protein structure prediction. Nat. Struct. Biol. **6**(2), 108–111 (1999)
52. Marti-Renom, M.A., et al.: Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. **29**, 291–325 (2000)
53. Arnold, K., Bordoli, L., Kopp, J., Schwede, T.: The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. **22**(2), 195–201 (2006)
54. Bates, P.A., Kelley, L.A., MacCallum, R.M., Sternberg, M.J.: Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins. Suppl. 5, 39–46 (2001)
55. Eswar, N., et al.: Tools for comparative protein structure modeling and analysis. Nucleic. Acids. Res. **31**(13), 3375–3380 (2003)
56. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**(3), 779–815 (1993)
57. Altschul, S.F., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic. Acids. Res. **25**(17), 3389–3402 (1997)
58. Wistrand, M., Sonnhammer, E.L.: Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics. **6**, 99 (2005)
59. McGovern, S.L., Shoichet, B.K.: Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J. Med. Chem. **46**(14), 2895–2907 (2003)

60. Rockey, W.M., Elcock, A.H.: Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? Curr. Protein. Pept. Sci. **7**(5), 437–457 (2006)
61. Tuccinardi, T., Botta, M., Giordano, A., Martinelli, A.: Protein kinases: docking and homology modeling reliability. J. Chem. Inf. Model. **50**(8), 1432–1441 (2010)
62. Blake, J.D., Cohen, F.E.: Pairwise sequence alignment below the twilight zone. J. Mol. Biol. **307**(2), 721–735 (2001)
63. Jennings, A.J., Edge, C.M., Sternberg, M.J.: An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein. Eng. **14**(4), 227–231 (2001)
64. Sanchez, R., Sali, A.: Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. **7**(2), 206–214 (1997)
65. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**(1), 243–257 (2001)
66. Marti-Renom, M.A., Madhusudhan, M.S., Sali, A.: Alignment of protein sequences by their profiles. Protein. Sci. **13**(4), 1071–1087 (2003)
67. von Ohsen, N., Sommer, I., Zimmer, R.: Profile-profile alignment: a powerful tool for protein structure prediction. Pac. Symp. Biocomput. 252–263 (2003)
68. Jaroszewski, L., Rychlewski, L., Godzik, A.: Improving the quality of twilight-zone alignments. Protein Sci. **9**(8), 1487–1496 (2000)
69. Sauder, J.M., Arthur, J.W., Dunbrack, R.L., Jr: Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins. **40**(1), 6–22 (2000)
70. Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., Sali, A.: Reliability of assessment of protein structure prediction methods. Structure. **10**(3), 435–440 (2002)
71. Eswar, N., Sali, A.: (2007) Comparative modeling of drug target proteins. In: Taylor J., Triggle D., Mason J.S., (eds.) Computer-Assisted Drug Design, Comprehensive Medicinal Chemistry II, vol. 4, pp. 215–236. Elsevier, Oxford, UK
72. Sanchez, R., et al.: Protein structure modeling for structural genomics. Nat. Struct. Biol. 7 Suppl. 986–990 (2000)
73. Eramian, D., Eswar, N., Shen, M.Y., Sali, A.: How well can the accuracy of comparative protein structure models be predicted? Protein. Sci. **17**(11), 1881–1893 (2008)
74. Baker, D., Sali, A.: Protein structure prediction and structural genomics. Science. **294**(5540), 93–96 (2001)
75. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug. Discov. Today. **14**(13–14), 676–683 (2009)
76. Diller, D.J., Li, R.: Kinases, homology models, and high throughput docking. J. Med. Chem. **46**(22), 4638–4647 (2003)
77. Oshiro, C., et al.: Performance of 3D-database molecular docking studies into homology models. J. Med. Chem. **47**(3), 764–767 (2004)
78. Kairys, V., Fernandes, M.X., Gilson, M.K.: Screening drug-like compounds by docking to homology models: a systematic study. J. Chem. Inf. Model. **46**(1), 365–379 (2006)
79. Ferrara, P., Jacoby, E.: Evaluation of the utility of homology models in high throughput docking. J. Mol. Model. **13**(8), 897–905 (2007)
80. Fan, H., et al.: Molecular docking screens using comparative models of proteins. J. Chem. Inf. Model. **49**(11), 2512–2527 (2009)
81. Barril, X., Morley, S.D.: Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J. Med. Chem. **48**(13), 4432–4443 (2005)
82. Birch, L., Murray, C.W., Hartshorn, M.J., Tickle, I.J., Verdonk, M.L.: Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. J. Comput. Aided. Mol. Des. **16**(12), 855–869 (2002)
83. Hillisch, A., Pineda, L.F., Hilgenfeld, R.: Utility of homology models in the drug discovery process. Drug. Discov. Today. **9**(15), 659–669 (2004)
84. Wishart, D.S., et al.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic. Acids. Res. **36**(Database issue), D901–906 (2008)

85. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**(4), 823–826 (1986)
86. O'Donovan, C., Apweiler, R., Bairoch, A.: The human proteomics initiative (HPI). Trends. Biotechnol. **19**(5), 178–181 (2001)
87. Park, S.J., Kufareva, I., Abagyan, R.: Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. J. Comput. Aided. Mol. Des. **24**(5), 459–471 (2010)

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

5. Discussion and Conclusions

5.1 Role of tRNA modifications in genome structure and codon usage

5.1.1 The appearance of two tRNA modification enzymes shaped the tRNA gene content and the codon usage bias

As discussed previously, the degeneracy of the genetic code implies that several 'synonymous' codons code for the same amino acid, although they are not used with equal frequencies. Codon usage bias is a critical determinant for gene expression and genome function, but why and how it differs between organisms remains poorly understood.

In this work we have tried to answer the following questions:

i) *Why are some codons preferred relative to others recognized by the same anticodon?*
ii) *How do tRNA gene pools evolve in terms of anticodon number and type?*
iii) *How do tRNA gene pools co-evolve with codon usage in relation to the optimization of the translation machinery and the maximization of growth?*

i) Why are some codons preferred relative to others recognized by the same anticodon?

There is abundant literature regarding codon usage biases, and sophisticated techniques profit from this information to infer adaptive evolution (Suzuki et al., 2001), horizontal transfer (Médigue et al., 1991), expression levels (Sharp and Li, 1987) and cellular localization (Chiapello et al., 1999). However, most works are focused on the analysis of codon frequencies from DNA sequences (Andersson and Kurland, 1990; Duret 2002). The complementary approach, understanding the tRNA gene number and anticodon trype, has been much less developed in the framework of comparative genomics, and has mostly been focused on the correlation between tRNA content and codon usage within one or a few species (Ikemura 1985; Kanaya et al., 1999; Duret, 2000).

From our tRNA gene content analysis, we find that tRNA gene content has evolved in a kingdom-specific manner, thanks to the appearance of two distinct tRNA modification enzymes (ADATs and UMs) that increase the translation efficiency of a subset of tRNA isoacceptors. These tRNA isoacceptors susceptible of being either ADAT- or UM- modified are found to be

enriched in gene copy number in their respective genomes, and thus are 'preferred' relative to others recognized by the same anticodon.

ii) How do tRNA gene pools evolve in terms of anticodon number and type?

It has been suggested that it is more favorable to have more tRNAs of the same type –several copies for a same tRNA isoacceptor-, because this allows the co-evolution of codon usage bias in highly expressed genes, which then creates a strong demand for these smaller sets of tRNAs (Curran and Yarus, 1989; Berg and Kurland, 1997). In this regard, we propose that tRNA gene content, and consequently, codon usage bias, have evolved to increase mostly those tRNA isoacceptors -and the codons read by them- susceptible to be ADAT- or UM-modifiable, given that they increase the translation efficiency.

We propose a critical factor that had not been previously taken into account in tRNA evolution or codon usage bias: how modification at the wobble position of the anticodon impacts codon utilization. Indeed, inclusion of this parameter leads to near-perfect correlations between codon usage and tRNA abundance, and is consistent across known extant major phylogenetic groups (**Figure 5.1**). This provides a compelling scenario for the diversification of genetic code usage in evolution driven by tRNA modifications.

iii) How do tRNA gene pools co-evolve with codon usage in relation to the optimization of the translation machinery and the maximization of growth?

Our results suggest that tRNA gene content has evolved to increase those tRNA isoacceptors are modifiable by ADATs or UMs. Importantly, these tRNA modifications expand the decoding capacity of the tRNA isoacceptors, suggesting that translation efficiency has been a major factor in determining the evolution of tRNA gene content across species. Furthermore, we find that codon usage bias in highly expressed proteins has evolved such that they are enriched in 'preferred' codons (those read by specific modified tRNAs), indicating that the activity of tRNA modification enzymes constitutes a novel post-transcriptional regulation mechanism of protein abundances.

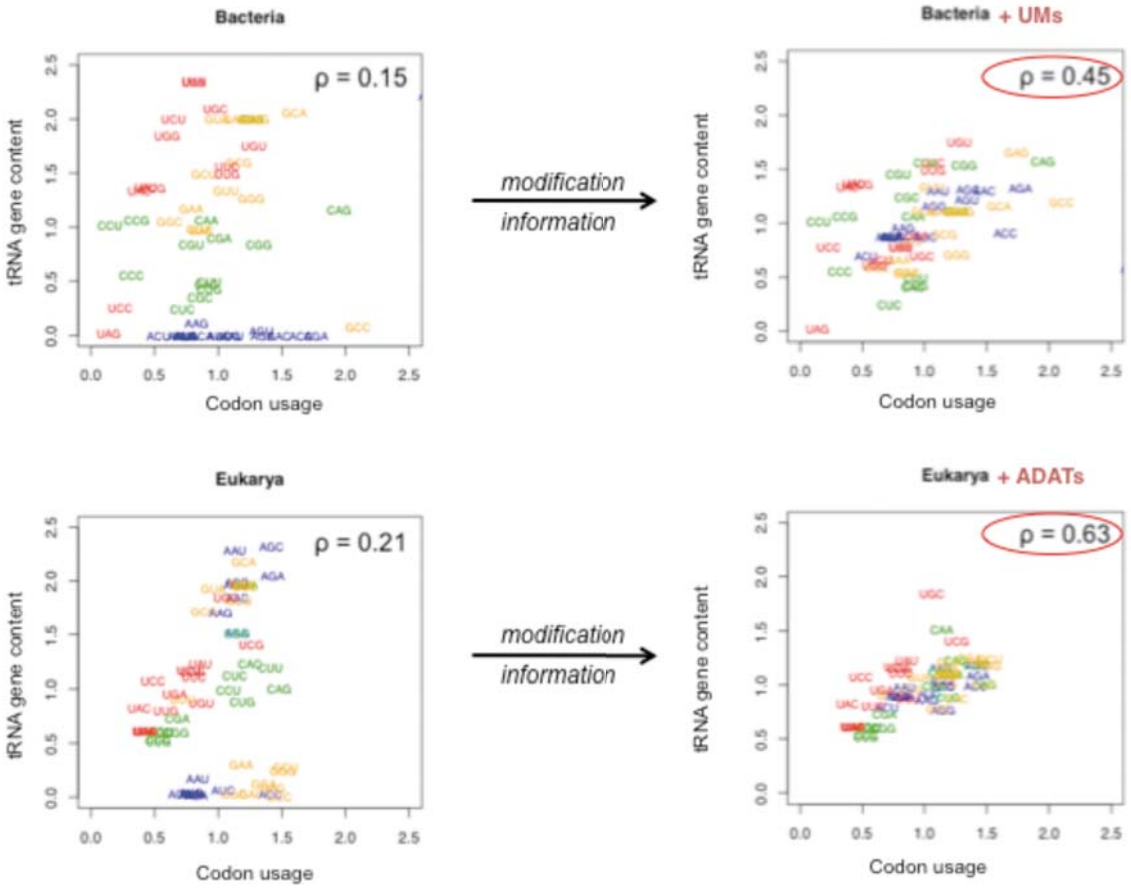


Figure 5.1. Quantitative correlation between tRNA gene content and codon usage. Each tRNA isoacceptor has been coloured according to its wobble base. Initial correlations have been computed taking into account the W-C codon-anticodon possible base-pairings. The inclusion of G-U wobble base-pairing did not substantially increase the Pearson correlations (data not shown). When including the modification information in Eukarya (ADATs) and Bacteria (UMs), the Pearson correlations were significantly increased.

5.1.2 tRNA gene content within each kingdom does not follow the tree of life

From our tRNA gene content analysis, we could see that the evolution of the tRNA gene content followed the evolution of the tree of life, in the sense that it clusters species from the same kingdom together, and coincides with the appearance of hetADATs in Eukarya and UMs in Bacteria (**Figure 5.2**). However, bacterial species seem to be widely spreaded in the phylogeny. What is causing the difference between the diverse Bacteria?

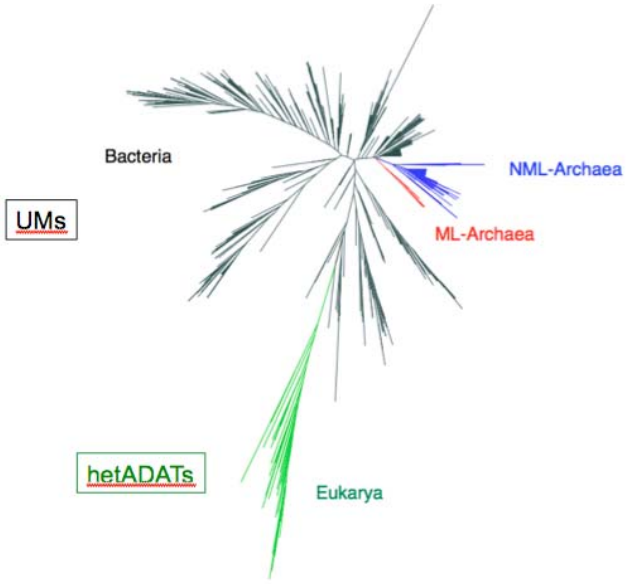


Figure 5.2. Phylogeny based on tRNA gene content. The four phylogenetic clades identified have been coloured accordingly, and the tRNA modification enzymes responsible for the separation are labelled accordingly: UMs in Bacteria (black), and hetADATs in Eukarya (green). Adapted from Novoa et al., 2012.

To decipher which variable/s explained the distribution of bacterial species, we first checked whether the separation of the species followed the rRNA canonical tree of life. However, this did not seem to be the case (**Figure 5.3**). Then, what is explaining the diversity in tRNA gene content found across bacteria? To answer this question, we investigated whether the species were clustered depending on other parameters, such as: salinity, oxygen requirement, temperature, pathogenicity, ability to live in multiple habitats, genome size or GC content. For each bacterial species, we annotated its diverse characteristics (e.g. aerobic, anaerobic, facultative or microaerophilic -in the case of oxygen requirement-) and searched for a possible correlation between its tRNA gene content and any of the analyzed parameters. For better visualization of the results, we performed a principal component analysis (PCA) of the tRNA relative gene frequencies (RGF) of each species, and plotted each species in a two dimensional plot. Each species was then coloured according to its characteristics, for each of the variables analyzed (**Figure 5.4**).

We found that none of the variables tested was explaining the variance found in the tRNA gene content. If the any of the variables would be explaining the differences between the species, we would expect that the species would be grouped by its color, which is not the case. Thus, with this data we cannot conclude what is causing the differences in tRNA gene content between bacterial species.

However, from our tRNA gene content analysis throughout the three kingdoms (**Figure 5.2** and **Publication 1**), we found that the appearance of UMs and hetADATs, and consequently, its differences in the strategies for maximizing translation efficiency, caused a clear separation of the three kingdoms in terms of tRNA gene content. Nevertheless, it is possible that within Bacteria, the major force driving tRNA gene content evolution was not selection, but instead, mutational drift.

The fact that the bacterial phylogeny based on tRNA gene content does not follow the tree of life is compatible with the idea that the separation of the species into its corresponding kingdoms (**Figure 5.2**) is a consequence of the appearance of two modification enzymes. If the tRNA phylogeny was only a mere mirror of the rRNA phylogeny, we should also see a taxonomic clustering within bacteria (**Figure 5.3** and **Figure 5.4a**).

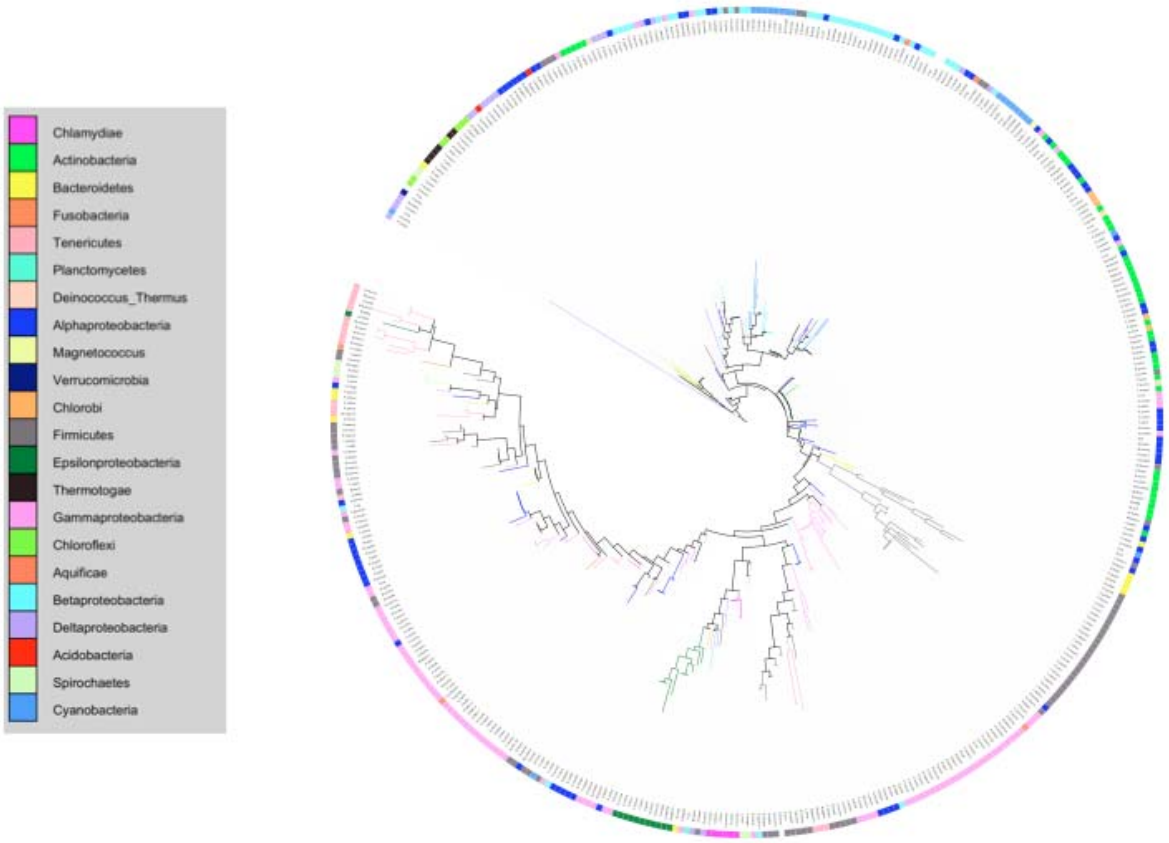


Figure 5.3. Bacterial phylogeny based on tRNA gene content. Each species has been coloured according to its corresponding phylum.

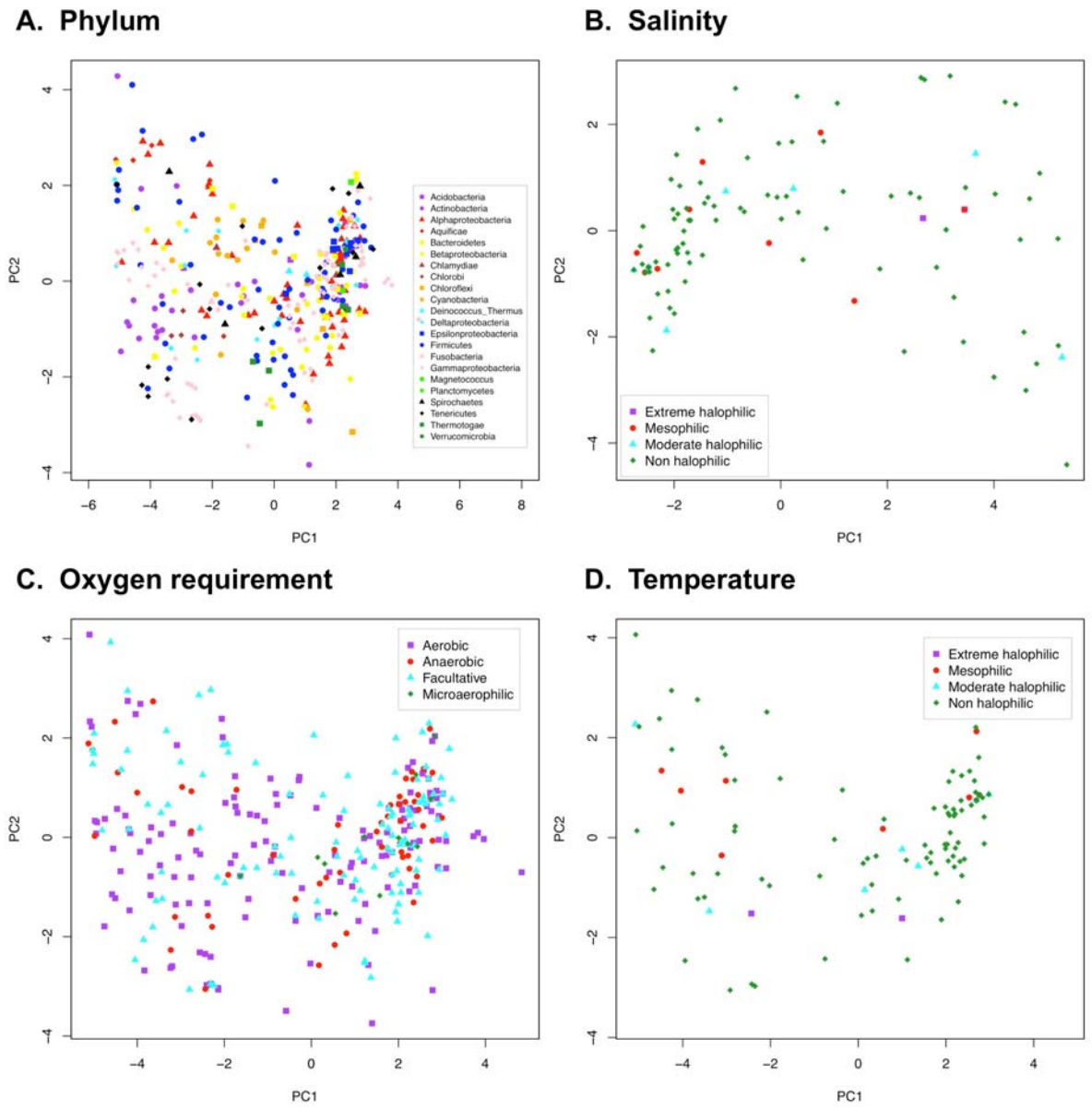
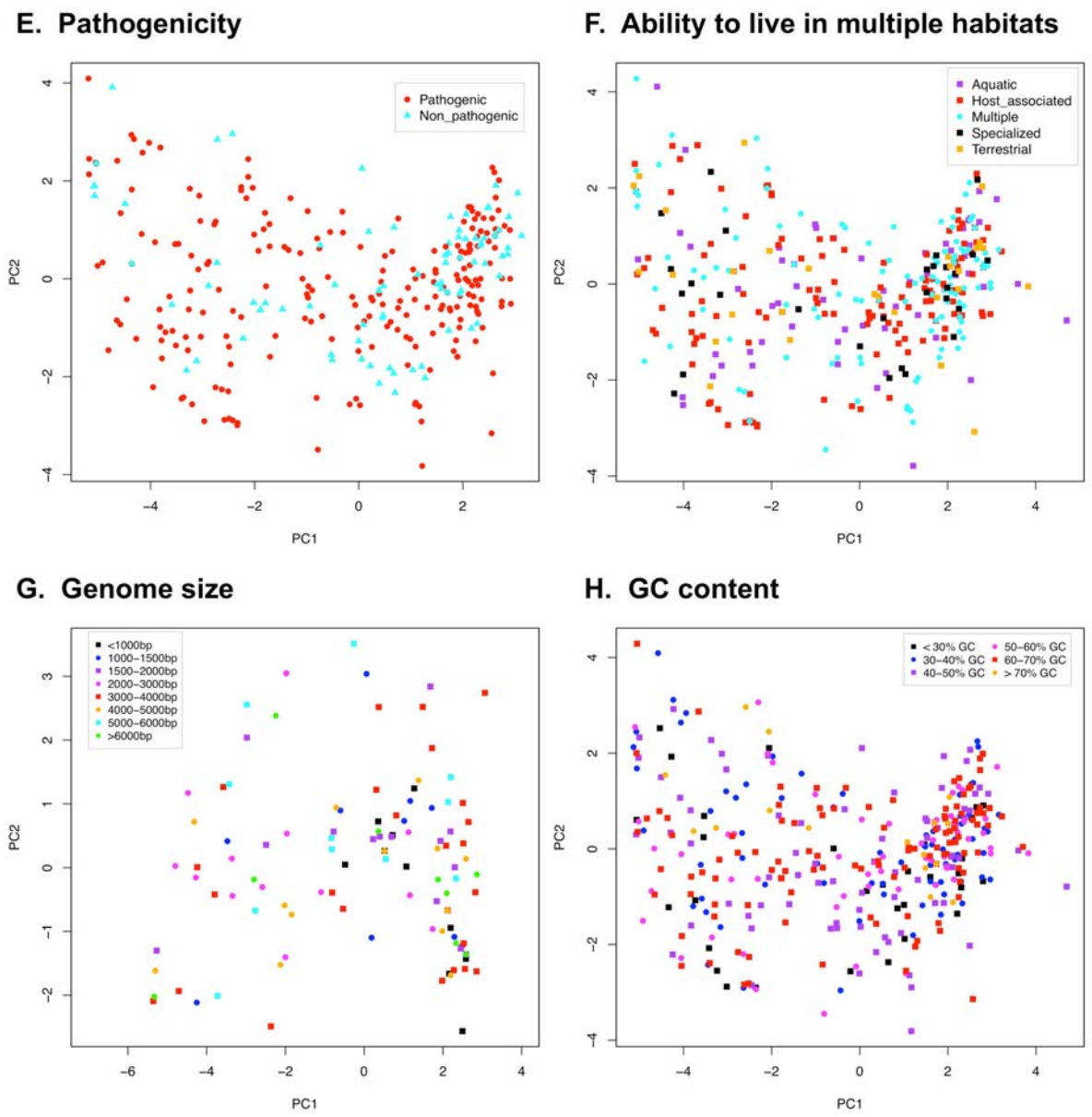


Figure 5.4. Correlation between tRNA gene content and other variables within bacterial species. A principal component analysis of the tRNA relative gene frequencies (RGF) has been performed on all bacterial species. Each dot represents a bacterial species, and has been coloured according to its phylum (A), salinity (B), oxygen requirement (C), temperature (D), pathogenicity (E), ability to live in multiple habitats (F), genome size (G) or GC content (H). It is important to remark that or each type of analysis, data were not available for all species, and therefore each analysis does not contain exactly the same subset of bacterial species.



Cont. **Figure 5.4 . Correlation between tRNA gene content and other variables within bacterial species.**

5.1.3 tRNA modifications as a novel mechanism for post-transcriptional regulation

As we begin to decipher some of the rules that govern codon usage and tRNA abundances it is becoming clear that both parameters are not just a means to increase gene expression, but also tools used by genomes to regulate the speed of protein translation, the efficiency of protein folding, and the expression of functionally related gene families.

The discovery of the importance of ADATs and UMs in translation efficiency opens new questions and opens novel research lines and biotechnological applications. More specifically, we are interested in answering the two following questions:

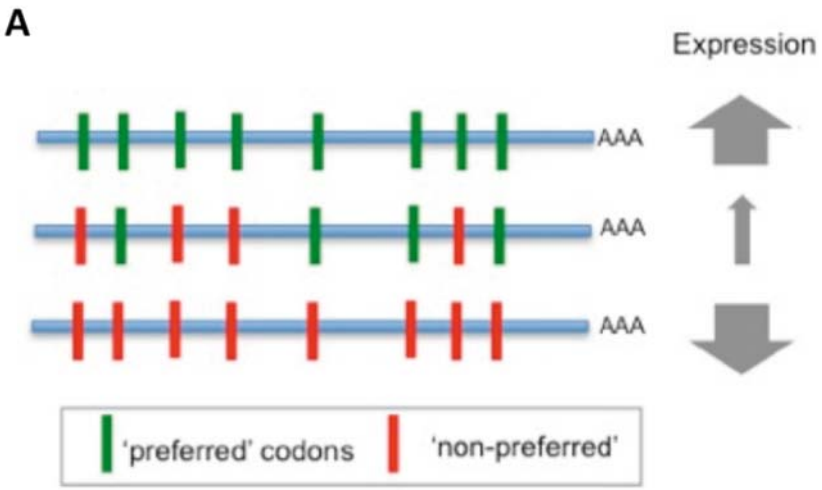
i) *Can we enhance the protein expression levels by increasing the relative abundance of codons read by ADAT- or UM-modifiable tRNAs?*
ii) *Do tRNA modification enzymes have a role in the changes in protein expression levels observed in cancer?*

i) Can we enhance the protein expression levels by increasing the relative abundance of codons read by ADAT- or UM-modifiable tRNAs?

The expression of functional proteins in heterologous hosts is a cornerstone of modern biotechnology. Proteins are often difficult to express outside their original context, due to factors such as the presence of codons that are rarely used in the desired host, or regulatory elements within their coding sequence. Differences in codon usage can impede translation due to the demand for one or more tRNAs that may be rare or lacking in the population (Kane, 1995; Goldman et al., 1995). Insufficient tRNA pools can then lead to translational stalling, premature translation termination, translation frameshifting and amino acid misincorporation (Kurland and Gallant, 1996). To overcome this limitation, codon-optimized synthetic genes can be used to enhance the expression of heterologous proteins.

Most codon-optimization algorithms attempt to mimic the codon usage bias of the host species in order to improve the protein expression levels of the heterologous protein. However, the discovery of kingdom-specific strategies based on tRNA modification enzymes to optimize translation efficiency (ADATs in Eukarya and UMs in Bacteria) opens new possibilities to further improve heterologous gene expression systems.

We suggest that, to mimic the translation efficiency strategy of the host species, *i.e.* to increase the abundance of codons read by ADAT- or UM- modifiable tRNAs (depending on whether the host species in an eukaryote or a prokaryote), can improve heterologous protein expression by helping gene compositions match the mature tRNA gene population of the host species (**Figure 5.5**).



B

First Base	U	C	A	G	Third Base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 5.5. Codon optimization strategy based on tRNA modifications. A) Representation of three mRNA sequences (blue) with different codon usage. 'Preferred' codons (green) correspond to those read by ADAT- or UM-modified tRNAs, whereas 'non-preferred' codons (red) correspond to those codons that cannot read by ADAT- or UM-modified tRNAs –and an alternative codon of the same family box would be capable of being read by an ADAT- or UM-modified tRNA. **B)** Genetic code table highlighting ADAT- 'preferred' and 'non-preferred' codons.

To experimentally verify this hypothesis, we have synthesized 2 GFP sequences with different codon usage, one of them containing codons recognizable by ADAT-modifiable tRNAs (GFP-ADAT), whereas the other contains codons non-recognizable by ADAT-modifiable tRNAs (GFP-nonADAT). The synthetic GFP sequences have been designed such that they have similar codon-adaptation index (CAI), GC-content, and codon autocorrelation (Canarozzi et al. 2010), but different percentage of ADAT-modifiable codons.

We expect to transfect each plasmid to HEK293T cells. The expression levels of GFP and ADAT will be measured both at the mRNA (qPCR) and protein level (WB and FACS). We predict that the GFP-ADAT sequence should be expressed at higher levels than the GFP-nonADAT. To verify that the differences observed in expression levels are due to ADAT, we will knockdown ADAT2. The knockdown should only decrease the levels of GFP-ADAT, but not of GFP-nonADAT (**Figure 5.6**).

Codon optimisation followed by gene synthesis is a useful but expensive for expressing problematic recombinant proteins. An alternative to this method is the use of certain strains such as Rosetta^{TM}, which contain plasmids with extra copies of certain tRNA isoacceptors. Previous works already showed that expression yields of proteins whose genes contain rare codons can be dramatlcally improved when the cognate tRNA is increased in copy number within the host, which is achieved by inserting the wild type tRNA gene on a multiple copy plasmid (Rosenberg et al., 1993; Seidel et al., 1992). For instance, the yield of human tissue plasminogen was increased 10-fold in a strain that contained $tRNA^{Arg}_{UCU}$ –which recognizes AGG/AGA codons- (Brinkmann et al., 1989). Following the results of the pRIG plasmid, which contained extra copies of the most rare tRNA genes in *E. coli* (Baca and Hol, 2000), Novagen added two tRNA genes to the pRIG plasmid to create the pRARE plasmid (**Figure 5.7**). The plasmids were transformed into various strains to create the Rosetta^{TM} series of expression hosts, and are now commonly used in many laboratories for the heterologous expression of proteins with “rare” codon usage biases.

Figure 5.6. Pipeline and expected results of the experiment. Each of the three vectors (GFP-WT, GFP-ADAT, and GFP-nonADAT) contain a GFP sequence with different codon usage bias, but similar CAI, GC-content and codon autocorrelation. After transfecting the cells, we expect to see similar levels of mRNA, but diverse levels of protein (higher in GFP-ADAT than GFP-nonADAT). After transfecting siRNA(ADAT2), we expect to see a decrease in the protein levels of GFP-ADAT and GFP-WT, but not in GFP-nonADAT. This would demonstrate that the differences in protein expression levels are due to the presence of codons recognizable by ADAT-modifiable tRNAs.

In this regard, we propose that instead of using a plasmid that adds extra copies of tRNA genes that are in lower copy number in the host, a better strategy would be the use of a plasmid that mimics the translation efficiency strategy of the original host. For instance, if we aim to express an eukaryotic protein in *E. coli*, we should mimic the eukaryotic strategy for increasing translation efficiency, which consists in the use of ADAT-modified tRNAs. We propose a novel vector, pADAT, which could be used to enhance heterologous protein expression of eukaryotic proteins in bacterial systems. This plasmid contains a bicistronic construct of the ADAT2-ADAT3 ORFs, and the tRNA isoacceptors that are used as substrates of ADAT.

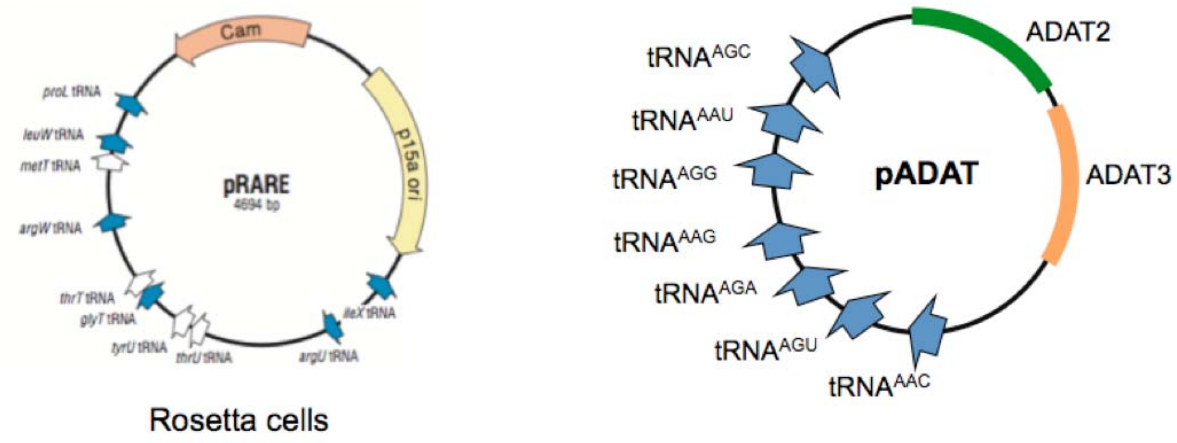


Figure 5.7. Biotechnological applications for heterologous protein expression. In the left, the structure of the pRARE plasmid (Novagen), which is found in the Rosetta™ competent cells. In the right, the proposed structure of the pADAT plasmid, which contains the ORFs encoding for ADAT2 and ADAT3 –the functional enzyme is an heterodimer-, and the tRNA genes used as ADAT2/3 substrates.

ii) Do tRNA modification enzymes have a role in the changes in protein expression levels observed in cancer?

In the last years it has become increasingly apparent that the mis-regulation of protein translation plays a critical role in cell transformation and tumorigenesis (White, 2004; Marshall and White, 2008; Chen et al., 1997). A wide variety of human tumors have been shown to depend on activated signal transduction pathways that control protein translation, *i.e.* PI 3-kinase and mTOR pathways. Indeed, inhibition of mTOR-driven protein translation by the drug rapamycin is now in clincal trials as cancer treatment.

The work presented here suggests that ADAT plays an essential role in translation efficiency. Given that protein translation is altered in cancer cells, it would be worth to characterise whether the proteins involved in these proteomic changes can be partly explained by changes in ADAT levels. Many efforts have been placed to identify the mis-translated proteins that contribute to oncogenesis. However, we suggest that besides focusing on the mis-translated proteins, it would be important to understand the potential proteins causing this mistranslation, and ADAT represents an interesting candidate for this role.

How would changes in ADAT levels affect protein translation? It would be reasonable to think that ADAT downregulation could cause a decrease in translation efficiency, and consequently, a change in the proteome composition, whereas ADAT upregulation could cause protein mistranslation, by modification of near-cognate tRNA isoacceptors that should not be modified by ADAT. Both characteristics (protein mistranslation and changes in proteome composition) are seen in cancerous cells.

To have an initial overview of the possible up/down regulation of ADAT levels in cancerous tissues, the Gene Expression Atlas was used. This database provides meta-analysis based summary statistics over a curated subset of the ArrayExpress Archive, which services queries for biologically interesting genes or samples (**Figure 5.8**). Interestingly, ADAT2 shows differential expression with respect to normal cells in many of the experiments analyzed, suggesting that it may play a role in the mis-regulation of protein translation in cancer cells.

Given that mRNA levels are not necessarily reflected in protein levels, ADAT2 levels in diverse cancer cell lines were analyzed (**Figure 5.9a**). Amongst the 9 different breast and lung cancer cell lines analyzed, we observe up to 12-fold differences in ADAT2 levels. Furthermore, when comparing Ras-transformed and non-Ras transformed mouse 3T3 fibroblasts, an 19-fold

difference in its ADAT2 protein levels is found (**Figure 5.9b**). These results suggest that the levels of ADAT2 proteins are modified in cancerous cells, and support the hypothesis of the role of ADAT2 in the mis-regulation of protein translation in cancer. These preliminar results support the hypothesis of a possible role of ADATs in protein translation mis-regulation in cancer cells. However, further work should be done to identify the precise roles of ADAT mis-regulation in cancer.

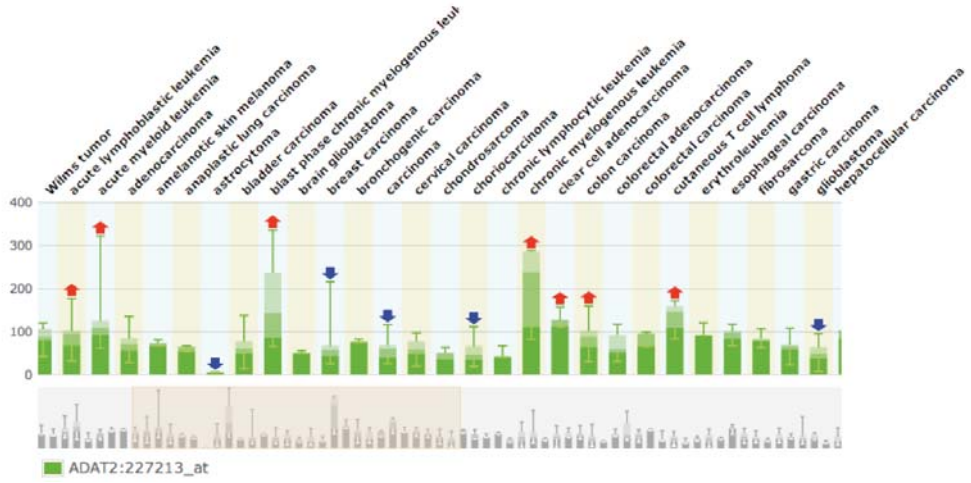


Figure 5.8. Transcription expression profiles of ADAT2 in diverse cancerous tissues. ADAT2 is found to be differentially expressed in 219 experiments, including tissues from 68 organismal parts and 166 cell lines. Red arrows indicate over-expression, whereas blue arrows indicate under-expression. Data taken from the Gene Expression Atlas (http://www.ebi.ac.uk/gxa/).

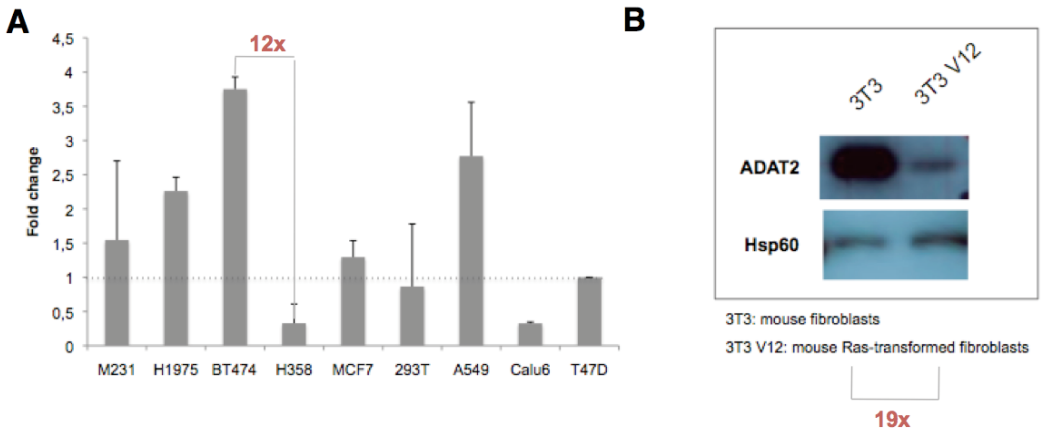


Figure 5.9. Protein expression levels of ADAT2 in cancer cells, as determined from Western Blot. A) Levels of ADAT2 (measured from WB) in diverse breast and lung cancer cell lines, normalized by tubulin levels. The units have been arbitrarily assigned (T47D cell line has been normalized to 1). B) Levels of ADAT2 in 3T3 mouse fibroblasts and Ras-transformed 3T3 mouse fibroblasts. In both A) and B), the total amount of protein was previously determined by Bradford assay in order to charge similar amounts of protein in each lane. The levels of loaded protein were checked by measuring tubulin levels and Hsp60 levels, respectively.

5.2 Aminoacyl-tRNA synthetases as antimalarial drug targets

5.2.1. Antimalarial drug discovery: old and new approaches

Resistance of malaria parasites to available drugs continues to grow, increasingly limiting our ability to control this serious disease (Nwaka, 2005). Indeed, endoperoxides are the only drug class for which clinically significant resistance has not been reported (Eastman and Fidock, 2009). Although our understanding of the parasite's biology has increased with the sequences of the *Plasmodium* genome (Gardner et al., 2002), few new drug targets or classes of drugs have been clinically validated (Munos, 2006). Major antimalarial efforts during the last years include strategies as diverse as: the use of combination therapy, the development of analogues of existing agents, the discovery of natural products, the use of compounds that were originally developed against other diseases, the evaluation of drug resistance reversers and the consideration of new chemotherapeutic targets (Rosenthal, 2003; Goodman et al., 2007; Dahl and Rosenthal, 2007; Dahl et al., 2006).

Amongst the latter, several high-throughput *in vitro* screenings against *P.falciparum* iRBCs have been recently published (Gamo et al., 2010; Guiguemde et al., 2010; Plouffe et al., 2008). A library of 2 million compounds from GlaxoSmithKline's chemical library was screened aginst *P. falciparum* cultures, from which 13.500 inhibited parasite growth and more than 8.000 also showed potent activity against a multidrug resistant strain (Gamo et al., 2010). The public availability of this large set of potent and drug-like antiplasmodial structures provides reasonable staring points for further drug development. In a similar fashion, chemical genetic approaches to assay more than 300.000 chemicals (Guiguemde et al., 2010) and more than 12.000 natural products (Plouffe et al., 2008) against *P. falciparum* iRBCs have been performed. While these approaches are extremely powerful to identify novel potent antimalarial scaffolds, the lack of knowledge of their respective drug targets constricts the hit-to-lead optimisation process in the drug development process. Thus, the major limitation at this stage of drug development is the identification of the drug targets for this large amount of chemical scaffolds that have now seen the light as novel antimalarial starting points.

Importantly, some of the positive hits provided by these high-throughput screens have been further characterised by to identify the most promising compounds (Meister et al., 2011), and finally discover its target using reverse genomic approaches (Hoepfner et al. 2012). The target identification strategy applied in this latter work (Giaver et al., 1999; Winzeler et al., 1999) constitutes a successful case in which the target of a phenotypic screen was successfully

defined by a combination of genetic and biochemical approaches. However, it is important to remark that although the employed strategy for deciphering the target was successful in this occasion, it presents serious limitations. In first place, the target must inhibit the yeast enzyme (but not the human counterpart if we want the drug to be selective). Furthermore, the confirmation of the target requires the development of resistant parasites, implying that the parasite is capable of finding in a relatively easy manner a strategy to develop resistance, which is an undesirable property for a potential drug target.

Altogether, although high-throughput screens are powerful tools for identifying novel antimalarial scaffolds, the inverse approach in drug discovery, starting from the target, is still extremely useful, and offers different advantages but also some other limitations. The major limitation for proper structure-based drug design in the case of most antimalarial targets is the lack of experimentally determined structures. Thus, we must rely on homology-based models for the initial *in silico* screenings. In this work we attempted to determine the reliability of homology models for docking purposes, in order to identify the limitations of the technique when using homology models with modest sequence identities. We find that homology models up to 30% sequence identity still provide reasonable enrichments and binding mode predictions, and thus can be used for docking purposes.

5.2.2 Our approach: aaRS as antimalarial drug targets

Plasmodial proteins can be difficult to characterize structurally using traditional *in vitro* approaches. These problems can be partially overcome using a number of *in silico* approaches. This work is a clear example which shows that the combination of both *in silico* an *in vitro* procedures can facilitate and accelerate the discovery of candidate hits.

aaRS are already the target of commercialized drugs (Bactroban, GlaxoSmithKline), and have been used for as drug targets in the search of novel antibacterials. However, until recently, the have remained unexplored as potential antimalarial drug targets. Here we have employed four different drug design strategies for discovering and developing novel antimalarials targeting plasmodial aaRS (**Figure 5.10**):

i) Analogues of the aaRS reaction intermediates
ii) Derivatives of known aaRS inhibitors
iii) High-throughput virtual screening methods for the search of novel scaffolds
iv) Inhibitors of the hinge movement required for tRNA binding

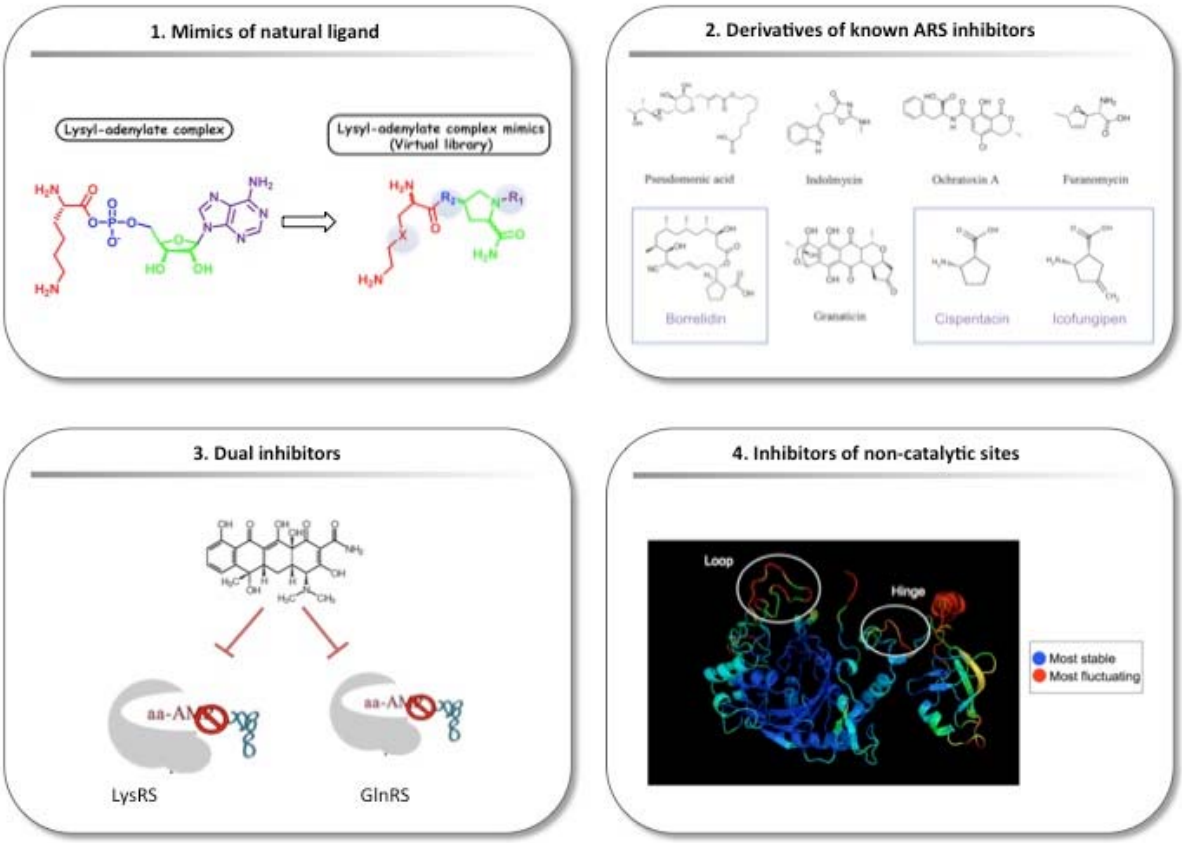


Figure 5.10. Outline of the diverse drug design strategies used in this work for the development of novel antimalarial drugs targeting *P. falciparum* aaRS. These approaches include the development of: **i)** Mimics of the aminoacyl-adenylate have been performed against PfKRS-2; **ii)** derivatives of known aaRS inhibitors targeting PfTRS; **iii)** of dual inhibitors targeting both PfKRS-2 and PfQRS identified using high-throughput virtual screens; and **iv)** inhibitors of non-catalytic sites such as the inhibition of hinge movement of PfKRS-2, which is required for proper aminoacylation.

i) Analogues of the aaRS reaction intermediates

Both using phylogenetic methods and structural comparisons, we show that PfKRS-2 and its human counterpart are distantly related. Taking advantage of the structural differences found between their catalytic sites, we designed, virtually screened, synthesized and tested a library of lysyl-adenylate analogues. Amongst the synthesized hits, we find two compounds that show clear inhibition of the PfKRS-2 enzyme. Importantly, these two compounds do not inhibit the HsKRS enzyme.

The major drawback of the two molecules that show both *in vitro* activity against PfKRS-2 and also selectivity versus its human homologue is their low potency – the IC_{50} values are 38.4 and 84.7uM, respectively-. Their low potency is probably the cause of the lack of *in vivo* activity in *P. yoelii*-infected mice. However, other plausible explanations would be a differential inhibitory

activity towards the *P. yoelii* enzyme compared to the *P. falciparum* enzyme, or the *in vivo* instability of the compounds in the blood stream, among others. It would be worth to characterise the precise failure of the drug *in vivo* to improve the potency of the compounds in a future second round of synthesis of this family of compounds.

Another limitation of this set of compounds is that their primary target is an apicoplastic enzyme. Apicoplastic proteins constitute interesting targets due to its cyanobacterial origin, making it a good choice for the development of selective drugs. Indeed, multiple antibiotics with antimalarial activity exert their effects by interfering with apicoplastic proteins, including apicoplastic ribosomes (tetracyclines, macrolides and lincosamides), apicoplastic RNA polymerases (rifampicin) or apicoplastic DNA gyrases (fluoroquinolones). However, the effects of these compounds are not seen immediately after drug treatment. Instead, they present a "delayed" death, in which the drug-treated parasites inherit a non-functional apicoplast, leading to a delayed but potent antimalarial effect during the second life cycle of the parasite, *i.e*. 98hours post-treatment. Nevertheless, doxycycline, clindamycin and azithromycin are effective and commonly used antimalarials, though slow acting. Thus, they are best used in combination with a more rapid acting drug (Borrmann et al., 2004; Noedl et al., 2006; Taylor et al., 2001).

In summary, our results validate PfKRS-2 as a druggable enzyme that can be selectively inhibited, and provide starting points for future antimalarial drug design and hit-to-lead optimisation for the generation of new drugs based on the reaction intermediate.

ii) Derivatives of known aaRS inhibitors

After initial *in vitro* screenings of a library of known aaRS inhibitors on cell cultures of *P. falciparum*-infected RBCs, we chose borrelidin as the most potent compound showing inhibitory activity, with an IC_{50} of 1nM. However, borrelidin does not show enough selectivity towards the plasmodial enzyme versus its human counterpart, and therefore cannot be used for clinical purposes. In this regard, we intended to find borrelidin derivatives that showed higher selectivity while maintaining their potency. Fortunately, some of the compounds show an increase of 10-fold in selectivity, whilst maintaining almost the same potency. Indeed, these compounds cleared the parasitemia of *P. yoelii*-infected mice using doses comparable to chloroquine, a commonly used antimalarial drug. Therefore, we suggest that these promising compounds should be further investigated and characterised –e.g. ADME properties or pharmacokinetics- as antimalarial drug candidates, in order to continue towards future clinical trials.

To further perform hit-to-lead optimisation, a reliable docking model of the drug in the active site of both PfTRS and HsTRS would be of great help. Unfortunately, the prediction of a reliable binding mode for these compounds could not be yet obtained. The binding site of the natural compound borrelidin also remains unknown, although previous studies on the *E. coli* enzyme have shown that borrelidin is a non-competitive inhibitor of threonyl-tRNA synthetase with respect to its natural substrates (Ruan et al., 2005). Thus, our future work on this project includes the prediction of the binding mode for these compounds, in order to explain the basis for their target selectivity, and use this knowledge to fasten the development of borrelidin analogues as antimalarial drugs.

5.2.3. Future work on aaRS as antimalarial drug targets

Aminoacyl-tRNA synthetases have been proposed as useful drug targets for many years (Kim et al., 2003; Schimmel et al., 1998; Ochsner et al., 2007). However, until recently, they have remained completely unexplored as antimalarial drug targets. This work aimed first to characterize the set of aaRS in *P. falciparum*, then select the best potential drug targets amongst the set of plasmodial aaRS, and finally design and evaluate candidate drugs which would selectively inhibit plasmodial aaRS.

From this work we find that plasmodial aminoacyl-tRNA synthetases are indeed druggable enzymes that can be used as antimalarial drug targets. Our results suggest that further characterization of the protein synthesis machinery in *Plasmodium falciparum* should be performed, and used for the development of new antimalarials. Importantly, we find that borrelidin analogues constitute interesting scaffolds as antimalarial drug targets, given that they are potent inhibitors both *in vitro* and *in vivo* with high selectivity towards the plasmodial enzyme versus its human counterpart.

Winzeler and colleagues (Hoepfner et al., 2012) recently identified the target of an antimalarial drug, cladosporin, which had been previously identified as inhibitor of intraerythrocytic parasites through high-throughput phenotypic screens (Plouffe et al., 2008). Interestingly, the target of cladosporin was shown to be the cytosolic lysyl-tRNA synthetase (PfKRS-1). Altogether, these findings validate aminoacyl-tRNA synthetases as drug targets for malaria and potentially in related parasitic diseases (e.g. toxoplasma, leishmania, trypanosome). The essential role of this family of enzymes in both liver and blood stages represents a tremendous opportunity for the discovery of the next generation of antimalarials.

Contents

Abstract

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

6. Summary (Spanish)

6.1. Resumen

La traducción de proteinas es un proceso central que ocurre en los tres dominios de la vida, en el cual el ARN mensajero (ARNm) es descodificado para producir un polipéptido específico, según las normas del código genético. Nuestro grupo de investigación estudia la traducción genética, y más específicamente, el mecanismo mediante el cual el ARN de transferencia (ARNt) es aminoacilado. En la reacción de aminoacilación, un amino acido particular es transferido a su ARNt específico. Las enzimas que catalizan esta reacción altamente específica son las aminoacil-ARNt sintetasas (aaRS), y son responsables de establecer el código genético. Las aaRS son el link entre los mundos proteico y de ácidos nucleicos. No es solo la relación entre la estructura y función lo que ha capturado la imaginación de los biólogos, pero tambien la posibilidad de que estas proteinas pudiesen desvelarnos los secretos del código genético. Entender el funcionamiento de estas enzimas es añadir una pieza de gran importancia al puzzle de lo que la célula es, y cómo funciona.

Este trabajo está centrado en el estudio y la caracterización de la maquinaria de traducción genética usando tanto aproximaciones *in silico* como *in vitro*, con un énfasis especial en los dos jugadores de la reacción de aminoacilación: las aaRS y los ARNt. Además, en este trabajo he caracterizado con mayor detalle la maquinaria de traducción genética de Plasmodium falciparum, la especie mñás mortal causante de la malaria, con el fin de diseñar y testear inhibidores que específicamente inhiban su maquinaria de traducción genética.

Esta tesis ha sido estructurada en tres secciones distintas, correspondientes a los diferentes proyectos que se han realizado relacionados con la caracterización de la maquinaria de traducción genética:

6.1.1. Caracterización de la maquinaria de traducción genética y su evolución en las especies

A pesar del paper central de los ARNt en la traducción de proteinas, las conexiones entre la dimámica de la población de genes de ARNt y la evolución de los genomas apenas han sido estudiadas. Además, no comprendemos por qué existen variaciones entre los pools de ARNt

entre las distintas especies, ni los principios que determinan las abundancias de ARNt o la composición de codones de los genomas.

Para entender las presiones evolutivas que dieron forma a la maquinaria de traducción genética, hemos analizado cientos de genomas desde el punto de vista de su contenido de genes de ARNt y su uso de codones. A través de nuestro analisis observamos que dos enzimas de modificación del tRNA específicas de reino contribuyeron en gran medida a la evolución de los genomas y a la aparición de los actuales usos de codones: las adenosina deaminasas dependientes de ARNt (ADATs) en Eukarya, y las uridina metiltransferasas (UMs) en Bacteria. Nuestros resultados sugieren que estas dos enzimas de modificación del tRNA ejercieron una selección positiva en sus respectivos genomas, causando una desviación hacia aquellos codones que podían ser leídos por estos ARNt modificados en genes altamente expresados. Por tanto, la abundancia de codones leidos por estos ARNt modificados en un gen correlaciona directamente con sus niveles de expresión. Esto sugiere no solo que la desviación en el uso de codones es una estrategia para regular los niveles de expresión génica, sino también que la modulación de la eficiencia de traducción tiene lugar a través del uso de modificaciones de ARNt específicas.

El descubrimiento de estrategias específicas de reino para optimizar la eficiencia de traducción abre nuevas posibilidades para mejorar la expresión heteróloga de proteinas. Además, resultados preliminares sugieren que estas modificaciones pueden tener potenciales papeles en ciertas enfermedades. Por tanto, las modificaciones de ARNt pueden no ser solo "decoraciones" de la función y la estructura de los ARNt, sino más bien toda una capa de regulación de los niveles de expresión génica.

6.1.2. Diseño de fármacos *in vitro* e *in vivo* contra la maquinaria de traducción genética de *Plasmodium falciparum*

La maquinaria de traducción genética representa una de las más útiles dianas para el desarrollo de nuevos antiinfectivos. Varias familias de antibióticos funcionan bloqueando la síntesis proteica. Y a pesar de ello, se conoce muy poco de la maquinaria de traducción genética en *Plasmodium*. En este trabajo pretendemos caracterizar la biología del ARNt en *Plasmodium falciparum*, y desarrollar screenings *in silico* e *in vitro* para seleccionar nuevos fármacos antimaláricos que inhiban la actividad de las aminoacil-ARNt sintetasas de *Plasmodium*, que son enzimas esenciales y dianas farmacológicas demostradas, y que por

tanto representan interestantes dianas nuevas en el descubrimiento de nuevos fármacos antimaláricos.

Hay tres diferentes reservorios genómicos que pueden ser traducidos en *P. falciparum*: el apicoplástico, el mitocondrial y el nuclear. Nuestros resultados predicen que hay un total de 37 aaRS codificadas en el genoma nuclear, que van dirigidas al citosol o al apicoplasto, obteniendo un set de aaRS completo en ambos compartimentos. De las 37 aaRS que se predicen, hemos decidido focalizarnos en dos de ellas como dianas antimaláricas: la lisil-ARNt sintetasa apicoplastica (PfKRS-2) y la glutaminil-ARNt sintetasa (PfQRS).

Las proteinas de *Plasmodium* son dificiles de caracterizar usando las tradicionales aproximaciones in vitro. Sin embargo, algunos de estos problemas pueden solventarse usando algunas aproximaciones in silico. Este trabajo es un claro ejemplo que demuestra que la combinación de estrategias *in silico* e *in vitro* puede facilitar y acelerar el descubrimiento de fármacos. Además, demostramos que las aminoacil-ARNt sintetasas de *Plasmodium* son enzimas que se pueden inhibir y que por tanto pueden ser usados como dianas antimalárias. En resumen, este trabajo demuestra que debe seguir caracterizandose la maquinaria de traducción genética en *Plasmodium falciparum*, y usar este conocimiento para el desarrollo de nuevos antimaláricos.

6.1.3. Desarrollo de métodos

Para desarrollar los proyectos mencionados antes, dos proyectos adicionales computacionales has sido realizados:

1. Desarrollo de un método para predecir proteinas relacionadas con patogenicidad a partir de su secuencia aminoacídica
2. Desarrollo de una estrategia de docking para determinar la predictibilidad, ratios de enriquecimiento y precisión de predicción de los modos de unión de los fármacos cuando se usan modelos de homología para hacer docking

2. Introducción

2.1. La traducción genética

La transición secuencial de información de ADN a ARNm a proteina constituye el dogma central de la biología molecular. Determina que esta información no puede ser transferida hacia atrás desde la proteina a la proteina o a ácido nucleico (Crick, 1970).

Traducir el código de 4 letras del ARN al alfabeto de 22 letras de las proteinas es una parte central de la célula. En el proceso de traducción genético, el ARNm es descodificado en el ribosoma para producir una cadena de amino acidos que después se plegará para dar lugar a una proteina activa. La maquinaria de traducción está dedicada a interpretar el código de ácidos nucleicos en un proceso que tiene dos fases. Primero, los amino acidos son unidos a sus correspondientes ARNt a través de una reacción catalizada por un grupo de proteinas conocidas como aminoacil-ARNt sintetasas (aaRS). Los aminoacil-ARNt (aa-ARNt) son llevados al ribosoma por factores de elongación, y en el ribosoma el anticodon del ARNt es encajado con el codon del ARNm, produciendose la transferencia del amino acido cargado en el ARNt a la cadena polipeptídica que se está sintetizando.

El código genético estandar, con alguna excepción, es el mismo en todas las especies. Esta compuesto de 64 tripletes distintos (codones), de los cuales 61 codifican amino acidos. Como hay más codones (64) que amino acidos (20), la mayoría de amino acidos están codificados por varios codones. Este fenómeno es conocido como degeneración del código genético. Las únicas excepciones a esta degeneración son la metionina y el triptófano.

El mecanismo mediante el cual un organismo puede leer los 61 codones fue hipotetizado por primera vez por Francis Crick en su “Hipótesis Bamboleante” (“Wobble Hypothesis”). Ahora se sabe que un ARNt tiene la capacidad de descodificar múltiples codones a través de la flexibilidad de emparejamiento entre la tercera posición (3’) del codón de ARNm y la primera posición (5’) del anticodón de ARNt, también conocido como la posición bamboleante (“wobble position”).

2.2. El ARNt

Los ARN de transferencia (ARNt) son las moléculas adaptoras que hipotetizó Crick hace más de 50 años (Crick, 1958). Como regla general, hay al menos un ARNt para cada uno de los 20 amino acidos. En muchos casos, hay múltiples isoaceptores de ARNt para un mismo amino acido, con cada uno de estos isoaceptores reconociendo diferentes o solapados subgrupos de codones para el mismo amino acido.

Los ARNt son relativamente cortos – de 75 a 95 nucleótidos- y exhiben una conservada estructura secundaria (Sprinzl et al., 1998). Esta estructura en forma de trebol contiene un brazo aceptor, un brazo D, un brazo T y un lazo T.

Los genes de ARNt suelen estar presented en multiples copias, con un número de copias de ARNt correspondiente para aquellos codones más altamente usados en el genoma. Por tanto, la distribución de isoaceptores no es uniforme, e incluso algunos isoaceptores (de los 64 posibles) estan ausentes.

En organismos unicelulares, el número de copias de un gen de ARNt correlaciona con el nivel intracelular de ARNt (Sorensen y Pedersen, 1991; Tuller et al., 2010). Por tanto, se cree que la expresión de los genes de ARNt es proporcional a su relativo número de copias en el genoma. Sin embargo, en organismos superiores, varios genes de ARNt no siguen esta regla, sugiriendo que otras variables como la epigenética pueden estar jugando en un papel en la expresión de niveles de ARNt.

Los transcritos de ARNt suren extensas modificaciones post-transcripcionales para dar lugar a un ARNt completamente funcional y maduro para ser usado en la traducción genética. Este proceso es conocido como edición de ARNt, y es esencial para la supervivencia de las células (Döring et al., 2001; Nangle et al., 2006). Hay más de 100 modificaciones post-transcripcionales identificadas en los ARNt. Algunas de estas modificaciones están distribuidas en los tres dominios de la vida, mientras que otras son específicas de dominio.

Las modificaciones de ARNt tienen diversas funciones, incluyendo:

i) Extensión y restricción de su capacidad de emparejamiento con bases
ii) Modificación de la estabilidad de la interacción codón-anticodón
iii) Mantenimiento de la pauta de lectura
iv) Efectos en la eficiencia de traducción

2.3. Las aminoacil-ARNt sintetasas

Las aminoacil-ARNt sintetasas (aaRS) son, conjuntamente con el ARNt, las protagonistas de la primera fase de la traducción de proteinas: la reacción de aminoacilación. Esta reacción ocurre en dos pasos. En el primero, el amino ácido es adenilado o "activado" por el ATP para formar aminoacil-adenilato (aa-AMP), liberando pirofosfato (PPi). Después, el amino ácido activado, que permanece unido a la enzima, es transferido al extremo 3' terminal (A76) del ARNt mediante un enlance covalente, liberándose AMP y aminoacil-ARNt.

Con notables excepciones, hay 20 aaRS en cada organismo, una para cada amino acido usado en el código genético. Estas 20 aaRS caen en dos distintas clases, según su arquitectura del centro activo (Eriani et al., 1990): las aaRS de clase I tienen una aquirectura basada en un plegamiento Rossman, mientras que las de clase II tienen una estructura de hoja beta antiparalela flanqueada por hélices alfa.

Las aaRS son una familia de proteinas multidominio. Tienen un dominio catalitico que lleva a cabo la función de aminoacilación, pero además pueden tener unidos varios dominios que llevan a cabo funciones como el incremento de la especificidad de sustrato o el incremento de la eficiencia de la aminoacilación. Además de estos dominios básicos, nuevos dominios y motivos han sido añadido progresivamente a las aaRS para expandir sus funcionalidades. Además, durante su evolución, las aaRS han experimentado numerosos eventos de duplicaciones, inserciones y eliminaciones de dominios. Las pseudo-ARS son proteinas que han resultado de estos eventos. Esta familia de parálogos de dominios de aaRS llevan a cabo muchas funciones que no siempre estan relacionadas con la traducción de proteinas (Martinis y Pang, 2007).

La mayoría de filogenias de aaRS no son consistentes con la filogenia de los organismos, es decir, que violan el patrón canónico filogenético que se encuentra en muchas otras enzimas, y que separa Archaea, Bacteria y Eukarya. Esto es debido a la abundance transferencia horizontal que ha ocurrido en la familia de aaRS.

La fidelidad de traducción de información del ARNm a la proteina es esencial para la función celular. Las sintetasas deben tener una especificidad de sustrato muy alta para evitar possibles errores, y esta se consigue mediante la especificidad de amino acido y la edición de amino acidos cargados erroneamente.

2.4. *Plasmodium falciparum*

La malaria causa 225 millones de casos clínicos al año y aproximadamente un millón de muertes anuales (OMS, 2010). La mayoría de estas muertes está causada por la especie *Plasmodium falciparum*, una de las cuatro especies de *Plasmodium* que afectan a los humanos. A pesar de que hay fármacos contra la malaria, no se ha llegado a una situación global. Los mecanismos de control actuales incluyen terapias profilácticas y terapéuticas, así como mecanismos de bloqueo de la transmisión del vector, el mosquito *Anopheles*.

El genoma de *Plasmodium falciparum* incluye ~5770 genes, de los cuales muchos no tienen homología con ningún otro organismo conocido (Gardner et al., 2002). Esto causa dificultades para describir sus funciones, pero a la vez son interesantes candidatos para identificar posibles dianas terapéuticas específicas de *Plasmodium*. El genoma de *P. falciparum* es extremadamente rica en A y T (más del 80% del genoma), causando que su uso de codones sea especialmente desviado, con un enriquecimiento extremo de aquellos codones acabados en A y T. Además, los genes de *P. falciparum* tienden a ser mucho más largos que sus homólogos en otras especies –hasta un 50% más largos- (Frugier et al., 2010). Estas inserciones corresponden en muchos casos a regiones de baja complejidad (LCRs), que están caracterizadas por repeticiones de un mismo amino ácido.

La maquinaria de traducción de *P. falciparum* es diana de importantes fármacos antimaláricos. Sin embargo, esta maquinaria está muy poco caracterizada. *P. falciparum* contiene tres compartimentos celulares con genomas: el núcleo, la mitocondria y el apicoplasto –un organelo esencial presente en todos los Apicomplexa-. Cada uno de estos compartimentos requiere su propia transcripción y traducción para su supervivencia (Jackson et al., 2011). Los tres genomas utilizan al menos una vez todos los amino acidos, y por tanto un set completo de aaRS es necesario para la traducción de estos. *Plasmodium falciparum* contiene 37 aaRS codificadas en el núcleo, y estas van dirigidas al citosol o al apicoplasto. Se cree que la mitocondria utiliza ARNt previamente aminoacilados en el citosol, que son importados y usados para la síntesis de proteinas mitocondriales.

2.5 aaRS como dianas farmacológicas

La aparición de resistencias a los antibióticos existentes requiere el desarrollo de nuevos agentes antimicrobianos dirigidos contra nuevas dianas. Las aaRS constituyen una familia de enzimas ancestrales y esenciales para la síntesis de proteinas, y por tanto, candidatas como dianas farmacológicas. El mayor requisito que debe cumplir cualquier fármaco que inhiba una aaRS es una alta selectividad hacia la enzima del patógeno en comparación con su homólogo humano. Esta selectividad es posilble, ya que las aaRS son diana de un fármaco comercializado, el ácido pseudomónico (comercializado como Bactroban, GlaxoSmithKline), que inhibe la isoleucil-ARNt sintetasa bacteriana, con una selectividad de 8000 veces en comparación con su homólogo humano.

Las estrategias de diseño de fármacos basadas en aaRS pueden ser clasificadas en:

1. Análogos de los intermediarios de la reacción de aminoacilación
2. Análogos de inhibidores naturales de aaRS
3. Fármacos que impiden la interaccion con el tRNA
4. Inhibidores de la actividad de edición de las aaRS
5. Screening virtual y inhibidores basados en estructura
6. Screening *in vitro* a alta escala

3. Objetivos

Capítulo 1: Caracterización de la maquinaria de traducción genética y su evolución en las especies

1. Caracterización de la evolución de genes de tRNA en las distintas especies

2. Identificación de las correlaciones potenciales entre el contenido de genes de ARNt y el uso de codones

3. Determinar los papeles potenciales de uso desigual de codones dentro de los genes de una especie y entre distintas especies

4. Aplicar el conocimiento adquirido para aplicaciones biotecnológicas y mejora de la comprensión de los defectos en la traducción genética en las enfermedades

Capítulo 2: Diseño de fármacos *in vitro* e *in vivo* contra la maquinaria de traducción genética de *Plasmodium falciparum*

1. Estudiar y caracterizar la maquinaria de traducción genética de *Plasmodium falciparum*, incluyendo la identificación de sus aminoacil-ARNt sintetasas y sus ARNt, uso de codones y identificación de su estrategia para maximizar su eficiencia de traducción.

2. Identificación y caracterización de nuevas dianas farmacológicas en *Plasmodium falciparum*, incluyendo el análisis filogenético, la caracterización estructural y las comparaciones entre distinas aminoacil-ARNt sintetasas y sus homólogos humanos, asi como la determinación *in vitro* de las localizaciones subcelulares de las dianas farmacológicas elegidas.

3. Investigar el uso de inhibidores de traducción genética en *Plasmodium falciparum*, generar nuevos compuestos y testearlos, usando para ello distintas estrategias de diseño de fármacos, que incluyen el diseño de fármacos basado en estructura, el screening virtual de alta escala y el uso de librerias combinatorias.

Capítulo 3: Desarrollo de métodos

1. Desarrollar un método para predecir la patogenicidad de una proteina basada en su secuencia aminoacídica

2. Cuantificar la capacidad de predicción de los modelos de homología para usos de docking

3. Desarrollar un protocolo para maximizar la predictibilidad de los modelos de homología para usos de docking

4. Publicaciones

***Publicación 1**: A role for tRNA modifications in genome structure and codon usage.*
Novoa EM, Pavon-Eternod M. Pan T and Ribas de Pouplana L.
Cell 2012, 149: 202-213

1. Papel de las modificationes de tRNA en la estructura del genoma y el uso de codones
El número de copias de genes de ARN de transferencia (ARNt) es una característica diferenciadora de los genomas, que contribuye a la eficiencia de la maquinaria de traducción, pero los principios que determinan el número de copias génicas de ARNt y la composición de codones todavía no se comprenden. En este trabajo determinamos que la aparición de dos enzimas específicas de modificación del ARNt moldearon la estructura y la composición de los genomas. A través del análisis de más de 500 genomas, identificamos dos enzimas de modificación de ARNt específicas como principales contribuyentes que causaron la separación de los genomas de Archaea, Bacteria y Eukarya en términos de su composición de genes de ARNt. Demostamos, contrariamente a observaciones anteriores, que el uso de codones y las frecuencias génicas de ARNt estan correlacionadas en todos los reinos si estas dos modificaciones son tenidas en cuenta, y que la presencia o ausencia de estas modificaciones explica los patrones de expresión génica observada en estudios previos. Finalmente, demostramos experimentalmente que los niveles de expresión de genes en humanos correlacionan con la composición genómica de codones si estas modificaciones son tenidas en cuenta.

***Publicación 2**: Speeding with control: codon usage, tRNA and ribosomes*
Novoa EM and Ribas de Pouplana L.
Trends in Genetics 2012 (in press)

2. Aumentando el control: uso de codones, ARNt y ribosomas
El uso de codones y la abundancia de ARNt son parámetros críticos para la síntesis de genes. Sin embargo, las fuerzas que determinan la desviación en el uso de codones dentro de los genomas y entre organismos distintos, al igual que las funciones de las composiciones con desviaciones de codones, están poco comprendidas. De forma similar, la composición y dinámica de las poblaciones maduras de ARNt en las células en términos de abundancias de ioaceptores, y la prevalencia y función de las modificaciones de nucleótidos de ARNt tampoco se comprenden. A medida que comenzamos a comprender las reglas que gobiernan el uso de codones y la abundancia de ARNt, es cada vez más evidente que estos parámetros no solo

son una forma de aumentar los niveles de expresión de proteinas, sino que también regulan la velocidad de traducción génica, la eficiencia de plegamiento de proteínas, y la expresión coordinada de familias de genes funcionalmente relacionadas. Aquí discutimos la importancia de las interacciones codón-anticodón en la regulación de la traducción, y remarcamos la contribución de las distribuciones no aleatorias de codones y de las modificaciones post-transcripcionales en esta regulación.

Publication 3: *Selective inhibition of an apicoplastic aminoacyl-tRNA synthetase from Plasmodium falciparum*

Hoen R*, **Novoa EM***, López A, Camacho C, Cubells L, Martin P, Bautista JM, Vieira P, Santos M, Cortes A, Ribas de Pouplana L and Royo M. (*equal contributors)

J Med Chem (under review)

3. Inhibición selectiva de una aminoacil-ARNt sintetasa apicoplástica de *Plasmodium falciparum*

La resistencia de los párasitos maláricos a los fármacos disponibles sigue creciendo, haciendo necesario el desarrollo de nuevas terapias antimaláricas. Las aminoacil-ARNt sintetasas (ARS) constituyen un conjunto de dianas prometedoras para el desarrollo de nuevos antimaláricos. Las ARS son enzimas esenciales y dianas antibacterianas demostradas, cuya naturaleza ancestral facilita el desarrollo de inhibidores específicos. El origen cianobacterial del apicoplasto, un orgánulo común a todos los Apicomplexa y que es esencial para *Plasmodium*, está reflejado en sus enzimas de tipo bacteriano (incluyendo las ARS). A pesar de su potencial para ser dianas farmacológicas, las ARS apicoplásticas permanecen inexploradas. Aquí demostramos que la inhibición selectiva de ARS apicoplásticas es posible, y describimos una serie de nuevos compuestos que presentan actividad antimalárica y que específicamente inhiben la lisil-ARNt sintetasa apicoplástica de *Plasmodium*.

Publicación 4*: Systematic study on Plasmodium falciparum aminoacyl-tRNA synthetases as antimalarial drug targets*

Camacho C, **Novoa EM**, Cubells L, Wilkinson B, Martin P, Bautista JM, Cortés A and Ribas de Pouplana L.

To be submitted

4. Estudio sistemático de las aminoacil-ARNt sintetasas de Plasmodium falciparum como dianas farmacológicas

La malaria sigue siendo un problema mayor de salud global, y la emergente resistencia a los

actuales fármacos resulta en una urgencia para el desarrollo de nuevos antimaláricos. Nuevos antibiótios para los cuales los parásitos todavía no hayan adquirido resistencia deben ser desarrollados. La traducción de proteinas es la diana de varios fármacos antimaláricos actualmente en suo. Para explorar el potencial de las aminoacil-ARNt sintetasas (ARS) como posibles dianas antimaláricas, hemos tratado cultivos de *Plasmodium falciparum* con una batteia de fármacos de ARS, y hemos comparado sus actividades. Entre los compuestos probados, la borrelidina, un inhibidor natural de la treonil-ARNt sintetasa (ThrRS) tiene un potente efecto antimalárico. A pesar de su prometedora actividad antimalárica, la borrelidina también inhibie la ThrRS humana, y es altamente tóxica para las células humanas. Para evitar este problema, hemos explorado las actividades antimaláricas de una librería de análogos de borrelidina, y hemos evaluado su citotoxicidad en células humanas. Encontramos que algunos de estos compuestos presentan mayor selectividad hacia la enzima de *Plasmodium*, mientras mantiene su actividad antiparasític tanto *in vitro* como *in vivo*. Proponemos que la borrelidina es un prometedor fármaco antimalárico que debería ser explorado en mayor profundidad en la búsqueda de nuevos fármacos antimaláricos.

Publication 5: *A genomics method to identify pathogenicity-related proteins. Application to aminoacyl-tRNA synthetase-like proteins.*
Novoa EM, Castro de Moura M, Orozco M and Ribas de Pouplana L.
FEBS Lett 2010, 584 (2): 460-466.

5. Un método genómico para la identificación de proteinas relacionadas con patogenicidad. Aplicacion a la familia de pseudo-aminoacil-ARNt sintetasas

Durante su larga evolución, las aminoacil-ARNt sintetasas (ARS) han experimentado numerosos eventos de duplicación, inserción y eliminación de dominios. Las pseudo-ARS son proteínas resultantes de estos eventos genéticos. Este grupo de polipéptidos llevan a cabo una variedad de funciones que no necesariamente tiene que estar relacionadas con la traducción de proteinas. Muchas de estas proteinas permanecen sin caracterizar. Al menos 16 diferentes pseudo-ARS han sido identificadas, pero sus funciones permanecen incomprendidas. Aquí revisamos la distribución filogenética individual de estas proteinas en bacterias, y aplicamos un nuevo método genómico para determinar su potencial implicación en patogenicidad.

Publicación 6: *Ensemble docking in homology models.*
Novoa EM, Ribas de Pouplana L, Barril X and Orozco M.
J Chem Theory Comput 2010, 6 (8): 2547-2557

6. Docking por conjunto en modelos de homología

En este trabajo presentamos una exploración sistemática de la calidad de estructuras proteicas derivadas del modelaje por homología, cuando son usadas como molde para docking a alta escala. Encontramos que las estructuras derivadas de modelaje por homología tiene amenudo una calidad similar para docking que las estructuras provinentes de cristalización, incluso en aquellos casos en los que el molde usado para crear el modelo de homología solo tiene una moderada identidad de secuencia con la proteina de interés. Hemos diseñado una estrategia de "docking por conjunto" basada en el uso de múltiples modelos de homología. El método produce resultado que son de mejor calidad que los obtenidos por una sola estructura experimental de rayos X. El uso de esta estrategia nos permite aumentar hasta cinco veces el universo de proteinas humanas que se pueden utilizar para fines de docking a alta escala, permitiendo cubrir alrededor del 75% de las dianas terapéuticas humanas.

.

Publication 7: *Small molecule docking from theoretical structural models*

Novoa EM, Ribas de Pouplana L and Orozco M.

In: "Computational Modelling of Biological Systems: From Molecules to Pathways".

Ed Springer, New York (USA) Vol 4, pp 75-96.

7. Docking de pequeñas moléculas a partir de modelos teóricos estructurales.

El docking molecular ha sido usado desde 1980 como técnica de simulación líder que facilita el diseño de fármacos y el proceso de descubrimiento de fármacos. A pesar de que el número de algoritmos de docking disponibles ha aumentado durante los últimos años, incluir la flexibilidad de proteinas y ligandos en estas simulaciones todavía es un reto pendiente. En este capítulo revisamos los actuales algoritmos de docking y sus usos, con un enfoque especial en el uso de modelos de homología para docking.

5. Discusión y conclusiones

5.1. Papel de las modificaciones de ARNt en la estructura del genoma y el uso de codones

De nuestro analisis de contenido de genes de ARNt, encontramos que el contenido de genes de ARNt ha evolucionado de una forma diferente en cada dominio de la vida, gracias a la aparición de distintas enzimas de modificación del ARNt (ADATs en Eukarya, y UMs en Bacteria). Estas modificaciones aumentan la eficiencia de traducción de un grupo de isoaceptores de ARNt, que son precisamente los que se encuentran aumentados en número de copia de genes. Debido a su aumentada eficiencia de traducción, los codones que son leidos por estos ARNt modificados son más abundantes en proteinas que tienen que expresarse a altos niveles (e.g. proteinas ribosomales), en comparación con aquellas proteinas que no se expresan a niveles tan altos.

Nuestros resultados sugieren que el contenido de genes de ARNt ha evolucionao para aumentar aquellos que son modificables por ADATs o UMs. Estas modificaciones expanden la capacidad de descodificar codones de los isoaceptores de ARNt, cosa que sugiere que la eficiencia de traducción ha sido un factor principal a la hora de determinar la evolución del contenido de genes de ARNt en las distintas especies.

El hecho de que las proteinas que tienen altos niveles de expresión tengan distintos enriquecimientos en codones que las que son expresadas en menores niveles, y que estos cambios esten relacionados con su potencial para ser leidos por ARNt modificados o no, sugiere que la actividad de las enzimas de modificación constituye un nuevo mechanismo de regulación post-transcripcional de los niveles de expresión de proteinas.

Nuestros resultados nos hicieron interesarnos por casos en los que la traducción de proteinas está alterada, como es en el caso de enfermedades como el cáncer. Es por ello que decidimos mirar si tal vez los distintos niveles de expresión de proteinas en células cancerosas podrian estar causados por cambios en los niveles de enzimas de modificación de ARNt. Resultados preliminares demuestran que las células cancerosas tienen los niveles de ADAT modificados con respecto a no cancerosas, y que estos niveles varian según el tipo de cáncer y linea celular. Estos resultados apoyan la hipótesis de un potencial papel de ADAT en la desregulación de la traducción de proteinas en cancer, y por tanto, se debería continuar con esta linea de trabajo para identificar el papel preciso de ADAT en células cancerosas.

5.2. Aminoacil-ARNt sintetasas de *Plasmodium falciparum* como dianas farmacológicas

Se necesitan nuevos antimaláricos con gran urgencia. Las aaRS son dianas de fármacos comercializados, y han sido estudiadas como dianas de muchos fármacos antibacterianos, pero no como dianas antimaláricas. En este trabajo se han utlizado distintas estrategias para el desarrollo de nuevos antimaláricos contra las aaRS de *P. falciparum*, incluyendo:

i) Análogos de los intermediarios de la reacción de aminoacilación (contra PfKRS-2)
ii) Derivados de productos naturales inhibidores de aaRS (contra PfTRS, entre otras)
iii) Screening virtual a alta escala (contra PfKRS-2 y PfQRS)
iv) Inhibidores del movimiento de hinge requerido para la unión del ARNt (contra PfKRS-2)

Recientmente, Winzeler y colaboradores (Hoepfner et al., 2012) han identificado la diana de un fármaco antimalárico, la cladosporina. Esta molécula había sido identificado como inhibidor de parásito intraeritrocíticos a través de screenings *in vitro* a gran escala (Plouffe et al., 2008), y a través de aproximaciones genómicas reversas determinaron que la diana farmacológica de esta molécula era la lisil-ARNt sintetasa citosólica (PfKRS-1).

En este trabajo demostramos que las aminoacil-ARNt sintetasas de *Plasmodium falciparum* son enzimas contra las cuales se pueden producir fármacos antimaláricos. Encontramos que los análogos de borrelidina consituyen potentes inhibidores tanto *in vitro* como *in vivo*, y que presentan una alta selectividad hacia el enzima malárico en comparación con su homólogo humano. Por tanto, debería seguirse caracterizando en mayor detalle la maquinaria de traducción genética de *P. falciparum*, y usar este conocimiento para el desarrollo de nuevos antimaláricos.

En resumen, estos resultados validan las aminoacil-ARNt sintetasas como dianas farmacológicas contra la malaria y potencialmente en enfermedades parasíticas relacionadas. El papel esencial de esta familia de enzimas tanto en la fases tanto de hígado como de sangre representa una gran oportunidad para el descubrimiento de la futura generación de antimaláricos.

Contents

Abbreviations

1. Introduction

2. Objectives

3. PhD Advisor Report

4. Publications

5. Discussion and Conclusions

6. Summary (Spanish)

7. References

7. References

Afonso, A., Neto, Z., Castro, H., Lopes, D., Alves, A.C., Tomas, A.M., and Rosario, V.D. (2010). Plasmodium chabaudi chabaudi malaria parasites can develop stable resistance to atovaquone with a mutation in the cytochrome b gene. Malar J 9, 135.

Aguero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., Carmona, S., Carruthers, I.M., Chan, A.W., Chen, F., et al. (2008). Genomic-scale prioritization of drug targets: the TDR Targets database. Nat Rev Drug Discov 7, 900-907.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136, 927-935.

Alff-Steinberger, C. (1969). The genetic code and error transmission. Proc Natl Acad Sci U S A 64, 584-591.

Ambrogelly, A., Frugier, M., Ibba, M., Soll, D., and Giege, R. (2005). Transfer RNA recognition by class I lysyl-tRNA synthetase from the Lyme disease pathogen Borrelia burgdorferi. FEBS Lett 579, 2629-2634.

Andersson, S.G., and Kurland, C.G. (1990). Codon preferences in free-living microorganisms. Microbiol Rev 54, 198-210.

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 100, 3889-3894.

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 100, 3889-3894.

Austin, J., and First, E.A. (2002). Potassium functionally replaces the second lysine of the KMSKS signature sequence in human tyrosyl-tRNA synthetase. J Biol Chem 277, 20243-20248.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14, 283-291.

Baca, A.M., and Hol, W.G. (2000). Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli. Int J Parasitol 30, 113-118.

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. Science 294, 93-96.

Barrera, L.O., and Ren, B. (2006). The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol 18, 291-298.

Berg, O.G., and Kurland, C.G. (1997). Growth rate-optimised tRNA abundance and codon usage. J Mol Biol 270, 544-550.

Bernier, S., Dubois, D.Y., Habegger-Polomat, C., Gagnon, L.P., Lapointe, J., and Chenevert, R. (2005). Glutamylsulfamoyladenosine and pyroglutamylsulfamoyladenosine are competitive inhibitors of E. coli glutamyl-tRNA synthetase. J Enzyme Inhib Med Chem 20, 61-67.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The mammalian epigenome. Cell 128, 669-681.

Bjork, G.R., Durand, J.M., Hagervall, T.G., Leipuviene, R., Lundgren, H.K., Nilsson, K., Chen, P., Qian, Q., and Urbonavicius, J. (1999). Transfer RNA modification: influence on translational frameshifting and metabolism. FEBS Lett 452, 47-51.

Borrmann, S., Adegnika, A.A., Matsiegui, P.B., Issifou, S., Schindler, A., Mawili-Mboumba, D.P., Baranek, T., Wiesner, J., Jomaa, H., and Kremsner, P.G. (2004). Fosmidomycin-clindamycin for Plasmodium falciparum Infections in African children. J Infect Dis 189, 901-908.

Boyce, J.M. (2001). MRSA patients: proven methods to treat colonization and infection. J Hosp Infect 48 Suppl A, S9-14.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. (2003). The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1, E5.

Brinkmann, U., Mattes, R.E., and Buckel, P. (1989). High-level expression of recombinant genes in Escherichia coli is dependent on the availability of the dnaY gene product. Gene 85, 109-114.

Brown, J.R. (2003). Ancient horizontal gene transfer. Nat Rev Genet 4, 121-132.

Brown, M.J., Mensah, L.M., Doyle, M.L., Broom, N.J., Osbourne, N., Forrest, A.K., Richardson, C.M., O'Hanlon, P.J., and Pope, A.J. (2000). Rational design of femtomolar inhibitors of isoleucyl tRNA synthetase from a binding model for pseudomonic acid-A. Biochemistry 39, 6003-6011.

Brown, M.V., Reader, J.S., and Tzima, E. (2010). Mammalian aminoacyl-tRNA synthetases: cell signaling functions of the protein translation machinery. Vascul Pharmacol 52, 21-26.

Brown, P., Best, D.J., Broom, N.J., Cassels, R., O'Hanlon, P.J., Mitchell, T.J., Osborne, N.F., and Wilson, J.M. (1997). The chemistry of pseudomonic acid. 18. Heterocyclic replacement of the alpha,beta-unsaturated ester: synthesis, molecular modeling, and antibacterial activity. J Med Chem 40, 2563-2570.

Brown, P., Richardson, C.M., Mensah, L.M., O'Hanlon, P.J., Osborne, N.F., Pope, A.J., and Walker, G. (1999). Molecular recognition of tyrosinyl adenylate analogues by prokaryotic tyrosyl tRNA synthetases. Bioorg Med Chem 7, 2473-2485.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. Nature 325, 728-730.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897-907.

Chan, C.T., Dyavaiah, M., DeMott, M.S., Taghizadeh, K., Dedon, P.C., and Begley, T.J. (2010). A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. PLoS Genet 6, e1001247.

Chan, C.T., Pang, Y.L., Deng, W., Babu, I.R., Dyavaiah, M., Begley, T.J., and Dedon, P.C. (2012). Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. Nat Commun 3, 937.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A 101, 3480-3485.

Chen, W., Heierhorst, J., Brosius, J., and Tiedge, H. (1997). Expression of neural BC1 RNA: induction in murine tumours. Eur J Cancer 33, 288-292.

Chiapello, H., Ollivier, E., Landes-Devauchelle, C., Nitschke, P., and Risler, J.L. (1999). Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. Nucleic Acids Res 27, 2848-2851.

Craig, I.R., Essex, J.W., and Spiegel, K. (2010). Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. J Chem Inf Model 50, 511-524.

Crick, F. (1970). Central dogma of molecular biology. Nature 227, 561-563.

Crick, F.H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-163.

Cubist Pharmaceuticals Inc (1998). Aminoacyl adenylate minics as novel antimicrobial and antiparasitic agents. US patent 5,726,195.

Curnow, A.W., Hong, K.W., Yuan, R., and Soll, D. (1997). tRNA-dependent amino acid transformations. Nucleic Acids Symp Ser, 2-4.

Curran, J.F., and Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol 209, 65-77.

Cusack, S. (1997). Aminoacyl-tRNA synthetases. Curr Opin Struct Biol 7, 881-889.

Cusack, S., Berthet-Colominas, C., Hartlein, M., Nassar, N., and Leberman, R. (1990). A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature 347, 249-255.

Dahl, E.L., and Rosenthal, P.J. (2007). Multiple antibiotics exert delayed effects against the Plasmodium falciparum apicoplast. Antimicrob Agents Chemother 51, 3485-3490.

Dahl, E.L., Shock, J.L., Shenai, B.R., Gut, J., DeRisi, J.L., and Rosenthal, P.J. (2006). Tetracyclines specifically target the apicoplast of the malaria parasite Plasmodium falciparum. Antimicrob Agents Chemother 50, 3124-3131.

Dale, T., and Uhlenbeck, O.C. (2005). Amino acid specificity in translation. Trends Biochem Sci 30, 659-665.

Desjardins, M.; Garneau, S.; Desgagnes, J.; Lacoste, L.; Yang, F.; Lapointe, J. and Chenevert, R. (1998) Glutamyl adenylate analogues are inhibitors of glutamyl-tRNA synthetase. Bioorg Chem 26, 1–13.

Deutscher, M.P. (1984). Processing of tRNA in prokaryotes and eukaryotes. CRC Crit Rev Biochem 17, 45-71.

Dittmar, K.A., Mobley, E.M., Radek, A.J., and Pan, T. (2004). Exploring the regulation of tRNA distribution on the genomic scale. J Mol Biol 337, 31-47.

Doring, V., Mootz, H.D., Nangle, L.A., Hendrickson, T.L., de Crecy-Lagard, V., Schimmel, P., and Marliere, P. (2001). Enlarging the amino acid set of Escherichia coli by infiltration of the valine coding pathway. Science 292, 501-504.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32, 5036-5044.

Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16, 287-289.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12, 640-649.

Eastman, R.T., and Fidock, D.A. (2009). Artemisinin-based combination therapies: a vital tool in efforts to eliminate malaria. Nat Rev Microbiol 7, 864-874.

El Yacoubi, B., Bailly, M., and de Crecy-Lagard, V. (2012). Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs. Annu Rev Genet.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347, 203-206.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49.

Esseiva, A.C., Naguleswaran, A., Hemphill, A., and Schneider, A. (2004). Mitochondrial tRNA import in Toxoplasma gondii. J Biol Chem 279, 42363-42368.

Evers, A., and Klebe, G. (2004). Ligand-supported homology modeling of g-protein-coupled receptor sites: models sufficient for successful virtual screening. Angew Chem Int Ed Engl 43, 248-251.

Feagin, J.E. (1992). The 6-kb element of Plasmodium falciparum encodes mitochondrial cytochrome genes. Mol Biochem Parasitol 52, 145-148.

Feng, Z.P., Zhang, X., Han, P., Arora, N., Anders, R.F., and Norton, R.S. (2006). Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Mol Biochem Parasitol 150, 256-267.

Fersht, A.R. (1981). Enzymic editing mechanisms and the genetic code. Proc R Soc Lond B Biol Sci 212, 351-379.

Fersht, A.R., and Kaethner, M.M. (1976). Mechanism of aminoacylation of tRNA. Proof of the aminoacyl adenylate pathway for the isoleucyl- and tyrosyl-tRNA synthetases from Escherichia coli K12. Biochemistry 15, 818-823.

Fichera, M.E., and Roos, D.S. (1997). A plastid organelle as a drug target in apicomplexan parasites. Nature 390, 407-409.

Finn, J.; Hill, J.; Ram, S.; Morytko, M.; Yu, X.; Gimi, R.; Silverman, J.; Stein, R.; Lim, A.; Mak, E.; Gallant, P.; Wendler, P.; Rose, S.; Stevens, A. and Keith, D. (2001) Novel antibacterial agents targeting methionyl-tRNA synthetase: a chemInformatic approach to convert HTS data into quality medicinal chemistry leads. In: Proceedings of the 41st Annual Interscience Conference on Antimicrobial Agents and Chemotherapy Chicago, Ill.

Finn, J., Mattia, K., Morytko, M., Ram, S., Yang, Y., Wu, X., Mak, E., Gallant, P., and Keith, D. (2003). Discovery of a potent and selective series of pyrazole bacterial methionyl-tRNA synthetase inhibitors. Bioorg Med Chem Lett 13, 2231-2234.

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A.M., Alako, B.T., Ehlgen, F., Ralph, S.A., Cowman, A.F., Bozdech, Z., et al. (2009). Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog 5, e1000569.

Forrest, A.K., Jarvest, R.L., Mensah, L.M., O'Hanlon, P.J., Pope, A.J., and Sheppard, R.J. (2000). Aminoalkyl adenylate and aminoacyl sulfamate intermediate analogues differing greatly in affinity for their cognate Staphylococcus aureus aminoacyl tRNA synthetases. Bioorg Med Chem Lett 10, 1871-1874.

Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F., and McFadden, G.I. (2003). Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. Science 299, 705-708.

Fraser, T.H., and Rich, A. (1975). Amino acids are not all initially attached to the same position on transfer RNA molecules. Proc Natl Acad Sci U S A 72, 3044-3048.

Frugier, M., Bour, T., Ayach, M., Santos, M.A., Rudinger-Thirion, J., Theobald-Dietrich, A., and Pizzi, E. (2010). Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Lett 584, 448-454.

Fujimura, S., and Watanabe, A. (2003). Survey of high- and low-level mupirocin-resistant strains of methicillin-resistant Staphylococcus aureus in 15 Japanese hospitals. Chemotherapy 49, 36-38.

Gamo, F.J., Sanz, L.M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.L., Vanderwall, D.E., Green, D.V., Kumar, V., Hasan, S., et al. (2010). Thousands of chemical starting points for antimalarial lead identification. Nature 465, 305-310.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419, 498-511.

Gebauer, F., and Hentze, M.W. (2004). Molecular mechanisms of translational control. Nat Rev Mol Cell Biol 5, 827-835.

Giaever, G., Shoemaker, D.D., Jones, T.W., Liang, H., Winzeler, E.A., Astromoff, A., and Davis, R.W. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. Nat Genet 21, 278-283.

Goldman, E., Rosenberg, A.H., Zubay, G., and Studier, F.W. (1995). Consecutive low-usage leucine codons block translation only when near the 5' end of a message in Escherichia coli. J Mol Biol 245, 467-473.

Goodman, C.D., Su, V., and McFadden, G.I. (2007). The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. Mol Biochem Parasitol 152, 181-191.

Gray, N.K., and Hentze, M.W. (1994). Regulation of protein synthesis by mRNA structure. Mol Biol Rep 19, 195-200.

Guiguemde, W.A., Shelat, A.A., Bouck, D., Duffy, S., Crowther, G.J., Davis, P.H., Smithson, D.C., Connelly, M., Clark, J., Zhu, F., et al. (2010). Chemical genetics of Plasmodium falciparum. Nature 465, 311-315.

Guo, M., Yang, X.L., and Schimmel, P. (2010). New functions of aminoacyl-tRNA synthetases beyond translation. Nat Rev Mol Cell Biol 11, 668-674.

Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R., Brooker, S., et al. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. PLoS Med 6, e1000048.

He, C. (2010). Grand challenge commentary: RNA epigenetics? Nat Chem Biol 6, 863-865.

Heacock, D.; Forsyth, C.J.; Shiba, K. and Musier-Forsyth, K. (1996) Synthesis and aminoacyl-tRNA synthetase inhibitory of prolyl adenylate analogs. Bioorg Chem 24, 273–289.

Hill, J.; Finn, J.; Wang, Z.; Silverman, J.; Oliver, N.; Gallant, P.; Wender, P. and Keith, D. (2001) Synthesis and activity of spirocyclic tetrahydrofurans as inhibitors of phenylalanine tRNA synthetase. In: Proceedings of the 41st Annual Interscience Conference onAntimicrobial Agents and Chemotherapy Chicago, Ill.

Hoepfner, D., McNamara, C.W., Lim, C.S., Studer, C., Riedl, R., Aust, T., McCormack, S.L., Plouffe, D.M., Meister, S., Schuierer, S., et al. (2012). Selective and specific inhibition of the plasmodium falciparum lysyl-tRNA synthetase by the fungal secondary metabolite cladosporin. Cell Host Microbe 11, 654-663.

Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R., and Zamir, A. (1965). Structure of a Ribonucleic Acid. Science 147, 1462-1465.

Hotokezaka, Y., Tobben, U., Hotokezaka, H., Van Leyen, K., Beatrix, B., Smith, D.H., Nakamura, T., and Wiedmann, M. (2002). Interaction of the eukaryotic elongation factor 1A with newly synthesized polypeptides. J Biol Chem 277, 18545-18551.

Hurdle, J.G., O'Neill, A.J., and Chopra, I. (2005). Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents. Antimicrob Agents Chemother 49, 4821-4833.

Ibba, M., and Soll, D. (2000). Aminoacyl-tRNA synthesis. Annu Rev Biochem 69, 617-650.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151, 389-409.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2, 13-34.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218-223.

Irwin, J.J., and Shoichet, B.K. (2005). ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model 45, 177-182.

Istvan, E.S., Dharia, N.V., Bopp, S.E., Gluzman, I., Winzeler, E.A., and Goldberg, D.E. (2011). Validation of isoleucine utilization targets in Plasmodium falciparum. Proc Natl Acad Sci U S A 108, 1627-1632.

Jackson, K.E., Habib, S., Frugier, M., Hoen, R., Khan, S., Pham, J.S., Ribas de Pouplana, L., Royo, M., Santos, M.A., Sharma, A., et al. (2011). Protein translation in Plasmodium parasites. Trends Parasitol 27, 467-476.

Jackson, K.E., Pham, J.S., Kwek, M., De Silva, N.S., Allen, S.M., Goodman, C.D., McFadden, G.I., de Pouplana, L.R., and Ralph, S.A. (2012). Dual targeting of aminoacyl-tRNA synthetases to the apicoplast and cytosol in Plasmodium falciparum. Int J Parasitol 42, 177-186.

Jackson, R.J., Hellen, C.U., and Pestova, T.V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 11, 113-127.

Jakubowski, H., and Fersht, A.R. (1981). Alternative pathways for editing non-cognate amino acids by aminoacyl-tRNA synthetases. Nucleic Acids Res 9, 3105-3117.

Jarvest, R.L., Berge, J.M., Berry, V., Boyd, H.F., Brown, M.J., Elder, J.S., Forrest, A.K., Fosberry, A.P., Gentry, D.R., Hibbs, M.J., et al. (2002). Nanomolar inhibitors of Staphylococcus aureus methionyl tRNA synthetase with potent antibacterial activity against gram-positive pathogens. J Med Chem 45, 1959-1962.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., et al. (1999). Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285, 1573-1576.

Kadonaga, J.T. (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116, 247-257.

Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143-155.

Kane, J.F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr Opin Biotechnol 6, 494-500.

Kapp, L.D., and Lorsch, J.R. (2004). The molecular mechanics of eukaryotic translation. Annu Rev Biochem 73, 657-704.

Karlberg, O., Canback, B., Kurland, C.G., and Andersson, S.G. (2000). The dual origin of the yeast mitochondrial proteome. Yeast 17, 170-187.

Kellner, S., Burhenne, J., and Helm, M. (2010). Detection of RNA modifications. RNA Biol 7, 237-247.

Kim, S., Lee, S.W., Choi, E.C., and Choi, S.Y. (2003). Aminoacyl-tRNA synthetases and their inhibitors as a novel family of antibiotics. Appl Microbiol Biotechnol 61, 278-288.

Kim, S.H., Sussman, J.L., Suddath, F.L., Quigley, G.J., McPherson, A., Wang, A.H., Seeman, N.C., and Rich, A. (1974). The general structure of transfer RNA molecules. Proc Natl Acad Sci U S A 71, 4970-4974.

Kim, S.Y., and Lee, J. (2003). 3-D-QSAR study and molecular docking of methionyl-tRNA synthetase inhibitors. Bioorg Med Chem 11, 5325-5331.

Kim, S.Y., Lee, Y.S., Kang, T., Kim, S., and Lee, J. (2006). Pharmacophore-based virtual screening: the discovery of novel methionyl-tRNA synthetase inhibitors. Bioorg Med Chem Lett 16, 4898-4907.

Kirillov, S., Vitali, L.A., Goldstein, B.P., Monti, F., Semenkov, Y., Makhno, V., Ripa, S., Pon, C.L., and Gualerzi, C.O. (1997). Purpuromycin: an antibiotic inhibiting tRNA aminoacylation. RNA 3, 905-913.

Kole, R., and Altman, S. (1979). Reconstitution of RNase P activity from inactive RNA and protein. Proc Natl Acad Sci U S A 76, 3795-3799.

Konishi, M., Nishio, M., Saitoh, K., Miyaki, T., Oki, T., and Kawaguchi, H. (1989). Cispentacin, a new antifungal antibiotic. I. Production, isolation, physico-chemical properties and structure. J Antibiot (Tokyo) 42, 1749-1755.

Konrad, I., and Roschenthaler, R. (1977). Inhibition of phenylalanine tRNA synthetase from Bacillus subtilis by ochratoxin A. FEBS Lett 83, 341-347.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44, 283-292.

Krab, I.M., and Parmeggiani, A. (2002). Mechanisms of EF-Tu, a pioneer GTPase. Prog Nucleic Acid Res Mol Biol 71, 513-551.

Kurland, C., and Gallant, J. (1996). Errors of heterologous protein expression. Curr Opin Biotechnol 7, 489-493.

Leberman, R., Hartlein, M., and Cusack, S. (1991). Escherichia coli seryl-tRNA synthetase: the structure of a class 2 aminoacyl-tRNA synthetase. Biochim Biophys Acta 1089, 287-298.

Lee, J., Kang, S.U., Kang, M.K., Chun, M.W., Jo, Y.J., Kwak, J.H., and Kim, S. (1999). Methionyl adenylate analogues as inhibitors of methionyl-tRNA synthetase. Bioorg Med Chem Lett 9, 1365-1370.

Lee, J., Kang, S.U., Kim, S.Y., Kim, S.E., Job, Y.J., and Kim, S. (2001). Vanilloid and isovanilloid analogues as inhibitors of methionyl-tRNA and isoleucyl-tRNA synthetases. Bioorg Med Chem Lett 11, 965-968.

Levitt, M. (1969). Detailed molecular model for transfer ribonucleic acid. Nature 224, 759-763.

Lockhart, D.J., and Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature 405, 827-836.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-964.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308-312.

Macarron, R., Mensah, L., Cid, C., Carranza, C., Benson, N., Pope, A.J., and Diez, E. (2000). A homogeneous method to measure aminoacyl-tRNA synthetase aminoacylation activity using scintillation proximity assay technology. Anal Biochem 284, 183-190.

Mahlab, S., Tuller, T., and Linial, M. (2012). Conservation of the relative tRNA composition in healthy and cancerous tissues. RNA 18, 640-652.

Mak, J., and Kleiman, L. (1997). Primer tRNAs for reverse transcription. J Virol 71, 8087-8095.

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet 39, 415-421.

Marshall, L., and White, R.J. (2008). Non-coding RNA production by RNA polymerase III is implicated in cancer. Nat Rev Cancer 8, 911-914.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29, 291-325.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99, 12246-12251.

Martinis, S.A., and Joy Pang, Y.L. (2007). Jekyll & Hyde: evolution of a superfamily. Chem Biol 14, 1307-1308.

Mata, J., Marguerat, S., and Bahler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. Trends Biochem Sci 30, 506-514.

McFadden, G.I., and Roos, D.S. (1999). Apicomplexan plastids as drug targets. Trends Microbiol 7, 328-333.

McFadden, G.I., and Roos, D.S. (1999). Apicomplexan plastids as drug targets. Trends Microbiol 7, 328-333.

McGovern, S.L., and Shoichet, B.K. (2003). Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46, 2895-2907.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol 222, 851-856.

Mehlin, C., Boni, E., Buckner, F.S., Engel, L., Feist, T., Gelb, M.H., Haji, L., Kim, D., Liu, C., Mueller, N., et al. (2006). Heterologous expression of proteins from Plasmodium falciparum: results from 1000 genes. Mol Biochem Parasitol 148, 144-160.

Meister, S., Plouffe, D.M., Kuhen, K.L., Bonamy, G.M., Wu, T., Barnes, S.W., Bopp, S.E., Borboa, R., Bright, A.T., Che, J., et al. (2011). Imaging of Plasmodium liver stages to drive next-generation antimalarial drug discovery. Science 334, 1372-1377.

Menard, R. (2005). Medicine: knockout malaria vaccine? Nature 433, 113-114.

Munos, B. (2006). Can open-source R&D reinvigorate drug research? Nature Rev Drug Discov 5, 723-729.

Nagel, G.M., and Doolittle, R.F. (1991). Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc Natl Acad Sci U S A 88, 8121-8125.

Nangle, L.A., Motta, C.M., and Schimmel, P. (2006). Global effects of mistranslation from an editing defect in mammalian cells. Chem Biol 13, 1091-1100.

Nass, G., Poralla, K., and Zahner, H. (1969). Effect of the antibiotic Borrelidin on the regulation of threonine biosynthetic enzymes in E. coli. Biochem Biophys Res Commun 34, 84-91.

Noedl, H., Krudsood, S., Chalermratana, K., Silachamroon, U., Leowattana, W., Tangpukdee, N., Looareesuwan, S., Miller, R.S., Fukuda, M., Jongsakul, K., et al. (2006). Azithromycin combination therapy with artesunate or quinine for the treatment of uncomplicated Plasmodium falciparum malaria in adults: a randomized, phase 2 clinical trial in Thailand. Clin Infect Dis 43, 1264-1271.

Nwaka, S. (2005). Drug discovery and beyond: the role of public-private partnerships in improving access to new malaria medicines. Trans R Soc Trop Med Hyg 99 Suppl 1, S20-29.

O'Donoghue, P., and Luthey-Schulten, Z. (2003). On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol Mol Biol Rev 67, 550-573.

Ochsner, U.A., Sun, X., Jarvis, T., Critchley, I., and Janjic, N. (2007). Aminoacyl-tRNA synthetases: essential and still promising targets for new anti-infective agents. Expert Opin Investig Drugs 16, 573-593.

Ogilvie, A., Wiebauer, K., and Kersten, W. (1975). Inhibition of leucyl-transfer ribonucleic acid synthetasymol. Biochem J 152, 511-515.

Painter, H.J., Morrisey, J.M., Mather, M.W., and Vaidya, A.B. (2007). Specific role of mitochondrial electron transport in blood-stage Plasmodium falciparum. Nature 446, 88-91.

Park, S.G., Schimmel, P., and Kim, S. (2008). Aminoacyl tRNA synthetases and their connections to disease. Proc Natl Acad Sci U S A 105, 11043-11049.

Pavon-Eternod, M., Gomes, S., Geslain, R., Dai, Q., Rosner, M.R., and Pan, T. (2009). tRNA over-expression in breast cancer and functional consequences. Nucleic Acids Res 37, 7268-7280.

Persson, B.C. (1993). Modification of tRNA as a regulatory device. Mol Microbiol 8, 1011-1016.

Peterkofsky, A., Litwack, M., and Marmor, J. (1971). Modified bases and transfer RNA function. Cancer Res 31, 675-678.

Pino, P., Aeby, E., Foth, B.J., Sheiner, L., Soldati, T., Schneider, A., and Soldati-Favre, D. (2010). Mitochondrial translation in absence of local tRNA aminoacylation and methionyl tRNA Met formylation in Apicomplexa. Mol Microbiol 76, 706-718.

Pisarev, A.V., Hellen, C.U., and Pestova, T.V. (2007). Recycling of eukaryotic posttermination ribosomal complexes. Cell 131, 286-299.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12, 32-42.

Plouffe, D., Brinker, A., McNamara, C., Henson, K., Kato, N., Kuhen, K., Nagle, A., Adrian, F., Matzen, J.T., Anderson, P., et al. (2008). In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. Proc Natl Acad Sci U S A 105, 9059-9064.

Pohlmann, J., and Brotz-Oesterhelt, H. (2004). New aminoacyl-tRNA synthetase inhibitors as antibacterial agents. Curr Drug Targets Infect Disord 4, 261-272.

Putz, J., Giege, R., and Florentz, C. (2010). Diversity and similarity in the tRNA world: overall view and case study on malaria-related tRNAs. FEBS Lett 584, 350-358.

Qin, H., Wu, W.B., Comeron, J.M., Kreitman, M., and Li, W.H. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 168, 2245-2260.

Quinn, C.L., Tao, N., and Schimmel, P. (1995). Species-specific microhelix aminoacylation by a eukaryotic pathogen tRNA synthetase dependent on a single base pair. Biochemistry 34, 12489-12495.

Raczniak, G., Ibba, M., and Soll, D. (2001). Genomics-based identification of targets in pathogenic bacteria for potential therapeutic and diagnostic use. Toxicology 160, 181-189.

Rao, V.R., Ramanjeneyulu, R., Rao, D.M., and Kumar, C.S. (2006). Comparative modeling of class 1 lysyl tRNA synthetase from Treponema pallidum. Bioinformation 1, 81-82.

Reuven, N.B., and Deutscher, M.P. (1993). Multiple exoribonucleases are required for the 3' processing of Escherichia coli tRNA precursors in vivo. FASEB J 7, 143-148.

Ribas de Pouplana, L., and Schimmel, P. (2001). Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. Cell 104, 191-193.

Ridley, R.G. (2002). Medical need, scientific opportunity and the drive for antimalarial drugs. Nature 415, 686-693.

Rock, F.L., Mao, W., Yaremchuk, A., Tukalo, M., Crepin, T., Zhou, H., Zhang, Y.K., Hernandez, V., Akama, T., Baker, S.J., et al. (2007). An antifungal agent inhibits an aminoacyl-tRNA synthetase by trapping tRNA in the editing site. Science 316, 1759-1761.

Rosenberg, A.H., Goldman, E., Dunn, J.J., Studier, F.W., and Zubay, G. (1993). Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. J Bacteriol 175, 716-722.

Rosenthal, P.J. (2003). Antimalarial drug discovery: old and new approaches. J Exp Biol 206, 3735-3744.

Rould, M.A., Perona, J.J., Soll, D., and Steitz, T.A. (1989). Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science 246, 1135-1142.

Rovira-Graells, N., Gupta, A.P., Planet, E., Crowley, V.M., Mok, S., Ribas de Pouplana, L., Preiser, P.R., Bozdech, Z., and Cortes, A. (2012). Transcriptional variation in the malaria parasite Plasmodium falciparum. Genome Res 22, 925-938.

Roy, A., Cox, R.A., Williamson, D.H., and Wilson, R.J. (1999). Protein synthesis in the plastid of Plasmodium falciparum. Protist 150, 183-188.

Ruan, B., Bovee, M.L., Sacher, M., Stathopoulos, C., Poralla, K., Francklyn, C.S., and Soll, D. (2005). A unique hydrophobic cluster near the active site contributes to differences in borrelidin inhibition among threonyl-tRNA synthetases. J Biol Chem 280, 571-577.

Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.C., and Moras, D. (1991). Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science 252, 1682-1689.

Ryckelynck, M., Giege, R., and Frugier, M. (2005). tRNAs and tRNA mimics as cornerstones of aminoacyl-tRNA synthetase regulations. Biochimie 87, 835-845.

Salinas, T., Duchene, A.M., and Marechal-Drouard, L. (2008). Recent advances in tRNA mitochondrial import. Trends Biochem Sci 33, 320-329.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8, 424-436.

Sander, C., and Scheider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 9, 56-68.

Sauerwald, A., Zhu, W., Major, T.A., Roy, H., Palioura, S., Jahn, D., Whitman, W.B., Yates, J.R., 3rd, Ibba, M., and Soll, D. (2005). RNA-dependent cysteine biosynthesis in archaea. Science 307, 1969-1972.

Schafferhans, A., and Klebe, G. (2001). Docking ligands onto binding site representations derived from proteins built by homology modelling. J Mol Biol 307, 407-427.

Schimmel, P., and Ribas de Pouplana, L. (1995). Transfer RNA: from minihelix to genetic code. Cell 81, 983-986.

Schimmel, P., and Schmidt, E. (1995). Making connections: RNA-dependent amino acid recognition. Trends Biochem Sci 20, 1-2.

Schimmel, P., Tao, J., and Hill, J. (1998). Aminoacyl tRNA synthetases as targets for new anti-infectives. FASEB J 12, 1599-1609.

Schuman, E.M., Dynes, J.L., and Steward, O. (2006). Synaptic regulation of translation of dendritic mRNAs. J Neurosci 26, 7143-7146.

Seidel, H.M., Pompliano, D.L., and Knowles, J.R. (1992). Phosphonate biosynthesis: molecular cloning of the gene for phosphoenolpyruvate mutase from Tetrahymena pyriformis and overexpression of the gene product in Escherichia coli. Biochemistry 31, 2598-2608.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15, 1281-1295.

Shatkin, A.J. (1976). Capping of eucaryotic mRNAs. Cell 9, 645-653.

Sissler, M., Eriani, G., Martin, F., Giege, R., and Florentz, C. (1997). Mirror image alternative interaction patterns of the same tRNA with either class I arginyl-tRNA synthetase or class II aspartyl-tRNA synthetase. Nucleic Acids Res 25, 4899-4906.

SmithKline Beecham PLC (1999). Quinolones used as MRS inhibitors and bactericides. WO 99/55677

SmithKline Beecham PLC (2000a). 2NH-pyridones and pyrimidones as MRS inhibitors. WO 00/71524 A.

SmithKline Beecham PLC (2000b). Benzimidazole derivatives and their use as methionyl-tRNA synthetase inhibitors. WO 00/71522 A1.

Sorensen, M.A., and Pedersen, S. (1991). Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J Mol Biol 222, 265-280.

Sprinzl, M., and Cramer, F. (1975). Site of aminoacylation of tRNAs from Escherichia coli with respect to the 2'- or 3'-hydroxyl group of the terminal adenosine. Proc Natl Acad Sci U S A 72, 3049-3053.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 26, 148-153.

Stoletzki, N., and Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol 24, 374-381.

Sukuru, S.C., Crepin, T., Milev, Y., Marsh, L.C., Hill, J.B., Anderson, R.J., Morris, J.C., Rohatgi, A., O'Mahony, G., Grotli, M., et al. (2006). Discovering new classes of Brugia malayi asparaginyl-tRNA synthetase inhibitors and relating specificity to conformational change. J Comput Aided Mol Des 20, 159-178.

Suzuki, Y., Gojobori, T., and Nei, M. (2001). ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics 17, 660-661.

Sylvers, L.A., Rogers, K.C., Shimizu, M., Ohtsuka, E., and Soll, D. (1993). A 2-thiouridine derivative in tRNAGlu is a positive determinant for aminoacylation by Escherichia coli glutamyl-tRNA synthetase. Biochemistry 32, 3836-3841.

Tamura, K., Himeno, H., Asahara, H., Hasegawa, T., and Shimizu, M. (1992). In vitro study of E.coli tRNA(Arg) and tRNA(Lys) identity elements. Nucleic Acids Res 20, 2335-2339.

Tanaka, K., Tamaki, M., and Watanabe, S. (1969). Effect of furanomycin on the synthesis of isoleucyl-tRNA. Biochim Biophys Acta 195, 244-245.

Tao, J., and Schimmel, P. (2000). Inhibitors of aminoacyl-tRNA synthetases as novel anti-infectives. Expert Opin Investig Drugs 9, 1767-1775.

Taylor, W.R., Widjaja, H., Richie, T.L., Basri, H., Ohrt, C., Tjitra, Taufik, E., Jones, T.R., Kain, K.C., and Hoffman, S.L. (2001). Chloroquine/doxycycline combination versus chloroquine alone, and doxycycline alone for the treatment of Plasmodium falciparum and Plasmodium vivax malaria in northeastern Irian Jaya, Indonesia. Am J Trop Med Hyg 64, 223-228.

Thompson, D.M., and Parker, R. (2009). Stressing out over tRNA cleavage. Cell 138, 215-219.

Tomita, K., and Weiner, A.M. (2001). Collaboration between CC- and A-adding enzymes to build and repair the 3'-terminal CCA of tRNA in Aquifex aeolicus. Science 294, 1334-1336.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141, 344-354.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput Biol 3, e248.

Vaughan, M.D., Sampson, P.B., Daub, E., and Honek, J.F. (2005). Investigation of bioisosteric effects on the interaction of substrates/ inhibitors with the methionyl-tRNA synthetase from Escherichia coli. Med Chem 1, 227-237.

Walhout, A.J. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Res 16, 1445-1454.

Wells, T.N., Alonso, P.L., and Gutteridge, W.E. (2009). New medicines to improve control and contribute to the eradication of malaria. Nat Rev Drug Discov 8, 879-891.

Werner, R.G., Thorpe, L.F., Reuter, W., and Nierhaus, K.H. (1976). Indolmycin inhibits prokaryotic tryptophanyl-tRNA ligase. Eur J Biochem 68, 1-3.

White, N.J. (2011). Determinants of relapse periodicity in Plasmodium vivax malaria. Malar J 10, 297.

White, R.J. (2004). RNA polymerase III transcription and cancer. Oncogene 23, 3208-3216.

Winkler, M.E. (1998). Genetics and regulation of base modification in the tRNA and rRNA of prokaryotes and eukaryotes. In: Modification and Editing of RNA, ed. H Grosjean, R Benne, pp. 441-69. Washington D.C.: ASM Press.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901-906.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74, 5088-5090.

Woese, C.R., Olsen, G.J., Ibba, M., and Soll, D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64, 202-236.

Yadavalli, S.S., Musier-Forsyth, K., and Ibba, M. (2008). The return of pretransfer editing in protein synthesis. Proc Natl Acad Sci U S A 105, 19031-19032.

Yi, C., and Pan, T. (2011). Cellular dynamics of RNA modification. Acc Chem Res 44, 1380-1388.

Yoshida, N., Oeda, K., Watanabe, E., Mikami, T., Fukita, Y., Nishimura, K., Komai, K., and Matsuda, K. (2001). Protein function. Chaperonin turned insect toxin. Nature 411, 44.

Yu, X.Y., Hill, J.M., Yu, G., Wang, W., Kluge, A.F., Wendler, P., and Gallant, P. (1999). Synthesis and structure-activity relationships of a series of novel thiazoles as inhibitors of aminoacyl-tRNA synthetases. Bioorg Med Chem Lett 9, 375-380.

Yu, X.Y., Hill, J.M., Yu, G., Wang, W., Kluge, A.F., Wendler, P., and Gallant, P. (1999). Synthesis and structure-activity relationships of a series of novel thiazoles as inhibitors of aminoacyl-tRNA synthetases. Bioorg Med Chem Lett 9, 375-380.

Zhang, G., Fedyunin, I., Miekley, O., Valleriani, A., Moura, A., and Ignatova, Z. (2010). Global and local depletion of ternary complex limits translational elongation. Nucleic Acids Res 38, 4778-4787.