UNIVERSITAT DE BARCELONA

Exploring structure-function relationship of the mitochondrial DNA packaging protein Abf2p and its dialogue with the DNA

Arka Chakraborty

Exploring structure-function relationship of the mitochondrial DNA packaging protein Abf2p and its dialogue with the DNA

ARKA CHAKRABORTY

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

Exploring structure-function relationship of the mitochondrial DNA packaging protein Abf2p and its dialogue with the DNA

ARKA CHAKRABORTY, 2016

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

PROGRAMA DE DOCTORAT

Biotecnología

Exploring structure-function relationship of the mitochondrial DNA packaging protein Abf2p and its dialogue with the DNA

Memòria presentada per Arka Chakraborty per optar al títol de doctor per la universitat de Barcelona

Dra. Maria Solà Vilarrubias	Arka Chakraborty	Dra. Josefa Badia Palacín
Director de tesi	Doctorand	Tutora de tesi

ARKA CHAKRABORTY, 2016

Don't you have time to think?

-Richard P. Feynman

Acknowledgements

I start by thanking space and time for existing as such with all its richness and symmetries and the gargantuan brains of the past and present in whose footsteps we follow in our quest for truth and harmony, before entropy does its job.

The last few years in Barcelona and the work that leads to this thesis are intertwined with contributions from several people, both scientific and humanitarian.

I would like to start by expressing my gratitude towards my thesis supervisor Dr. Maria Sola, who has bestowed her guidance and patience on me during the doctoral years. Whatever little I have contributed is clearly a function of the freedom I received from her.

I would like to thank my tutor Dr. Josefa Badia for her immense help and kind disposition.

I extend my heartfelt gratitude to Dr. Ramon Eritja for his down to earth nature and extensive help.

A special thanks to Dr. Josep Vilardell for being a great mentor and collaborator; to Dr. Federica Battistini (Fede) for collaboration and gossips; Dr. Rafel Prohens for help with ITC experiments and collaboration; Dr. Raimundo Gargallo for his kind disposition and help with CD experiments and Giorgio Medici (Masters student) for his strong efforts during his stay at the lab.

I have imbibed a lot from other group leaders at the Structural Biology department at IBMB. I specially thank Dr. Isabel Uson for always answering my doubts and her appreciation. I thank Dr. Xavier Gomis-Ruth from whom I indirectly learned a lot and whom I hold in high respect; Dr. Ignacio Fita for being an inspiration and being one of the most down to earth persons I have ever known and Dr. Nuria Verdaguer who didn't mind the incessant beeps of the centrifuge left unattended.

I am heavily indebted to the people of my lab who not only showed me the path to tread on but were a family far away from home. Specially, my seniors Pablo and Anna Rubio who are a big brother and sister to me and the emotional support and friendship that they offered me (and still do). I feel lucky that I met people of such great character and humanitarian qualities; Seb (Dr. Sebastien Lyonnais) whom I hold in high respect for his amiable nature and broad vision as a scientist and who provided the much needed support and encouragement during times of crisis; Reicy (Dr. Reicy Brito), one of the sweetest persons on earth and has always been there when I needed empathy and help; Cuppi (Anna Cuppari-Siciliana-lost in Kaluza dimension) has been a close friend and confidant; Claus (the crazy German) with time became easier to get along (read humour) and is a good friend; Cris (Cristina Silva) and Javi (Javier Bermudes Morales) who extended their friendship and without whose organization the lab would have been lost; last but not the least Aleix, the new one,who is a very bright kid who will contribute greatly to science in the years to come and in the short time that he has been with the lab, has become a close friend and cherished companion.

My colleagues from other laboratories have always been a great help, some are close friends and some are seniors I was blessed to have. With the risk of forgetting a few names I would like to extend my gratitude to Roeland (specially for his skiing lessons), Albert, Montse, Rosa (for allowing me unplanned

purifications on her Akta), Cristina (Cri1), David, Oriol, Marietta, Damia, Cristina (Cri6), Laia, Monica, Salvadorre, Luca (Cri2), Sergio, Pedro, Pablo Nuevo, Irene, Theo, Tibi (who loves to hug the calf), Inyaki, Laura I, Mariana, Laura II, Diego, Anne, Mireia and Jorge (for their friendship and excellent yeast genomic DNA samples), Tiago, Mads (not for breaking the Cartesian robot), Massimo, Claudia, Alfonso (for all those soft-wares he installed for me), Giovanna, Rafel (the Brazilian who showed us chu-chu), Luca (Cri4). Specially, I would like to thank Mar for being a supportive friend and helping me out in wading through the bureaucracy related to the thesis.

I cannot thank enough people from the Crystallography and Protein Purification platform: Joan (Dr. Joan Pous), Sonia (a close friend), Xandra, Roman, Isabel.

There are people around me who extend their help everyday e.g. the gentlemen at the security, the ladies at the different house-keeping facilities etc. without whom conducting experiments swiftly and safely will be impossible. A huge gratitude towards all of them.

Finally, I would like to thank the European Commission for the Marie Curie ITN RAPID fellowship that funded my PhD research and provided an excellent opportunity for interaction with researchers from different countries, thus fostering my scientific development.

Contents

Abstract

ABSTRACT (ENGLISH)

Mitochondria are intracellular double-membrane bound organelles in eukaryotic cells that act as the major suppliers of adenosine triphosphate (ATP). They possess their own DNA (mtDNA) that codes for components of the oxidative phosphorylation (OXPHOS) pathway. mtDNA is assembled into nucleo-protein structures called nucleoids and maintained differently compared to histone mediated packaging of nuclear DNA. The molecular basis of mtDNA packaging and maintenance remains poorly understood.

In *Saccharomyces cerevisiae* (budding yeast) mtDNA is a ~80kb linear molecule, packaged by Abf2p, a double-HMG-box DNA binding protein. Abf2p interacts with DNA in a non-sequence-specific manner, but displays a distinct and yet unexplained 'phased-binding' at specific AT-rich DNA stretches containing poly-adenine tracts (A-tracts). Molecular details of DNA binding and maintenance by this protein as well as the mechanism behind its 'phased binding' behavior remain to be elucidated. In this doctoral thesis, crystal structures of Abf2p in complex with mtDNA derived fragments bearing A-tracts are presented. That reveal that Abf2p binds and induces 180^{o} U-turn bends in the DNA. Additionally, it avoids binding to A-tracts, giving rise to a unique 'dual binding' phenomenon where a single protein molecule binds two DNAs simultaneously. To probe the functional roles played by the different protein structural parts in vitro and in vivo assays were carried out that revealed that a 12-residue N-terminal helix, unique to this protein, is crucial for its DNA binding activity. The dynamics of the protein and protein-DNA complex were probed via in-solution (Small Angle X-ray Scattering or SAXS) and simulation (Molecular dynamics or MD) techniques that revealed key mechanisms pertaining to the DNA binding event. Additional computational analysis of Abf2p binding on A-tract containing DNA revealed a DNA-structure mediated protein positioning mechanism. The said mechanism would play a key role in orchestrating global nucleoid architecture, given that *S. cerevisiae* mtDNA has a high percentage of A-tracts. Additionally, the crystal structures disclose an inherent capability of the protein to bind separate DNA strands, that would facilitate DNA packaging by this protein and form an essential mechanistic feature of the process. The findings reported here thus advance our understanding of mtDNA packaging in the yeast mitochondria.

RESUM (CATALAN)

Els mitocondris posseeixen un ADN (ADNmt) que codifica components de la via de la fosforilació oxidativa. L'ADNmt es compacta en unes estructures nucleo-proteiques, els nucleoides, que s'estructuren de manera diferent a l'ADN nuclear. La base molecular de l'empaquetament de l'ADNmt és desconegut.

A *Saccharomyces cerevisiae* l'ADNmt és una molècula lineal d'uns 80kb empaquetada per la proteïna Abf2p, que conté dos dominis HMG-box d'unió a ADN. Abf2p contacta l'ADN de forma no específica, però també mostra una unió en fase en regions riques en poli-adenina (regions poly-A). Els detalls moleculars d'aquests dos tipus d'unió encara no s'han dilucidat. En aquesta tesi doctoral es presenten les estructures cristal·logràfiques de l'Abf2p en complex amb fragments d'ADNmt derivats de l'ADN de llevat, que demostren que Abf2p uneix i indueix una curvatura de 180º a l'ADN. A més a més, en els cristalls, l'Abf2p evita la unió a una regió poly-A induïnt un fenomen únic de d'unió d'una molècula de proteïna a dues molècules d'ADN. Per investigar la funció dels diferents dominis d'Abf2p en la unió ADN hem dut a terme assajos *in vitro* i *in vivo* amb fragments i amb la proteïna sencera que mostren que una hèlix de 12 residus N-terminal, única per aquesta proteïna, és crucial per a la unió. A més a més hem estudiat la dinàmica del complex proteïna-ADN en solució per mètodes biofísics (SAXS) que demostren la flexibilitat de la proteïna i que corroboren el condicionament de la regió poly-A en la unió. Finalment, per dinàmica molecular (MD) hem descobert que l'ADN utilitzat per cristal·litzar té unes propietats estructurals en les regions poly-A, amb un solc menor molt estret, que condicionen el posicionament d'Abf2p. Aquest fenomen és clau en l'organització de l'arquitectura global del nucleoide, atès que en *S. cerevisiae* l'ADNmt té fins al 30% de regions poly-A, atípic en altres genomes. A més, les estructures cristal·lines mostren la capacitat inherent de la proteïna per unir molècules d'ADN independents, que podrien facilitar el seu empaquetament. Els resultats aquí presentats són un avenç en la nostra comprensió de l'empaquetament de l'ADNmt en el llevat.

Introduction

1. Mitochondria and Endosymbiosis

Mitochondria are intracellular double-membrane bound organelles in eukaryotic cells that act as the major suppliers of adenosine triphosphate (ATP) [1, 2]. ATPs are synthesized via participation of the tri-carboxylic acid (TCA) cycle and the electron-transport system (ETS) [3] housed in the mitochondria and serve as a source of chemical energy for driving cellular processes. Thus, mitochondria have been aptly named the 'power house' of the cell. In addition, they are the major suppliers for the intracellular electron carrier NADH (reduced nicotinamide adenine dinucleotide), are involved in pyrimidine and lipid biosynthetic pathways, regulation of metabolites and amino acids, metabolism of metals such as heme and in iron-sulfur (Fe-S) cluster synthesis [2]. They regulate calcium (Ca^{2+}) homeostasis and flux and thus are involved in neurotransmitter release in the neurons, neuronal plasticity and neurogenesis. The TCA cycle intermediates are utilized for synthesis of gamma amino butyric acid (GABA) and glutamate neurotransmitters that function in neuronal signaling [4,5]. Other functions of the mitochondria include regulation of membrane potential [6], reactive oxygen species (ROS) mediated signaling [7], programmed cell death or apoptosis [8], steroid synthesis [9] and hormonal signaling [10]. Considering the plethora of cellular functions that mitochondria are associated with, it is not surprising that mitochondrial mal-function is the cause of several disease states including Kearns-Sayre syndrome [11], MELAS syndrome[11], Parkinson's disease [12], Alzheimer's disease, schizophrenia, bipolar disorder, dementia and epilepsy [13], stroke, cardiovascular disease, chronic fatigue syndrome, retinitis pigmentosa, and diabetes mellitus [14].

The evolutionary origin of the mitochondria is still uncertain and several explanations have been put forward [15]. In all cases, mitochondria have been designated as descendants of an α-proteobacterium, the differences lying in the stage at which they were incorporated into another cell. In one case, it is proposed that the α-proteobacterium was engulfed by a proto-eukaryote, the latter being likely evolved from archaea and already possessing nucleus, endomembranes and phagocytic capability [15]. Another explanation is that this mitochondrial predecessor got involved in metabolic endosymbiosis with an archaeon [16]. A third hypothesis proposes that an archaeon got involved in metabolic endosymbiosis within a bacterium (different from the predecessors of mitochondria) and later engulfed an α-proteobacterium [17]. Subsequent to the incorporation, the mitochondria transferred essential genes encoded in its genome to the nucleus (endosymbiotic gene transfer), thus entering into an obligatory symbiosis. This is illustrated by the fact that α-proteobacteria possess genome sizes ranging from 1.3

mega base pairs (Mbp) to >9 Mbp (as found in *Pelagibacter* and *Rickettsia* species)[15] while mitochondrial genomes are much smaller (~80 kbp in the yeast *Saccharomyces cerevisiae* and16.5 kbp in humans) [18–20].

2. Structural Components of the Mitochondria

A mitochondrion consists of 5 major structural parts:

2.1 Outer membrane

A 60-75 Å thick outer membrane encloses the mitochondrion (Figure I1). It harbors integral membrane proteins called **porins** that permit the free diffusion of molecules of size 5000 Da or lesser across the membrane. Additionally, large multi-subunit complexes called **translocases** actively transfer larger proteins that possess mitochondrial signaling/targeting sequences [21]. Among them, the major translocase system is the translocase of the outer membrane (TOM) complex [22]. The outer membrane is also the locale of enzymes involved in fatty acid elongation, epinephrine oxidation and degradation of tryptophan. Permeabilization of the outer membrane triggers cell death due to leakage of cytochrome C into the cytosol and activation of the caspase mediated apoptotic pathway [23]. Such mitochondria mediated apoptosis is essential in embryonic development and in tissue homeostatis [24]. Finally, the outer membrane associates with the endoplasmic reticulum (ER) at MAM (mitochondria associated ER membrane) that are important in ER-mitochondria Ca^{2+} signaling and exchange of lipids between the two organelles [25].

2.2 Intermembrane Space

The intermembrane space, also known as the perimitochondrial space, is the space between the inner and the outer membrane (Figure I1). Due to the porins in the outer-membrane, the concentration of small solutes like ions and sugars are the same as that in the cytosol. However, protein composition is different from the cytosol since only proteins that possess a signaling sequence can access to this compartment [22].

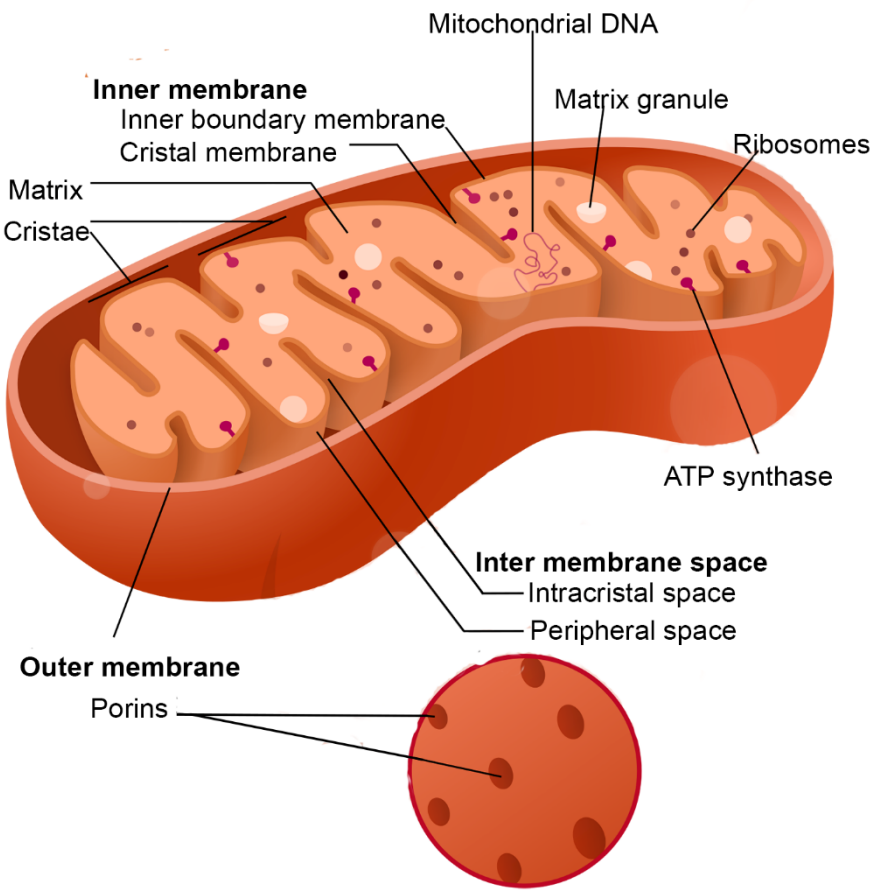


Figure I1. Schematic representation of the structural parts of a mitochondrion (Figure adapted and modified from Wikipedia: the free encyclopedia- https://en.wikipedia.org).

2.3 Inner membrane

The inner membrane (Figure I1) hosts more than 150 different polypeptides and contains 1/5th of the total protein content of the mitochondrion. These include proteins belonging to the electron transport system (ETS) including the ATP synthase [3] that are involved in ATP synthesis, transport proteins involved in metabolite and protein transport and proteins involved in mitochondrial fission and fusion [26]. Thus, it has a very high protein to phospholipid ratio (> 3:1 by weight). In addition, it is rich in **cardiolipin**, a phospholipid that might be playing a role in making the inner membrane impermeable. Due to the absence of porins, the inner membrane is thus highly impermeable to all molecules and special transporters are required to transport ions and other molecules across it. The translocase of the inner

An interspecies comparison of mitochondrial nucleoids

Species	Cytological appearance	Size	Number per cell	Number of mitochondrial genomes per nucleoid	Size of mitochondrial genome	Visualization tools
Saccharomyces cerevisiae	Globular foci	~0.2–0.4 μm in aerobic and ~0.6–0.9 μm in anaerobic cells (diameter)	~40–60 in aerobic and ~7.6 in anaerobic cells	~1–2 in aerobic and ~20 in anaerobic cells	75–80 kb	DAPI, GFP tagging
Physarum polycephalum	Rod shape	Up to ~1.5 μm in length	~15	~40–80	63 kb	DAPI, ethidium and thionine staining, light or electron microscopy
Crithidia fasciculata	Disk shape	~1.0 μm × ~0.35 μm	1	Several thousand mini circles and a few dozen maxi circles	0.5–10 kb for mini circles and 20–40 kb for maxi circles	DAPI and ethidium staining, immunofluorescence and GFP tagging, light or electron microscopy
Humans	Globular foci	~0.068 μm (diameter)	466–806 in cell lines	~2–10	16.5 kb	Ethidium and PicoGreen staining, immunocytochemical staining with DNA-specific antibodies, bromodeoxyuridine labelling, GFP tagging

DAPI, 4′,6-diamidino-2-phenylindole.

Figure I2. Comparison of mtDNA and nucleoids from selected species. Chen and Butow. *Nat. Rev. Genet.* **6,**815–25 (2005).

membrane (TIM) complex and the Oxa1 transports proteins across the inner membrane into the mitochondrial matrix [27].

2.4 Cristae

The inner mitochondrial membrane shows invaginations called cristae (Figure I1) which protrude into the mitochondrial matrix. These membrane folds increase the surface area of the inner membrane and host the electron transport system (ETS) and the mitochondrial ATPase [3]. Thus, by increasing the membrane surface area, they increase the ATP producing ability of the mitochondria, and cells with a greater ATP demand, e.g. muscle cells, tend to have more cristae.

2.5 Matrix

The inner membrane encloses a space called the matrix (Figure I1). It contains a concentrated mixture of macromolecules. The matrix harbors components of metabolic pathways that are essential for cell life in aerobic eukaryotes such as the TCA cycle (Krebs cycle). It also contains enzymes of the fatty acid oxidation pathway and the ornithine (urea) cycle [28]. In addition, it contains components required for protein synthesis such as mitochondrial ribosomes and tRNA, and several copies of the mitochondrial genome [29].

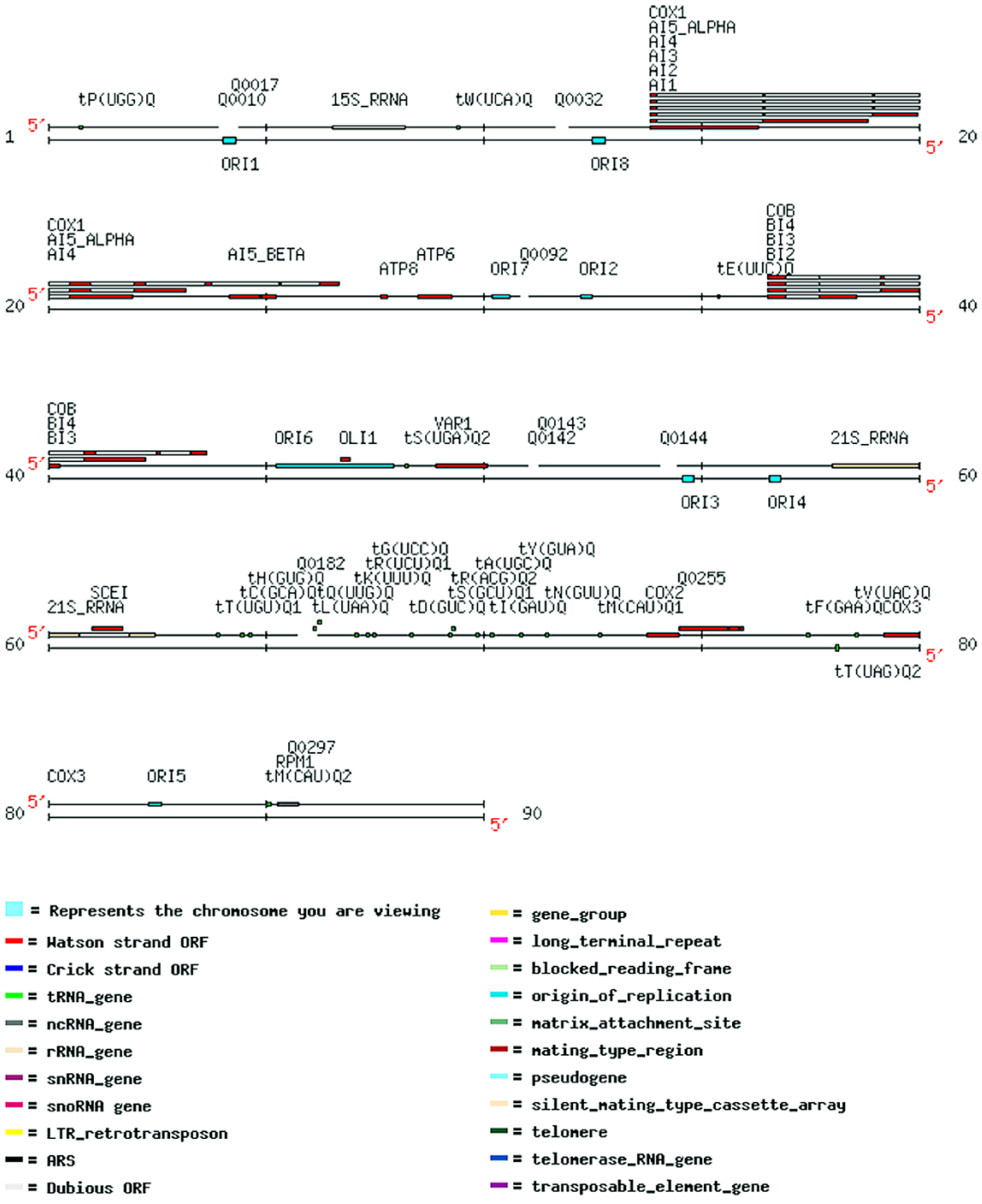


Figure I3. Genetic map of the yeast mtDNA. The key explains the different features. (map adapted from the Saccharomyces Genome Database (SGD)- http://www.yeastgenome.org/)

3. Mitochondrial DNA (mtDNA) in *Saccharomyces cerevisiae*

After incorporation into a proto-eukaryotic cell/archaeon the α-proteobacterial ancestor of the mitochondrion transferred genes to the nucleus during endosymbiosis. However, it still retained important genes. Across species, the mtDNA molecule differs in size and composition. In humans it is a 16.5 kbp circular molecule [29]. In contrast, in the unicellular eukaryote *S. cerevisiae* (budding yeast), the mitochondrial DNA (y-mtDNA) is 70-85 kilo base pairs (bp) long [18, 19] (Figure I2, I3) with extensive stretches of A-T rich intergenic regions. However, in all cases mtDNA codes for components of the oxidative phosphorylation (OXPHOS) pathway [28]. In *S. cerevisiae* it codes for the cytochrome *c* oxidase subunits cox1, cox2 and cox3, ATP synthase subunits atp6, atp8 and atp9, apocytochrome *b* (*cytb*), *var1* (a ribosomal protein), multiple intron related open reading frames (ORFs), 24 tRNAs, 15S and 21S ribosomal RNAs, the 9S RNA of the RNA processing enzyme RNase P and 8 replication origin like (*ori*) elements (Figure I3). In addition, some of the introns of *cox1* and *cytb* produce maturases, site-specific endonucleases and reverse transcriptases by independent or in frame translation with respect to their upstream exons [19, 30, 18].

S. cerevisiae are facultative aerobes that can survive in both fermentable (e.g. glucose) and non-fermentable (e.g. glycerol) media [31]. They are classified as ρ (rho)$^{+}$, ρ^{-} or ρ^{0} depending on whether they retain the entire mtDNA (termed rho factor), a part of it or lose it completely [32]. Additionally, the ρ^{-} type has a hyper-suppressive (hs) subtype which, in crosses with a ρ^{+} strain, completely replaces the wild type mtDNA with its ρ^{-} mtDNA (Figure I4). In some of these hs ρ^{-} yeasts, which have lost substantial portions of the mitochondrial DNA, the remaining DNA is repeated in tandem to maintain a total amount of y-mtDNA similar to that in ρ^{+} (wild-type) cells [33]. Historically, y-mtDNA was believed to be circular like its mammalian counterpart [34, 32] . This notion was guided by electron microscopy (EM) visualizations of mtDNA from birds and mammals. However, subsequent EM studies on y-mtDNA failed to find circular species. Pulsed-field-gel-electrophoresis (PFGE) studies revealed that majority of y-mtDNA exists as linear tandem arrays of 75-150 kbp with minute fractions of circular species [32,34]. Linear mtDNA in yeast is not an exception and several other species possess the same [32] (Table I1). This raises the logical question regarding DNA end maintenance. There is no evidence for maintenance of the free ends in a fashion similar to telomeres. However, it has been speculated that constant recombination might be

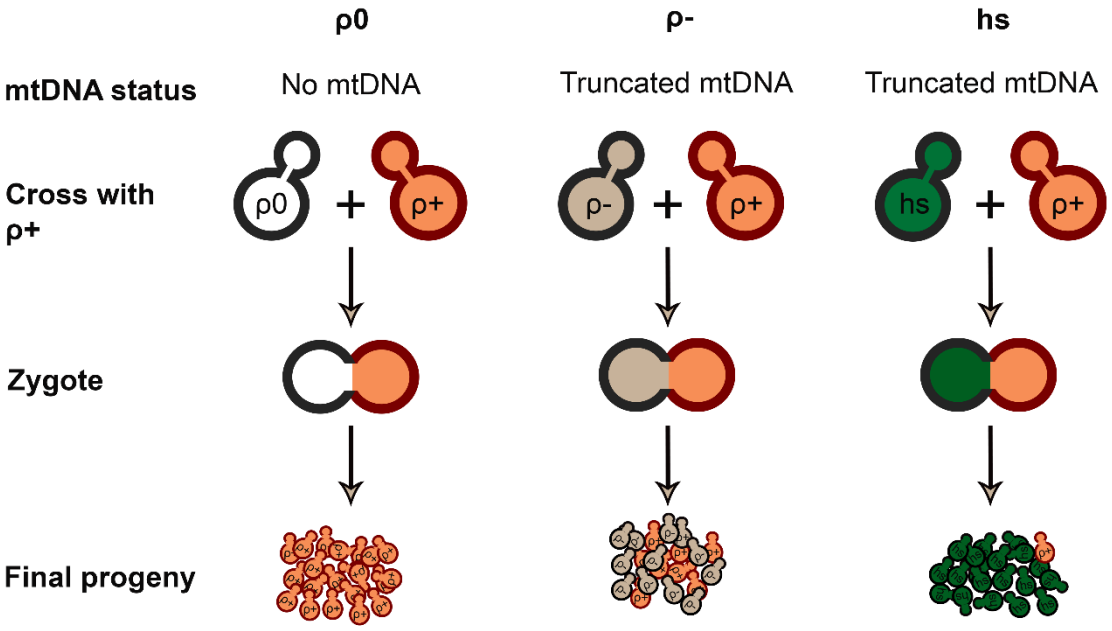


Figure I4. **Different yeast-subtypes**. The classification is based on their mtDNA status and tendency to replace wild type (ρ^+) mtDNA (suppressivity). ρ^+ yeasts possess wild type mtDNA and produce larger colonies. ρ^0 sub-types have lost their mtDNA entirely and in crosses with ρ^+ cells produce solely ρ^+ progeny. ρ^- cells have lost significant portions of their mtDNA and are moderately suppressive. Hyper-suppressive (hs) ρ^- cells are characterized by a strong tendency to replace ρ^+ mtDNA. ρ^0, ρ^- and ρ^- hs strains all produce small colonies. The above representations are entirely schematic.

another way of maintaining the ends [32]. The difference in topology between mammalian and yeast mtDNA also indicates that the mode of DNA replication might also differ. Several replication origin-like sequences (termed reps/oris) have been identified by analyzing mtDNA sequences in hs ρ^- yeasts [18,19,35–37]. However, there is no definitive proof till date that these rep/ori sequences actually serve as origins of DNA replication [32]. Although, for one of these rep sites, a promoter was identified which could initiate RNA primed replication [38], the fact that replication can occur without RNA priming in ρ^- yeast and that mtDNA without ori sequences are stably maintained [33], argues that RNA primed replication is not the dominant mechanism in yeast [32]. On the other hand, y-mtDNA has been shown to be constantly undergoing recombination inside the mitochondria [39]. This has led to speculations that the reps/oris, instead of acting as origins of replication, might be facilitating DNA replication via recombination (recombination driven replication or RDR) by acting as recombination hotspots[32]. The hypersuppressivity of hs ρ^- strains could then be explained by their content of large number of copies of

Organisms with linear mtDNA

Organism	Genome size (kbp)
Fungi	
Torulopsis glabrata	19
Saccharomyces cerevisiae	75-85
Schizosaccharomyces pombe	19
Aspergillus flavus	~33
Aspergillus nidulans	33
Fusarium oxysporum	~52
Neurospora crassa	63
Nectria haematococca	>30
Leptosphaeria maculans	>30
Saprolegnia ferrax	>30
Phytophora vignae	>30
Phytophora megasperma	45
Penicillium sp.	~22
Apicomplexans	
Plasmodium yoelli	6
Plasmodium falciparum	6
Plasmodium gallinaceum	6
Eimeria tenella	~6
Euglenoid flagellate	
Euglena gracilis	<19
Plant	
Chenopodium album	270

Table I1. **Presence of linear mtDNA in different organisms and the length of their mitochondrial genome**. Table adapted from Williamson, *Nat. Rev. Genet.***3,** 475–81 (2002).

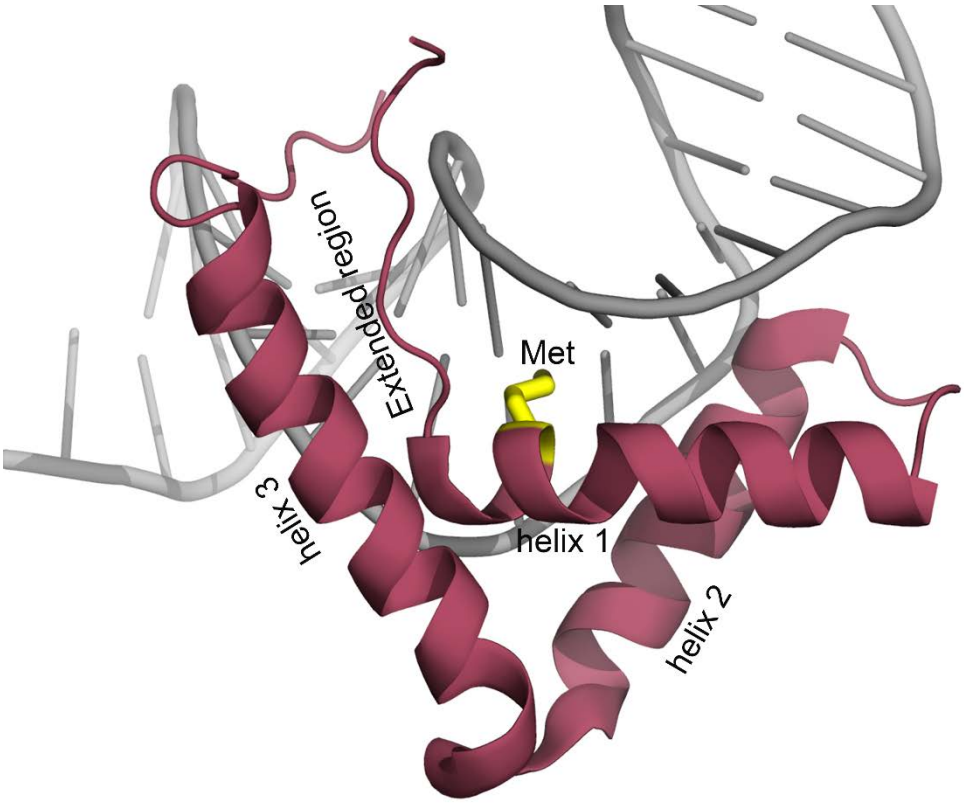


Figure I5. HMG-box protein Sox18 from *Mus musculus* bound to DNA (PDB ID: 4Y60). The typical L shape is prominent along with the characteristic extended region followed by helix1, loop, helix2, loop and helix3. The protein is bound to the DNA minor groove, inducing a severe DNA bend, a general characteristic of HMG-box proteins. The inserting methionine (Met) residue is shown in yellow.

the ori sequences which would allow efficient recombination and replication and thus out-compete the ρ^+ mtDNA [32]. Rolling circle replication (RCR) [40], on the other hand has been suggested to give rise to the circular mtDNA molecules coming from petites that have been detected in minute fractions along with linear mtDNA. Linearity of mtDNA and recombination driven replication (RDR) as the major or sole mtDNA replication mode has been demonstrated in the fungi *Candida albicans* [41, 42]. Furthermore, linear and branched mtDNA in *Schizosaccharomyces pombe* indicate the same. Thus the difference in topology (linear vs circular) between yeast and mammals is accompanied by a difference in the way the mtDNA is propagated. In this context *S. cerevisiae* could be employing mechanisms similar to other eukaryotes with linear mtDNA such as the malarial parasite *Plasmodium falciparum* where replication is recombination dependent (Table I1). It has additionally been speculated that the RDR mechanism in *S. cerevisiae* is similar to that of T4 phages [32]. However, the exact mode of DNA replication in yeast mitochondria is still under investigation.

4. mtDNA packaging

The mitochondria harbor the Krebs cycle, the Electron Transport System (ETS) and the ATP synthesis machinery and are thus a hotspot for generation of reactive oxygen species (ROS) [43]. In addition, iron (Fe) liberated from iron-sulphur (Fe-S) proteins that are part of the electron transport system (ETS) generate hydroxyl radicals via Fenton chemistry [44]. Thus, mtDNA is at a much higher risk of damage than nuclear DNA that is protected from oxidative damage by histone mediated packaging [45]. Added to this is the general packing requirement for mtDNA so that it can be maintained and propagated properly [46]. Similar to the nuclear case, a mechanism for protecting mtDNA from ROS mediated damage is to package it with specific proteins [47]. However, the mechanism of mtDNA packaging differs from the histone mediated packaging of nuclear DNA into nucleosomes [46]. mtDNA is assembled into nucleoprotein structures called nucleoids that contain proteins involved in mtDNA transcription, replication and maintenance, together with proteins that compact the DNA and thus regulate its accessibility [46]. The nucleoids have been seen to be associated with the inner mitochondrial membrane [48, 46] and their dimensions and content of mtDNA copies varies across species (Figure I2). The ~25 μm long (70-85 kbps) *S. cerevisiae* mtDNA is packaged into ~0.3 μm nucleoids and contains 1-2 copies per nucleoid under aerobic conditions [29,49]. In contrast, in humans, recent studies using super-resolution techniques have shown that the nucleoids are smaller in size (~0.1 μm) and contain ~1.4 mtDNA per nucleoid on average [50] (compared to previous report of 2-10 copies per nucleoid [29,51], Figure I2). The differences in mtDNA length and mitochondrial protein content between species point to a probable difference in mtDNA organization. In both human and yeast (*S. cerevisiae*), mtDNA packaging is mediated chiefly by nucleus encoded dedicated proteins belonging to the high mobility group (HMG) box family [52–56]. These are a class of proteins that show high divergence in terms of amino acid sequence but tend to preserve tertiary structure features.

4.1 HMG-box proteins

HMG-box proteins belong to the high mobility group (HMG) B superfamily and are members of the HMG class of proteins, the latter owing its name to high mobility of the proteins of this class in poly-acrylamide

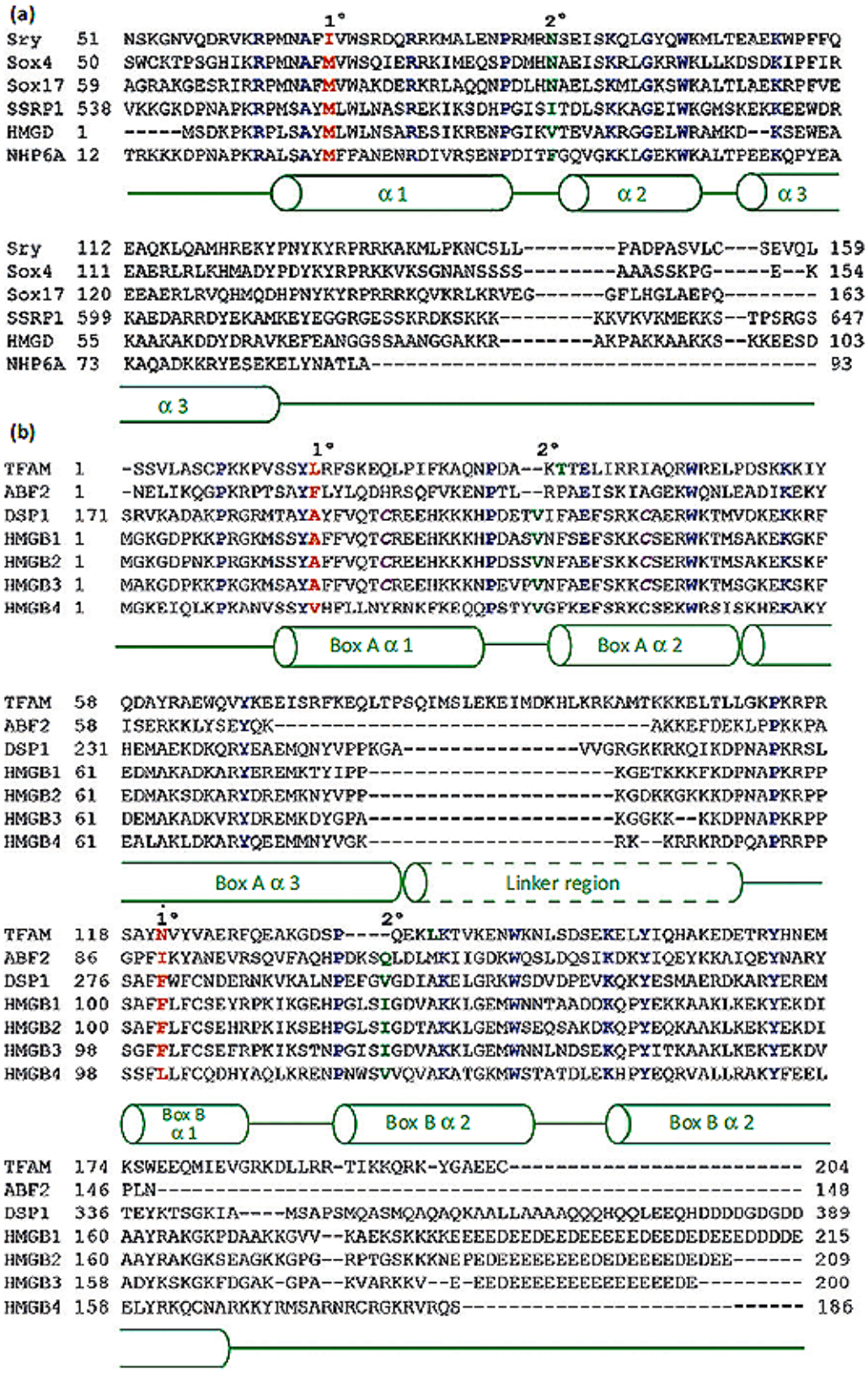


Figure I6. Sequence alignment for selected single (a) and double (b) HMG-box proteins. Intercalating/inserting residues in the 1st and 2nd HMG-boxes are indicated in red. Conserved residues are indicated in blue. Malarkey *et al*. *Trends Biochem. Sci.***37,** 553–62 (2012).

electrophoresis gels (PAGE) [57]. There are two other superfamily of HMG proteins, namely HMGA (HMG A-T hook) superfamily and HMGN (HMG nucleosome-binding superfamily) that differ from HMGB in sequence, molecular mass and mode of binding to the DNA [57].

The HMGB proteins typically bind DNA at the minor groove and cause significant distortion of the DNA. They contain characteristic L-shaped HMG-box domains (Figure I5) constituted by three α-helices: two short helices 1 and 2 form the short L-arm, whereas the extended region and helix 3 form the long L-arm [56]. The HMG-box domain consists of approximately 70 residues. Further, the HMGB proteins consist of two subfamilies according to the number of HMG-boxes they contain, DNA sequence recognition specificity and evolutionary relationships [57] (Figure I6). The first subfamily nests proteins with multiple HMG-boxes that are present in all cell types and show little or no DNA sequence specificity. These include the nuclear HMG1 protein involved in organization of nuclear DNA and in transcription and contains two HMG-boxes followed by a C-terminal tail [58]; HMG2 that is implicated in DNA break repair and V(D)J recombination [59]; the RNA-polymerase1 transcription factor UBF which contains 6 HMG-box domains and binds DNA as dimers [60]; the human mitochondrial transcription factor A (TFAM) that also functions as mtDNA packaging protein and thus shows both sequence specific and non-specific DNA binding properties [52–55,61,62] and finally the yeast (*S. cerevisiae*) mtDNA packaging protein ARS-binding factor 2 [29,63-72] (Abf2p; see below). The HMG-box domains within each protein differ in their DNA binding properties. This suggests that in context of the full-length protein, the overall DNA binding properties and affinity are a function of intra-molecular interactions within the protein.

The second subfamily includes proteins with a single HMG-box, are restricted to specific cell types and bind to specific DNA sequences [73]. The members of this second subfamily include the sex determining factor SRY that acts as a transcription factor; the SOX group (A-H) of proteins that are transcription factors involved in embryonic development [74]; the lymphoid enhancer-binding factors LEF1 and fungal regulatory proteins Mat-Mc, Mat-a1, Ste11 and Rox1 [75]. However, in this subfamily, non-specific single HMG-box proteins are also found e.g. HMGD in *Drosophila melanogaster* that functions in chromatin organization [76].

Therefore, the human and yeast mtDNA packaging proteins belong to multi-domain HMG-box protein subfamily.

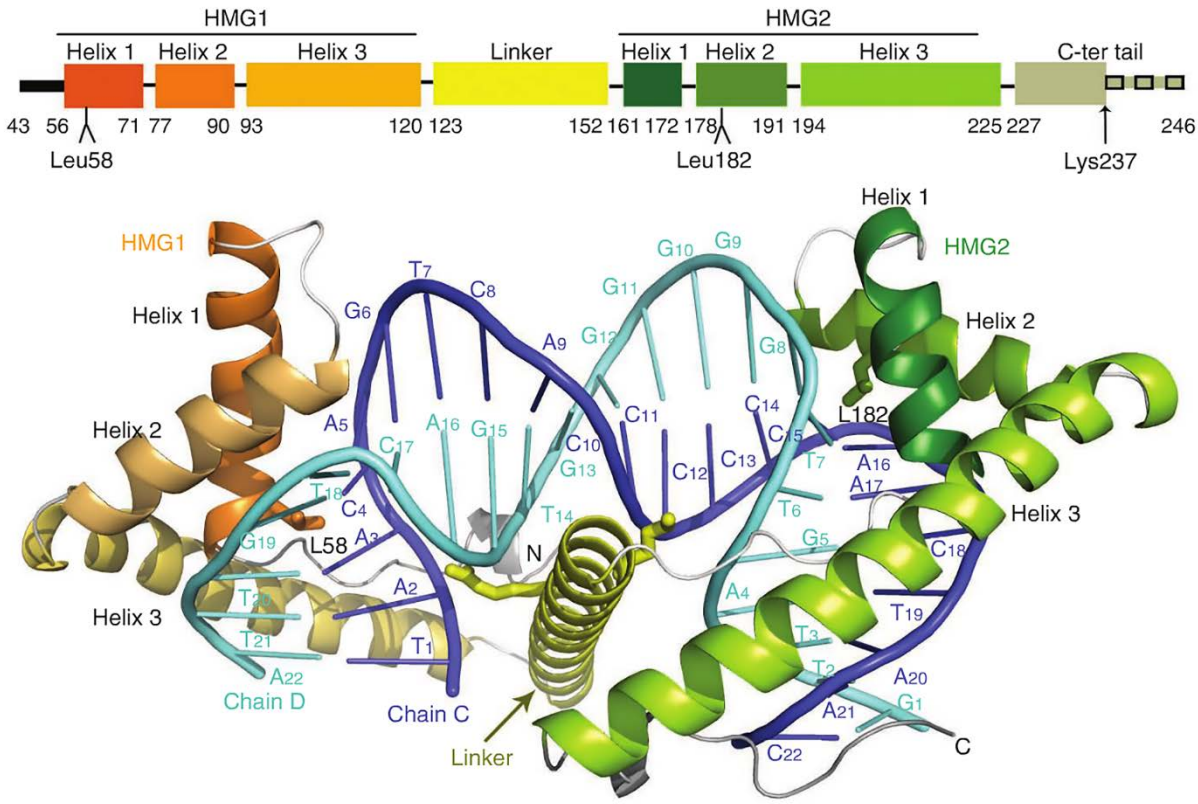


Figure I7. **Crystal structure of TFAM bound to a 22 bp DNA fragment containing the mtDNA light strand promoter (LSP)**. Rubio-Cosials, A. *et al. Nat. Struct. Mol. Biol.* **18,**1281–9 (2011).

4.2. DNA binding and bending by human mitochondrial transcriptional factor A, TFAM

The reported crystallographic structures of TFAM in complex DNA are the only structures available for a mtDNA packaging protein, till date [52–55]. In binding DNA, each of its two HMG-boxes distort DNA at the minor groove and induce a ~90° bend, thus resulting in an overall 180° U-turn of the DNA (Figure I7). Each HMG-box contains a key amino acid (aa) residue (Leu58 for box1 and Leu182 for box2) that inserts between DNA base-steps and thus causes the distortion [52,54]. The C-terminus of the 30 aa linker between the two HMG-boxes makes additional contacts with the DNA and thus helps to mitigate electrostatic repulsion between the DNA phosphate groups that are brought closer due to the DNA bend (Figure I7). Additionally, it allows TFAM to wrap around the DNA and increases DNA binding efficiency of HMG-box2 that has a low intrinsic affinity for DNA. Thus the linker helps in coordinating the actions of the two HMG-boxes [53,54,77]. TFAM also possesses a C-terminal tail that is responsible for interaction with the transcription initiation complex and thus plays an important role in transcription initiation [78].

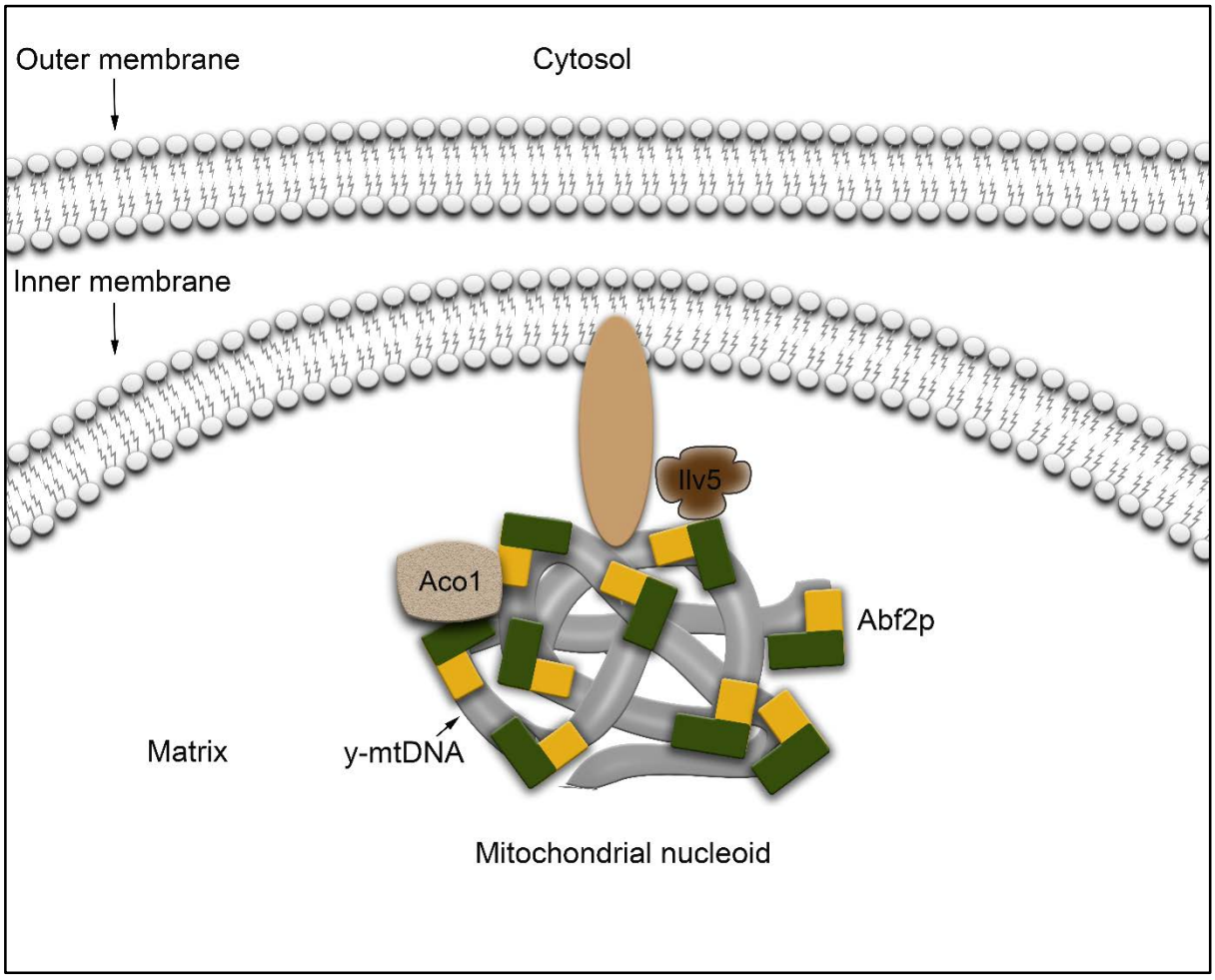


Figure I8. Schematic representation of mitochondrial nucleoid in *S. cerevisiae*. The chief mtDNA packaging protein Abf2p along with additional nucleoid associated proteins Aco1 and Ilv5 are depicted. The nucleoid is shown tethered to the inner mitochondrial membrane by a putative tethering protein. Figure concept adapted and modified from Chen and Butow. *Nat. Rev. Genet.* **6,** 815–25 (2005)

4.3 mtDNA packaging and maintenance in *S. cerevisiae*

The 70-85 kbp long linear y-mtDNA in wild type ρ^+ yeasts is packaged and maintained chiefly by the protein ARS (Autonomously Replicating Sequence) binding factor 2 or Abf2p (Figure I8). The protein was identified in 1991-1992 by Diffley and Stillman as a mitochondrial DNA binding protein based on immunofluorescence studies [70,71]. Yeast (*S. cerevisiae*) cells possess around 50-100 copies of the full length mtDNA [68]. Yeast can make use of both fermentable (e.g. glucose) and non-fermentable (e.g. glycerol) carbon source, and studies on composition of nucleoids have suggested that the details of the mtDNA packaging mechanism differ in the two situations [29,66,69,79-81]. In fermentable media y-mtDNA is

not essential for survival, as the mitochondrial oxidative phosphorylation linked ATP synthesis is not functional. Under these conditions, y-mtDNA is packaged and maintained by Abf2p and loss of the Abf2 gene leads to loss of mtDNA in yeasts growing in fermentable media [66]. This renders them unfit to subsequently survive in non-fermentable media, where operation of the respiratory chain requires presence of functional and transcriptionally active y-mtDNA [70]. In non-fermentable media (e.g. glycerol) Abf2p is still required to maintain y-mtDNA although the ratio of Abf2p to y-mtDNA is lower. The latter is achieved by an increase in y-mtDNA copy number (~2 fold) in glycerol medium while the number of available Abf2p molecules remains the same [29]. This results in a more open structure of the mitochondrial nucleoids that allows access of the y-mtDNA by transcriptional machineries, permitting expression of genes required to support respiration [29]. In both types of media, y-mtDNA copy number increases by 100 % on average with increase in Abf2p levels of up to 2-3 fold, while a 8-10 fold overexpression causes rapid loss of y-mtDNA and generation of ρ^0 (petite) mutants [66]. The former effect hints towards a contribution of Abf2p towards increased DNA replication (either direct or indirect and in fermentative and non-fermentative conditions) while the latter effect can be attributed to excessive compaction of the mitochondrial DNA at high Abf2p levels, making it inaccessible to the replication machinery [66]. Under respiratory conditions other proteins have been implicated to be additionally involved in protecting y-mtDNA [66,81,82]. These include the mitochondrial aconitase Aco1 that converts isocitrate to citrate in the Krebs cycle and the acetohydroxyacid reductoisomerase Ilv5 that is involved in branched chain amino acid biosynthesis. [29,69,81,82]. These proteins serve a bifunctional role and help to synchronize y-mtDNA packaging and maintenance with metabolic conditions and thus with the environment. (Figure I8, Table I2). Aco1 binds to both double-stranded (ds) and single stranded (ss) DNA and has been shown to protect mtDNA from point mutations and ssDNA breaks [29,81]. Its overexpression can prevent y-mtDNA loss in cells lacking Abf2p and also prevents mtDNA instability in cells lacking the mitochondrial helicase Pif1p [29], a protein involved in mtDNA maintenance under genotoxic stress conditions [83]. Ectopic expression of Aco1 at levels prevalent under respiratory conditions can prevent mtDNA loss in Abf2Δ cells [29]. Thus Aco1 can provide additional protection to mtDNA under respiratory conditions where ROS production increases [29]. Ilv5 is recruited to the mitochondria during amino acid starvation [84] and plays a role in mtDNA transmission [85]. Although mutants of Ilv5 causing y-mtDNA instability have been described, it is not known whether Ilv5 performs

Table 1 | **mtDNA-associated proteins in yeast, mammals and *Xenopus laevis***

Protein*	Primary function	mtDNA stability in mutant
Yeast		
Abf2‡§‖	mtDNA packaging	ρ^0 or ρ^-
Aco1§	Citric acid cycle	ρ^0
Arg5,6¶	Arginine biosynthesis	Stable
Ald4§	Ethanol metabolism	Stable
Atp1§	ATP synthesis	ρ^0-lethal
Cha1§¶	Catabolism of hydroxy amino acids	Stable
Idh1§	Citric acid cycle	Moderate instability
Idp1§	Oxidative decarboxylation of isocitrate	Stable
Ilv5§	Biosynthesis of Val, Ile and Leu	Moderate instability
Ilv6§	Biosynthesis of Val, Ile and Leu	Stable
Kgd1‡§	Citric acid cycle	Moderate instability
Kgd2‡§	Citric acid cycle	Moderate instability
Lpd1‡§	Citric acid cycle, catabolism of branched-chain amino acids	Moderate instability
Lsc1§	Citric acid cycle	Stable
Mgm101§‖	mtDNA maintenance or repair	Unstable ρ^+ and *ori*-lacking ρ^-
Mip1§‖	mtDNA replication	ρ^0
Mnp1‡	Putative mitochondrial ribosomal protein	Stable
mtHsp60§	Mitochondrial chaperonin	Unstable *ori*-containing ρ^-
mtHsp10§	Mitochondrial chaperonin	Unknown
mtHsp70§	Protein import	Unknown
Pda1§	Oxidation of pyruvate	Moderate mtDNA instability
Pdb1§	Oxidation of pyruvate	Moderate mtDNA instability
Rim1§	mtDNA replication	ρ^0
Rpo41§	mtDNA transcription	ρ^0 or ρ^-
Sls1§	Coordination of transcription and translation	ρ^0 or ρ^-
Yhm2§	Mitochondrial carrier	Stable
Mammals		
TFAM‡#	mtDNA transcription and packaging	mtDNA depletion
Twinkle‖	mtDNA replication	Multiple mtDNA deletions
mtSSB‡#	mtDNA replication	Unknown
Polymerase γ#	mtDNA replication	Multiple mtDNA deletions
BRCA1#	Tumour suppressor	Unknown
PRSS15**	Protein degradation	Unknown
Xenopus laevis		
mtTFA‡	mtDNA transcription and packaging	Unknown
mtSSB‡	mtDNA replication	Unknown
PDC-E2‡	Oxidation of pyruvate	Unknown
BCKAD-E2‡	Catabolism of branched-chain amino acids	Unknown
Prohibitin 2‡	Protein folding	Unknown
ANT1‡	ADP–ATP exchange on inner membrane	Unknown

Table I2. Mitochondrial DNA associated proteins in yeast (*S. cerevisiae*), mammals and *Xenopus laevis*, The primary functions of the proteins and the effect of their deletion on mtDNA stability are stated. Chen and Butow. *Nat. Rev. Genet.* **6,**815–25 (2005).

Figure I9. **Secondary structure prediction for the mature Abf2p protein** (without the mitochondrial targeting sequence) generated using the PSIPRED server. The 16 residue N-terminal segment prior to HMG-box1 is boxed in green and the six residue linker (estimated by end of the 3rd helical segment of HMG-box1 and beginning of HMG-box2) is boxed in grey.

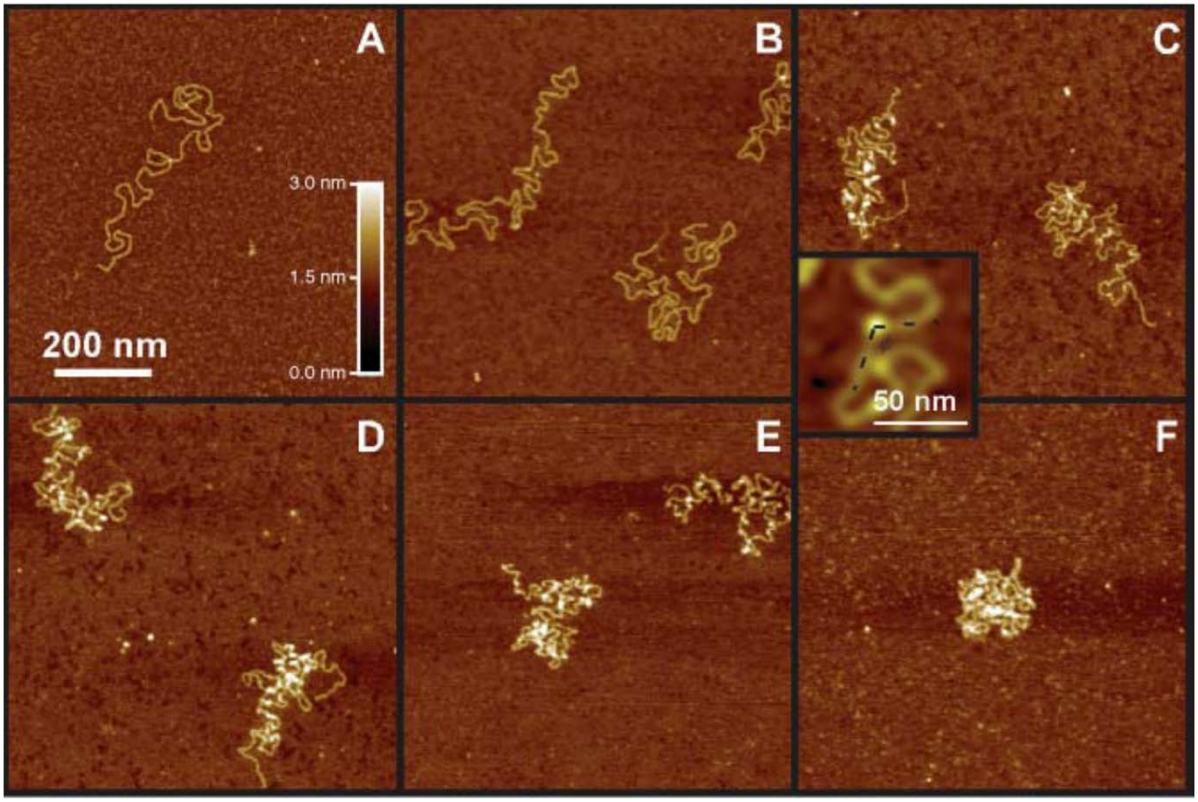


Figure I10. Atomic force microscopy (AFM) analysis of DNA packaging by Abf2p at different Abf2p/bp ratios. DNA used is linearized pBR322. A: no Abf2p; B: 1:20; C: 1:8; D: 1:4; E: 1:2; F: 1:1. The inset shows the DNA bend induced by Abf2p (~78º). Friddle *et al*. *Biophys. J.* **86,** 1632–9 (2004).

its functions via direct interaction with the DNA or by protein-protein interactions with other components of the nucleoid. A third protein, the heat shock protein Hsp60, a mitochondrial chaperonin, is also found associated with nucleoids under conditions of glucose repression and seems to regulate mtDNA transmission [86], although the specifics of the regulatory process are not known. Still other proteins such as Ald4, Idh1, Idh2 and Kgd1 have been shown to be involved in physical interactions with Abf2p in mitochondrial nucleoids [29,69] (Table I2). The mitochondrial nucleoids are thus dynamic assemblies that undergo reshaping in terms of both compaction and protein content according to the metabolic status of the cells. Abf2p is the central player in its packaging and maintenance.

4.4. Abf2p and its mechanism of mtDNA packaging and maintenance

Abf2p is a HMG-box protein with two tandem HMG-boxes, joined by a predicted 10 amino acids (aa) linker [70] (Diffley and Stillman, 1991). However, prediction with the PSIPRED server [87], shows a 6 residue

linker (Figure I9). The length of the linker will determine the way in which the two HMG-boxes are coordinated and thus determine its function. However, the actual linker length can only be verified in light of an experimentally determined structure, the latter not being available till now. Additionally, the structure and function of the 16 residue N-terminal segment of Abf2p remains to be elucidated (Figure I9). Unlike its human counterpart, TFAM, Abf2p does not possess a 30 aa long linker and lacks a C-terminal tail that has been implicated in interaction with the transcriptional machinery and transcription activation [62]. Indeed, it has been demonstrated that Abf2p does not play any significant role in transcription [70], although attaching the TFAM C-terminal tail to Abf2p enables it to activate transcription from the human mitochondrial light strand promoter (LSP) [88]. Thus, in the yeast mitochondria, Abf2p predominantly plays the role of a packaging protein. An older estimate of the number of Abf2p molecules per cell was 250,000 [70,71] while more recent investigations report 3810 [89] and 860 [90] molecules. Thus considering 50-100 y-mtDNA molecules of 85 kbp each per cell, and a binding site of 26 bp [29,70] (from DNA footprinting, Diffley and Stillman, 1992) there would be a total 160000-320000 Abf2p binding sites. This suggests that much of the y-mtDNA in the yeast mitochondria is not covered by Abf2p, implicating that nucleoids are relatively loosely packed in yeast. This is corroborated by atomic force microscopy (AFM) studies that reveal that at Abf2p ratios of 1protein per 8-20 bps the DNA is not tightly compacted [64]. Only at 1Abf2p:1bp ratio tight compaction was observed (Figure I10F). At the local and molecular level, however, very little detail is known about the interaction of Abf2p with DNA. Like other non-specific HMG-box proteins, Abf2p is expected to bind DNA at the minor groove and bend it. AFM studies report that Abf2p induces a bend of around 78^{o} (Figure I10), although low resolution did not permit assessment of contribution of the individual HMG-boxes and whether Abf2p binds DNA as a monomer or a dimer [64].

4.5 Role of Abf2p in mtDNA recombination

DNA replication in yeasts is proposed to be predominantly recombination dependent [32,39](see above), and Abf2p shows a high affinity for recombination intermediates (DNA 4-way junctions) [68] and influences the level of recombination intermediates is *S. cerevisiae* [65]. Recent electron microscopy studies have shown interaction of Abf2p with 4-way junctions [68] (Figure I11) and electrophoretic mobility shift assays

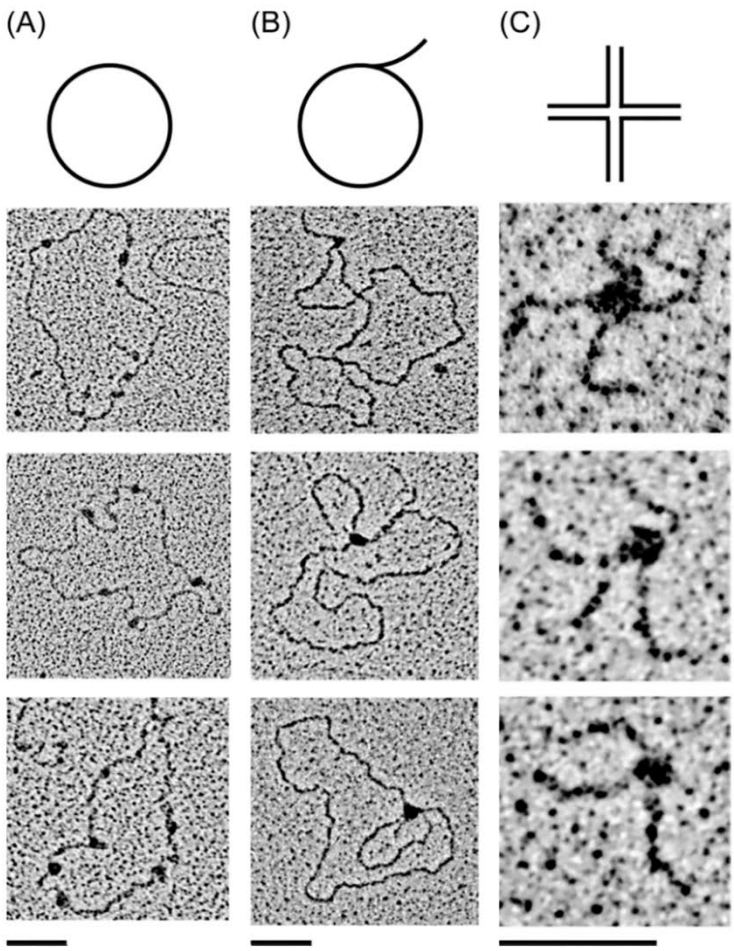


Figure I11. **Binding of Abf2p to circular ds DNA (A), replication forks (B) and 4-way junctions (C) studied by electron microscopy (EM)**. Concentration of Abf2p and DNA are 15 ng/µl and 2 ng/ µl in each case. Bakkaiova *et al*. *Biosci. Rep.* **36,** (2015).

(EMSA) have demonstrated a high affinity of Abf2p for 4-way junctions [68]. Thus, Abf2p could also play a direct role in mtDNA replication via RDR, although no direct evidence is available yet.

5. Abf2p and phased DNA binding

Although Abf2p is a non-sequence specific DNA binder, it shows 'phased binding' at AT rich y-mtDNA sequences that occur as tandem repeats in hs ρ^- yeast. Fangman *et al* [33], 1918, reported that the hs ρ^- strain HS3324 (whose mtDNA consisted of a 963bp repeat containing the rep2/ori5 sequence) produced deletion mutant sub-strains whose mtDNA consisted of short 100% A-T sequences [33,36, 70] repeated in tandem. These sequences were stably maintained and were amplified such that the total amount of mtDNA matched that present in ρ^+ cells (Figure I12a). One of the shortest of such DNA sequences

a

Strain	% Suppressive[a]	Repeat length (bp)	Step size (bp)	% Total DNA[b]	Copies/ cell
[*rho*$^+$]				13.4	29
HS3324	96.0	963	981.0 ± 118[c]	13.7	2,300
Reduced suppressive					
1a	2.1	92	91.2 ± 0.9	20.1	38,000
S5	35.2	89	91.3 ± 2.8	20.0	39,000
S4	1.8	70	71.0 ± 1.6	18.3	45,000
4a	17.1	64	62.6 ± 0.8	15.6	40,000
5a	36.0	35	34.6 ± 1.0	11.9	54,000

b

Figure I12. The 64 and 35 bp sequences and Abf2p phased binding (a). Table from Fangman et al, *Mol. Cell. Biol.* **9,** 1917–21 (1989), showing the different deletion mutant sub-strains derived from HS3324. The mtDNA repeat length, amount relative to total DNA and number of copies per cell are shown for each sub-strain. The sub-strains 4a and 5a, possessing mtDNA consisting of tandem repeats of 64 bp and 35 bp respectively are boxed in red. **(b)**. General organization of the y-mtDNA ori5/rep2 origin of replication, including the location of ARS consensus sequences (ARS-c, 11/11 matches to 5'-(A/T)AAA(T/C)ATAAA(A/T)-3', in orange) and the downstream 64bp region (in yellow). Below, the 64bp sequence is shown, the near matches (ARS-m) to ARS-c are depicted in orange and the position of the 35bp sequence is demarcated by vertical bars.

which have been observed in ρ$^-$ strains are a 64 base pair (bp) sequence that occurs downstream of the RNA polymerase binding site at the rep2/*ori5* replication origin in ρ$^+$ mtDNA (19,459-19,522 on y-mtDNA, Figure I12b) [18]. Its 35 bp derivative (19,475-19,509), is another such sequence (Figure I12a,b). Diffley and Stillman (1991) showed that Abf2p, while having a general non-specific binding profile on *rep2*, contained "phased binding sites" [70] inside the 64 bp sequence. Specifically, Abf2p did not bind to 9/11 or 10/11 matches of the consensus 5' A/TAAAYATAAAA/T 3' (where Y stands for pyrimidine) found in Autonomously Replicating Sequences (ARS) [70]. ARSs occur in the yeast nuclear genome and function as DNA replication origins [91-94]. They consist of four regions A, B1, B2 and B3. Region A is highly conserved and contains the above consensus. This "phased binding" of Abf2p was additionally shown to be a general phenomenon, observed *in vitro* across other origins of replication including the nuclear

ARS1 [70]. Interestingly, it was observed that at the 5' end of the 35 bp repeat (which is the smallest repeat found in ρ^- yeasts) a near match of the ARS consensus sequence occurs (Figure I12b), followed by 26 additional bps, the latter matching very closely to the footprint of Abf2p detected by them [33, 70]. Moreover, ARS consensus sequence near matches, though not as efficient as exact matches in initiation of DNA replication, when present in enough numbers can maintain similar replication activity [36]. Thus the phased binding of Abf2p at these sequences points towards a possible role of the protein in such events.

The present state of knowledge about the *S. cerevsiae* mtDNA packaging protein Abf2p lacks molecular details of its interaction with the DNA, the role played by the two HMG-boxes and the involved dynamics along with in-depth understanding of the global architecture of mitochondrial nucleoids. Furthermore, mechanistic details of the 'phased binding' of Abf2p at specific sequences and their functional implications remain to be elucidated.

Objectives

1. Elucidation of crystal structure of the mitochondrial DNA packaging protein Abf2p in complex with functionally relevant DNA fragments in order to obtain structural and functional insights into the molecular mechanism of DNA binding
2. Investigate the roles played by the protein domains in DNA binding by *in vitro* and *in vivo* assays
3. Understand the dynamics of the free protein and the protein/DNA complex by in solution and simulation techniques to obtain information on the modes of interaction between the protein and the DNA counterparts
4. Understand how DNA properties alter protein binding and thus seek explanation of phased binding by Abf2p
5. Investigate thermodynamics of the protein-DNA interactions as a function of DNA structure
6. Assimilate acquired knowledge to obtain a coherent picture of DNA packaging mechanisms in the mitochondria and find clues to fathom how the properties of packaging proteins and mitochondrial DNA affect global nucleoid architecture

Materials and Methods

1. Protein Expression and Purification

The yeast Abf2p gene was cloned from genomic DNA using standard Polymerase Chain Reaction (PCR) with a proofreading DNA polymerase (Pfu Ultra, from *Agilent*) and inserted into the pCri7a [95] expression vector to produce a 6-His-tagged fusion protein (Forward primer: 5'ATCACCATGGCTCATCATCATCATCATCATAAGGCTTCCAAGAGAACG3', Reverse primer:5'TAGATCTCGAGCTAGTTGAGAGGGTAGC3'). This construct encodes residues 27-183 (Saccharomyces Genome Database ID. S000004676) corresponding to full-length mature Abf2p (without the N-terminal mitochondrial signalling sequence, residues 1-26). The plasmid was transformed into BL21 (*pLys*) *Escherichia coli* (*Merck Millipore*) strain. Cells were grown in LB medium for 2 h at 37°C until the optical density (O.D) at 600 nm reached 0.6. After subsequent induction with 1mM isopropyl β-d-1-thiogalactopyranoside (IPTG), the culture was grown for 4 h at 37°C and shaking at 225 rpm. The cells were pelleted down by centrifugation at 5000 rpm and 4^{o} C for 30mins, flash frozen in liquid nitrogen and stored at -80^{o} C.

Cells were sonicated for 10 min (cycle 2s on/6s off) on ice in 50mM Tris-HCl pH 7.5 and 1M NaCl. Since the protein lacked cysteine residues, reducing agents were not used except for 1 mM β-mercaptoethanol at the lysis step. The lysate was injected into a Ni-NTA-affinity column (HisTrap HP, *GE Healthcare*) mounted on an AKTA Purifier (*GE Healthcare*) system. The protein was eluted with a linear gradient of 20 column volumes (100ml for 5ml column) from Buffer A (20mM imidazole, 50mM Tris-HCl pH 7.5, 750mM NaCl) to Buffer B (500mM imidazole, 50mM Tris-HCl pH 7.5, 750mM NaCl). Quality of the protein fractions was assessed by SDS-PAGE and the purest fractions pooled and concentrated for gel filtration chromatography using a Superdex 75 10/300 column (*GE Healthcare*) pre-equilibrated with running buffer (50mM Tris-HCl pH 7.5, 750mM NaCl). The peak fractions were collected and purity was analyzed by SDS-PAGE (Figure R1). The yield obtained at the end of the size exclusion chromatography step was between 8-10 mg/liter of bacterial cell culture.

1.1 Seleno-methionine Derivatives

Non-auxotrophic (i.e. capable of methionine production) BL-21 pLys bacterial strain was used; the cells were grown at 37^{o} C to an optical density of 0.6 and methionine synthesis inhibition was carried out prior to induction with 1 mM IPTG (Base medium, amino acid nutrient mix and seleno-methionine were

purchased from *Molecular Dimensions*. Inhibition mix was prepared with amino acids purchased from *Sigma*). The cells were grown at 22^{o} C post-induction for 16-18 hours prior to pelleting and flash freezing. The Leu-52-SeMet derivative was purified identically as the wildtype with inclusion of 1mM and 5mM β-mercaptoethanol at the Ni-affinity and gel filtration steps respectively.

1.2 Deletion Mutants

Deletion mutants (see Results) were generated by 'round the horn' PCR [96] with KOD DNA polymerase (*Novagen*). All mutants were designed to have a C-terminal 6-histidine tag. As a quality check, subsequent DNA sequencing was performed for all constructs. Expression and purification protocols were identical to that of the wild-type protein.

2.Characterization of Protein-DNA Binding: Electrophoretic Mobility Shift Assay (EMSA)

2.1 Theoretical Bites

Electrophoretic mobility shift assay (EMSA) is an experimental technique widely used to analyze macromolecular interactions, including but not limited to protein-protein and protein-DNA interactions [97]. Gel electrophoresis makes use of two physical properties of macromolecules: size and charge. The set up consists of a porous gel (usually a poly-acrylamide or agarose gel) immersed in a buffer solution (Tris-acetate-EDTA or Tris-borate-EDTA) and containing wells into which experimental samples can be loaded. Smaller, more compact and/or more negatively charged species will migrate faster (towards the positively charged electrode) whereas larger, extended and/or positively charged species will run slower, thus enabling separation and visualization. In an EMSA, a putative complex is loaded into the gel along with suitable controls in parallel lanes. Complex formation will lead to a change in size and charge of the molecular species and this will produce a change in migration and an accompanied shift of the visible band [97]. In certain cases, if required, a setup can be done with reversed polarity.

EMSA is ideally suited for characterizing protein-DNA interactions and for detecting the associated stoichiometry. In combination with radioactive labelling of the DNA oligo-nucleotides, EMSA can also be effectively used for calculating dissociation constants (or equivalently binding affinities) for protein-DNA interactions.

2.2 EMSA with Long DNA

M13mp18 dsDNA plasmids (7249 bps; *Bayou Biolabs*) were linearized with BsrB I (*New England Biolabs*) and purified with the *illustra GFX* PCR DNA purification kit (*GE Healthcare*). Two-fold serial dilutions of proteins were performed on ice in 20mM Tris-HCl pH 7.5 and 750mM NaCl. DNA binding experiments contained 2ng·µL-1 linearized M13mp18 and were performed at 25°C in a binding buffer with a final composition of 20mM Tris-HCl pH 7.5 and 100mM NaCl. Reactions were initiated by addition of proteins as appropriate and incubated for 30 minutes. Subsequently, samples were resolved on 25cm-long, 1.2 % (w/v) agarose gels (*SeaKem* LE Agarose, *Lonza*) in 0.5x TBE (*Sigma*). Gels were run in a *Sub-Cell GT* cuvette (*BioRad*) for 17-18h at room temperature at 3V/cm, stained with *SybrSafe* (*Molecular Probes*) and scanned for fluorescence using a *Typhoon 8600* scanner (*GE Healthcare*).

2.3 EMSA with short DNA-fragments

Oligo-nucleotides were purchased from *Sigma*. Duplex DNA fragments were prepared by annealing complementary oligonucleotides in 20mM Tris-HCl, pH 7.5, 100mM NaCl at a concentration of 0.1mM. The mixtures were heated at 85^{o} C for 30 minutes followed by slow cooling to room temperature overnight. The DNA concentration was kept fixed at 200-400 nM while the protein concentration was varied. For each stoichiometry an initial mixture with the DNA was prepared in 50 mM Tris-HCl pH 7.5, 75 mM NaCl. In order to avoid aggregation or precipitation of the protein, it was serial diluted in its high salt purification buffer (50 mM Tris pH 7.5, 750 mM NaCl) to 10X of the final concentration required for the experiment. Subsequently the serial diluted protein was added to the initial DNA-buffer mixture (the final salt concentration in the mixture after protein addition was 150 mM), incubated for 30 mins and loaded in a 10% 10 cm poly-acrylamide gel (running buffer 0.5X TBE). Post-run staining and scanning of the gels were performed identically as mentioned above for EMSA with long DNA.

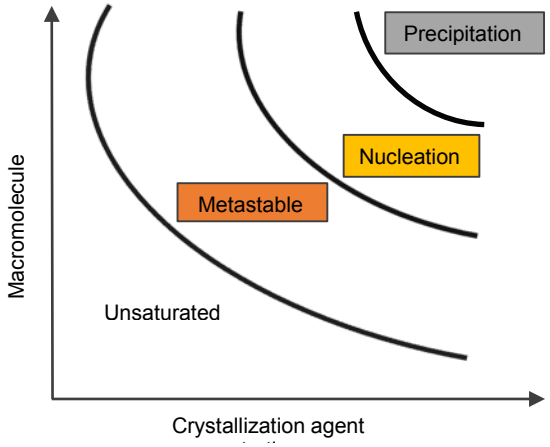


Figure M1. Crystallization phase diagram showing different zones related to crystal formation and growth.

3. Crystallization

3.1 Theoretical Bites

Crystallization [98] is a multi-parameter process where a protein (a macromolecule or small molecule in general) in solution undergoes a phase transition and forms ordered arrays (crystals), induced by supersaturation and chemical conditions. In order for crystals to be formed an energetic barrier must be overcome (for nucleation to occur). Considering other parameters constant, a phase diagram can be constructed, varying the concentration of the crystallizing agent on the horizontal axis and the protein concentration on the vertical axis (Figure M1). If both the variables have high values, as in the precipitation zone, the protein will form amorphous precipitate instead of ordered crystal lattices. In the nucleation zone nuclei can form but they cannot grow. Already formed nuclei can grow in the metastable zone. So the ideal case would be to move sufficiently into the nucleation zone to promote nucleation; as the nuclei form there will be a drop in protein concentration as the protein molecules leave the solution. The system will then move to the metastable zone where the nuclei will grow in size. Different crystallization methods approach this in different ways. In the vapor diffusion technique, a drop is created by mixing appropriate amounts of the macromolecular and crystallization solutions (the latter consisting of a precipitant or crystallizing agent, a buffer and additives) and left to equilibrate over a well solution containing the precipitant. The entire setup is sealed off. The system starts at an unsaturated state and as the solvent molecules leave the crystallization drop, the drop undergoes saturation followed by super-saturation. At appropriate concentrations of the precipitant the super-saturation is just right to enter the nucleation zone. Other methods like the micro-batch method start directly at or near the nucleation zone,

whereas strategies combining the two methods are also available where drops are set-up in micro-batch mode but covered with a low density oil which allows slow evaporation of solvent from the drop [98]. Still other methods such as capillary counter-diffusion and dialysis are also available. Temperature is an additional parameter which can be varied to facilitate the crystallization process. Crystallization is an extensively elaborate science and just a brief overview is provided above.

3.2 Method

Duplex DNA fragments for crystallization were prepared by annealing complementary oligonucleotides: (5'AATAATAAATTATATAATATAA3' and 5'TTATATTATATAATTTATTATT3' for Af2_22; 5'AATAA5-BrUAAATTATATAATATAA3' and 5'TTATATTATATAATTTATTATT3' for Af2_Br22 and 5'TTATATAATATAAAATAATAAA3' and 5'TTTATTATTTTATATTATATAA5' for Af2_shift22) in 20mM Tris-HCl pH 7.5, 100mM NaCl at a concentration of 0.1mM. The mixtures were heated at 85^{o} C for 30 minutes followed by slow cooling to room temperature overnight. Protein-DNA complex for crystallization was prepared by mixing Abf2p and DNA fragment at a molar ratio of 2:1 (protein concentration 0.5 mg/ml) and performing stepwise overnight dialysis to reach a final buffer composition of 50mM Tris-HCl, pH 7.5, 20mM NaCl. (3500 Da cutoff dialysis membrane -pre-wet for 10-15 minutes). The salt concentration was reduced in three steps to 500 mM, 250 mM and finally to 20 mM NaCl. At each step the incubation time was 2 hours at 4 ^{o}C, except overnight incubation for the last step. Subsequently the complex was concentrated to 10-12 mg/ml and initial crystallization trails were setup with crystallization condition screens prepared at the Automated Crystallography Platform located at the Science Park, Barcelona. Such sparse matrix screens include PAC1 (Crystal Screens I and II from Hampton Research), PAC2 (Wizard Screens I and II), PAC3 (Index), PAC 4 (Salt RX) and PAC10 (Protein-DNA Screen) in 96 well sitting-drop vapor-diffusion format at 20 ^{o}C. Crystals for Abf2p/22, Abf2p/Br22 and Abf2p/shift22 were obtained in conditions of 21-25 % w/v PEG 4000, 0.1 M Tri-sodium citrate pH 4.5 and 0.2 M Ammonium acetate by addition of oil (*Al's Oil* from *Hampton Research*- 1:1 v/v mixture of Silicon and Paraffin oil) on top of the well solution (400 µL) or on top of the drop (2 µL). Crystals were cryo-protected with a solution of 15% glycerol, 21-25% PEG 4000, 0.1 M Tri-sodium citrate pH 4.5, 0.2 M Ammonium acetate and vitrified in liquid nitrogen. Data collection of intermediate crystals and optimized ones was performed at synchrotron ALBA (Cerdanyola del Vallès, Spain) and European Synchrotron Radiation Facility, ESRF (Grenoble, France).

4. Crystallographic Structure Solution

4.1 Theoretical Bites: Crystalline arrangement, Diffraction Physics and Fourier synthesis

A crystal comprises of a **unit cell** [99,100] which is the smallest building block that when translated in three dimensions generates the crystal. A unit cell in turn comprises of an **asymmetric unit** [99] which is the smallest unit (comprising of one or more copies of the crystallized molecule), multiple copies of which are arranged inside the unit cell according to the symmetry inherent to the crystal. The asymmetric unit itself can possess symmetry between its components and is referred to as **non-crystallographic symmetry**.

The goal of an X-ray diffraction experiment is to construct a three dimensional density describing the contents of the unit cell from measurements of intensities of X-rays diffracted by the crystalline matter.

In a diffraction experiment one obtains intensity measurements only when **Bragg's law** [101,102] is satisfied**:**

$$2dsin\theta = n\lambda$$

where d is the inter-planar distance between a set of parallel crystal lattice planes, θ is the angle at which the X-ray is diffracted, λ is the X-ray wavelength and n is a positive integer. Thus the condition for constructive interference (formation of a diffraction spot on the detector) is that the path length difference between two diffracted waves ($2dsin\theta$) has to be an integral multiple of the X-ray wavelength (William Lawrence Bragg and William Henry Bragg, 1913) [101].

The intensities $I(\boldsymbol{h})$ recorded on a detector in a diffraction experiment are related to the 'structure factor' $F(\boldsymbol{h})$ by the equation

$$|F(\boldsymbol{h})| = \sqrt{\frac{kI(\boldsymbol{h})}{Lp}}$$

where $|F(\boldsymbol{h})|$ is the magnitude of the structure factor, L is the Lorentz factor (related to the amount of time a crystal remains in position to allow diffraction from a set of parallel lattice planes), p the polarization factor (related to the way the X-rays are monochromated in a particular setup) and k is a

constant (depends on intensity of the X-ray beam, size of the crystal etc and is used as an overall scaling parameter for all reflections in a diffraction dataset).

The calculation of accurate values for the structure factor amplitudes is crucial because electron density ($\rho(x,y,z)$) of the unit cell is related to the structure factor (which is modelled as a vector in the complex plane and thus consists of a magnitude $|F(\boldsymbol{h})|$ and a phase angle Φ: $F(\boldsymbol{h}) = |F(\boldsymbol{h})|\boldsymbol{e}^{-\mathrm{i}\Phi}$) according to the equation:

$$\rho(x,y,z) = \frac{1}{V}\sum_{\boldsymbol{h}} F(\boldsymbol{h})\boldsymbol{e}^{-\boldsymbol{2\Pi i(hx+ky+lz)}}$$

$$or$$

$$F(\boldsymbol{h}) = \int_{\boldsymbol{0}}^{\boldsymbol{c}}\int_{\boldsymbol{0}}^{\boldsymbol{b}}\int_{\boldsymbol{0}}^{\boldsymbol{a}} \rho(x,y,z)e^{\boldsymbol{2\Pi i(hx+ky+lz)}}dx\,dy\,dz$$

where V is the unit cell volume, a, b, c are the dimensions of the unit cell and h, k, l are three numbers defining parallel sets of lattice planes for the crystal under investigation. Thus electron density of the unit cell can be reconstituted from knowledge of the $F(\boldsymbol{h})$ values (both amplitude and phase) in a Fourier synthesis.

While the former (amplitudes) are readily available from a X-ray diffraction dataset, the latter (phases) are not recorded and this corresponds to the "**phase problem**" in crystallography [103]. To overcome this, several strategies are available (see below).

4.2 Theoretical Bites: Data Processing Prior to Structure Solution

In order to obtain a three dimensional (3D) map of the molecule of interest, the acquired diffraction data first needs to be processed [104] . A typical diffraction data processing protocol involves the following steps:

a. Data Reduction

Diffraction spots on a subset of 2D diffraction images are first located and their coordinates are converted to approximate 3-D scattering vectors in the reciprocal space to attribute h, k, l values to

individual diffraction spots (indexing). Subsequently an initial unit cell (triclinic) is defined from which a '**reduced unit cell**' (i.e. with the shortest cell edges and angles closest to 90^{o}) is obtained by transformations of the initial cell to each of the 44 characteristic lattices belonging to the 14 Bravais lattices and subsequently applying a penalty scheme to select the one with the least amount of deviation. This process is called **cell reduction**. Subsequently, the unit cell parameters and diffraction geometry parameters (detector distance, beam stop position etc) are refined to obtain better estimates.

The last step in the data reduction process is integration and involves calculating intensity values for each reciprocal lattice point corresponding to which spots are recorded in the dataset. This is done by assessing the background and the spot regions and summing the pixel counts for the spot region accompanied by background subtraction. Integration can be done via 2D (*iMosflm*) [105] or 3D (*XDS* [106]) integration schemes. The final goal is to obtain intensity estimates for all recorded reciprocal lattice points along with error estimates for the same (σ(I)). The output intensities at the end of integration are no longer raw intensities but have been corrected for polarization of the X-ray beam and for Lorentz factor.

b. Symmetry Detection

The true symmetry of the crystal (symmetry by which asymmetric units are repeated inside the unit cell) is a hypothesis until the structure has been solved. However, analysis of the symmetry of the diffraction pattern (Laue symmetry) allows to deduce the point group (**rotational symmetry**). Monitoring of **systematic absences** (absence of intensities, due **screw symmetry, glide symmetry or lattice centering**) allows to select the most probable space group (described by combination of rotational and translational symmetry existing in 3 orthogonal directions).

c. Scaling, Merging and Truncation

A diffraction dataset will typically contain multiple recordings of the same **reflection** (diffraction intensity from the same set of parallel lattice planes) as well as symmetry related reflections which ideally should have the same intensity values. The process of scaling thus attempts to minimize the discrepancy between individual reflection intensities and the weighted mean of the intensities from all symmetry related reflections. This process helps to make the diffraction data internally consistent and puts it on a common scale.

Merging combines partially recorded observations to produce complete reflections (this is true for 2D integration. In case of 3D-integration the merging of partially recorded reflections into complete ones

happens as part of the integration step). Additionally, symmetry related reflections are averaged to produce unique **reflections**. For native datasets the **Friedel** pairs (reflection pairs hkl and the 180^0 symmetry mate –h-k-l) are additionally merged. This is due to centro-symmetry of the diffraction pattern (**Friedel's law**-see below). In case of anomalous datasets, they are kept separate as presence of anomalous scatterers breaks Friedel symmetry.

The last step in the data processing workflow involves generating structure factor amplitudes from the estimated intensity values. This is done according to the algorithm of French and Wilson, 1978 [104].

d. Data Quality Estimators

In order to assess the overall quality of the processed data, a number of estimators are available. The most relevant ones are mentioned below.

$CC_{1/2}$ and CC*

$CC_{1/2}$ [107] is a Pearson correlation coefficient between two random half-datasets and thus monitors the agreement between intensity averages (X and Y) calculated from the two halves:

$$CC_{1/2} = cov\frac{(X,Y)}{\sigma_X \sigma_Y}$$

and

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

where σ stands for standard deviation. CC^* provides the extent of agreement between experimental data and the underlying true signal. These two estimators can be used to assess the extent of significant signal or information in the data and thus to demarcate the resolution of the dataset.

R_{meas} or $R_{r.i.m}$

This estimator is the redundancy independent merging R factor [108,109] and provides the precision of individual recorded intensities based on unmerged data. It is defined as:

$$R_{meas} = \frac{\sum_{\boldsymbol{h}} \left\{\frac{n}{n-1}\right\}^{1/2} \sum_i I_i(\boldsymbol{h}) - \langle I_i(\boldsymbol{h})\rangle}{\sum_{\boldsymbol{h}} \sum_i I_i(\boldsymbol{h})}$$

where n is the number of observations or redundancy of the refection $I_i(\boldsymbol{h})$, i denoting its i[th] observation.

4.3 Theoretical Bites: Anomalous Scattering and Single Wavelength Anomalous Diffraction

When an X-ray photon of insufficient energy (away from the characteristic absorption edge(s) of the atom) hits an electron in the atom, it is scattered but cannot trigger electronic transitions (jumping of electrons to higher energy levels). However, at or near the absorption edge, three principal kinds of interactions happen: some photons are just scattered, some are absorbed and re-emitted with a lower energy and some are absorbed and re-emitted immediately at the same energy. This third type lag behind a normally scattered photon (in the classical theory parlance this is interpreted as a change in phase of the anomalously scattered wave with a gain in the imaginary component (imaginary here refers to complex number theory). This shift in amplitude and phase is termed anomalous scattering [110] and can be used to circumvent the problem that phase information is not an observable in a X-ray diffraction experiment and only amplitudes are recorded. There are two components involved in anomalous scattering, a dispersive component f' and an anomalous component f'' [110]. The anomalous component is 90[o] out of phase with respect to the dispersive component. f' and f'' are related by the Kramers-Kronig equation [111]

$$f'(\omega) = \frac{2}{\pi} \int_0^{\omega} \frac{\omega' f''(\omega') \partial \omega'}{\omega^2 - \omega'^2}$$

where ω is the complex variable.

The shift in phase due to anomalous scattering causes **Friedel's law** ($F(\boldsymbol{h}) = F(-\boldsymbol{h})$) to break down (where $F(\boldsymbol{h})$ is the structure factor corresponding to the point $\boldsymbol{h}$ in reciprocal space and $F(-\boldsymbol{h})$ is its complex conjugate, corresponding to the structure factor for the centrosymmetric point $-\boldsymbol{h}$). Thus from the knowledge of the difference in the amplitudes of Friedel pairs (which is a direct observable in the diffraction experiment as difference in the corresponding intensities due to the relation

$$|F(\boldsymbol{h})| = \sqrt{\frac{kI(\boldsymbol{h})}{Lp}}),$$

along with an appropriate available model of the heavy atom causing the anomalous diffraction, Harker diagrams can be constructed and the phase problem resolved [110].

To enable anomalous scattering experiments, heavy atoms (with absorption edges reachable by X-ray sources such as synchrotrons or in-house sources) are introduced into the crystals, either by post-crystallization soaking in harvesting solution consisting of the crystallization condition with ~5% higher precipitant concentration and the metal of choice (Hg, Pt, Au, Gd etc) or by incorporation into the protein primary structure (substitution of methionines with seleno-methionines and cysteines with seleno-cysteines). Alternatively, selenium derivatized DNA (selenium introduced into DNA bases or sugars) can be used [112,113]. As the f' and f'' values for heavy atoms depend on their chemical environment and do not generally agree with standard values determined in vacuum, a fluorescence energy scan is typically performed prior to anomalous diffraction data collection to identify the peak absorption energy and concomitantly f' and f''. For a single wavelength anomalous diffraction experiment (**SAD**) [114–117] diffraction data are collected at the peak or on the high energy side of the peak as f'' decays slowly on the high energy side (this is particularly useful for heavy atoms for which the exact absorption edge cannot be approached due to technical limitations of synchrotron sources).

In contrast to **SAD**, in Multiple Wavelength Anomalous Diffraction (**MAD**) [115–117], datasets are collected from the same or isomorphous crystals (having identical or near identical unit cell dimensions i.e. less than 5% difference) at the peak, the inflexion and optionally at low and high energy remote wavelengths. Harker diagrams are constructed in a manner similar to SAD. However, the two-fold phase ambiguity inherent in **SAD** is not present in **MAD** due to intersection of the phase circles at a unique point [110].

4.4 Theoretical Bites: Molecular Replacement

When a model is available which closely resembles the structure to be solved, phase information can be obtained from it [118]. However, this requires defining the position of the molecule inside the unit cell of the crystal under consideration. This position is specified by three rotational angles related to the orientation of the molecule inside the unit cell along with three translational vectors which specify the position of the molecule. This search is most effectively performed in the **Patterson** space. A Patterson map is a map of all interatomic vectors and is essentially centrosymmetric (as a vector can be drawn from atom A to B or from B to A). The interatomic vectors are of two classes: intramolecular interatomic

vectors which are distributed closer to the origin and intermolecular interatomic vectors which are distributed away from the origin. The orientation of a molecule is specified only by its intramolecular vectors and thus a rotational search in Patterson space is done first as it involves only a part of the Patterson map near the origin (and thus is computationally inexpensive). Once the orientation is known, the intermolecular interatomic vectors (which depend on both orientation and position) are utilized to place the molecule in the unit cell. Therefore, a six-dimensional search (3 angles of rotation and 3 vectors of translation) is converted into two 3-dimensional searches, which contributes to save computational time [118]. Present day molecular replacement programs such as ***Phaser*** [119] use maximum-likelihood based functions for the rotation and translation searches with additional packing criteria (that monitors appropriate packing of the placed model in the unit cells).

4.5 Theoretical Bites: Structure Refinement and Validation

Once an initial Fourier map is obtained, manual or automated model building can be performed to construct an initial model of the crystallographic asymmetric unit (and thus of the unit cell by application of symmetry) which typically consists of the macromolecule itself, a solvent model, along with other associated ligands that are visible in the map. However, this model needs refinement of the atomic positions in order to represent the experimental data (diffraction data) at its best. The atomic positions are related by chemical bonds, angles, charge interactions, van der Waals contacts, etc. All these parameters do not have a single value but oscillate within a range of values determined experimentally, from which dictionaries are available to restrain the relative positioning of atoms. In addition, during refinement the temperature factors or atomic displacement factors (ADPs) (which indicate static and dynamic disorder in the crystal) are also adjusted. The ultimate goal is to generate a model which best fits the data and prior knowledge, while avoiding overfitting. At present, almost all refinement programs employ a maximum likelihood based function [120] which maximizes the probability of the data (given the model) by adjusting model parameters. The exact implementations differ from program to program (***Phenix*** [119], ***Refmac*** [121], ***Buster*** [122] etc). A typical structure solution trial involves iterative cycles of model building and refinement until the disagreement between the model and the data cannot be reduced any further or are within an acceptable range. The most frequently used indicator of data-model agreement is the R factor:

$$R = \sum(||\boldsymbol{F}_0| - |\boldsymbol{F}_C||)/\sum|\boldsymbol{F}_0|$$

where F_0 and F_C are the structure factor amplitudes corresponding to the data and the model respectively. The R factor alone is not a good evaluator of data-model fit as it is affected by overfitting. For this purpose, an R_{free} value is calculated where the disagreement between a set of diffraction data points (which are assigned arbitrarily in most cases and are not used in refinement) and the corresponding values calculated from the model are monitored. A high difference between R and R_{free} indicates excessive overfitting of model to data, whereas a close agreement (difference below 5-6 %) indicates acceptable overfitting.

At the end of refinement, the final model has to be checked for physical, chemical and biological sense. Several validation metrics are available including steric clash scores, Ramachandran plot (monitoring a proper distribution of backbone torsion angles), ADP checks, amino acid side chain rotamer outliers etc [123] . At the end of validation, the model is ready for interpretation.

4.6 Method: Preparation of heavy-atom derivative crystals for Experimental Phasing

Heavy atom derivatives were prepared by post-crystallization soaking (concentrations of heavy atoms ranging from 0.5 mM to saturated solution, in combination with different soaking times: 5 mins to 48 hrs) as well as co-crystallization (0.1 to 10 mM concentration in the drop) using the following compounds: $HgCl_2$, $Hg(OAc)_2$, $AuCl_3$, and $PtCl_4$ (covalent binders); $Pb(OAc)_2$, $Yb(OAc)_3$, NaBr, and NaI (non-covalent binders) as well as Ta6Br(2+)12 (cluster compound). Derivatization with 5-Amino-2,4,6-triiodoisophthalic acid (I3C or *JBS Magic Triangle* from *Jena Bioscience*) was performed by co-crystallization with 10 mM I3C.

4.7 Method: Single wavelength Anomalous Diffraction (SAD) Data Collection

Abf2p-SeMet/22 crystals were diffracted at ID-23 1 beamline at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. Test diffraction patterns were collected with a transmission of 100 %, exposure time of 0.037 sec and 1^o oscillation at a beam flux of 1.62 e+12 photons/sec. Data collection strategy included collection of Friedel pairs in the same image by using the mini-kappa device

and setting the kappa orientation manually. Three consecutive datasets, each with 360° rotation and an oscillation range of 0.15°, were collected at 10% transmission at the peak wavelength from the same crystal. The datasets were processed using *XDS* [106] and put on a common scale with the first data set as reference using *XSCALE* [106]. Anisotropic scaling was performed before heavy atom search in *Shelx D* [124]. Initial phasing was performed with *Shelx E* [124] without density modification and autotracing. The resulting map was used to perform density modification using solvent flattening in *DM* from the *CCP4* suite [125].

Iterative rounds of manual model building and refinement were performed using *Coot* [126] and *phenix.refine* [119] respectively. Only the first of the three collected datasets was used for this purpose. The positions of the selenium atoms were used to guide protein sequence assignment.

4.8 Method: Native Data Collection and Structure Solution

X-ray diffraction datasets for the native Abf2p-Af2_22 protein-DNA crystals were obtained at ID14-4 (ESRF, Grenoble, France) on a ADSC QUANTUM 315r charge-coupled device (CCD) detector at 12.658 keV. The data were collected in two passes- a low resolution pass with an oscillation range of 1.5° and a nominal resolution of 2.6 Å; a high resolution pass with an oscillation range of 0.55°, exposure time of 1.1 sec at 74 % transmission (flux 4.89e+11 photons/sec) with a nominal resolution of 2.18 Å. Datasets were processed using *XDS* [106] and combined and scaled using *XSCALE* [106]. The partial model from experimental phasing was truncated to a poly alanine model and all B factors were set to 20 Å^2. The partially built DNA was retained in the search model. Molecular replacement trials were carried out using *Phaser* [119] (in *Phenix* suite). The Abf2p/shift22 dataset was collected at beamline ID29 (ESRF, Grenoble, France) and the structure was solved by a similar strategy.

4.9 Method: Confirmation of DNA Sequence Register-Abf2p/Af2_Br22 crystals

Crystals of Abf2p in complex with the modified sequence (Af2_Br22) were obtained as discussed in the **Crystallization** section. X-ray diffraction dataset was collected at ID23 1 beamline (ESRF, Grenoble, France) up to a nominal resolution of 3.37 Å. The X-ray energy for data collection was set to 13489 eV The dataset was processed using *XDS* and merged using *Aimless* [127]. Molecular replacement was performed with a poly-alanine version of the refined Abf2p/22 model with all B factors set to 20 Å^2.

Subsequent iterative model building and refinement were carried out using *Coot* [126] and *phenix.refine* [119]. As the dataset showed presence of weak anomalous signal, the refined model (*Rfree* 0.30) was used in conjunction with unmerged data to generate an anomalous density map with *AnoDe* [128]. The procedure involves calculating the native phases ϕ_{native} from the macromolecular atomic coordinates and using them to calculate a Fourier synthesis with coefficient F_A (where A stands for the anomalous scatterers) and $\phi_A = \phi_{native} - \alpha$, where α is estimated from the anomalous differences.

5. Small Angle X-ray Scattering (SAXS)

5.1 Theoretical Bites

Small angle X-ray Scattering (SAXS), in the context of macromolecular characterization, is a solution scattering technique where elastic scattering of X-rays at low angles (0.1 to 10°) is used to extract 'low-resolution' structural information (down to 20 Å) [129,130]. Since macromolecules in solution undergo rotational tumbling and their orientations and positions are not correlated, the scattering intensities in a particular direction sum up (due to absence to inter-particle interference). The scattering pattern from the ensemble of molecules in the solution is thus continuous, isotropic and is proportional to the spherical average of scattering from a single particle in all possible orientations. If the sample additionally possesses conformational heterogeneity, the extent to which each conformation contributes to the total scattering is dictated by its frequency in solution. One of the most important requirements of SAXS is homogeneity in terms of the species present in the solution to be characterized, though this can be alleviated to some extent by coupling to size exclusion chromatography (SEC). However, conformational heterogeneity often cannot be avoided for macromolecules.

For macromolecules, SAXS measurement involves separately measuring the scattering from the macromolecular solution and that from the solvent. If the solvent is modelled as having a constant scattering density of ρ_s, the Fourier transform $\boldsymbol{F}$ of the difference scattering density $\Delta\rho(\mathbf{r})$ defines the difference scattering amplitude from a single particle relative to an equivalent solvent volume:

$$A(\boldsymbol{s}) = \boldsymbol{F}[\Delta\rho(\mathbf{r})] = \int \Delta\rho(\mathbf{r})\exp(\mathrm{i}\mathbf{s}\mathbf{r})d\boldsymbol{r}$$

where **r** is the position vector in real space and **s** is the reciprocal space vector denoting the difference

between the wave vectors of the incident and the scattered beams [129]. The integral is over the volume of the particle. The spherically averaged scattering intensity from a particle (or equivalently the scattering intensity of the ensemble is solution) can be described as:

$$I(\boldsymbol{s}) =< A(\boldsymbol{s})A*(\boldsymbol{s}) >_{\Omega} =< \int\int \Delta\rho(\mathbf{r})\Delta\rho(\mathbf{r}')\exp\{(i\mathbf{s}(\mathbf{r}-\mathbf{r}')\}d\boldsymbol{r}d\boldsymbol{r}' >_{\Omega}$$

where $A*(\boldsymbol{s})$ is the complex conjugate of $A(\boldsymbol{s})$ and Ω denotes spherical average [129].

a. The Guinier plot

At very small (**s** tending to zero) and very high (**s** tending to infinity) values of the reciprocal space vector (or equivalently of momentum transfer s=$4\pi\lambda^{-1}sin\theta$ where 2θ is the scattering angle), the scattering intensity is directly related to overall particle parameters. At s<1.3/Rg, where Rg is the radius of gyration, the Guinier approximation [129] becomes valid, i.e.

$$I(\boldsymbol{s}) = I(0)\exp(-\frac{1}{3}\mathrm{R}g^2s^2)$$

Thus from a plot of ln $I(\boldsymbol{s})$ over s^2 (Guinier plot), the zero angle scattering intensity $I(0)$ can be determined from the y intercept and Rg can be determined from the slope. As $I(0)$ is related to the mass of the molecule, from the knowledge of $I(0)$, the molecular mass can be determined using data on a relative scale (e.g. the BSA method where scattering from bovine serum albumin or BSA is used as a reference) or on an absolute scale (using scattering of water as a reference) [129,131].

b. The Kratky plot

The degree of compactness of a protein can be analysed from SAXS data using the Kratky plot [129] where $s^2I(\boldsymbol{s})$ is plotted against s. For globular proteins the plot shows a bell shaped appearance whereas for Gaussian chains or extended/non-globular conformations it plateaus at large values of s.

c. The Distance Distribution Function

The scattering intensity $I(\boldsymbol{s})$ can be written as:

$$I(\boldsymbol{s}) = 4\pi \int_0^{Dmax} r^2\,\gamma(r)\,\frac{\sin sr}{sr}\,dr$$

$$and\ \gamma(r) = < \int \Delta\rho(\boldsymbol{u})\Delta\rho(\boldsymbol{u}+\boldsymbol{r})\,du >_r$$

Here $r^2\gamma(r) = p(r)$ describes the distribution of distances within the particle and can be calculated from experimentally observed scattering intensity as an inverse Fourier transform [129] :

$$p(r) = \frac{r^2}{2\pi^2}\int_0^{\infty} s^2\,I(s)\,\frac{\sin sr}{sr}\,dr$$

This resulting real space plot reveals information regarding the shape of the particle. For example, a spherical particle has a characteristic p(r) function which is different from that of a rod shaped particle. Thus crude information regarding the shapes of macromolecules under study can be obtained in a simple intuitive manner [130].

d. Porod Volume

The Porod approximation [129] states that the scattering intensity $I(\boldsymbol{s})$ can be written as:

$$I(\boldsymbol{s}) = K * s^{-4}$$

where K is a constant. Additionally,

$$\frac{K}{Q} \propto \frac{S}{V} \text{ where } Q = \int_{s=0}^{\infty} s^2 I(\boldsymbol{s})ds,$$

$\frac{S}{V}$ being the shape to volume ratio. Q is called the Porod invariant. For homogeneous particles (i.e. particles with no significant variation in intra-particle contrast),

$$Q = 2\Pi^2(\Delta\rho)^2V,$$

and given that $I(0) = (\Delta\rho)^2V^2$, the excluded (Porod) volume

$$V = 2\Pi^2 I(0)Q^{-1}.$$

In general particle inhomogeneity leads to deviation from the Porod approximation, which can however be taken care of by subtracting a constant from the data [129]. The Porod volume can be used to estimate the molecular mass (MM) of the particle under investigation by applying the rule of thumb [131]

$$\frac{V}{2.0} \leq MM \leq \frac{V}{1.5}$$

The estimated molecular mass does not depend on normalization of $I(0)$ by dividing it by the concentration of the sample and thus provides a better way of MM estimation in comparison to reference based MM estimation methods for SAXS (e.g. BSA method).

e. Ensemble Optimization Method

Ensemble optimization method [132,133] allows to create ensembles of conformations (starting from atomic structures obtained experimentally or through homology modelling) and to fit their calculated average scattering intensity profile with experimentally observed SAXS profile. The domains are specified as pdb files and are treated as non-flexible bodies that are connected by flexible regions. For the latter, residue coordinates are reduced to alfa-carbons which preserve a regular standard distance between them but are free to rotate in space. A sequence file is required as an input which serves to designate the domain regions and flexible parts. A large pool of conformations is generated (typically 10000), from which sub-ensembles are chosen based on a genetic algorithm. The average intensity profile for each sub-ensemble is matched to the experimental one(s). Finally, the sub-ensemble with the best fit to the experimental data is selected. The fit is monitored based on a Chi squared value as:

$$\chi^2 = \frac{1}{K-1}\sum_{j=1}^{K}\left[\frac{cI(s_j) - I_{exp}(s_j)}{\sigma(s_j)}\right]^2$$

where $I(s_j)$ and $I_{exp}(s_j)$ are the calculated average intensity from the sub-ensemble and the experimental values respectively for the jth data point, K is the number of data points, $\sigma(s_j)$ is the standard deviation and c is a scaling factor [131].

EOM thus offers an opportunity to study conformational changes of macromolecules from experimental data obtained in solution. A distribution of the radii of gyration (Rg) can be plotted for the final selected sub-ensemble and this gives a picture of the conformational freedom accessible to the macromolecule under study. However, the coordinates belonging to the sub-ensemble do not depict exact

conformational states adopted by the macromolecule and the R(g) distribution is more meaningful [133]. Nonetheless, representation of the sub-ensemble models gives an intuitive picture of the conformational space explored by the molecule in solution.

5.2 Method: SAXS Sample Preparation

For SAXS measurements, purified Abf2p was dialyzed overnight in 50mM Tris-HCl pH 7.5 and 500mM or 150mM NaCl. The protein samples were subsequently concentrated. The filtrate was used as a blank in all cases to obtain maximal buffer match between the blank and the sample. The Abf2p/22 complex was prepared identically as for crystallization. Protein concentration was measured using absorbance values obtained from NanoDrop 1000 Spectrophotometer and extinction coefficients calculated from the amino acid sequence (EnCorBio server: http://encorbio.com/protocols/Prot-MW-Abs.htm). For protein-DNA complex, concentrations were estimated using the Bradford method, *Biorad Protein Assay, Biorad*).

5.3 Method: SAXS Data Collection and Processing

SAXS measurements were performed at BM29 beamline at the European Synchrotron Radiation Facility (ESRF). Measurements were performed at 20^{o} C, using 100% beam transmission and 30 to 50 µL sample volume per injection in flow mode. The X-ray wavelength utilized was 12.5 keV with a detector distance (Pilatus 1M) of 2.867 m. Exposure time per frame was set to 2 sec. Guinier approximation was used to calculate the forward scattering (*I(0)*) and radius of gyration (*Rg*). Pair distribution (*p(r)*) plot and maximum particle dimension (*Dmax*) were calculated from scattering data using *GNOM* [131]. Data analysis was performed using *Primus* [131] in *ATSAS* package. Molecular weight of the protein was estimated from the Porod Volume. Ensemble analysis was performed using *EOM 2.0* [133] with standard parameters for the genetic algorithm and for *Crysol* [131]. For *EOM*, the N-assembly and HMG-box2 were treated as rigid domains while allowing flexibility to the linker (113-Lys-Leu-Pro-115, residue 113 added to include the unwinding of HMG-box1 helix3 C-terminus as observed from MD simulations- see Results), the N-terminal 6-His tag and the last two C-terminal residues.

6. Molecular Dynamics Simulations

6.1 Theoretical Bites

Molecular dynamics or MD simulations [134] comprise a set of computer simulation techniques which allow us to analyze the physical movements of a system of interest based on the assumption that once the initial position and velocities of the particles involved are known, their future positions and velocities are completely determined by the laws of classical mechanics (Newton's laws of motion). All-atom MD simulation pertains to solving the N-body problem and achieves that by numerically solving Newton's equations of motion. It requires a set of initial conditions (positions and velocities) and a potential energy function (which depends on the positions of all the particles (atoms) constituting the system. The force $\boldsymbol{F}_i$ on each particle is estimated as a gradient of this potential energy function U :

$$\boldsymbol{F}_i = -\boldsymbol{\nabla} r_i U(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots \dots \boldsymbol{r}_n)$$

where $\boldsymbol{r}$ stands for position vector. Thus, an MD simulation run consists of defining the positions and velocities of the constituent particles, calculate the force on each particle at current time t_n, solving equations of motion for all the particles over a time step Δt and write desired properties to output files. These steps are repeated in a loop until the end of the simulation time is reached.

MD simulation allows estimation of macroscopic or average properties of a system from its microscopic manifestations. In statistical mechanics such averages are referred to as **ensemble averages** where *an ensemble is defined as the set of all possible states of a system which have different microscopic properties but the same macroscopic or thermodynamic properties*[134]. For example, for a protein molecule it can refer to the set of different conformations that are present in a sample at a point of time. The assumption that MD simulations can be used to estimate ensemble averages, rests on the **ergodic hypothesis** which states that *time average equals ensemble average*[134,135]. Thus for ergodic hypothesis to hold, the simulation time must be long enough so that adequate portion of the conformational space is sampled.

6.2 Simulation Setup

All-atom MD simulations for the protein, DNA and protein-DNA complexes were performed using the *Amber 14* force-field with particle-mesh-ewald (PMEMD) for long range electrostatics. All simulations

were performed with explicit water (TIP3P) and ions (150mM NaCl) using a truncated octahedron periodic box. For simulations consisting of DNA or protein-DNA complexes additionally *parmbsc1* force field was used for improved parameterization of DNA [136]. Models were prepared for simulation using *tleap* in *AmberTools 15*.

6.3 Minimization

Energy minimization was performed using a combination of steepest descent and Truncated Newton Conjugate Gradient (TNCG) method with single point energy calculation [137]. Minimization was performed in 5 steps, reducing the restraint weight on the macromolecular atoms from 25 to 5 kcal/mol-Å^2 from the first step to the last (reduction by 5 at each step). At each step, minimization method was changed from steepest descent to conjugate gradient after 500 cycles and a maximum total number of 1000 cycles was allowed at each step.

6.4 Thermalization

Thermalization was performed for 100 ps to bring the system from 0 K to 298.15 K using the weak coupling algorithm [138] and a constant volume periodic box, with 2 fs time step. Bonds involving hydrogen atoms were constrained. Non-bonded cut-off was set to 9.0 Å. A restraint weight of 5 kcal/mol-Å^2 was applied to all macromolecular (protein &/or DNA) atoms to hold them near to their initial position during the thermalization.

6.5 Equilibration

Equilibration was performed at constant mass, temperature and pressure (NPT) using a Berendsen thermostat (weak coupling) [138] and isotropic position scaling as implemented in *Amber 14*, with 2 fs time step. Reference pressure was set to 1.0 bar with a time constant of 2.5 psec. Temperature was maintained as in thermalization with a time constant of 2.5 psec. No restraints were applied and non-bonded cut-off was set as in thermalization.

6.6 Production Run

Production run was performed in 20 batches of 25 ns each, amounting to a total simulation time of 500 ns. Time constant for pressure and temperature coupling were set to 5 psec. Other parameters were set identically to those in equilibration.

6.7 Analysis of Results

Analysis of trajectories were performed with *cpptraj* [139]. Visualization was performed using *vmd* [140]. Analysis of DNA structural parameters were performed using *Curves+* [141,142]*, R* and *MS-Excel* were used for plotting.

6.8 DNA stiffness and deformation energy calculations

DNA structures can be described in terms of base-pair and base-step parameters that consist of three translations (shift, slide and rise) and rotations (tilt, roll and twist) and the DNA deformability along these six directions can be described by the associated stiffness constant matrix [143,144]. From the ensemble of MD simulations, the covariance matrix describing the deformability of the helical parameters for a given DNA fragment (e.g. a dinucleotide step) is computed and is inverted to generate the 6x6 stiffness matrix for each fragment. Pure stiffness constants corresponding to the six parameters mentioned above (kshift, kslide, krise, ktilt, kroll and ktwist) are extracted from the diagonal of the matrix while the total stiffness (Ktot) is obtained as a product of these six constants and provides a rough estimate of the flexibility of each base pair step. The normalized Boltzmann-like probability distribution,
exp(-ΔEdef/kbT), was derived from the deformation energy (ΔEdef) and was used to determine the relative probability of positioning the HMG-boxes on a given DNA segment. The deformation energy was calculated using a mesoscopic energy model [143,144], which is based on a harmonic approximation to describe deformability along DNA helical parameters. The equilibrium geometry and stiffness force constants were extracted from a dataset built from long all-atoms MD simulations of short DNA fragments in water using the *parmbsc1* force field.

7. Circular Dichroism

7.1 Theoretical Bites

A material is termed **dichroic** if it absorbs light of different polarization by different amounts or which causes visible light to split up into distinct beams of different wavelengths [145]. If a molecule contains chiral chromophores it absorbs left and right handed circularly polarized light (L-CPL and R-CPL) to different extents. Here it needs to be mentioned that L-CPL and R-CPL refer to two different spin angular momentum states for the photon. In the classical electromagnetic theory parlance, in L-CPL the electric field vector of the electromagnetic radiation undergoes left handed rotation around the wave propagation

vector and in case of R-CPL it undergoes a right handed rotation. Thus the tip of the electric field vectors (which stay constant in magnitude) trace out a helix along the direction of wave propagation.
Thus in a typical circular dichroism (CD) experiment the dichroic sample is irradiated with equal amounts of L-CPL and R-CPL and the difference in absorbance between the two is measured as:

$$\Delta A = \Delta \varepsilon C l$$

where C is the molar concentration, l is the path length in cm and Δε is the difference in the molar extinction coefficients for L-CPL and R-CPL and the molar circular dichroism. The measurements are usually reported in terms of molar ellipticity θ (milli degrees) where

$$\tan\theta = (E_R - E_L)/(E_R + E_L) \text{ and } \theta = 3298.2\Delta\varepsilon$$

E_R and E_L being the L-CPL and R-CPL electric field vector magnitudes.
Thus a CD spectrum for a molecule of interest typically consists of a scan across a wavelength range and measures the extent of dichroism as a function of wavelength. For proteins and peptides ultra violet (UV) CD spectra in the far UV region (180-260 nm) is typically done to analyze secondary structural features and UV CD spectra analysis in the near UV region (260-300 nm) for tertiary structural features.

7.2 Method: Sample Preparation

Circular dichroism experiments were performed in the far UV region (260-185 nm) to check for stability and secondary structure content of wild type Abf2p and its deletion mutants that were used to perform functional assays. Protein purification for mutants were performed identically to that for the wild type. The protein samples were dialyzed overnight in 10 mM potassium phosphate pH. 7.5, 100 mM ammonium sulphate.

7.3 Method: Data Collection and Analysis

A quartz sample cell of 1 mm path length was used for data collection on a *Jasco* J-815 CD spectrophotometer, with a scan rate of 50nm/min, data pitch of 0.5 nm resulting in 150 data points between 260-185 nm. Protein concentration was kept constant at 0.1 mg/ml for all samples.

8. Isothermal Titration Calorimetry (ITC)

8.1 Theoretical Bites

Isothermal titration calorimetry (ITC) [146] is a quantitative technique that allows estimation of enthalpy changes (ΔH), stoichiometry (n) and binding affinity (K_a) pertaining to interactions between two or more molecules. It additionally provides information on the change in entropy and Gibbs free energy due to the relation:

$$\Delta G = -RTlnK_a = \Delta H - T\Delta S$$

where ΔG is Gibbs free energy change, R is the universal gas constant, T the absolute temperature and ΔS the entropy change. Thus, ITC allows analyses of the thermodynamics involved in intermolecular interactions. It can be used effectively to obtain information on protein-ligand, protein-protein or protein nucleic-acid interactions.

An ITC calorimeter consists of two identical cells made of a chemically inert metal that has high thermal conductivity (such as Hastelloy or gold). The **reference cell** is filled with buffer or water and the **sample cell** contains the macromolecule of interest. A sensitive circuitry detects temperature differences between the two cells. Prior to the start of the experiment (ligand addition), a constant power is applied to the sample cell that maintains a constant temperature difference between the two cells. The ligand is added in known aliquots and the corresponding generation or absorption of heat is measured as a time dependent input of power required to maintain the two cells at constant temperature. The resultant raw data thus consists of spikes of power per injection. These peaks are then integrated to obtain estimates of heat exchange per injection. The resulting data series can then be analyzed and fitted to a suitable equation (**binding model**) to obtain the information on the involved affinity and thermodynamics.

8.2 Method: ITC sample preparation and data collection

Samples for ITC were prepared by simultaneously dialyzing the protein and DNA in the same buffer (25mM Hepes pH 7.5, 150mM NaCl) to obtain maximal buffer match between samples. Titrations were performed with a *VP ITC* instrument with 1400 µL cell volume and 300 µL syringe volume, at a temperature of 25^{o} C. The general setup was designed with protein in the cell and DNA in the syringe. The protein was used at a concentration of 8-16µM and the DNA at a concentration of 100 to 150µM.

DNA into buffer titrations were performed as control and since the profile was flat, subtraction of the control was not performed. The resulting data were analyzed using *Origin (Origin Lab, Northampton, MA)*.

9. Abf2p *in vivo* assays

Diploid ABF2 (YMR072W) hemizygous strain Y26205 (BY4743; MATa/MATα; ura3Δ0/ura3Δ0; leu2Δ0/leu2Δ0; his3Δ1/his3Δ1; met15Δ0/MET15; LYS2/lys2Δ0; YMR072w/YMR072w::KANMX4) was obtained from the European Saccharomyces cerevisiae Archive for Functional Analysis (*Euroscarf*). A plasmid-borne ABF2 was made by cloning a PCR fragment including Abf2 CDS plus 500nt on both sides into pRS316 [147] (forward primer: 5' ACTAACCaagCTTGGATTATACTAATGATAC 3' with HindIII restriction site; reverse primer 5' GGTATTTCtagAAAAAGATAACTTCAAGTTTTCACC 3' with Xba1 restriction site). This construct was introduced in Y26205 diploid cells, which were subsequently sporulated. The tetrad sacs were disintegrated with *zymolyase* (*Ecogen*). Kan+ (ABF2 disrupted) Ura+ (Abf2p-pRS316 transfected) haploid cells were isolated and the haploid condition was confirmed by PCR (forward primer: 5' AGTCACATCAAGATCGTTTATGG 3' common for both mating types a (Mat a) and α (Mat α); reverse primer 5' ACTCCACTTCAAGTAAGAGTTTG 3' produces product of 544 bp if mating type a; reverse primer 5' GCACGGAATATGGGACTACTTCG 3' produces product of 404 bp if mating type α). Additionally, their genotype was verified by PCR[148] (forward primer: 5' ACTAACCAAGCTTGGATTATACTAATGATAC; 3' reverse primer: 5' GGGCCTCCATGTCGCTG 3'). ABF2 truncations were made by PCR which included the 500nt flanks, and were cloned into pRS315 (a plasmid providing Leucine prototrophy [147]). To test ABF2 truncations, we followed a plasmid-shuffling strategy using 5-fluroorotic acid (5-FOA), toxic for Ura+ cells[149,150]. The mitochondrial DNA integrity of the corresponding Kan+ Ura- Leu+ cells was tested by growth in glycerol (respiration-only conditions).

Results

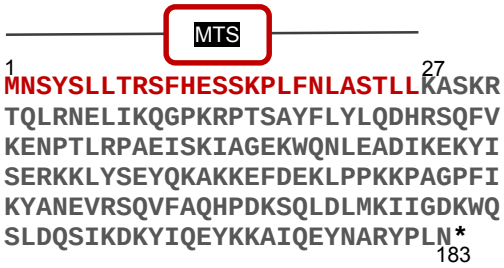


Figure R1. **Amino acid sequence of Abf2p**. The 26 residue mitochondrial targeting sequence (MTS) is shown in red. The mature protein starts at amino acid 27.

1. Protein Purification

Abf2p was cloned from genomic DNA using standard PCR. The primers were designed to add a six-Histidine tag at the N-terminus of the protein. Abf2p has a 26 residue mitochondrial targeting sequence [70] (Figure R1) which is not present in the mature protein and thus was not included in the construct. Maximum solubility of the protein was observed in 50 mM Tris-HCl, pH 7.5, 750 mM NaCl (assessed by SDS-PAGE) and thus this buffer composition was used in all subsequent stages of purification. The final protocol for Abf2p purification consisted of a Ni-affinity step followed by size-exclusion chromatography. Wild type Abf2p (and its seleno-methionine derivative; see below) was purified to >98% purity (Figure R2). The purified protein was used for crystallization in complex with DNA (Methods). For this purpose, several DNA fragments were tried, all derived from the mtDNA 64bp sequence (See introduction; Figure R3).

2. Crystallization

Well-diffracting crystals were obtained for native Abf2p bound to the A-T only double-stranded (ds)DNA, Af2_22 (5'-AATAATAAATTATATAATATAA-3', Figure R3, Figure R4), spanning 22 of the 35bp sequence that is amplified as concatamers in ρ- yeast mitochondria (position 19477 to 19511bp in y-mtDNA, see Introduction). Based on EMSA assays a 2:1 protein:DNA stoichiometry was used for crystallization as at this ratio most of the free DNA was bound to the protein (Figure R5). The DNA sequence contains a near match of the ARS consensus sequence (ARSc) that has been reported to prevent binding of Abf2p to DNA [70,71] (Figure R12a,b,d). Initial crystallization screening trials did not return any direct hit. However, a strong phase separation was observed for condition A9 of 96 well sitting drop PAC1 screen (30% w/v PEG 4000, 0.1 M Tri-sodium citrate pH 5.6, 0.2 M ammonium acetate).

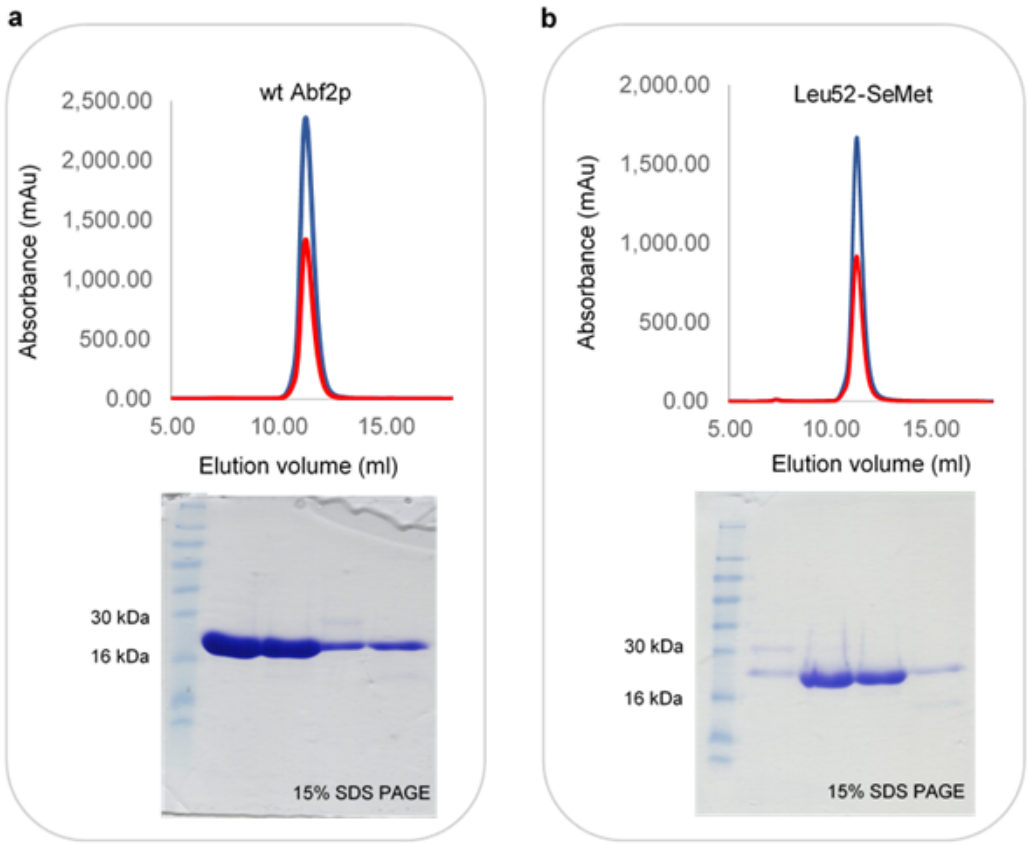


Figure R2. Size exclusion chromatography profile and analysis of the purified fractions by SDS-PAGE for wild type (wt) Abf2p and Leu-52-Semet derivative.

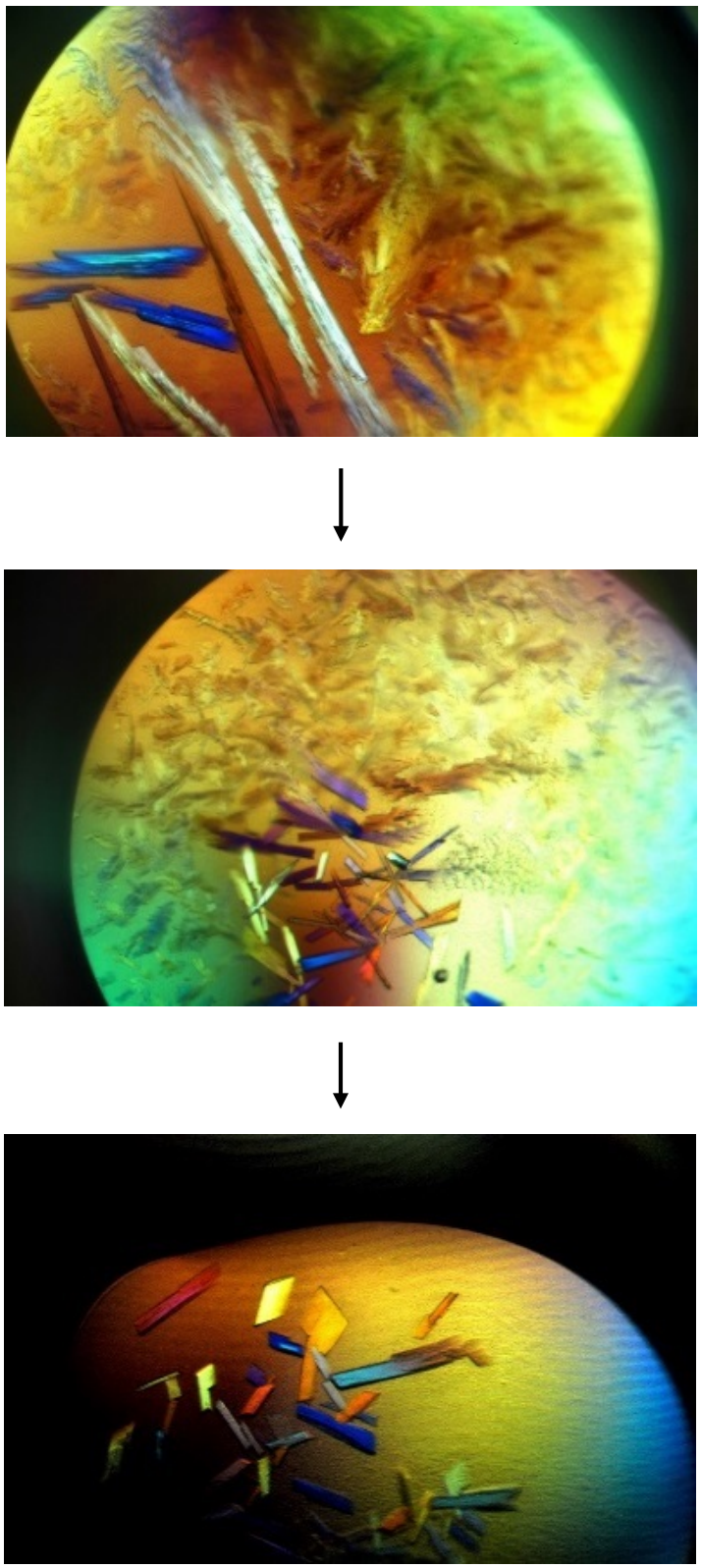


Figure R3. **Steps in obtaining diffraction quality crystals for Abf2p/22.** The rate of equilibration of the crystallization drops ' was slowed down by addition of 1:1 silicon:paraffin oil on top of the drops to obtain well-diffracting crystals.

dsDNA fragments	Crystals?	Diffraction
Af2_22a 5′AATAATAAATTATATAATATAA3′ Af2_22b 3′TTATTATTTAATATATTATATT5′	Yes	Yes (2.18 Å)
Af2_22.ovhng1A 5′TAATAATAAATTATATAATATAA 3′ Af2_22.ovhng1B 3′ TTATTATTTAATATATTATATTA5′	No	-
Af2_22.ovhng2a 5′AATAATAAATTATATAATAT 3′ Af2_22.ovhng2b 3′ ATTATTTAATATATTATATT5′	No	-
22Trunc1a 5′AATAATAAATTATATAATAT3′ 22Trunc1b 3′TTATTATTTAATATATTATA5′	No	-
22Trunc2a 5′AATAATAAATTATATAATATAAT3′ 22Trunc2b 3′TTATTATTTAATATATTATATTA5′	No	-
Af2_A4_22a 5′TATATAAAATAATAAATTATAT3′ Af2_A4_22b 3′ATATATTTTATTATTTAATATA5′	No	-
Af2_shift22a 5′TTATATAATATAAAATAATAAA3′ Af2_Shift22b 3′AATATATTATATTTTATTATTT5′	Yes	Yes (2.6 Å)
35bpseq_28bp1a 5′AATTATATAATATAATATATATATATATAT3′ 35bpseq_28bp1b 3′TTAATATATTATATTATATATATATATATA5′	Yes	Poor (~10 Å)
35bpseq_28bp2a 5′TAAAATAATAAATTATATAATATAATAT3′ 35bpseq_28bp2b 3′ATTTTATTATTTAATATATTATATTATATA5′	Yes	Poor (~15 Å)
GC_22a 5′GAAGATATCCGGGTCCCAATAA3′ GC_22b 3′CTTCTATAGGCCCAGGGTTATT5′	Yes	Poor (~15 Å)
35bp_noTract22a 5′TATATAATATAATATATATATATA3′ 35bp_noTract22b 3′ATATATTATATTATATATATAT5′	Yes	Poor (~15 Å)

Figure R4. Double stranded (ds) DNA fragments used for crystallization trials. Poly-adenine tracts (A-tracts; see later) are indicated in red when present. Success in crystallization trials for each of them and diffraction quality of obtained crystals are indicated.

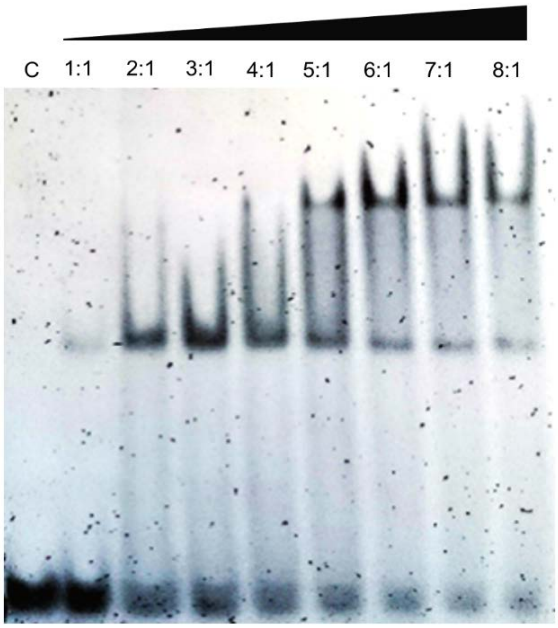


Figure R5. EMSA assay for Af2_22 DNA in 10cm 10% native poly-acrylamide gel. C stands for the free DNA control. The DNA is kept fixed at 200 nM and the protein concentration is increased from left to right. The protein to DNA molar ratio is indicated at the top of each well.

The condition was scaled up to 24 well sitting drop setup with 1 µL:1 µL protein/DNA complex to precipitant ratio and grid screens were setup varying precipitant concentration along horizontal axis and buffer pH along the vertical axis. This approach rendered crystals after two days of drop setup. However, diffraction quality was poor (4-5 Å) with streaky spots. Equilibration rate of the drops was subsequently reduced by addition of Al's oil (see Material and Methods), yielding the best diffracting crystals that appeared 3-4 days after drop setup (Figure R3). This crystal and the corresponding structure will be referred to as Abf2p/22 in all subsequent sections.

3. Structure Solution

X-ray diffraction data for optimized native Abf2p/22 crystals were collected at ID-14-4 beamline of the European Synchrotron Radiation Facility (ESRF), Grenoble, France. The best dataset was collected in two passes: a low resolution pass at 2.6 Å and a high resolution pass at 2.1 Å. Molecular replacement trials with X-ray diffraction data from native crystals using homology models did not yield any clear solution. Therefore, experimental phasing was attempted. To this end, several derivatives were prepared by using $HgCl_2$, $Hg(OAc)_2$, $AuCl_3$, and $PtCl_4$ (covalent binders); $Pb(OAc)_2$, $Yb(OAc)_3$, NaBr, and NaI (non-covalent binders) as well as Ta6Br(2+)12 [151,152] (cluster compound). However, the derived crystals diffracted poorly in comparison with native crystals and did not show significant anomalous signal (judged by anomalous correlation at different resolution shells after data processing in *XDS*). The 'Magic Triangle' (5-Amino-2,4,6-triiodoisophthalic acid) derivative crystals yielded usable anomalous signal, though the resulting Fourier maps were not interpretable. Subsequently, seleno-methionine incorporation was attempted in order to use the anomalous scattering from selenium atoms for obtaining experimental phase information in single or multiple wavelength anomalous diffraction (SAD/MAD) experiments. However, the native protein contains only one methionine residue (M147; out of 157 residues). Thus, in order to obtain sufficient anomalous signal for phasing (abiding by the rule of one selenium per 100 amino acids) [153] methionines (Met) were engineered into the protein by making amino acid substitutions which have been reported to be the least disruptive to native protein structures [153] (Figure R6). Selection of mutable residues was based on their similarity in terms of volume with methionine. Additionally, mutations of alanines to methionines were carried out. (Figure R6). The resulting single-site mutants were tested for solubility and the mutant Leu-52-Met was selected for

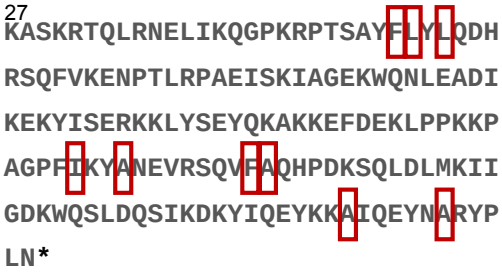


Figure R6. Amino-acid mutations to introduce additional methionines for Abf2p-SeMet derivatives

seleno-methionine incorporation (Materials and Methods). The best Abf2p-SeMet/22 crystal produced diffraction spots till 3.0 Å. Previous to data collection a fluorescence energy scan was performed. The peak of the absorption edge was determined to be 12.6623 keV with f' and f'' of -7.52 e and 6.53 e respectively. The inflection point was located at 12.6594 keV with f and f'' of -10.65 e and 3.04 e respectively. This crystal was aligned with the mini-kappa goniometer in order to collect both members of a Friedel pair on the same image and hence to minimize differences in noise and radiation damage between them. Thus Friedel pairs were recorded at the same point in time, preserving the maximal amount of mutual information. Shutter-less data collection was performed at the peak wavelength on a PILATUS 6M-F (Dectris) detector. Three datasets from the same crystal were collected and processed (see Materials and Methods). 8 SeMet atom positions were determined by SAD, and subsequent phasing yielded an interpretable, albeit noisy, Fourier map (Figure R7). The positions of the selenium atoms served as a guide to assign regions of the map to HMG-box1 or 2, as the positions of the seleno-methionine residues differed in the two HMG-boxes. Model building for the SeMet derivative dataset was difficult in certain regions due to very poor Fourier map density. Model building and refinement with this dataset was carried out until a considerable portion of the protein and DNA were built and an *R free* of 0.33 was reached (Table R1). Subsequently, this partial model was used for molecular-replacement (MR) with native Abf2p/22 dataset. A MR solution was obtained with *Phaser* in the *Phenix* suite. Analysis of the crystal packing for the solution indicated its validity (Figure R8) and subsequent iterative model building and refinement led to a well-refined and validated structure (Figure R9).

3.1 Confirmation of Sequence Register

Iterative manual model building and automated refinement of the Abf2p/DNA structure revealed that Abf2p binds two different DNA molecules via its two HMG-box domains (hereafter referred to as dual

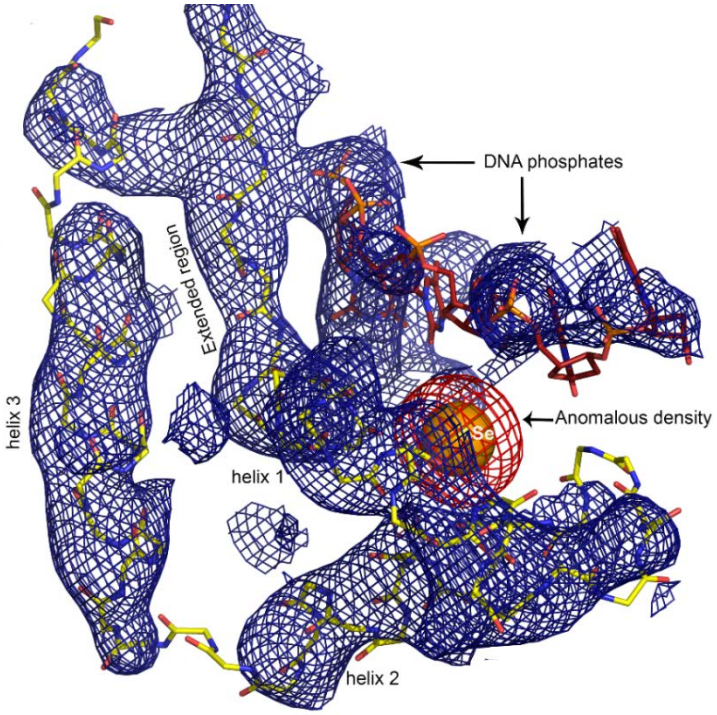


Figure R7. SAD phasing for Abf2p-Semet/22. Fourier map from experimental SAD phasing (blue; 1.0 r.m.s.d) along with built partial model (protein in yellow and DNA in red). The anomalous density corresponding to the selenium (Se) atom (5.0 r.m.s.d) is shown in red and the Se atom position as orange sphere. The recognizable secondary structure elements of HMG-box2 are indicated along with the DNA phosphates.

Table R1

Wavelength	0.9792
Resolution range	47.8 - 3.3 (3.418 - 3.3)
Space group	C 1 2 1
Unit cell	82.733 126.116 138.945 90 93.33 90
Total reflections	147248 (14974)
Unique reflections	21267 (2114)
Multiplicity	6.9 (7.1)
Completeness (%)	0.99 (0.99)
Mean I/sigma(I)	11.46 (2.31)
Wilson B-factor	92.60
R-merge	0.1303 (0.8182)
R-meas	0.141 (0.8831)
CC1/2	0.998 (0.806)
CC*	0.999 (0.945)
Reflections used in refinement	21267 (2114)
Reflections used for R-free	1058 (106)
R-work	0.2713 (0.3407)
R-free	0.3344 (0.4038)
CC(work)	0.922 (0.565)
CC(free)	0.901 (0.564)
Number of non-hydrogen atoms	7223
Macromolecule atoms	7223
Protein residues	607
RMS(bonds)	0.010
RMS(angles)	1.63
Ramachandran favored (%)	73
Ramachandran allowed (%)	14
Ramachandran outliers (%)	13
Rotamer outliers (%)	16
Clashscore	23.73
Average B-factor	125.14
B-factor macromolecules	125.14

* Statistics for the highest-resolution shell are shown in parentheses.

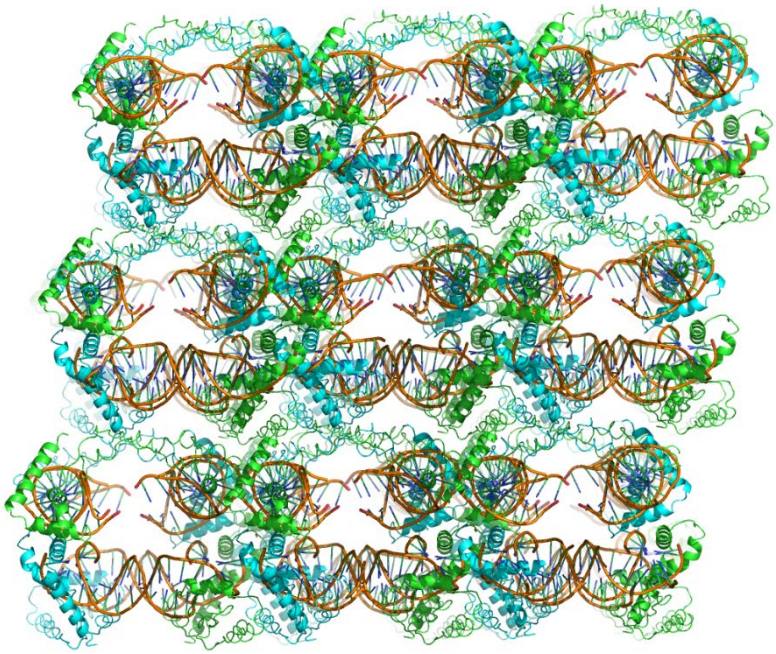


Figure R8. Crystal packing for Abf2p/22 molecular replacement solution obtained with *Phaser* in *Phenix* suite.

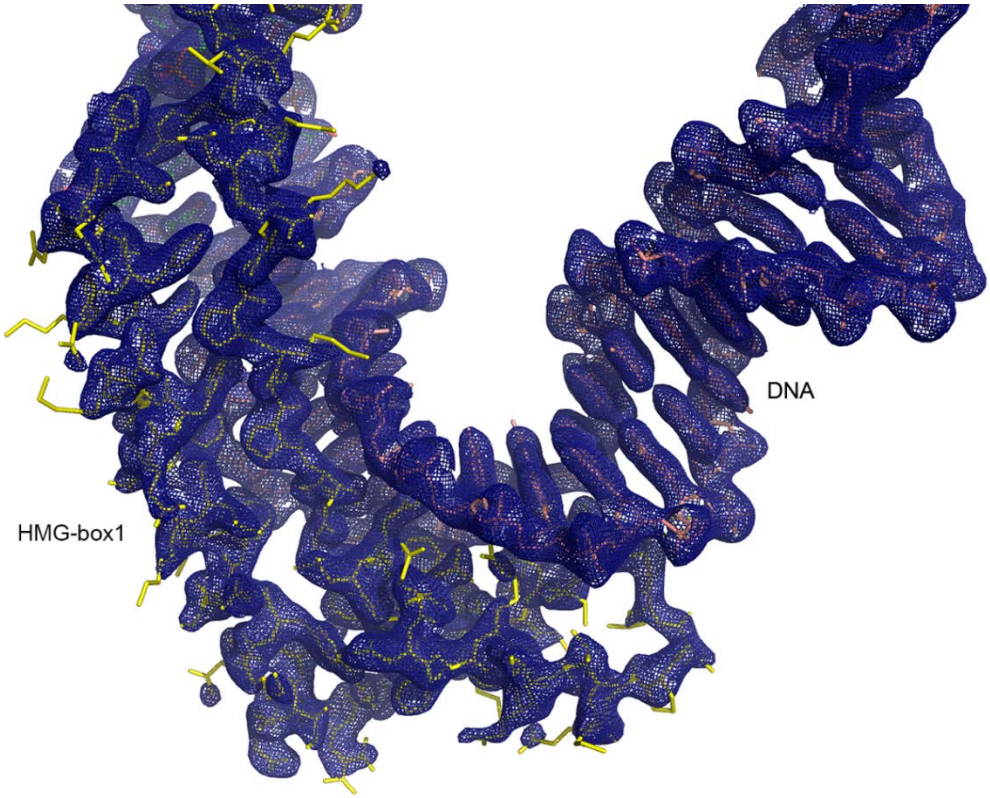
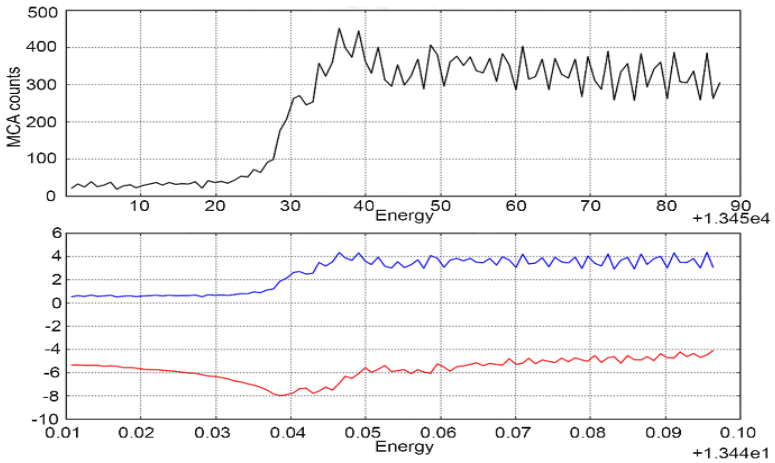


Figure R9. **Map and model quality for Abf2p/22 crystal structure**. A region of the 2mFo-DFc double difference Fourier map (shown as blue density; 1.0 r.m.s.d) and the final Abf2p/22 refined model (shown as yellow sticks). The visible protein and DNA features are indicated.

a



b

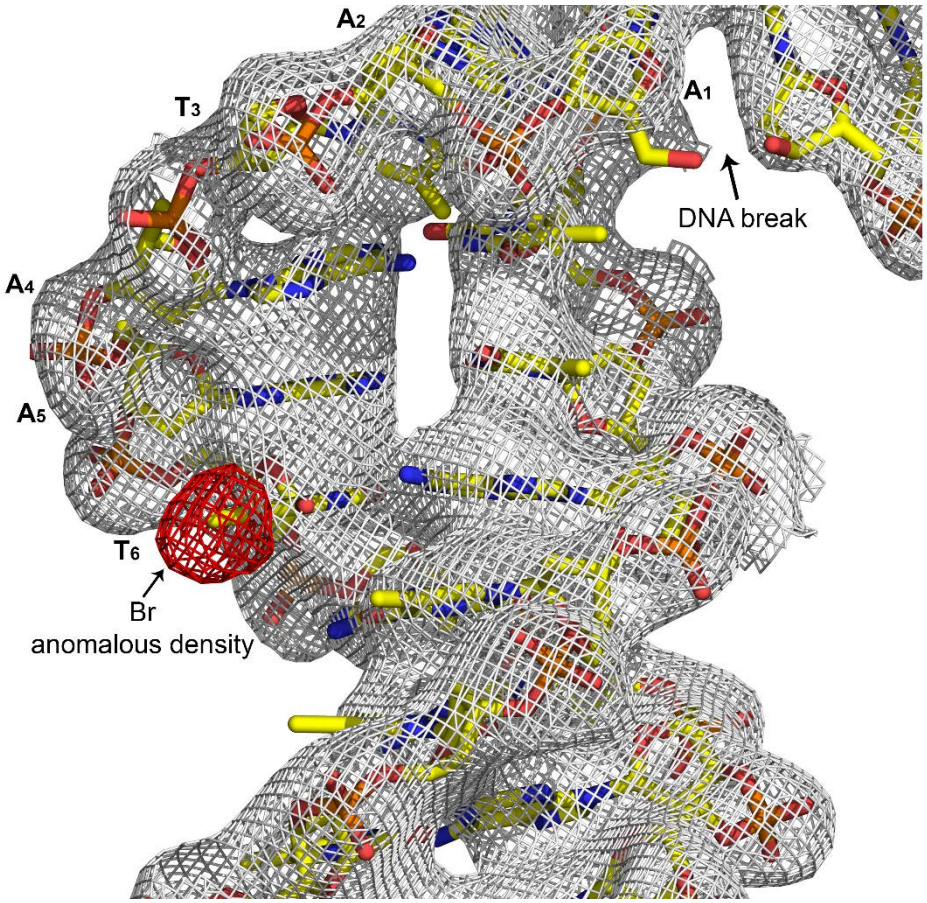


Figure R10. DNA sequence register. (**a**). X-ray fluorescence scan for Abf2p/Af2_Br22 complex crystals showing signal from Br. **b**. Confirmation of sequence register for Abf2p/22 from Abf2p/Af2_Br22 crystals. The anomalous density corresponding to Br is shown in red (5 .0 r.m.s.d) whose position coincides with methyl carbon of thymine 6 (T6). The latter had been replaced by 5-Br-Uracil in the Af2_Br22 DNA. The 2mFo-Fc double difference Fourier map (1.0 r.m.s.d) is shown in grey with the DNA shown as yellow sticks. The visible DNA break is indicated. The protein density is not shown for clarity.

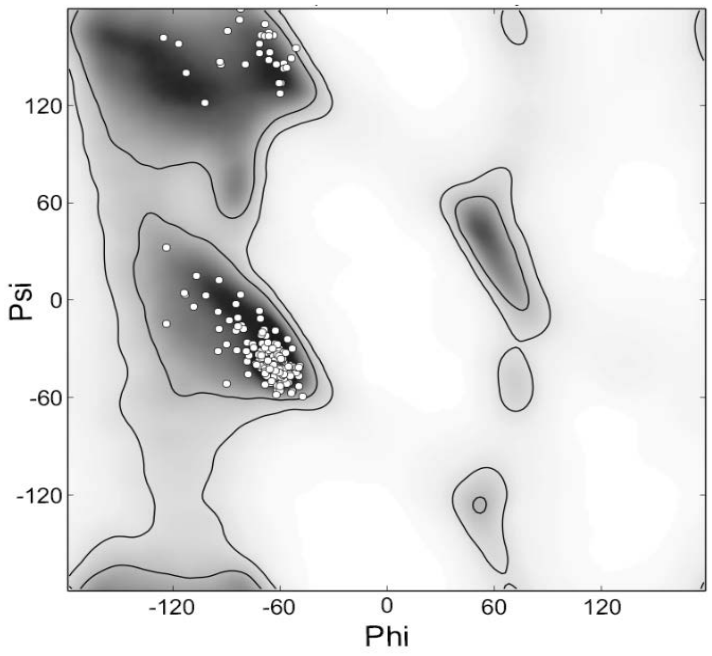


Figure R11. **Ramachandran plot for Abf2p/22 crystal structure**.

binding; see below; Figure R12b,d). To confirm this, portions of the built DNA were deleted and omit maps were synthesized. Additionally, the quality of the Fourier density was good enough to clearly indicate the sequence register of the DNA. However, to consolidate this, a variant of Af2_22 DNA was designed with thymine 6 replaced by 5-bromo uracil (5-BrU). The intention behind this modification was to use anomalous signal from the bromine atom to pin down the sequence register of the DNA and thereby confirm the dual binding. To this end, crystals with brominated DNA (Af2_Br22) were generated and, prior to diffraction data collection, an energy scan was performed. The X-ray energy for data collection was set manually to 13489 eV since the energy scan peak profile was noisy and the automatically detected peak wavelength was not reliable (Figure R10a). Data collection, processing and scaling were performed (see Material and Methods) and subsequently the structure solved by molecular replacement using a poly-alanine search model derived from the refined Abf2p/22 structure. With the refined model and unmerged data, an anomalous density map was calculated (Methods). The map showed peaks (at 5 r.m.s.d.) that corresponded to the position of the Br atoms and thus the DNA sequence register could be ascertained (Figure R10b).

Table R2

Wavelength	0.97950
Resolution range	43.48 - 2.18 (2.258 - 2.18)
Space group	C 1 2 1
Unit cell	88.49 113.41 67.71 90 103.803 90
Total reflections	126763 (6246)
Unique reflections	33305 (3075)
Multiplicity	3.8 (2.0)
Completeness (%)	0.98 (0.92)
Mean I/sigma(I)	12.01 (1.99)
Wilson B-factor	49.68
R-merge	0.06649 (0.5319)
R-meas	0.07359 (0.6935)
CC1/2	0.998 (0.632)
CC*	1 (0.88)
Reflections used in refinement	33305 (3074)
Reflections used for R-free	1881 (164)
R-work	0.2090 (0.3061)
R-free	0.2364 (0.3213)
CC(work)	0.956 (0.754)
CC(free)	0.938 (0.772)
Number of non-hydrogen atoms	4520
Macromolecule atoms	4445
Protein residues	313
RMS(bonds)	0.008
RMS(angles)	0.95
Ramachandran favored (%)	99
Ramachandran allowed (%)	0.65
Ramachandran outliers (%)	0
Rotamer outliers (%)	0.7
Clashscore	8.45
Average B-factor	58.93
B-factor macromolecules	59.04
B-factor solvent	52.14

* Statistics for the highest-resolution shell are shown in parentheses.

4. Structural organization of Abf2p

The crystal structure Abf2p/22 (PDB ID 5JH0; Figure R11, Table R2; Figure R12b,d) reveals an asymmetric unit (a.u) composed of two Abf2p molecules (polypeptide chains Abf2p-A and -D) and two double-stranded (ds) DNAs (chains-BC and -EF, 22bp each) (Figure R12b,d). Abf2p is a two-domain α helical protein belonging to the HMG-box superfamily and consisting of two tandem HMG-boxes. At the N-terminus of the protein, three residues in extended conformation (27-Lys-Ala-Ser-29, henceforth referred to as 'N-flag') and a 12 amino-acid (aa) N-terminal helix (N-helix) (Figure R12c,d and R13a) are found. Notably, such a helix is not present in any structure containing HMG-boxes reported so far. The N-helix leads to HMG-box1 (42-113) (Figure R12c, 13a), which displays a typical HMG-box 'L-shape' consisting of an extended segment (42-48), helix1 (49-64, with a kink at position 57), a four-residue turn,

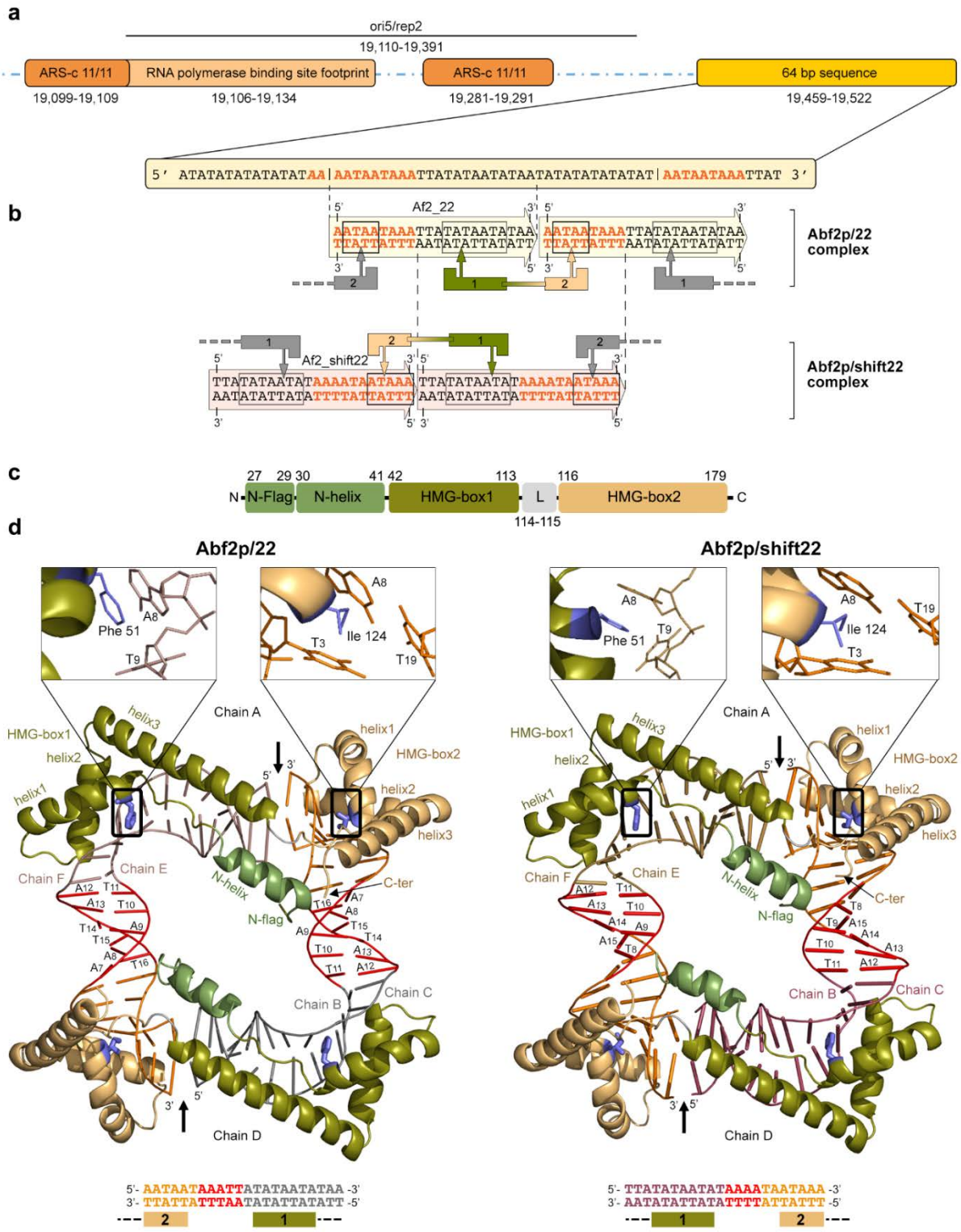


Figure R12. DNA design and Abf2p/DNA complex crystal structures (a) General organization of the y-mtDNA ori5/rep2 origin of replication, including the location of ARS consensus sequences (ARS-c, 11/11 matches to 5'-(A/T)AAA(T/C)ATAAA(A/T)-3', in orange) and the downstream 64bp region (in yellow) in y-mtDNA. Below, the 64bp sequence is shown, the near matches (ARS-m; 9/11) to ARS-c are depicted in orange and the position of the 35bp sequence (see Introduction) is demarcated by vertical bars. Vertical dashed lines indicate the DNA sequence 'Af2_22' within the 35bp fragment. **(b)** The dsDNA sequences used for crystallization are depicted inside arrows that show the head-to-tail arrangement found in the crystals; the DNA regions contacted

by Abf2p are framed. Abf2p is represented as two connected L-shaped boxes 1 and 2 (HMG-box1 and 2, respectively) with arrows representing DNA-inserting residues. The grey HMG-boxes represent the second protein molecule in the asymmetric unit, the dotted lines indicate the domains are connected. From the two 'Af2_22' sequences in tandem, a continuous 'Af2_shift22' dsDNA molecule was derived (vertical dashed lines) that yielded the Abf2p/shift22 structure, where the protein again contacts two DNAs. **(c)** Schematic representation of the Abf2p domains. 'L' stands for Linker, 'N-flag' and 'N-helix' stand for N-terminal flag and helix, respectively. **(d)** The crystal structures of Abf2p/22 (left) and Abf2p/shift22 (right) are shown, with secondary structure elements and domains labeled. The intercalating residues Phe51 and Ile124 are highlighted in blue and their insertion sites are shown in insets. The DNA head-to-tail arrangement is indicated with a black arrow. The ARS-m-like regions (see main text) within Af2_22 and Af2_shift22 are shown in orange, the A-tract sub-region in red. Underneath, the crystallized dsDNA sequences and contacting protein domains are shown; note that the DNA sequences are in the same orientation as in **(b)** while the orientation of Abf2p domains are opposite.

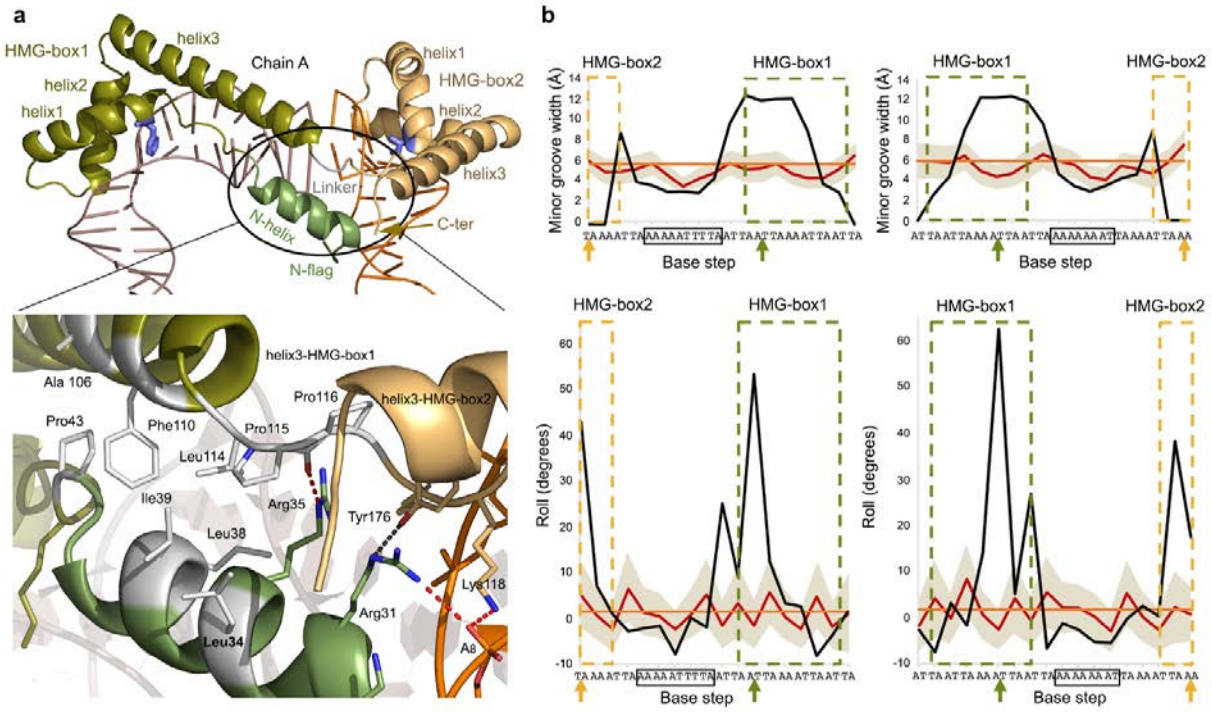


Figure R13. Interactions within Abf2p and DNA dynamics. (a) Above, side view of Abf2p (chain A) from the Abf2p/22 crystal. The encircled secondary structure elements (N-helix, helix3 of HMG-box1, the Linker and helix3 of HMG-box2) participate in the interactions that connect the Abf2p N- and C-terminal regions, shown in detail below. Hydrophobic side chains are shown in white, electrostatic contacts are shown in red (the weak ones in gray). **(b)** The DNA parameters 'minor groove width' (top graphs) and 'roll' (bottom) are shown for each base-pair step along the Af2_22 (left) and Af2_shift22 (right) sequences. Values that correspond to the crystal structures (in black), to the MD simulations for unbound Af2_22 and Af2_shift22 DNAs (averaging over the individual trajectories, in red; spread of the values, in grey), or to an ideal B-DNA (in orange) are shown. Values corresponding to the DNA contacted by Abf2p is framed in olive green for HMG-box1 and light orange for HMG-box2. The arrows colored according to the HMG-boxes indicate the insertion sites. The A-tracts are framed in black. Note the switch of the relative orientations of DNA and protein between complexes (also shown in Figure 12b,d).

helix2 (70-83, which, together with helix1 forms the short L-arm), a short two-residue connecting loop and, finally, helix3 (86-113, antiparallel to the extended segment and forming the long L-arm of the HMG-box L-shape). A Linker of two residues (Leu-Pro) follows, which is much shorter than the predicted length of 10aa [70] (Diffley and Stillman, 1991) or 6 aa (PSIPRED, see Introduction). Similar to HMG-box1, HMG-box2 (116-179) consists of an extended segment (116-121), helix1 (122-137), a five-residue turn, helix2 (143-156, with a kink at position 130) and a two-residue connecting loop leading to helix3 (159-179, Figure R12c,d and R13a). Finally, the last four residues 180-Tyr-Pro-Leu-Asn-183 form a short extended C-terminal segment.

5. Abf2p binds to two separate DNA strands via its two domains

In the a.u. of the Abf2p/22 crystal, two curved dsDNA molecules of 22bp are arranged head-to-tail by two protein molecules, forming a pseudo-continuous DNA circle exhibiting stacking interactions between respective 3'-5' DNA ends (Figure R12d). Thus, each protein holds together and bends two different DNA molecules via its HMG-boxes, much like a staple (hereafter referred to as dual-binding) (Figure R13a). Specifically, HMG-box1 from Abf2p-A contacts dsDNA EF chains (from T13 to A19 of chain-E), while HMG-box2 contacts the other DNA molecule BC (A2-A5 chain-B), with the DNA ends participating in perfect stacking interactions between HMG-box binding sites (Figure R12b,d, R13a). A similar scheme applies for the second protein molecule in the a.u. where HMG-box1 and HMG-box2 contact dsDNA chains BC and EF respectively. Each HMG-box engages in, mostly electrostatic interactions with the DNA minor groove (Figure R14, R15), opens it up to 12 Å (ideal B-DNA standard minor groove width 5.9Å, Figure R13b) and concomitantly bends the DNA by ~90°. Additionally, each HMG-box inserts hydrophobic residues between base steps, thus causing considerable alterations of DNA parameters (Figure R13b, R16). Specifically, HMG-box1 Phe51 inserts at step $A_8T_9/A_{14}T_{15}$ (chains F/E), the phenyl ring stacking with the six-membered ring of A8 (Figure R12b,d). This base-step shows a high positive roll (Figure R16) and the involved base pairs show a prominent buckle (Figure R12d, R13b, R16). In comparison, Ile124 from HMG-box2 inserts partially at $T_3A_4/T_{19}A_{20}$ (B/C), and induces a positive roll (Figure R12d, R13b, R16). Superposition of both HMG-boxes indicates high structural similarity (r.m.s.d. of 1.1Å for 66 Cα), both Phe51 and Ile124 being at equivalent positions in helix1. Thus, the two DNAs that are contacted by a single protein, together form a U-turn (or a half DNA circle, Figure R13a). Importantly, the Abf2p mutant K44A, R45A, K117A and K118A causes 80% of yeast cells to lose

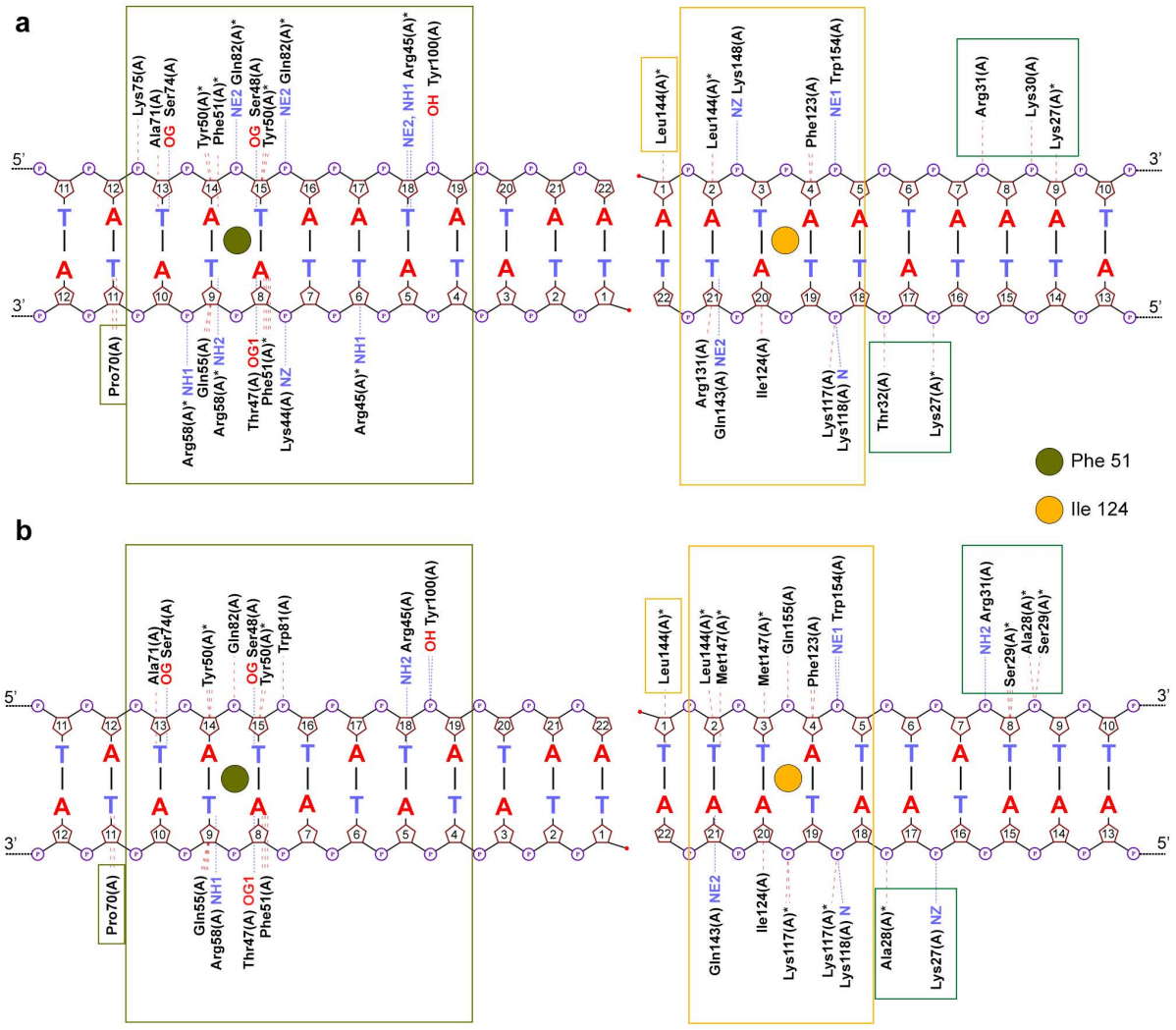
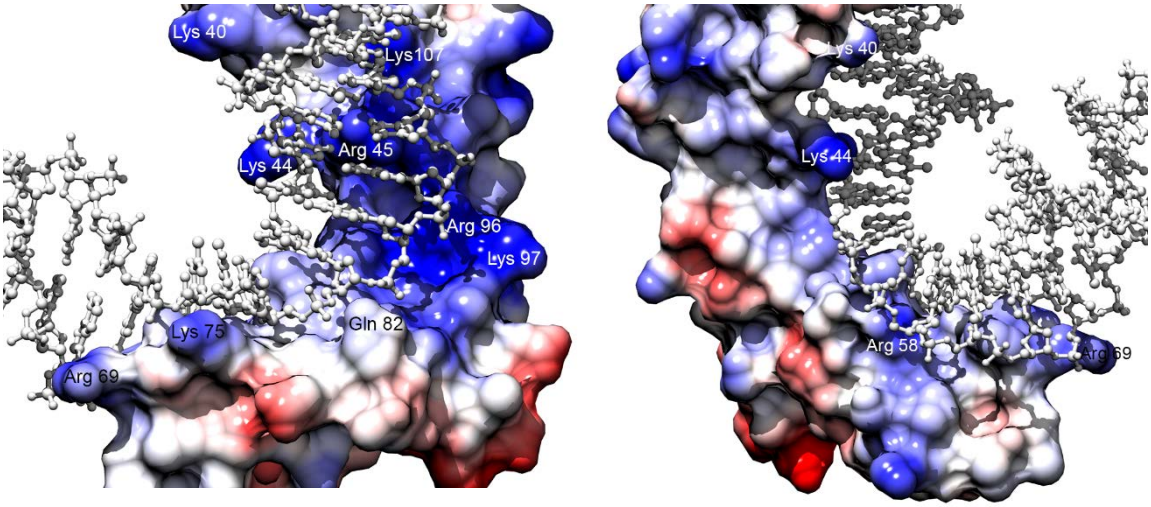


Figure R14. Protein-DNA interactions in the crystal structures. (a) Contacts of one Abf2p molecule with two Af2-22 dsDNAs in Abf2p/22 crystal structure. **(b)** Contacts of Abf2p with Af2-22shift DNA in Abf2p/shift22 crystal structure. Hydrogen bonding interactions are shown as blue dotted lines. Non-bonded contacts are shown as red dotted lines. The regions contacted by N-flag+N-helix (deep green), HMG-box1 (olive green), HMG-box2 (orange), all belonging to the same protein molecule, are demarcated by rectangular boxes. Note the break between DNA sequences contacted by the HMG-box domains of a single protein. Olive green and orange dots indicate the sites of insertion for HMG-box1 and HMG-box2 respectively. Black dotted lines to the left and right indicate the continuation of the DNA strands. Red dots at the 5'end of the DNA molecules represent the missing phosphate in purchased, synthetic oligonucleotides.

a



b

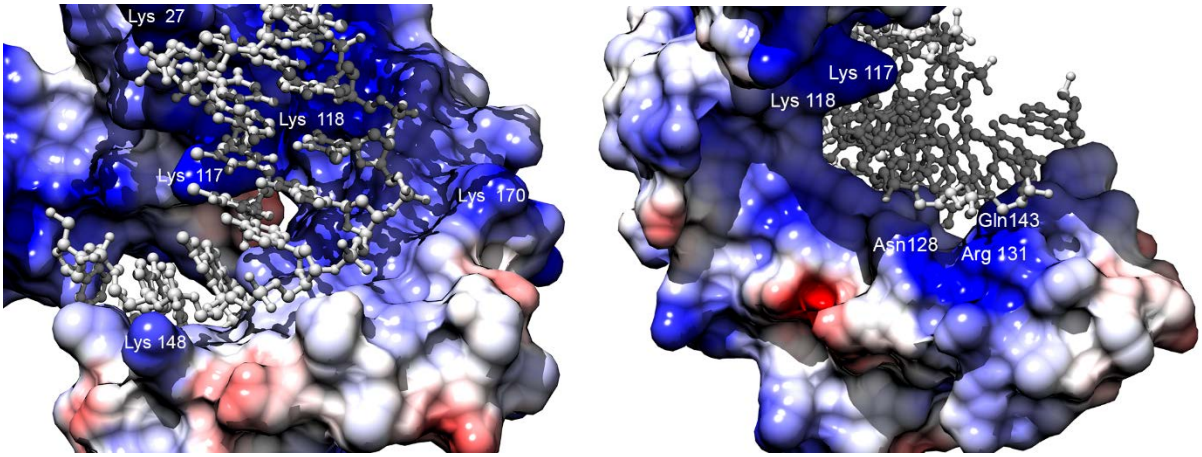


Figure R15. Electrostatic surface representation of HMG-box1 (a) and HMG-box2 (b) showing positively charged amino acid residues at the protein-DNA interface for Abf2p/22 crystal structure. Note that the residues indicated include those that form direct electrostatic contacts (Figure R14) as well as those that are not directly involved in electrostatic contacts but are present in proximity to the DNA.

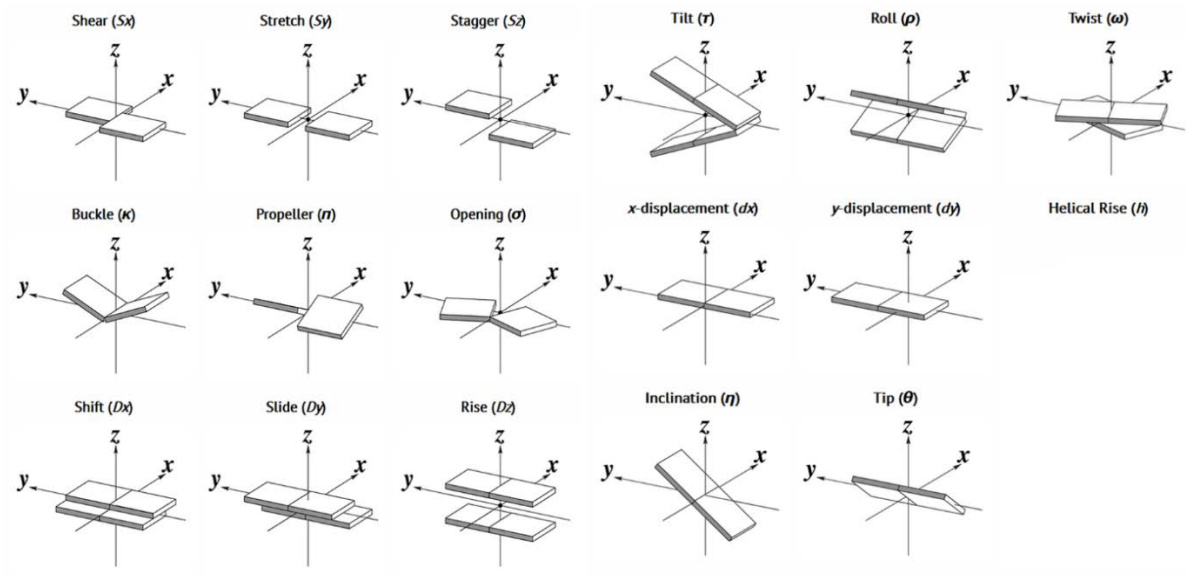


Figure R16. DNA base-pair and base-step parameters for describing DNA conformation. The base pair parameters describe conformation at a particular DNA base pair due to relative orientations of the involved bases, whereas the base step parameters describe conformations resulting from different relative orientations of two adjacent base pairs. The base-pair parameters include translational parameters: shear, stretch, stagger, x-displacement and y-displacement and rotational parameters: buckle, propeller twist, opening, inclination and tip. Similarly, the base step parameters include shift, slide and rise as translational parameters and tilt, roll and twist as rotational parameters.

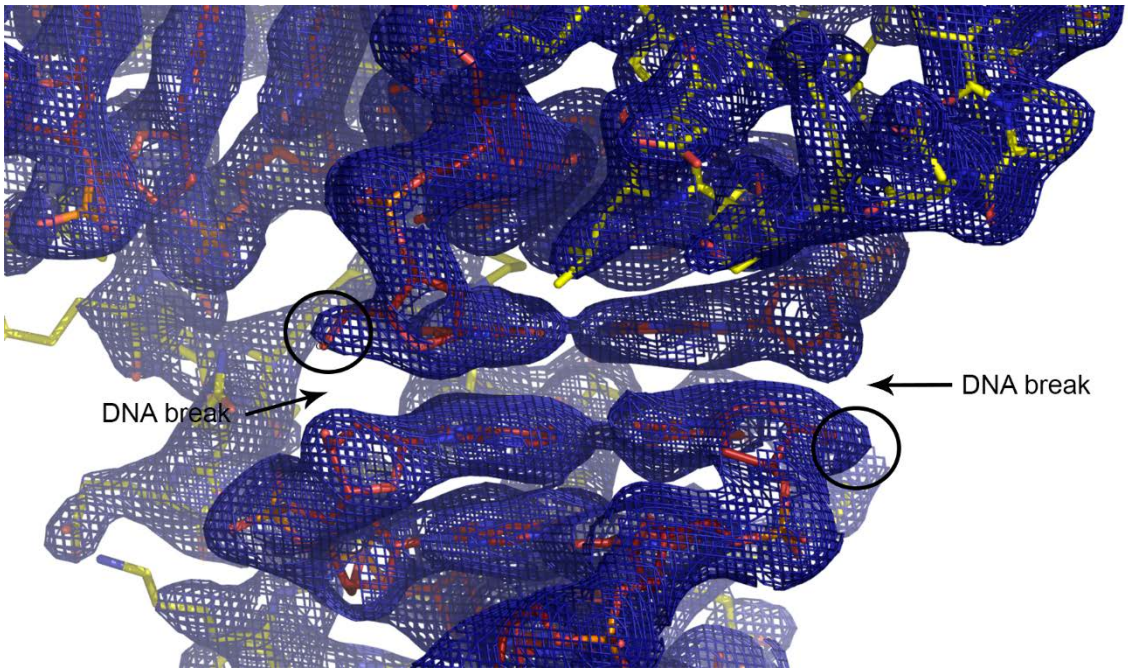


Figure R17. Abf2p/shift22 crystal structure and confirmation of dual-binding. The 2mFo-Fc double difference Fourier map (1.0 r.m.sd.) is shown as blue density. The protein and DNA model are shown as yellow and red sticks respectively. The clear DNA break and the absence of density for 5' phosphate at the DNA 5' ends are indicated by black arrows and circles respectively.

mtDNA [66] (and thus only survive in fermentable conditions). These amino acid residues correspond to a conserved Pro-B-B-Pro (B, basic aa) motif found in the N-terminal extended segment of either HMG-box domains, 43-Pro-Lys-Arg-Pro-46 in HMG-box1 and 116-Pro-Lys-Lys-Pro-119 in HMG-box2. The structure shows the NZ at the tip of these lysine side-chains being involved in interactions with the DNA phosphate backbone, while the Arg45 guanidinium is positioned close to the O2 base atoms of T6 (chainF) and T18 (chainE) at the minor groove, causing a negative roll (Figure R13b, FigureR14). Thus, absence of these contacts likely weakens DNA binding by Abf2p, eventually leading to y-mtDNA loss. DNA binding induces contacts between protein regions that are distant in the sequence. These include the N-helix, the HMG-box1 C-terminus, and the Pro-Leu Linker, which together form a hydrophobic core ('N-hydrophobic core', Figure R13a). Further, the N-helix forms electrostatic contacts with the protein C-terminus (NE of Arg31 with Tyr176 OH, Figure R13a). In this arrangement, residues from the N-flag are close to the minor groove region immediately downstream of HMG-box2 DNA binding site. Therefore, the N- and C-terminal regions are positioned in close proximity in the protein/DNA complex. We reasoned that if the dual binding mode was caused by a specific preference of the HMG-boxes for the contacted DNA sequence patches, then, joining both patches into a single 22bp DNA molecule should result in a complex of one protein bound to a single, continuous-DNA. Accordingly, we designed the sequence 'Af2_shift22' (5'- TTATATAATATAAAATAATAAA-3') (Figure R4, R12b). Surprisingly, the resulting structure (hereafter referred to as Abf2p/shift22; PDB ID 5JGH; Figure R17, R18 Table R3) again showed dual binding in an arrangement highly similar to the previous Abf2p/22 complex (Figure R12b,d). In this case, the dual binding was confirmed by generating double difference Fourier maps that clearly show the DNA break and lack of density for the 5' phosphate of the terminal base (the purchased oligonucleotides lack 5' phosphate) (Figure R17). To explain this, a thorough analysis of both DNA sequences was performed, and the presence of an ARS-m-like sequence containing poly-adenine tracts (A-tracts) was noticed in both DNAs (chainB (A_7)$A_8A_9T_{10}T_{11}$ in Af2_22, and chainC $A_{12}A_{13}A_{14}A_{15}$ in Af2_shift22, Figure R12d). Noticeably, the A-tracts are not contacted by the protein in either crystal. Moreover, crystallization trials with additional DNA fragments devoid of poly-A tracts (see methods), yielded poorly diffracting crystals, suggesting high disorder. Thus, the A-tracts are presumably delimiting the potential protein binding sites on the DNA and thus promoting formation of crystals with sufficient

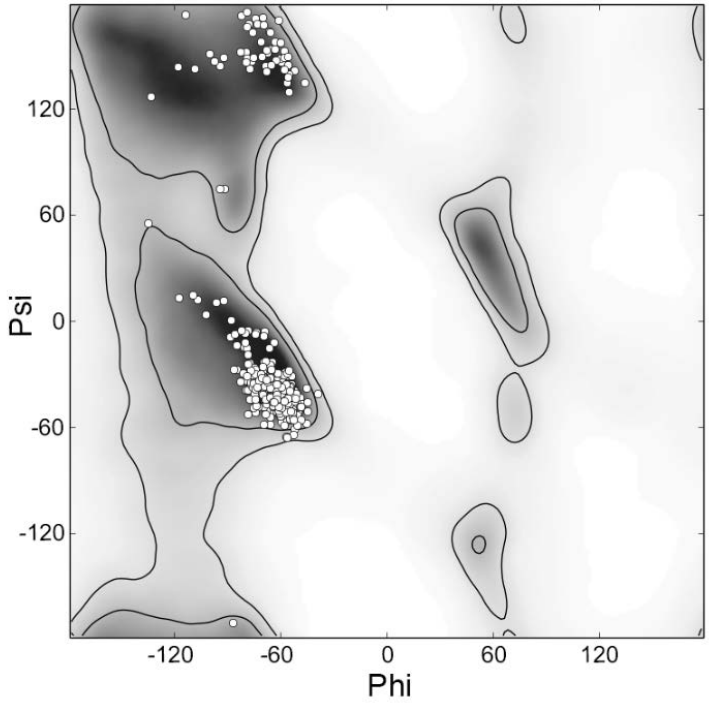


Figure R18. **Ramachandran plot for Abf2p/shift22 crystal structure**.

Table R3

Wavelength	0.97625
Resolution range	48.05 - 2.6 (2.693 - 2.6)
Space group	P 1 21 1
Unit cell	71.989 131.991 71.994 90 103.211 90
Total reflections	88589 (9083)
Unique reflections	38396 (3889)
Multiplicity	2.3 (2.3)
Completeness (%)	0.95 (0.97)
Mean I/sigma(I)	6.09 (1.05)
Wilson B-factor	55.15
R-merge	0.1022 (0.8523)
R-meas	0.1298 (1.085)
CC1/2	0.995 (0.207)
CC*	0.999 (0.585)
Reflections used in refinement	38396 (3885)
Reflections used for R-free	1828 (166)
R-work	0.2267 (0.3656)
R-free	0.2601 (0.3623)
CC(work)	0.948 (0.446)
CC(free)	0.907 (0.339)
Number of non-hydrogen atoms	8906
Macromolecule atoms	8772
Protein residues	618
RMS(bonds)	0.008
RMS(angles)	1.00
Ramachandran favored (%)	98
Ramachandran allowed (%)	1.6
Ramachandran outliers (%)	0
Rotamer outliers (%)	1.3
Clashscore	12.79
Average B-factor	61.38
B-factor macromolecules	61.53
B-factor solvent	50.10

* Statistics for the highest-resolution shell are shown in parentheses.

diffraction quality. The above findings are strongly reminiscent of the phased binding pattern associated with Abf2p (see Introduction). Such patterns have been speculated to be due to the presence of A-tracts [71] and could have relevant functional implications. Thus, we decided to characterize the properties of the protein and the DNA and their possible role in modulating DNA binding by Abf2p.

6. Role of different domains of Abf2p in DNA binding

To discern the role of the different Abf2p regions (N-flag, N-helix, HMG-box1 and HMG-box2) in binding DNA we produced deletion mutants, and analyzed their assembly on linearized M13mp18 dsDNA plasmids (5435bp) by electrophoretic mobility shift assays (EMSA). We constructed mutants Mut1 (no N-flag), Mut2 (HMG-box1+HMG-box2), Mut3 (N-helix+HMG-box1), Mut4 (HMG-box1), Mut5 (HMG-box2) and Mut6 (N-flag+N-helix+HMG-box1) (Figure R19a). Circular-dichroism experiments confirmed that all mutants were properly folded (Figure R20). The EMSA analyses (Figure R19b) showed that, under our experimental conditions, increasing concentrations of the full-length protein progressively up-shifted the band of DNA (kept at constant concentration). Removal of the N-flag in Mut1 did not modify DNA binding, although DNA migration showed some smearing effect, suggesting some destabilization of the nucleoprotein complex. In contrast, removal of the entire N-terminal segment (N-flag and N-helix in Mut2) strongly decreased DNA binding efficiency, showing that the Abf2p HMG-box tandem alone is not sufficient per-se for an optimized DNA binding. Accordingly, neither, HMG-box1 (Mut4) nor HMG-box2 (Mut5) showed any appreciable binding in this assay. However, HMG-box1, in presence of N-flag+N-helix (Mut6), surprisingly showed a DNA binding efficiency similar to the full-length protein. Dissecting the latter, Mut3 (N-helix+HMG-box1) behaved identically to Mut6 suggesting that the N-flag does not have a predominant functional role in DNA binding. Thus we conclude that the N-helix is a key factor that stabilizes the protein/DNA complex in Abf2p. EMSA assays with the 22 bp Af2_22 DNA showed similar trends as seen with M13 DNA (Figure 21a). The DNA binding efficiency was judged by the protein:DNA stoichiometry at which the first band shift was observed. While Mut1 retained DNA binding activity similar to full length, Mut2 didn't show any appreciable binding. Mut6 DNA binding, though more efficient than Mut2, was poorer than the full length Abf2p. The latter could be due to the short 22 bp DNA used for which the binding of the N-helix+HMG-box1 module was presumably not as efficient as when presented with long DNA having multiple binding sites.

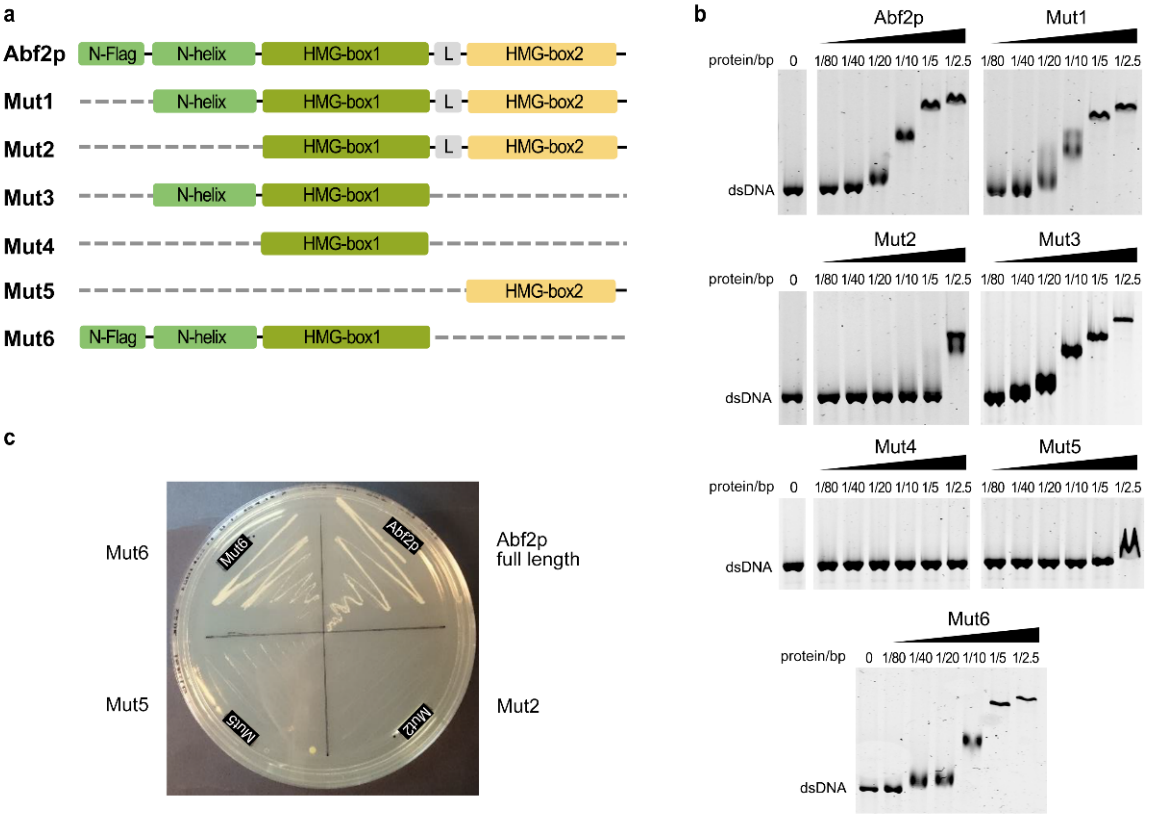


Figure R19. Abf2p truncations and their effect on DNA binding. (a) Schematic representation of the Abf2p deletion mutants. The domains are designated as in the text except for the linker, indicated with an 'L'. Dashed lines indicate deleted regions. **(b)** Titration of Abf2p and Mut1 to Mut6 constructs binding on linearized M13mp18 dsDNA (400pM) at 25°C, by EMSA. The protein/bp ratios are indicated and correspond respectively to 0.37, 0.75, 1.5, 3, 6 and 12 pmoles of Abf2p. **(c)** Effect of Abf2p truncations *in vivo*. A plate containing glycerol as carbon-source ('YPG plate') shows the effect of different Abf2p truncations on the ability to maintain y-mtDNA and thus on capability of yeast (*Saccharomyces cerevisiae*) cells to use non-fermentable carbon source. Full-length Abf2p and Mut6 (Abf2p without HMG-box2) are capable of protecting mtDNA whereas Mut2 (Abf2p without N-flag and N-helix) and Mut5 (HMG-box2 alone) are not functional.

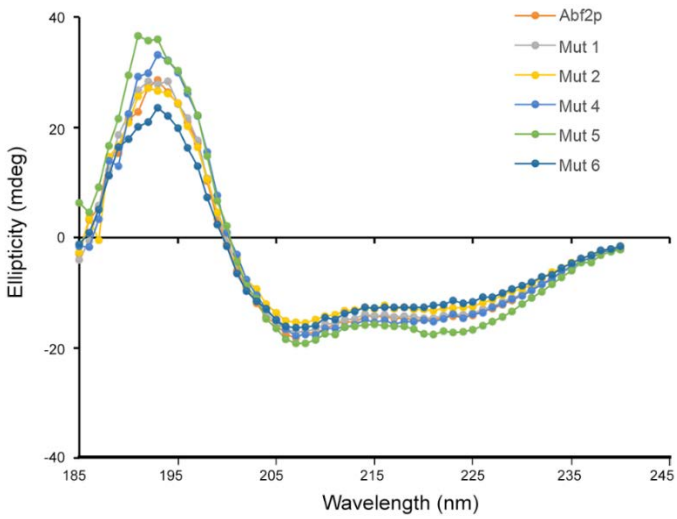
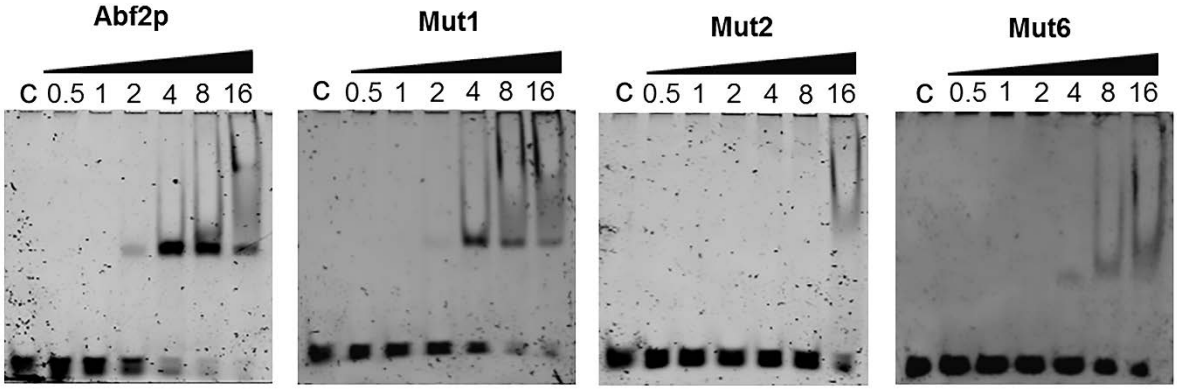


Figure R20. **Stability of Abf2p deletion mutants**. CD spectra for full-length Abf2p and the deletion mutants Mut1 to Mut6 used in the EMSA assays. The spectra show that the mutants are properly folded, having a profile similar to the wild-type protein.

a



b

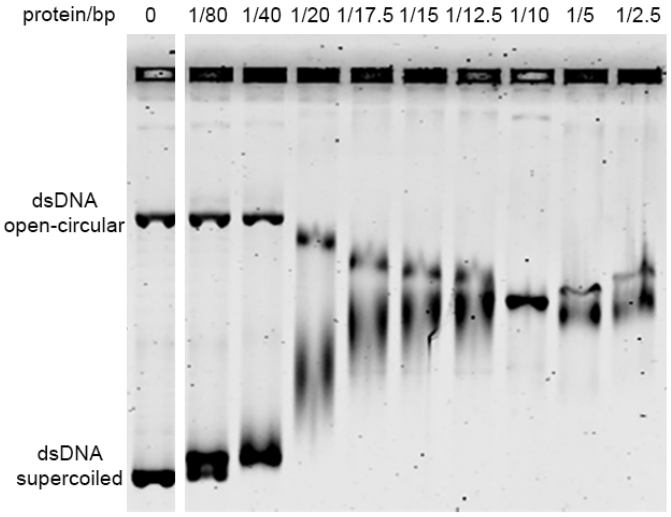


Figure R21. Binding of truncation mutants to 22bp Af2_22 DNA and of full length Abf2p to circular M13 DNA (a). EMSA assays with Af2_22 DNA and different Abf2p deletion mutants in 10cm 10% native acrylamide gels at 25º C. The DNA concentration was kept fixed at 200 nM while the protein concentration was increased from left to right. C stands for control free DNA. The protein:DNA molar ratio is indicated at the top of each well. **(b)**. EMSA assay of full length Abf2p with circular M13 DNA in 1.2 % agarose gels at 25º C. The protein/bp ratios are indicated and correspond respectively to 0.37, 0.75, 1.5, 1.7, 2, 2.4, 3, 6 and 12 pmoles of Abf2p.

EMSA assays were additionally performed with circular M13mp18 DNA and full length Abf2p (Figure R21b). The circular M13 DNA comprised of a supercoiled and an open circular (nicked) subspecies. Increasing ratios of Abf2p/bp of DNA caused the open-circular form to migrate faster, presumably due to compaction. On the other hand, the super-coiled form ran slower on Abf2p addition. The latter could be due to increase in size of the complex or due to reduction of super-coiling. Interestingly, at a Abf2p/bp ratio of 1:10, the bands corresponding to Abf2p complex with the open-circular and supercoiled DNA ran together on the gel, indicating a similar compaction state.

7. Effect of Abf2p truncations *in vivo*

In order to observe effects of the above deletions *in vivo*, we subsequently generated haploid yeast Abf2Δ cells, bearing plasmids encoding the full-length or mutant Abf2p constructs described above. Diploid strain Y26205 lacking one out of two copies of the Abf2p gene and with the genetic composition BY4743; MATa/MATα; ura3Δ0/ura3Δ0; leu2Δ0/leu2Δ0; his3Δ1/his3Δ1; met15Δ0/MET15; LYS2/lys2Δ0; YMR072w/YMR072w::kanMX4 were used. This strategy was necessary as haploid Abf2Δ cells lose mtDNA very quickly and thus are unsuitable for testing effect of Abf2p truncations on mtDNA maintenance. Cells transformed with the pRS316[147] plasmid bearing the full length Abf2p gene were selected on a SDC-Ura (Synthetic Dextrose(D-glucose) medium lacking uracil) plate. The selection was based on the principle that since the cells are ura3Δ0/ura3Δ0, only cells that have incorporated the pRS316 plasmid (possessing the ura3 gene) will survive on SDC-Ura plates. URA3 is a gene on chromosome V in *S. cerevisiae* (yeast) that encodes Orotidine 5'-phosphate decarboxylase (ODCase), an enzyme involved in the synthesis of pyrimidine ribonucleotides. The transformants were induced to sporulate by growing them on sporulation plates. Sufficient haploid formation was confirmed by visualization under the microscope. Subsequently the transformed and sporulated cells were treated with *zymolyase* to digest the spore wall and grown back on SDC-Ura plates. Colonies from this plate were sequentially grown for a second time in SDC-Ura (Figure R22a) and then in YPD-G418 (yeast extract peptone and dextrose with G418 or Geneticin) to ensure that the cells were Abf2Δ and retained the Abf2p-pRS316 plasmid (Figure R22b; G418 is an aminoglycoside antibiotic used to select cells using the KanMX selectable marker). Colonies that grew on both SDC-Ura and YPD-G418 were grown in YPG (yeast extract, peptone, glycerol) to ensure respiratory capability and thus functional mtDNA (Figure

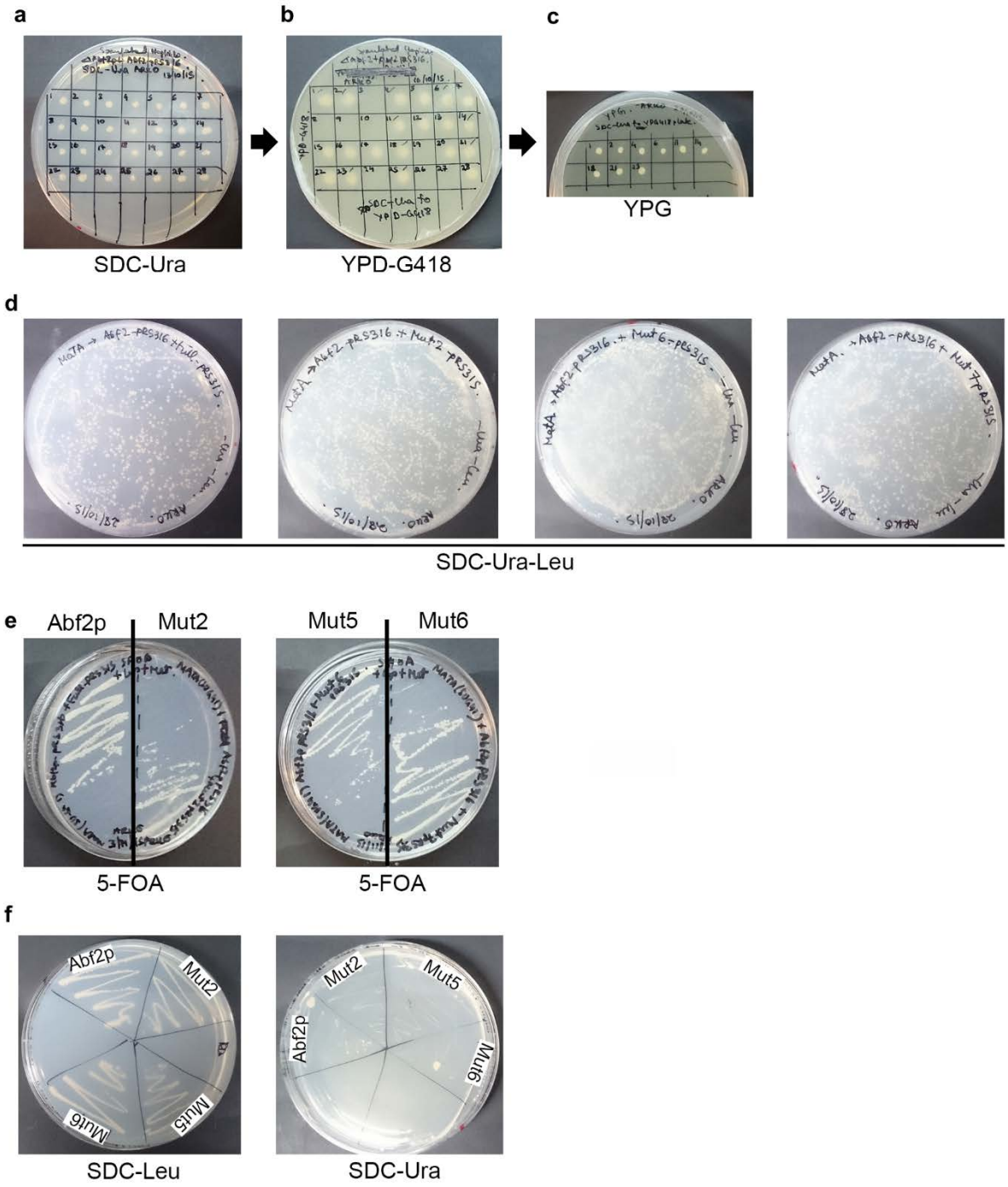


Figure R22. Steps in *in vivo* analyses of functions of Abf2p domains. (a). Haploid Abf2PΔ+ Abf2p-pRS316 colonies. **(b)**. Colonies from **a.** regrown in YPD-G418**. (c).** Cross-check of colonies from **b.** in YPG to check status of mtDNA. **(d)**. Haploid Abf2PΔ+ Abf2p-pRS316+ X-pRS315 (X representing any of the Abf2p truncation constructs) cells grown on SDC-Ura-Leu to check for incorporation of both plasmids. **(e).** Deletion of Abf2p-pRS316 by growth on 5-FOA for Abf2p, Mut2, Mut5 and Mut6 constructs. **(f)**. Check of colonies for the different mutants from FOA plates for X-pRS315 maintenance (SDC-Leu) and Abf2p-pRS316 deletion (SDC-Ura).

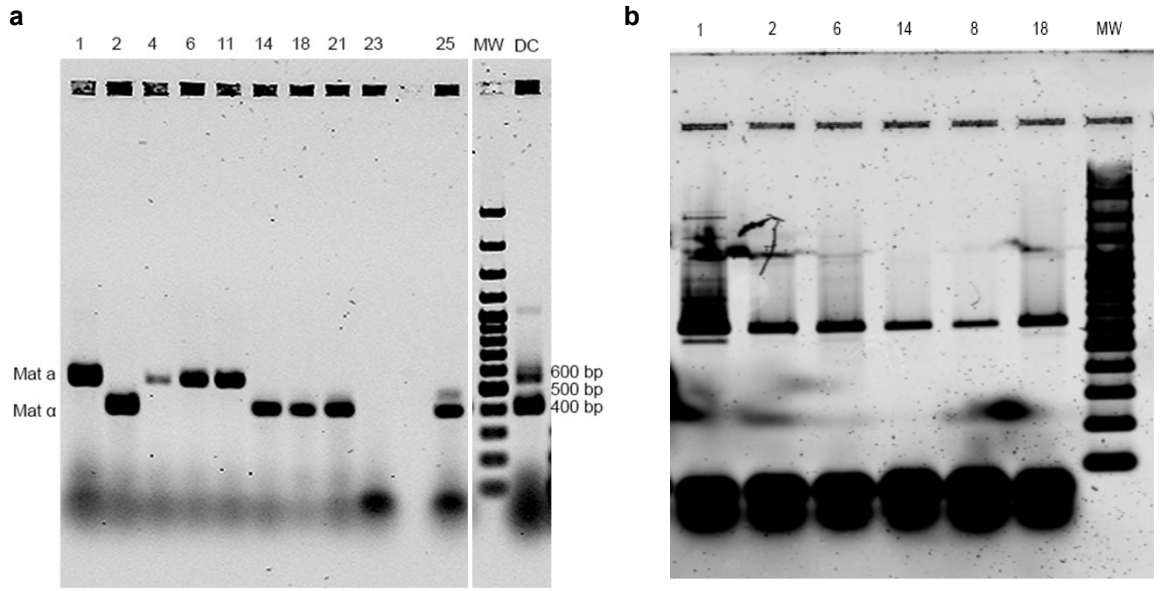


Figure R23. PCR based crosschecks for *in vivo* analyses. (a). Crosscheck of haploids by PCR and agarose gel electrophoresis. Numbers at the top of each well indicate identification number of colonies as in Figure 22a. DC stands for control PCR with the diploid strain. **(b)**. Check for selected haploid colonies for deletion of the Abf2 gene by PCR and agarose gel electrophoresis.

R22c). Cells from these selected colonies were cross-checked for haploid condition by a PCR based method [148] (see Methods) where mating type a (Mat-a) and mating type α (Matα) will result in 544 bp and 404 bp PCR products respectively while a diploid strain will produce both (Figure R23a). Additionally, selected haploids were double-checked for the kanamycin (G418) resistance gene (KanMX4) and thus Abf2p gene deletion by PCR using suitable primers (see Methods; Figure R23b). Subsequently, PCR generated constructs of different truncations of Abf2p with N-terminal MTS and 500bp upstream and downstream regions (which contained the promoter and terminator regions respectively-see Methods) were cloned into pRS315 [147], a plasmid carrying the *Leu* gene involved in leucine biosynthesis. The resulting transformants were selected for retention of both Abf2p-pRS316 and X-pRS315 (where X stands for one of the truncation constructs) by growing them on SDC-Ura-Leu plates (Figure R22d). Next, for each truncation type, the Abf2p-pRS316 plasmid was disposed of by growing the cells on 5-fluroorotic acid (5-FOA) plates [149,150]. Single surviving colonies from the 5-FOA plates (Figure R22e) were grown back on SDC-Leu plates (Figure R22f) to select for cells which had lost the Abf2p-pRS316 plasmid but retained X-pRS315. The same colonies were simultaneously checked for complete removal of the Abf2p-pRS316 plasmid by streaking them onto SDC-Ura plates where no growth was observed (Figure R22f). Single colonies from these plates were streaked onto

YPG (yeast extract-peptone-glycerol) plates to analyze which of the truncated Abf2p mutants can protect y-mtDNA and thus allow the cells to grow in YPG (proteins coded by the y-mtDNA are involved in the respiratory chain). Cells expressing full-length, wild-type Abf2p grew normally on non-fermentable carbon source (glycerol), indicating a functional respiratory chain (Figure R19c). Strikingly, cells with Abf2p lacking the N-flag+N-helix (Mut2) were able to ferment glucose but fail to grow on glycerol. This indicates the relevance of this fragment for y-mtDNA stability and maintenance and is in agreement with our in vitro data. In addition, and as previously shown [72], cells carrying the N-flag+N-helix+HMG-box1 construct (Mut6) grew normally on glycerol indicating that this Abf2p mutant is sufficient for maintaining y-mtDNA.

8. Abf2p dynamics provides clues to DNA binding mechanism

The EMSA and *in vivo* assays with different deletion mutants pointed to the relevance of the N-helix in mtDNA binding and metabolism. On the other hand, the crystal structure shows a detailed conformation of the final DNA bound state but does not provide mechanistic information about the involved binding process. To obtain this information, we performed molecular dynamics (MD) simulations on the isolated protein and the complex. A 500ns MD simulation of the free protein shows that, in absence of bound DNA, Abf2p adopts a rather extended conformation where the two HMG-boxes separate from each other and the contacts between the N- and C-termini are lost (Figure R24b). This movement is allowed by the flexibility of the short two-residues Linker (114-Leu-Pro-115) and can be aided by occasional unwinding of the C-terminal end of HMG-box1 helix3 (specifically Lys-113), which undergoes helix-turn-coil transitions (Figure R25a). In addition, the N-helix shows a great variability in its orientation (backbone r.m.s.d 6.38+/- 1.25 Å; Figure R25b) and turns along its main axis resulting in Leu34, Leu38 and Ile39 side chains being exposed to the solvent (Figure R24b). This results in disruption of the N-hydrophobic core. The calculated root mean square fluctuation (r.m.s.f) along the simulation demonstrates that the N-helix is highly flexible (Figure R24c). In the crystal structure the helix is aligned to one side of HMG-box1 helix 3, completing the hydrophobic core, which also provides opportunity for residues from the N-flag and N-helix to contact the DNA minor groove immediately adjacent to the DNA patch contacted by HMG-box2. This could be a phenomenon triggered during the DNA binding event. Since HMG-box1 possesses a much higher DNA binding efficiency than HMG-box2, DNA binding by Abf2p most probably involves a sequential process where binding of HMG-box1 completes the

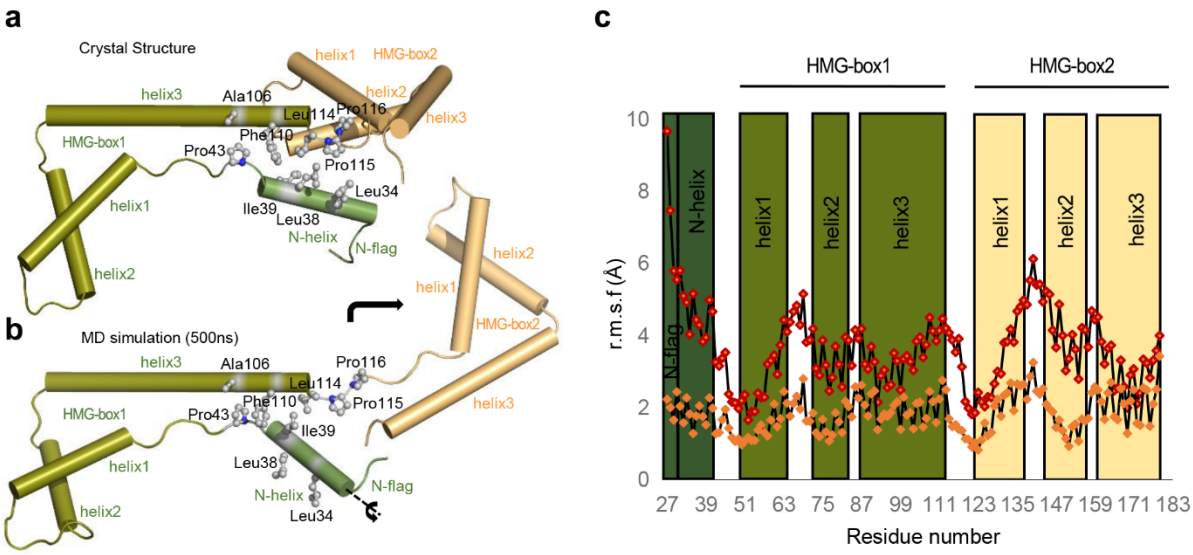


Figure R24. **Abf2p flexibility and dynamics**. **(a)** The Abf2p domains in the Abf2p/22 crystal structure are shown (same color-code as in Figure 1, the DNA molecule is not displayed for clarity) together with the amino-acid side-chains (depicted in gray ball-and-sticks and labeled) that participate in the N-hydrophobic core that stabilizes the contact between the N-helix, the C-terminal region of HMG-box1 helix3 and the linker. **(b)** A frame from the 500ns MD simulation of Abf2p that shows the hydrophobic residues from the N-helix exposed to the solvent, the changes in the orientation of N-helix and HMG-box2 (HMG2) (relative to HMG-box1 in the crystal structure above) indicated by arrows. **(c)** Root mean square fluctuation (r.m.s.f, in Å) of individual residues during MD simulation (500ns) of Abf2p alone (red dots) and bound to DNA (orange dots). The different regions of the protein are schematically represented along the sequence. Note that the Abf2p flexibility reduces in the protein/DNA complex.

hydrophobic core, positioning the N-flag and first turn of the N-helix in contact with minor groove of the downstream DNA. Thus, a conformation is locked which ideally positions HMG-box2 to contact DNA at the minor groove, as seen in the crystal structure.

In contrast, a 500ns all-atom MD simulation of Abf2p/DNA complex (1protein:2DNA) shows that the protein remains tightly bound to the DNA (Figure R24c and Figure R26a), with overall lesser fluctuation compared to the free protein. In this bound state, the N-helix, the N-hydrophobic core, and the loops between helices 1 and 2 from both HMG-box domains all show reduced mobility (Figure R24c). The stacking between the DNA ends observed in the crystal structure is preserved, while the other two free ends fluctuate considerably (Figure R26a). The insertions of Phe51 and Ile124 are maintained, further demonstrating the stability of the protein/DNA complex (Figure R26b). In conclusion, the MD simulations show that DNA binding brings together the N- and C-termini of Abf2p, and the N-helix functions as a pin-lock that consolidates the hydrophobic core.

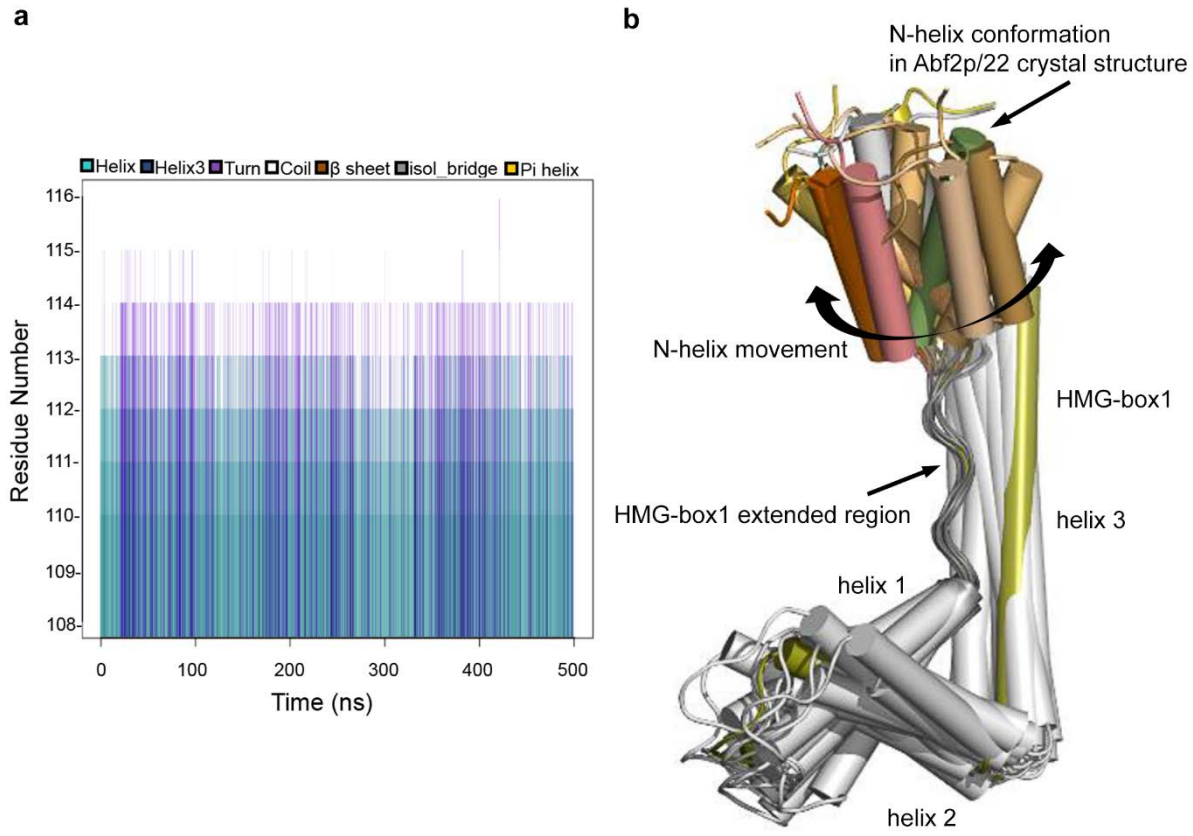


Figure R25. Abf2p dynamics: Conformational variability at the Linker and movement of the N-helix. (**a**) Secondary structure variability of the free protein during the MD simulation at and around the Linker (residues 108-116). This region includes the end of HMG-box1 helix3, the Linker and the beginning of HMG-box2 extended region (see text). Note that the conformation at Lys113 occasionally transits from helix (cyan) to turn (violet) to coil (white) while the amino acids before it, maintain helical conformation. (**b**) 15 frames from the MD simulation of free Abf2p (extracted every 25 ns of the simulation and superposed by the HMG-box1 extended region) demonstrate the extensive motion of the N-helix. HMG-box1 is shown in gray. The protein model from Abf2p/22 crystal structure is shown in olive green and its N-helix in deep green. HMG-box2 is omitted for clarity.

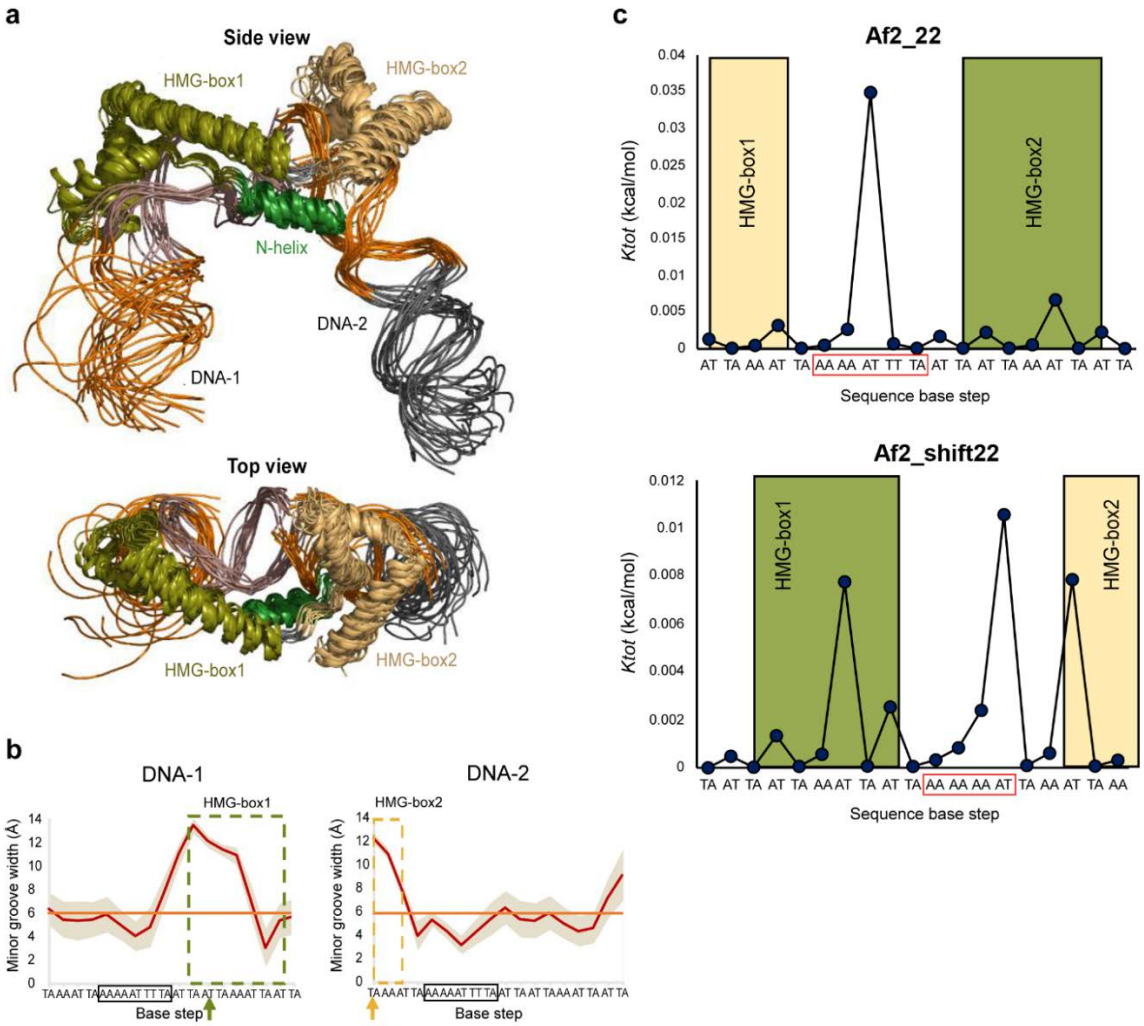


Figure R26. Visualization of conformational features of Abf2p-DNA complex by MD simulations and Stiffness of poly-adenine tracts (**a**) Superposed frames (15, extracted every 25 ns of the simulation) show that the protein conformational changes are constrained on binding to the DNA. Especially, the N-helix (in deep green) is locked in a conformation similar to that found in the crystal structures. The free DNA ends however, show considerable flexibility. The top view shows that the stacking of the DNA ends at the junction between DNA1 (in grey) and DNA2 (in orange), in between the binding sites of the HMG-boxes, is preserved. The ARS-m-like sequences are depicted in orange. (**b**) DNA minor groove variations along the two protein-bound DNA fragments in (**a**). The red line denotes average values calculated from the MD simulation with the spread of the values shown in gray. The orange line shows the value corresponding to an ideal B-DNA. Rectangular boxes show values corresponding to the DNA regions contacted by HMG-box1 (olive green) and HMG-box2 (orange). The insertion sites for the two boxes are indicated by arrows. The A-tract in the DNA sequence is boxed. (**c**) Total stiffness constant *Ktot* for Af2_22 and Af2_shift22 DNA. The regions contacted by HMG-box1 (olive green) and HMG-box2 (orange) are shown. The A-tracts are boxed in red.

9. Abf2p compacts upon DNA binding in solution

The inter-domain flexibility suggested from MD simulations was further experimentally demonstrated by small angle X-ray scattering (SAXS), a technique that allows extraction of information about the structure (~20Å) and dynamics of macromolecules in solution [129,130]. Data from a concentration series (3.8 mg/ml, 5.4 mg/ml, 7.0 mg/ml, 9.5 mg/ml and 13.5 mg/ml) was collected at 50mM Tris-HCl, pH 7.5, 500mM NaCl. Molecular weight of the protein was estimated from the Porod volume. The estimated molecular masses are 14.3-19.0 kDa, 16.3-21.8 kDa, 15.1-20.1 kDa, 16.1-21.5 kDa and 16.3-21.8 kDa respectively (the lower and upper estimates for each concentration are obtained by dividing the Porod volume by 2 and 1.5 respectively). These values match closely with the 6-histidine-tagged monomer of Abf2p (19.5kDa). The radii of gyration (*Rg*), calculated from the Guinier plot are 27.2 +/- 0.1, 27.3 +/- 0.1, 27.4 +/- 0.1, 27.3 +/- 0.1 and 27.8 (+/-0.1) Å, respectively. Additional data was collected at 50mM Tris-HCl, pH 7.5, 100mM NaCl (0.36 mg/ml; no reliable data could be obtained at other concentrations) for which the calculated molecular mass and *Rg* are 12.6-16.8 kDa and 28.4 (+/-0.6) Å respectively. The curve corresponding to the highest concentration (13.5 mg/ml) was used for all other subsequent analyses as it showed lesser noise at higher angles. The calculated radius of gyration (Rg=27.8 Å) is larger than that expected for a ~20 kDa protein (17Å according to Flory equation [133] Rg = 3 x Ne0.33) and agrees well with that calculated from MD simulation of the free protein (Rg=29.9Å). Furthermore, the pair- wise distribution function (P(r)), which reflects the distribution of intra-particle distances [130], yielded a large maximal particle dimension Dmax of 130Å (Figure R27a). These results are consistent with an extended state of the protein. Additional information regarding the conformational variability of macromolecules can be obtained from the Kratky representation [130]. In this case, the Kratky plot showed a flat profile that is typical for a flexible particle (Figure R27b), as indeed shown by MD simulations. To further describe Abf2p flexibility and the expanse of its conformational variability, we applied the ensemble optimization method [132,133] (see Methods) which, from a large pool of static structures (thousands, generated from an initial starting model), selects a sub-ensemble of conformations that best describes the SAXS data. Initially EOM was performed using protein coordinates from the Abf2p/22 structure as the starting point. Fitting the calculated average scattering profile from the sub-ensemble to the experimental SAXS data, yielded a fitting with χ^2=1.379. As an alternative, we used a frame from the MD simulation (extracted from the simulation run after the r.m.s.d. had converged) as a starting

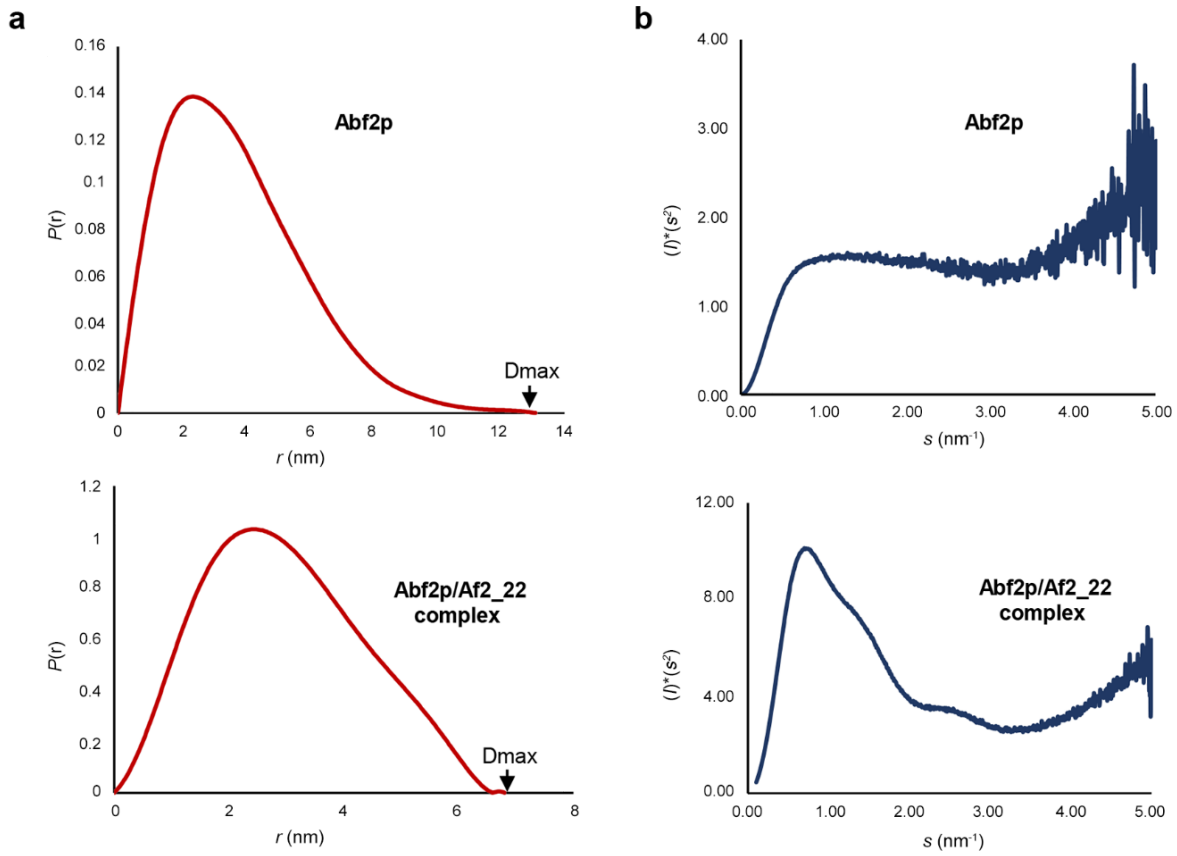


Figure R27. Conformational flexibility of free Abf2p and Abf2p/DNA complex in solution (a) Distance distribution (*P(r)*) plots for free Abf2p (top) and Abf2p/Af2_22 complex (bottom) from SAXS data. The maximum intra-molecular distances (Dmax) show that the protein alone is in a much more extended conformation compared to the protein/DNA complex. **(b)** Kratky plots for free Abf2p (top) and Abf2p/Af2_22 complex (bottom). For Abf2p, the curve plateaus and then shows increase at high *s*, thus indicating a non-globular conformation of the protein. In contrast, for the Abf2p/Af2_22 protein-DNA complex the plot shows a bell shaped feature at low *s* values and flat profile at high *s* values, indicating a more compact species.

model for EOM (Figure R24b). Using the same assignment of flexible regions, EOM generated a subset of five conformations, which, collectively, were in better agreement (χ^2=0.957) with the experimental data (Figure R28a), and showed a remarkable variability in the relative orientations of the HMG-boxes (Figure R28b), further confirming results from MD. The Rg estimates for the sub-ensemble (Figure R28c) show a bimodal distribution, suggesting two alternative predominant conformations of the free protein in solution, one being more compact than the other. The Abf2p/Af2_22 complex, in contrast to the free protein, showed characteristics of a more compact species, as reflected in the respective radii of gyration and maximal particle dimensions (Rg=13.8Å vs 27.8Å; Dmax=65Å vs 130Å, Figure R27a). This is supported by the corresponding Kratky plot (Figure R27b), which shows a bell-shaped profile,

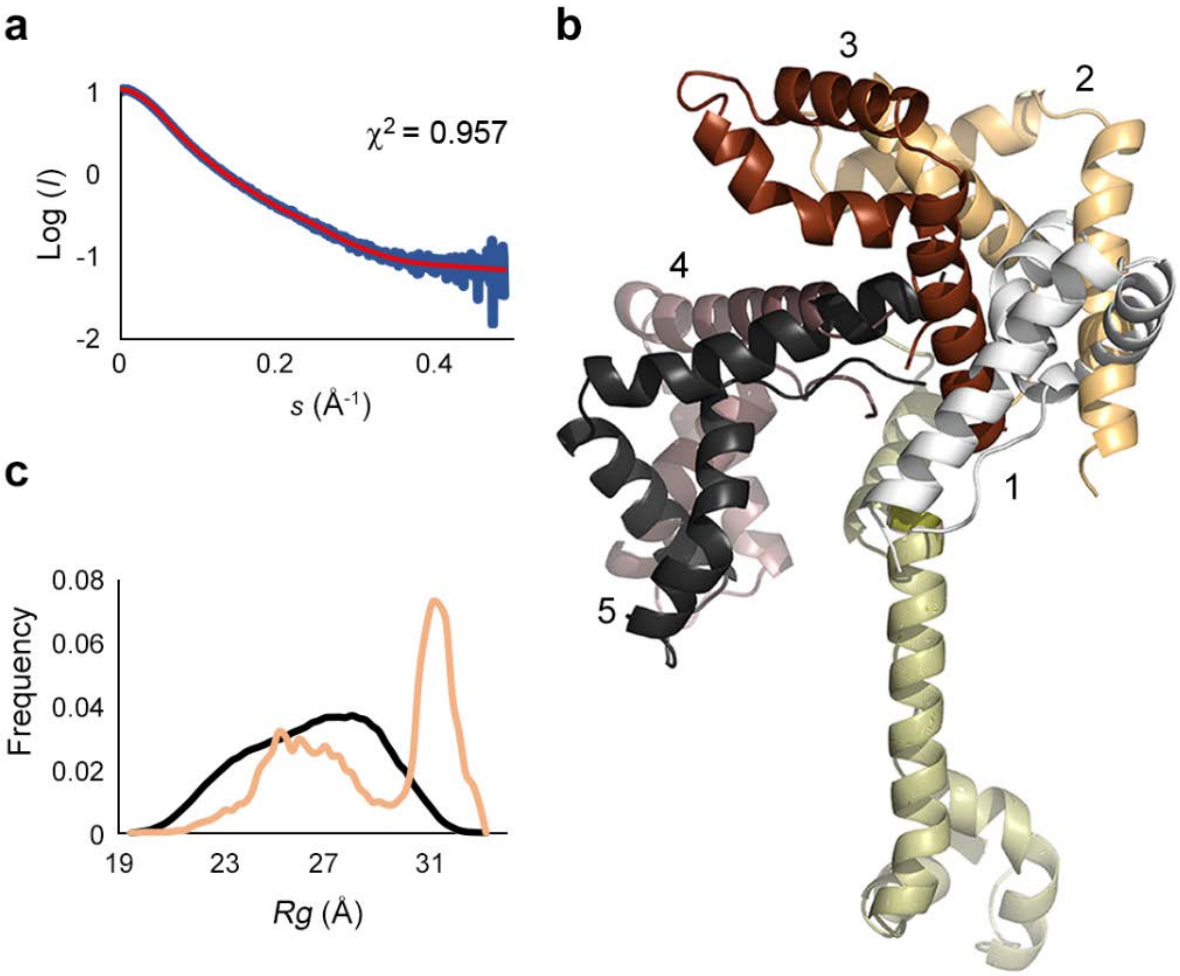


Figure R28. Conformational flexibility of Abf2p in solution. (a) Plot of the experimental SAXS intensity (*Log(I)*) along the momentum transfer *s* (blue curve) and fit (χ^2=0.957) of the calculated average intensity from the final selected sub-ensemble (red curve) **(b)**. Final sub-ensemble of 5 models selected by EOM, superposed by the HMG-box1 domain and showing the variability of the HMG-box2 position in the different models. **(c)** Radius of gyration (Rg) distribution of the EOM selected pool (light brown) compared to the initial random pool (black). The former shows a bimodal distribution.

characteristic of a compact particle [130]. Therefore, free Abf2p is highly flexible and extended in solution while DNA binding reduces its conformational freedom leading to a more compact particle.

10. Structural Features of the A-tracts prevent Abf2p binding

To investigate the structural basis of exclusion of Abf2p binding from near matches of the ARS consensus sequence, we analyzed properties of the DNA sequences by all-atom MD simulations, starting from standard B-DNA (Methods). From the simulation trajectories, averages of the intra- and inter-base pair parameters and groove parameters over the entire simulation were obtained which immediately revealed DNA features distinct from standard B-DNA and characteristic of adenine tracts (A-tracts) [143,144,154].

10.1 The case of Af2_22 DNA sequence

A-tracts generally come in two flavors: asymmetric A-tracts are a contiguous stretch of adenine bases (A_n, where n≥4) [154] and show a progressive narrowing of the DNA minor groove going from the 5' to the 3' end; symmetric A-tracts [154] are of the form A_nT_n, where n≥2 and exhibit narrowest minor groove at the middle of the tract. The Af2_22 DNA sequence has a A_2T_2 tract. The MD trajectory was analyzed to identify characteristic features of A tracts for Af2_22 DNA.

A-tracts influence the conformation of bases at their junctions, especially at the 3' end. They show buckling at the junctions with B-DNA. For the Af2_22 sequence, a reversal of buckling was observed at the centre of the AAATT region with the central A having a buckle close to zero and positive and negative values to the left and right respectively (Figure R29). This corroborates the data from Hizver *et al* [155] where the buckle goes from positive to negative in the 5' to 3' direction for the A_2T_2 tract. The A-tract influences the buckling of the bases adjacent to the 5' and 3' end of it before the buckle returns to near zero values. Propeller twisting and consequent bifurcated hydrogen bonds are another characteristic of A-tracts, predominantly found in asymmetric A-tracts of 4 adenines or longer. In case of the present A_2T_2 tract the bases become more negatively propeller twisted from the 5' to the 3' end, starting from -10°, except at the last T where it returns back to -10°. The DNA exhibits a similar extent of propeller twisting for the next 5 bases downstream (Figure R29). A-tracts have been additionally observed to possess negative roll and tilt and positive values occur at the flanking B-DNA regions. The Af2_22 A_2T_2 A-tract shows similar trends (Figure R13b, R29). Notably, the MD simulation shows that the symmetric

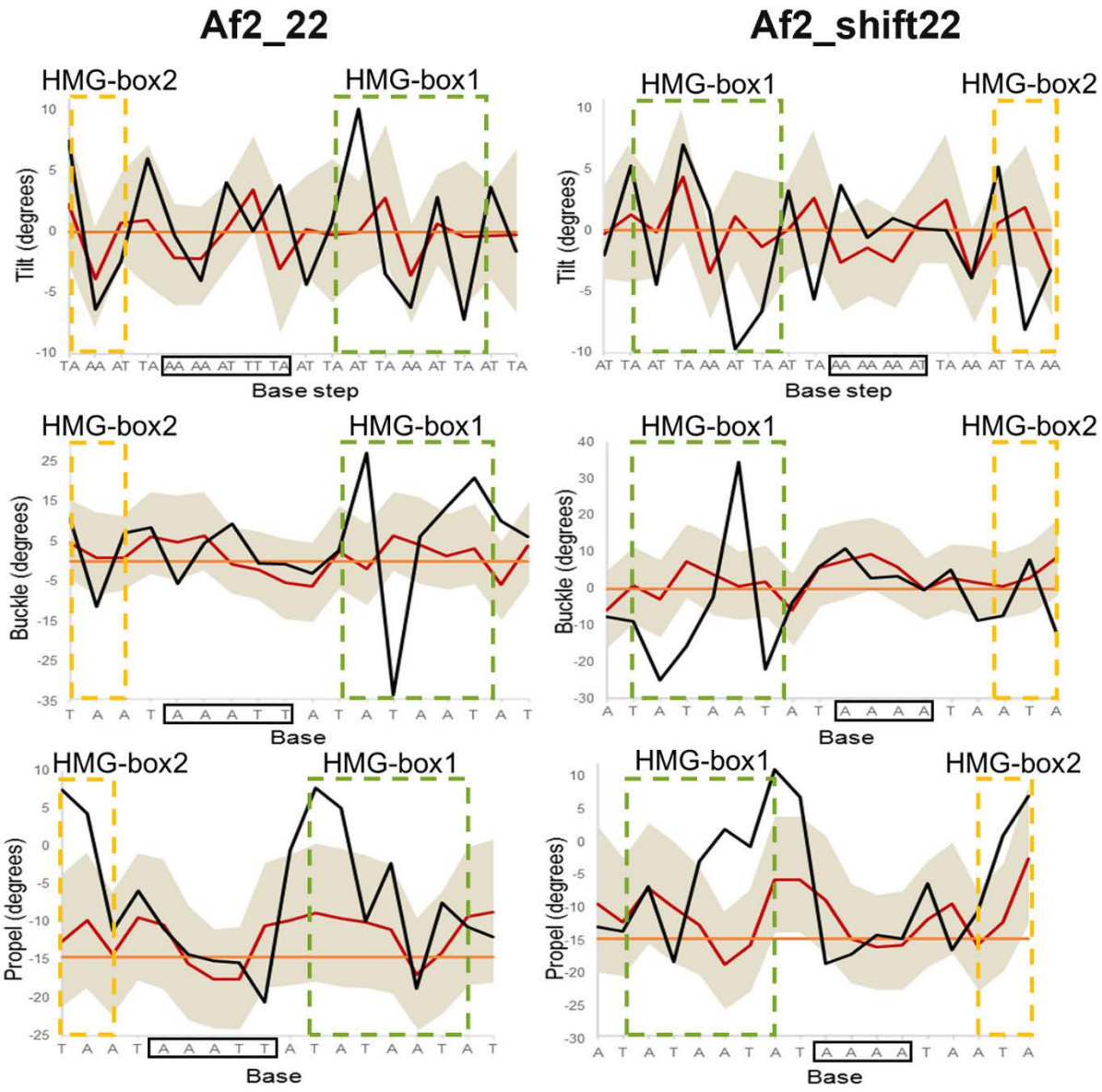


Figure R29. **Af2_22 and Af2_shift22 DNA parameters calculated from MD simulations**. Values that correspond to the crystal structures (in black), to the MD simulations for unbound Af2_22 and Af2_shift22 DNAs (averaging over the individual trajectories, in red; spread of the values, in grey), or to an ideal B-DNA (in orange) are shown. DNA contacted by Abf2p and the corresponding values are framed in olive green for HMG-box1 and light orange for HMG-box2. The A-tracts are framed in black.

A-tract at Af2_22 narrows from both ends to a minimum width of 3.75 Å (compared to the 5.91Å in an ideal B-form), precisely at the A_9T_{10} junction (Figure R13b).

10.2 The case of Af2shift22 DNA sequence

The Af2shift22 sequence possesses an asymmetric A_4 tract at positions 12-15. Significant positive buckle (compared to ideal B DNA buckle of 0^o) was observed within the A_4 tract as well as change in the direction of bucking at the 5' and 3' end of the tract (Figure R29). The thymine 5' to the A-tract retains a similar buckling direction as the adjacent A tract, presumably to preserve base stacking. Propeller twist values remain close to the average B-DNA value of -14.6^o except at the first adenine where it is less negative (-8.8^o). The flanking bases maintain a less negative propeller twisting relative to the A_4 tract. Indeed, high propeller twist has been observed in asymmetric A-tracts longer than 4 bases and it has been speculated that high propeller twist, permitting formation of bifurcated hydrogen bonds, might be responsible for stabilization of longer A-tracts [154]. Af2_shift22 additionally shows negative roll and tilt in the A-tract region (Figure R13b, R29). Most importantly, A_4 tract shows a prominent decrease in the minor groove width from the 5' to 3' end, reaching a minimum of 4.0 Å (Figure R13b).

Thus, the DNAs from both crystal structures possess prominent A-tracts with structural features distinct from standard B-DNA. Notably, a MD simulation of the Abf2p/22 complex (1protein:2DNA) showed that the A-tract maintains the structural characteristics of the unbound form (Figure R26b) and is not affected by the DNA distortions due to protein binding at adjacent non-A-tract regions. Furthermore, the crystallographic structures show remarkably narrow minor grooves precisely at the A-tracts (3.08 Å and 2.8 Å for Abf2p/22 and /shift22, respectively; Figure R13b). Thus, it can be deduced that Abf2p, a protein that binds to and opens up DNA minor grooves, finds A-tracts inhospitable. Values for the DNA stiffness descriptor [143] (*Ktot*, see Methods, Figure R26c, Figure R30), revealed that the above A-tract regions possess high intrinsic rigidity that make them difficult to adopt the required distorted conformation. All these observations suggest that the unusually narrow minor groove of A-tracts together with their inherent rigidity impede Abf2p binding, diverting the protein towards more permissive DNA sequences. This presumably causes the unexpected dual-binding observed in both the crystals. Specifically, in the crystal structures, HMG-box2 binds inside the ARS-m-like sequence, but avoids the A-tracts (Figure R12b,d), further indicating that the A-tracts are the key elements responsible for positioning Abf2p on DNA.

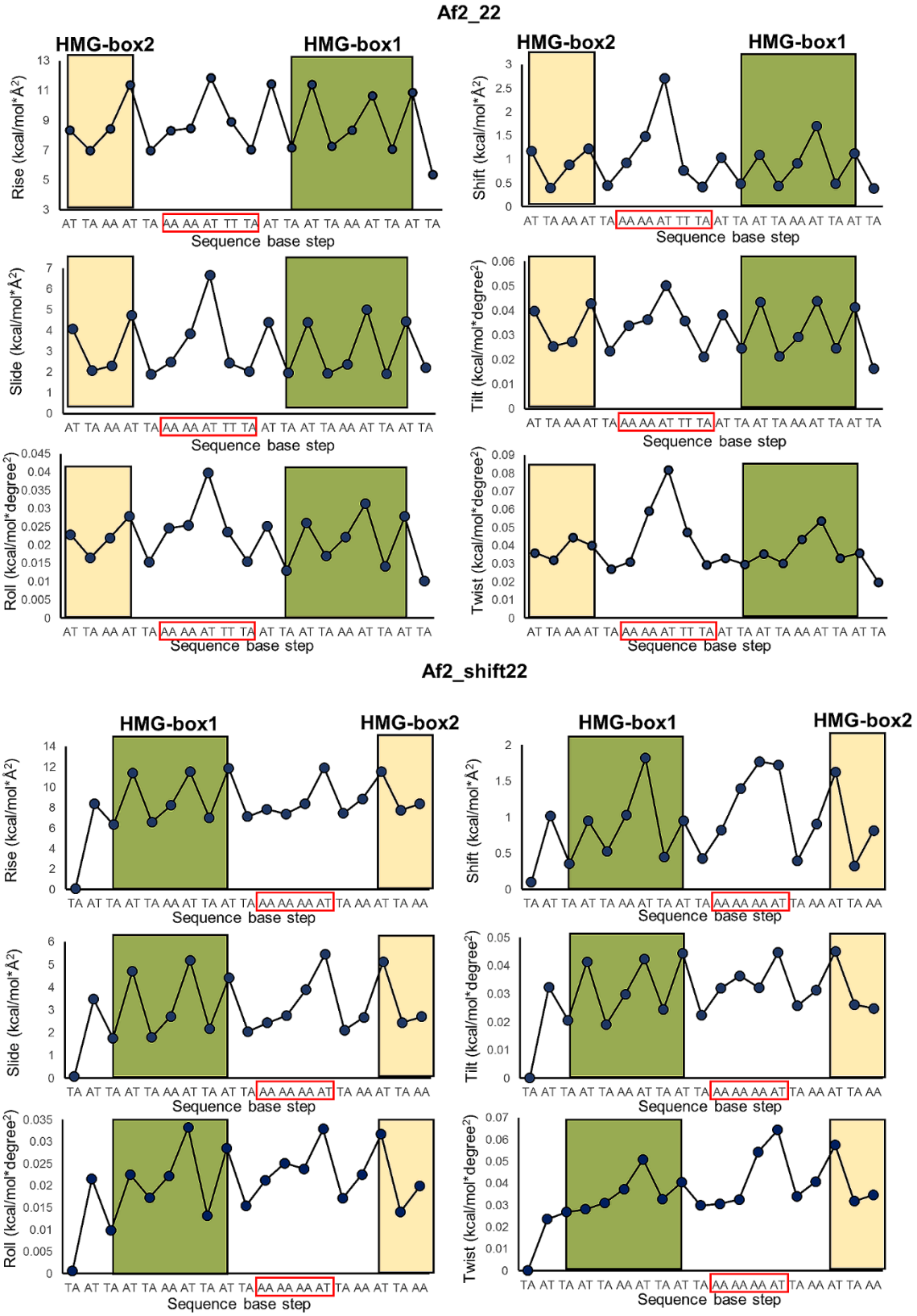


Figure R30. **Stiffness constants corresponding to the six DNA parameters rise, shift, slide, tilt, roll and twist for Af2_22 and Af2_shift22 DNA.** The A-tracts on the DNA sequences are boxed in red. The binding sites for HMG-box1 (olive green) and HMG-box2 (light orange) are shown with rectangles.

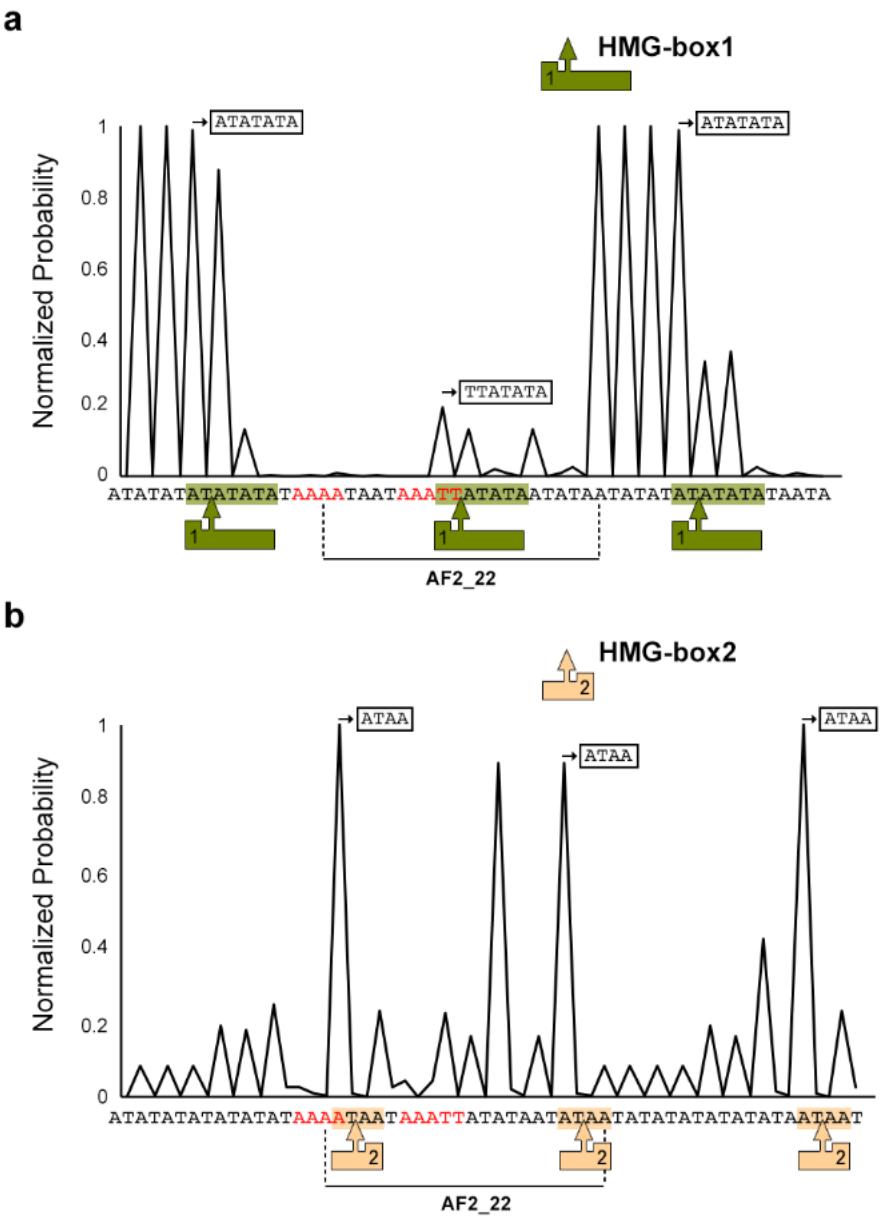


Figure R31. Binding probability of HMG-boxes on the 64bp sequence. (a) Normalized probability for HMG-box1 to bind and distort a sequence stretch (see Methods). Each peak corresponds to the probability of HMG-box1 binding to the 7bp stretch starting from the peak position. The said sequence stretches are indicated in boxes above each relevant peak and additionally highlighted in green on the 64bp sequence. HMG-box1 is represented by a green L shape (1) and the site of insertion is indicated by an arrow in each case. The location of the Af2_22 sequence within the 64bp sequence is demarcated. The A-tracts within the 64bp sequence are shown in red. **(b)** Normalized probability for HMG-box2. Features are represented similar to **(a)**. The 4bp HMG-box2 binding sites are shown for representative peaks.

Table R4. Abundance of symmetric and asymmetric A tracts of different lengths in y-mtDNA.

	A_2T_2	A_3T_3	A_4T_4	A_5T_5	A_6T_6	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
No.	4128	300	22	1	1	1201	530	158	52	38	33	15	6	2	2

10.3 A high deformation energy for A-tracts prevents Abf2p binding

In order to explore how the DNA conformational properties affect Abf2p binding, an Elastic Deformation Energy model [143,144] (see Methods) was used to calculate the energy required to contort segments of the natural y-mtDNA 64bp sequence into the conformations induced by Hmg-box1 or Hmg-box2. (see Methods). Based on the rationale that binding propensity will be higher for those segments that require less energy to be distorted, we derived a sequence-dependent probability of binding (see Methods, Figure R31). The results show exclusion of the HMG-boxes from the A-tracts. Additionally, HMG-box1 binding probability (Figure R31a) shows a more restrained pattern compared to HMG-box2 (Figure R31b), indicating that the former has more stringent sequence requirements and a higher tendency to avoid A-tracts. The binding probability of full-length Abf2p can be expected to be a combination of that observed for the individual boxes. In summary, the narrow minor groove and rigidity of the A-tracts make them unsuitable as binding sites for Abf2p, while the adjacent sequence stretches are well suited for insertion and are favored for positioning the HMG-boxes. This points to a mechanism of protein positioning mediated by DNA structure, which somehow resembles that suggested for nucleosome placements in nuclear chromatin [156-160]. Interestingly, we calculated that 33.3% of the y-mtDNA is composed of A-tracts [18] (symmetric and asymmetric; Table R4 and Appendix A), consistent with previous estimates [161]. According to the above results, Abf2p DNA binding would follow a similar trend at other y-mtDNA regions possessing A-tracts.

11. The thermodynamics of Abf2p DNA binding is altered by A-tracts

Analysis of the thermodynamics of Abf2p DNA recognition by isothermal titration calorimetry (ITC) showed that the presence of A-tracts modifies the mode of binding. Titration of Abf2p with Af2_22 DNA (protein in cell, DNA in syringe), produced a thermogram with an exothermic phase until a DNA:protein molar ratio of 1:1 was reached. Notably, a second endothermic phase followed (Figure R32a, top left). The isotherm was fitted to a model for two independent sets of sites (Figure R32a, bottom left), which would correspond to two binding sites on the protein. For site1, a stoichiometry of 0.86 (+/- 0.09) was obtained corresponding to a ~1:1 DNA:site1 ratio (Table R5). For site2, a binding stoichiometry of 0.49 (+/- 0.1) was obtained, corresponding to a 1:2 (DNA:site2) ratio (Table R5). In addition, the derived affinity for site1 (Kd=52.1 nM) was higher than for site2 (Kd=166.4 nM). Our EMSA assays demonstrate that the Nhelix+HMG-box1 construct has a much higher DNA binding efficiency than HMG-box2. Additionally, the crystal structures and the predicted probability of binding to the 64 bps sequence indicate that the Nhelix+HMG-box1 module is constrained to be positioned outside the A-tracts and has a single preferred binding site on the Af2_22 sequence (in grey in Figure R32b top, Figure R31). In contrast, HMG-box2 possesses two putative binding sites on Af2_22. Therefore, we rationalized that the first exothermic phase is mostly attributable to the binding of N-helix+HMG-box1, until a 1:1 (DNA:site1) molar ratio is reached (Figure R32b, top). After the Nhelix+HMG-box1 module is saturated, the HMG-box2 domains have access to two sites on each new injected DNA molecule, yielding a 1:2 (DNA:site2) stoichiometry (Figure R32b, top). Thus, the two binding events together lead to an overall molar ratio of 3:2 (DNA:protein).

Additional ITC experiments were performed with a 22bp DNA sequence that did not contain any A-tract (see Methods and Figure R32a, right). These showed an exothermic phase until a molar ratio close to 1:1 was reached, followed by a slightly endothermic phase (Figure R32a, top right). Fitting the data to the two-sites model yielded stoichiometry values of 0.55 (+/- 0.01) and 0.50 (+/- 0.03) for sites 1 and 2 respectively (Figure R32a, bottom right; Table R5). Thus, both sites correspond to a 1:2 (DNA:site) molar ratio. In this case we propose that the absence of A-tracts on GC_22 allows the Nhelix+HMGbox1 module (site 1) to bind freely, giving a 1:2 (DNA:site1) complex (Figure R32b, bottom). On near-saturation of the Nhelix+HMGbox1 fragment, HMG-box2 (site 2) binding becomes predominant, giving

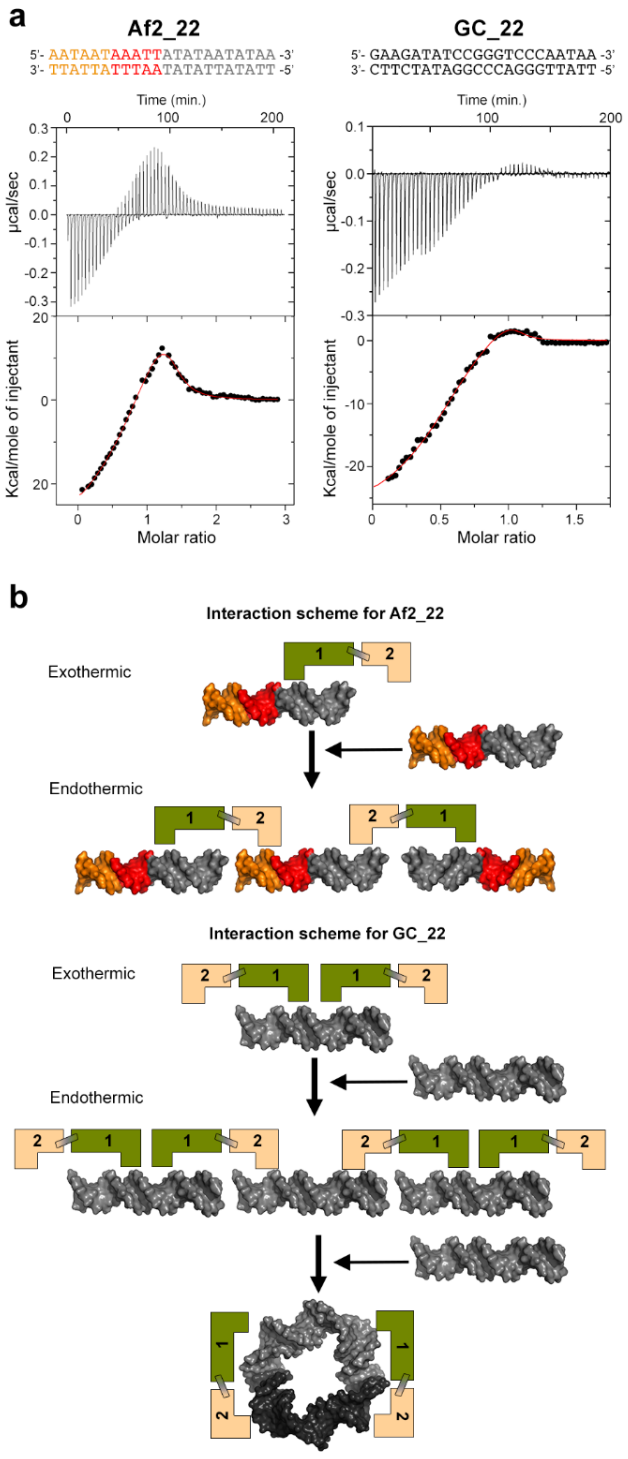


Figure R32. Thermodynamics of interaction of Abf2p with DNA sequences with and without A-tracts (a) Isothermal titration calorimetry thermograms (above) and fitting of the binding isotherms (below) to a model with two independent binding sites for Af2_22 (left) and GC_22 (right) DNA. The corresponding DNA sequences are shown at the top of each thermogram. **(b)** Schemes of the binding model with two independent binding sites for Af2_22 (top) and GC_22 DNA (bottom). Abf2p is depicted schematically with the HMG-boxes 1 and 2 represented in green (1) and orange (2) respectively. The non A-tract part of the ARSm-like sequence in Af2_22 is represented in orange whereas the A-tract is depicted in red. At the very bottom, the 1:1 stoichiometry is represented as a plausible circle in order to show its compatibility with the crystal structure (the two dsDNA molecules in the circle are depicted in different gray tones).

Table R5. Thermodynamic parameters obtained from ITC experiments

ITC Experiment	N1	ΔH1 (kCal/mole)	ΔS1 (Cal/mole)	Kd1 (nM)	N2	ΔH2 (kCal/mole)	ΔS2 Cal/mole	Kd2 (nM)
Abf2p/Af2_22	0.859 +/- 0.0873	-37.37 +/- 4.65	-91.9	52.08 +/- 202	0.494 +/- 0.100	57.76 +/- 25.7	225	166.39 +/- 1262.6
Abf2p/GC_22	0.554 +/- 0.0158	-29.5 +/- 1.8	-61.7	7.25 +/- 8.77	0.505 +/- 0.0332	11.21 +/- 2.66	71.6	36.76 +/- 47.85

a 1:2 ratio for site 2. Therefore, for the GC_22 complex the overall molar ratio (DNA:protein) is 1:1, whereas for the Af2_22 it is 3:2. Both for Abf2p/Af2_22 and Abf2p/GC_22 experiments, the enthalpy and entropy changes are negative for site 1 and positive for site 2 (Table R5). Thus, considering $\Delta G = \Delta H - T\Delta S$ (see Methods), where an event is spontaneous if ΔG is negative, site 1 binding is an enthalpy (ΔH or heat exchange) driven process while site2 is an entropy (measure of randomness) driven process. Additionally, in both cases, site 1 has a higher binding affinity Ka (thus lower dissociation constant Kd) compared to site 2 (Table R2). This supports the interaction scheme proposed in Figure R32b. Thus, the ITC experiments provide further experimental evidence that the A-tracts dictate the mode of DNA binding by Abf2p.

Discussion

Abf2p functions like a staple on DNA with each of the two HMG-boxes bending the DNA by 90 degrees, thus generating a U-turn. In this arrangement, the N- and C-terminus of Abf2p are positioned in close proximity. This feature is absent in the free protein, suggesting that the protein conformation observed in the crystal is induced by DNA binding. Additionally, the HMG-boxes alone have very low DNA binding efficiency, and the presence of the N-helix is crucial for maintaining DNA binding activity, both *in vitro* and *in vivo*. Our findings suggest a sequential DNA binding model where the N-helix+HMG-box1 module initiates DNA binding. In this event, the N-helix functions as a pin-lock that consolidates the N-hydrophobic core between the N-helix, HMG-box1 helix3, and the Linker. This arrangement might facilitate interaction of HMG-box2 with downstream DNA and consequent bending. The N-flag and the first turn of the N-helix can now make contacts with DNA minor groove further downstream, potentially contributing to the stability of the U-turn. The final outcome is a stable protein/DNA complex where the N- and C-termini of Abf2p are in close proximity.

Abf2p and its human counterpart TFAM are the best characterized mtDNA packaging proteins. Both show universal features of mtDNA bending and compaction through their HMG-box domains, but present very low sequence similarity (Figure I6). We don't know if this reflects different roles played by the two proteins in nucleoid organization and segregation or possibly, differences in mtDNA base composition and topology. Indeed, striking differences are found between respective mtDNAs: human mtDNA is a small G-C rich circular molecule of 16.5 kbp, whilst yeast genome is an A-T rich linear molecule of 80 kbp. Both proteins generate a U-turn on the DNA [52,54] (Figure D1a,b). However, TFAM, by virtue of its long and flexible 30aa linker, intertwines the DNA, positioning the HMG-boxes at opposite faces on the DNA (Figure D1b). In contrast, Abf2p possesses a short linker and thus the two vicinal HMG-boxes interact with the DNA from one side (Figure D1a). As a result, the N- and C-termini are distant in space in TFAM, but come close in Abf2p. Additionally, HMG-box1 is oriented in opposite directions in the two proteins (Figure D1c,d). HMG-box2 of TFAM, on the other hand superposes well with HMG-box2 of Abf2p.Further differences exist in terms of the insertions made by the two proteins in the DNA, relative to the U-turn. While both proteins maintain a 11 bps distance between the insertions made by box1 and box2, the insertions of TFAM are displaced by 1bp relative to those of Abf2p (Figure D1c). Additionally, the HMG-box2 insertion for Abf2p comes from Ile124 located on helix1, whereas that

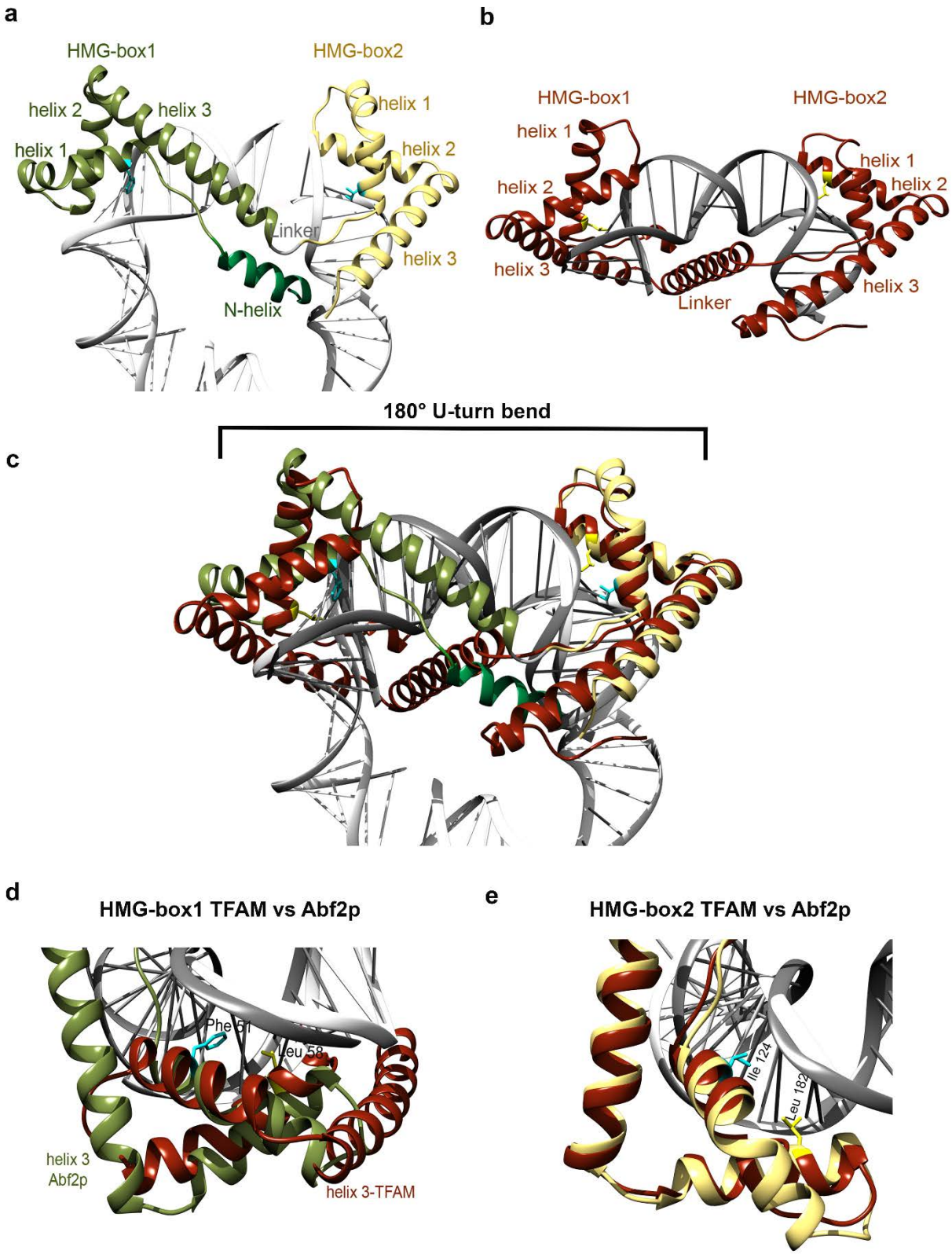


Figure D1. Similarities and differences in DNA bending by Abf2p and TFAM. **a**. DNA bending by Abf2p. The different protein structural parts are colored as in other figures. The inserting residues Phe51 and Ile124 are shown as sticks (cyan). **b**. DNA bending by TFAM (PDB ID 3TQ6). The protein is depicted in red. The inserting residues Leu58 and Leu182 are shown as sticks (yellow). **c**. Superposition of the U-turns for Abf2p and TFAM. The high similarity of the 180° bends is notable. **d**. Interaction and insertion of HMG-box1 from Abf2p and TFAM with the DNA. The respective insertions are shown. helix3 of Abf2p and TFAM are in opposite orientations at the bend. **e**. HMG-box2 interactions for Abf2p and TFAM. The inserting residues are shown. The Abf2p and TFAM box2 are in the same orientation relative to the DNA bend and are well superposed.

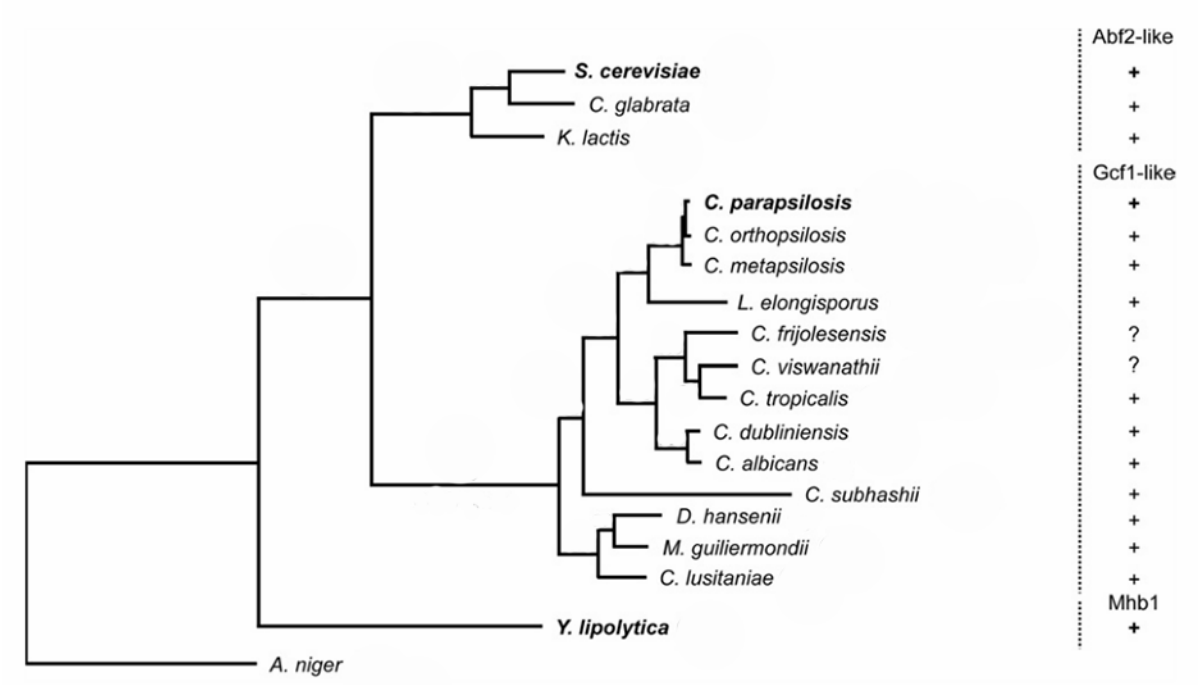


Figure D2. **Phylogenetic relationship between mitochondrial HMG-box proteins from different yeast species**. Bakkaiova *et al*. *Biosci. Rep.* **36,** (2015).

of TFAM comes from Leu 182 located on helix2 (Figure D1e). Therefore, the two homologs, constrained by their respective structural characteristics, utilize different binding strategies to achieve a similar DNA bending. This suggests that U-turns constitute the basic structural unit of mitochondrial nucleoid organization in both systems.

The EMSA assays (Figure R19b) reveal that Abf2p binds DNA templates at a ratio of 1Abf2p:10-20bps, which agrees with the number of DNA bps (~19bps) contacted by the protein in the crystal structures. Additionally, previous atomic force microscopy studies showed that Abf2p induces compaction at a 1Abf2p:20bps ratio [64], being more prominent at 1:10. At these ratios TFAM showed higher compaction, which correlated with a reduction in *in vitro* DNA replication and transcription [162]. A similar effect was observed for Abf2p, where Abf2p overexpression resulted in loss of y-mtDNA, suggesting excessive DNA compaction and consequent reduced access for the replication machinery [66]. Thus, at ratios of 1Abf2p:10-20bps and higher, considerable compaction of the DNA would occur and therefore relative abundance of Abf2p with respect to available DNA binding sites will dictate the extent of compaction in mitochondrial nucleoids.

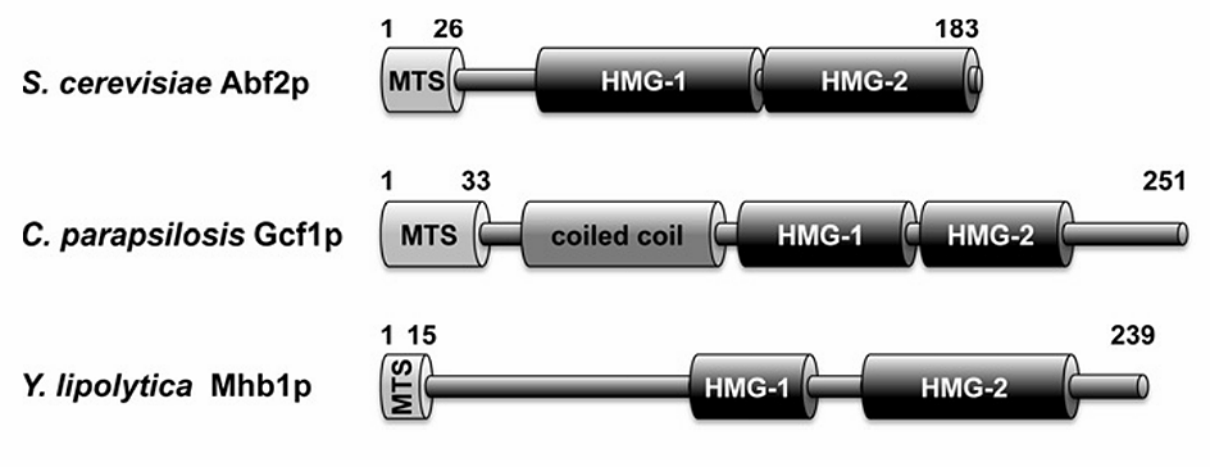


Figure D3. Predicted secondary structure features for the three proteins Abf2p, Gcf1p and Mhb1p. Bakkaiova *et al*. *Biosci. Rep.* **36,** (2015).

The dual-binding found in the crystal structures and the relative conformational freedom of the two HMG-boxes observed from SAXS-EOM analyses demonstrates the capability of the protein to bind two DNA molecules simultaneously. This ability would be compatible with 'cross-strand binding' or DNA looping by Abf2p, as also described for TFAM [55]. Thus, both DNA bending and cross-strand binding might be underlying the nucleoid packaging mechanism of the two proteins.

In extension of the above comparison between Abf2p and TFAM it is interesting to compare the former with putative mtDNA packaging proteins from other yeast species. In this regard, Bakkaiova *et al*, 2016 [68] have performed phylogenetic comparison of mitochondrial HMG-box proteins from different yeast species. Based on the said comparison they identify three main sub-types of HMG-box proteins in yeasts: *S. cerevisiae* type (Abf2p), *C.parapsilosis* type (Gcf1p) and *Y. lipolytica* type (Mhb1p) (Figure D2). The three HMG-box proteins differ in the arrangement and size of the boxes (Figure D3). Gcf1p possesses a coiled coil extension N-terminal to HMG-box1 and both is HMG-boxes are shorter than Abf2p. Additionally HMG-box1 region for Gcf1p was assigned based on weak sequence conservation with HMG-box1 in orthologues from *Candida lusitaniae*, *Candida subhashii*, *Debaryomyces hansenii* and *Meyerozyma guilliermondii.* The actual tertiary structure of this region in Gcf1p remains to be validated by structural studies. Mhb1p on the other hand is predicted to have an unstructured region at the N-terminus and its predicted HMG-box1 is much shorter than Abf2p. Thus there is considerable structural variation between the three proteins. The Abf2p structures presented here show an N-terminal

helix that could not be predicted by bioinformatics. This raises the question whether Gcf1p and Mhb1p also possess such N-terminal helices and if they play a similar functional role as in Abf2p. However, these questions can only be definitively answered in light of experimentally determined structures.

Nevertheless, functionally these proteins differ in their binding affinities for dsDNA and to non-canonical DNA structures like 4-way junctions and replication forks [68] and this might reflect functionally distinct roles played by the three proteins.

The presented results demonstrate that the A-tracts direct positioning of Abf2p on DNA. Crystallization trials with DNA sequences not containing A-tracts yielded poorly diffracting crystals, suggesting high disorder. Presumably, by specifying positioning of the protein, the A-tracts induced a complex that formed the ordered crystals reported here. Accordingly, previous competition experiments showed that binding of Abf2p to an ARS1 probe could not be competed by a poly(dA)·poly(dT) double-stranded homo polymer [70,71], indicating the incapability of this protein to bind to poly-adenine DNA. A-tract guided nucleo-protein interactions have also been proposed in the nucleus, where biochemical studies have shown that A-tracts help in positioning nucleosomes [157-160]. This has been further corroborated by computational approaches [144,156]. Additionally, other biochemical experiments have demonstrated that the A-tracts are involved in DNA synthesis termination in the human immunodeficiency virus [163,164] and in regulation of transcription in pathogenic bacteria [165,166,167]. Thus, A-tract mediated control of protein/nucleic acid interactions appears to be a general strategy across species. We and others detected that y-mtDNA has a high content of A-tracts [18,161]. Thus, location of these tracts might determine strategic positioning of Abf2p, suggesting a DNA-mediated control of nucleoid architecture and of mtDNA accessibility.

Another novel and unexpected feature observed in the crystals is DNA end-joining by Abf2p. This could be relevant for cellular processes where DNA ends are available, such as in DNA breaks or recombination. Notably, Abf2p levels correlate with the level of recombination intermediates in mitochondria [65,66] and, additionally, y-mtDNA replication is proposed to be recombination-dependent [32,42]. Thus, Abf2p might be involved in maintenance of y-mtDNA copy number via interaction with recombination intermediates [65,66]. However, any potential biological implication of Abf2p DNA end-joining in recombination remains to be analyzed by further studies.

Assimilation of the obtained results thus provides a molecular mechanism for Abf2p/DNA interaction and helps construct a putative model for global nucleoid architecture in yeast. The basic structural unit of DNA packaging seems to be a U-turn. This could be augmented by DNA looping or 'cross-strand' binding by Abf2p and possibly by other protein-protein interactions. The binding of Abf2p to DNA is modulated by the positioning of poly-adenine tracts and in view of the abundance of A-tracts in mtDNA of yeast, this poses a DNA structure guided orchestration of nucleoid architecture and packaging in conjunction with U-turns. The presented results thus demonstrate how the intrinsic structural properties of DNA can play a key role in directing localization of non-specific DNA binding proteins that are essential for mitochondrial maintenance, and thus for cell life.

Conclusions

1. Abf2p binds DNA from one side, makes two insertions into the DNA separated by 11bp and causes a 180^{o} DNA bend. Thus it uses a mechanism distinct from its human counterpart TFAM to achieve a similar U-turn. This suggests that U-turns are the basic structural units for mtDNA packaging in human and yeast.
2. *In vivo* and *in vitro* assays reveal that the two HMG-boxes of the protein possess low DNA binding efficiency on their own. A N-terminal helix, unique to this protein is crucial for protein-DNA interaction and results in significantly higher DNA binding efficiency of the N-helix-HMG-box1 module compared to HMG-box2. The crystal structures show that the N-helix forms a hydrophobic core in the protein/DNA complex and MD simulations reveal that this core forms on DNA binding. Thus a sequential binding model is proposed in which the N-helix-HMG-box1 module binds DNA first with concomitant consolidation of the N-hydrophobic core, followed by HMG-box2 binding.
3. In-solution biophysical analyses (SAXS) and MD simulations reveal that Abf2p is an intrinsically flexible protein with high relative conformational freedom of the two HMG-box domains. This inherent flexibility enables the protein to bind separate DNA molecules, as seen in the crystal structures. Such a capability could be important in mtDNA packaging by allowing DNA looping by Abf2p. On binding DNA, Abf2p forms a compact and stable complex.
4. Structural properties of poly-adenine tracts (A-tracts) prevent Abf2p binding, as observed from the crystal structures, MD simulations and related computational analyses. Extrapolation of these results to yeast mtDNA points to a DNA structure mediated protein positioning mechanism.
5. A-tract mediated Abf2p exclusion and difference in DNA binding efficiency between the N-helix-HMG-box1 module and HMG-box2 lead to distinct DNA binding phenomena and thermodynamic manifestations in case of A-tract containing DNA compared to DNA devoid of A-tracts. This allows

identification of an exothermic binding event for N-helix-HMG-box1 module and an endothermic binding event for HMG-box2.

6. The high abundance of A-tracts in the yeast mtDNA and the exclusion of Abf2p from A-tracts points to a DNA guided mechanism for orchestration of global nucleoid architecture and packaging, thus coordinating functional transactions of mtDNA.

Appendix

Appendix A: A-tracts in y-mtDNA

A-tracts, both symmetric and asymmetric (see main text), are demarcated in bold non-capital lettering.

TTCAT**aattaatttttt**ATATATATATTATATTATAATA**ttaatt**TATAT
TAT**aaaaa**TAATATTTATTA**ttaaaa**TATTTATTCTCCTTTCGGGGTTCC
GGCTCCCGTGGCCGGGCCCCGG**aatt**A**ttaattaa**TAAT**aatt**ATTA**tta**
aT**aatt**ATTTATTA**tttt**ATCA**ttaaaa**TATATAAAT**aaaaaa**TA**ttaaa**
aaGAT**aaaaaaaa**TAATGTTTATTCTTTATAT**aatt**ATATATATATATAT
aattaattaattaattaattaattaaTAAT**aaaaa**TAT**aatt**ATAAATAA
TATAAATATTATTCTTTA**ttaa**TAAATATATATTTATATATTAT**aaaa**GT
ATC**ttaattaa**T**aaaaa**TAAACAT**ttaa**TAATATG**aatt**ATATATTATTA
TTATTA**ttaa**Ta**aatt**A**ttaa**TAATAATCAATATG**aattaa**T**aaaaa**TCT
TAT**aaaaaa**GTAATGAATACTCC**ttttttaaaaa**T**aaaaa**GGGGTTCGGTC
CCCCCCCTTCCGTATACTTACGGGAGGGGGGTCCCTCACTCCTTC**ttaat**
taattATC**ttaattaatt**ATC**ttaattaatt**ATC**ttaattaatt**ATC**tta**
attaattATC**ttaattaattaaaa**GGGGACTTTATATTTATAAAGT**aatt**
ATATTATTATTATTATTATTATTTATTTA**ttttt**A**ttttt**ATTA**ttttt**ATT
ATATATATTATATA**ttaa**TACAGATAGAAGCC**aaaa**GGTCAGGCGCTTTC
TTTGGGAGAAAGACCTAGTTAGTTCGAGTCTATCCTATCTGATAAT**aatt**
taattaaCCA**ttaaaaaaaa**GTATATATATTTATCATAATATA**ttaattt**
tATTACATTACAAATGAACAC**tttt**ATTTATATTTAT**aaaaa**TATGAACT
CCTTCGGGGTCCGCCCCGCGGGGGCGGGCCGGACTCCATATTATTATTAT
TAT**aatt**ATTATTAT**aatt**ATTATTAT**aatt**ATTATTAT**aatt**ATTATTA
T**aattaa**AGAG**tttt**GGATACCAATATGATATAATATGATATAGGACCGA
AACCCCTCA**tttt**ATCATTTATTTATAATATTATAAAT**aaaaaaaaa**TAT
TATATATTATAATa**aattaa**TATCATAATATATTATATTATATATTATAT
TATATATATATATATATATATTC**tttt**AT**aaattt**ATATTCTTCTTA**tta**
aattaaaaaGGGAGCGGAC**ttttaatt**ATAT**ttaatt**ATAG**ttttttaa**TC
ATTGGTTGAGATTTC**aaaa**TAAGGTATAATATTTATATTATTCT**ttaa**CA
AATATTATATTAT**aaaaaaa**GATATAATATTTATATTATTCT**ttaa**CAAA
TATTATATTAT**aaaaaa**GATATAATATTTATATATTATTA**ttaa**TATTA**t**
ttttaaGTTCCGAAAGGAGAAACTTAT**aatttttt**ATATCATTATTTATTA
TTA**ttttttaatt**TCAACTCC**tttt**AGGTATTTCCAT**ttaa**CTTTCAGCAG
AGACTTTCT**aatt**AT**aatt**ATATATATATA**aatttaaa**TACATTTAT**aaaa**

aaGTATATAATAT**aatt**ATATTATATATAATAATATTA**ttaa**ATGAAGTA
TTCTTTATTA**ttaatt**ATAGGATATCTGGGGTCCA**ttaa**T**aatt**ATTATT
GTAAATAATAATAAGGACCCCCCCCATTATCT**aattaa**TAAATATATAAA
TAATCA**ttaa**TAAATATA**ttaa**T**aatt**A**ttaa**TAAATATATAAATAATCA
ttaaTAAATATATAAATAATATAATATATTAT**aaaaa**TATAATAATAA**Ta**
attTATTA**ttaaaa**TATAAT**aatt**TATTAT**aaaaa**TATAAT**aatt**TATTA
T**aaaaa**TATAATAATAACTCCTTTCGGGGTTCACACCTTTATAAATAATA
AATAATAAATAATAAATAATAAATAATAAATATTAGTATTCACTAATAT**a**
aaaTAAT**aatt**AT**aaaaa**TAATCATTA**ttaaaaa**TATTA**ttaatt**A**ttaa**
ttaaATAC**aattaa**TAT**aatt**TAGTTGTTTATAT**aattttaaa**TAATGTT
TATATC**aatttaa**Ta**attaatt**tATAGTTCCGGGGCCCGGCCACGGGAG
CCGGAACCCCGAAAGGAGTTTATCTATATATTATAATAACTATATG**aatt**
taattA**ttaaaaa**TAAT**aaaaa**TAAGG**aattttaa**TAAGAAGTAATATTT
ATTATATAATATAT**aaaaaaaa**TATATATATATATAT**aaaaa**TATATATA
ATAAG**tttt**ATTATAATATATA**ttaattaatt**ATTATGAGGGGTTCGGTC
CCTTTCCGGGCCCC**aatt**CATCTCATCTCA**tttt**ATTTCATTTCAATATC
ATCTAATCTCATTTCTTTATAGA**tttt**ACATATATATAAATATAAATATA
AGATATTCACATTTATATATAATATAATATAATATAATAGATATTCATTC
CTCTTTGA**ttaa**ACTAAT**aattaa**T**aattaa**T**aattaa**T**aattaa**T**aatt**
aaT**aatt**ATTCAGTAGAACTCCTTC**ttaaaaa**GGGGTTCGGTCCCCCTCC
CATTAGTATAGTATAGGGAGGGGTCCCTCACTCCTTCGGGGTCCGCCCCG
CAGGGGGCGGGCCGGACTATTA**ttaa**AT**aatt**TAT**aatt**TATTATTTA**tt**
aaTATATTTATATAATATAATATAATATAATATTATTCATAC**ttttt**A**tt**
aaTATAATATAATATAATATTA**ttaa**TACTTTCTCCTTTCGGGGTTCCGG
CTCCCGTGGCCGGGCCCCGGAACTA**ttaa**TATAAAG**aaaa**GAGTTTC**aat**
tATTTATTTATTTATTTA**tttttt**AT**aaaaa**TAAGTCCCCGCCCCGGCGG
GGACCCCGAAGGAGTA**ttaatttaaa**T**aatt**TAT**ttaa**TG**aatt**A**ttaat**
tATAAAT**aaaaa**TAAT**aatttttaaa**GATGTAATAT**aaaaa**TAAATATAA
TAT**aatt**TAGGAT**aatt**ATAT**aaaa**TATTTATTATATATAG**ttttt**ATAA
AGAG**ttttaaaa**GTGATAATATAATATATAATATTTATAAGTTCCGGGGC
CCGGCCACGGGAGCCGGAACCCCGAAAGGAGTTATTTATATATATAT**aat**
tATAATCTTA**ttaatt**ATTTATATATATAT**ttaa**TATTA**ttttt**ATAT**aa**
tttATA**ttaa**AGTATTAT**aatt**ATATAT**ttaa**TATTA**ttttt**ATAT**aat**
tttATATTATTTATTTATTTATTTATTTA**tttaaaaa**TATTATAATCATA
TAT**ttaa**TATTAT**ttaa**TATA**tttt**ATATATTATATC**tttt**ATTGATTTA
TATATATATAGAT**ttaa**TAAATATATATATATATATATATATAAATATTC
ATTATATATTTATTATTATTATTATTTATTACTA**tttttt**ATTATAT

A**ttaa**TAATATATATATTATTAGTTATGGGTATCCTAATAGTATATTATT
A**ttttttaa**TAAT**aatt**TATGATTTATGTATAATAAATAAGTAGGGAATCG
GTACGAATATCGAAAGGAGTTATATATTA**ttaatt**ATTTAT**aatt**A**tttt**
ATATATTA**ttaatt**ATTTAT**aatt**A**tttt**ATATATTTAT**aatt**A**tttt**AT
ATAGATAGGTTAGATAGGATAGATAGTATAGATAGGGGTCCCATTTATTA
TTTACAATAAT**aatt**A**ttaa**TGGGACCCGGATATCTTATTGTTA**ttaatt**
TATATATTATTCATTATTA**ttaa**TATATAT**ttaa**TAT**aattaa**ATATTAT
ATTATATTATATTATATTATTTA**ttaaaaaaaaa**TCTATTACTTA**ttttt**
tttA**ttaa**TATAT**aatt**ATTTATAT**aatt**TATCA**ttttt**ATTTATATATT
ATTA**tttttt**ATATAT**aattaa**TATATATATATATTATATATAC**tttttt**
ttttATAATATATCTATATATATAAATAAATATATTATATTATA**ttttt**A
TATAATATATTA**ttaatt**ATTA**ttttaattttt**CTATTCTATTGTGGGGGT
CCC**aatt**ATTA**tttt**CAATAAT**aatt**ATTATTGGGACCCGGATATCTTCT
TGTTTATCATTTATTA**tttt**A**ttaatt**tATTATTA**ttttttaatt**TATATT
TATATTATAT**aattaatt**ATATCGTTTATACTCCTTCGGGGTCCCCGCCG
GGGCGGGGACTTTATA**tttt**ATTATATAATATATTATATTCTTATAATAT
ATTTATTGATTATGTTAT**aaattt**ATTCTATGTGTGCTCTATATATAT**tt**
aaTATTCTGGTTATTATCACCCACCCCCTCCCCCTATTACGTCTCCGAGG
TCCCGGTTTCGTAAGAAACCGGGACTTATATATTTATAAATATAAATCTA
AC**ttaattaa**T**aatttaaa**TAATATACTTTATA**tttt**ATAAAT**aaaaa**T**a**
attATAACC**tttttt**AT**aatt**ATATATAATAATAATATATATTATCAAAT
aattATTATTTC**tttttttt**CT**ttaattaattaattaattaa**TA**tttt**AT
aaaaaTATATTTCTCCTTACGGGGTTCCGGCTCCCGTAGCCGGGGCCCGA
AACTAAAT**aaaa**TATATTA**ttaa**TAATATTATATAATATAATAATAATAT
AAT**aattttt**ATATAAATATATATTTATATA**ttaattaatt**AT**aattttt**AT
TATGa**aatt**ATATC**tttttttt**ATA**ttttt**ATATAAT**aaaaa**TATGTTAT
ATATATA**ttaa**TAAT**aaaa**GGTAGTGAGGA**ttaa**AT**aatt**ATATAAT**aat**
tATAACTC**ttaatt**AT**aaaa**TAAATATATATATATATATAAGTATCCATT
TCCATATAATC**ttttaa**TAAATA**ttaa**TAAATA**ttaaaaaaaaa**TAATAT
TATAATA**tttt**AGTATAT**aatt**CAATa**aatt**CATTGGAGGGGTAAATAAT
AAT**aatt**TACTAATGGCAAGTTATAGTC**ttaa**AGG**ttttt**A**tttttttt**A
ttaattaaT**aaaa**TAATAATACCATTTATATATTCCATTATATATATATA
T**ttaa**T**aaaaa**TAATAATATCATTTATATA**tttt**ATTATATATTATATAT
A**tttt**ATAT**aaaa**TAATAATAAT**aatt**tATA**ttttt**ATATATTATTA**tta**
aATAATAATAATATAAATAACTCCTTCGGGGTTCGGTCCCCACGGGTCCC
TCACTCCTTC**ttaa**GAAT**aaaaa**GGGGTTCGGTCCCCCTCCCGTTAGTAC
ACGGGAGGGGGTCTCTCACTCCTTC**ttaaaaaa**T**aaaaa**GGTGGAAGGAC

TAATAT**aattttaaa**TAAT**aattaa**TACT**ttaa**TAAT**aatt**TGTATTTCT
TTATTA**ttaa**TATA**ttaa**ATATAATAAT**aattaa**TAT**aatt**ACAATATA**t**
taaTATTATCAAATA**ttaa**TAAATATAC**tttttt**ATAT**aatt**TATTTATT
TATTTA**ttttttttttt**A**ttaa**ACT**aatt**AT**aatt**GT**aatt**TCG**aaaa**GGG
GGTGGGAGTAAACATATAT**aatt**TATAATCTATATATATATATATAT**aat**
ttttaaTAAATA**ttaa**TAAATATTTAT**aaaaa**GAAT**aatt**TATATTTAT
AATATAT**aatt**TATATA**tttt**A**ttttt**ATTATAC**aattaa**TAT**aaaa**TAT
aaaaTA**ttaa**ATA**ttaa**ATA**ttaa**ATA**ttaa**ATA**ttaa**ATA**ttaattttt**
ATAGGGGTTATATAAT**aatt**ATATTTAT**aatt**ATATAATA**ttaaaaa**GGG
TA**ttttt**AT**aatt**ATTACA**ttttt**A**tttt**ATTTAT**aaaaa**TA**ttaatttt**
aaTAAGTATTGAATACTTTATATAATATAAATA**ttaatt**ACAT**aattaa**T
aattaaATAATAT**ttaa**TAATATTAT**ttaatt**tATTATTTAT**aatt**ATTT
ATTTATa**aatt**CTA**ttttt**ATTATTATTA**ttttt**A**tttt**ATTA**ttaa**AGA
ttaaTATAAT**aatt**A**ttaa**TATA**ttaaaaa**TC**tttt**ATTATA**ttaa**TATT
TAT**aaaaaa**GTAT**ttaa**T**aaaaaa**GATGTAT**aatt**tAT**aatt**ATATAATA
TTA**ttaatt**TATATAATAATAATATTATAACTTTGTGATTGTC**aatt**TAG
ttaaTCATTGTTA**ttaa**TAAAGGAAAGATAT**aaaaaa**TATTCTCCTTC**tt**
aaaaaGGGGTTCGGTTCCCCCCCGTAAGGGGGGGGTCCCTCACTCCTTTG
GTCGGACTCCTTCGGGGTCCGCCCCGCGGGGGCGGGCCGGACT**aatttaa**
C**ttttaa**TA**ttaa**TA**ttaa**TATTATTTATA**ttttt**A**ttaa**TATAT**aaaaa**TAA
AT**aattttA**t**ttttA**t**taa**TAGTATATTATATAAACAAT**aaaa**TAGTA**tt**
aattATAT**aaattt**ATAT**aaaa**TATATAT**aatt**tATTATATATATATATA
ttaaTA**ttttaa**TAAAG**ttttt**ATTAT**aatt**tATTTATTTATTTATTATA
ATA**ttaa**T**aatt**TATTTATTATTATATAAGTAATAAATAATAG**tttt**ATA
TAATAATAATAATATATATATATATATATTATTATATTAGTTATATAATA
AGG**aaaa**GT**aaaaattt**ATAAGAATATGATGTTGGTTCAGA**ttaa**GCGCT
AAATAAGGACATGACACATGCGAATCATACGTTTATTATTGATAAGATAA
TAAATATGTGGTGTAAACGTGAGT**aatttt**ATTAGG**aattaa**TGAACTAT
AGAATAAGCTAAATAC**ttaa**TATATTATTATAT**aaaaa**T**aatt**TATATAA
T**aaaaa**GGATATATATATAATATATATTTATCTATAGTCAAGCCAATAAT
GGTTTAGGTAGTAGGTTTA**ttaa**GAG**ttaa**ACCTAGCCAACGATCCATAA
TCGATAATGAAAGTTAGAACGATCACGTTGACTCTGAAATATAGTCAATA
TCTATAAGATACAGCAGTGAGGAATATTGGACAATGATCGAAAGATTGAT
CCAGTTACTTATTAGGATGATATAT**aaaaa**TA**tttt**A**tttt**ATTTATAAA
TA**ttaa**ATATTTATAATAATAATAATAATAATATATATATAT**aatt**GA**tt**
aaaaaT**aaaa**TCCATAAAT**aattaaaa**TAATGATA**ttaatt**ACCATATAT
A**ttttt**ATATGGATATATATA**ttaa**TAATAATA**ttaattttt**ATTATTA**tt**

aaTAATATA**ttttaa**TAGTCCTGACTAATATTTGTGCCAGCAGTCGCGGT
AACACAAAGAGGGCGAGCG**ttaa**TCATAATGG**tttaaa**GGATCCGTAGAA
TG**aatt**ATATATTAT**aatt**TAGAG**ttaa**T**aaaa**TAT**aattaa**AG**aatt**AT
AATAGTAAAGATGAAATAATAATAAT**aatt**ATAAGACTAATATATGTG**aa**
aaTA**ttaattaa**ATA**ttaa**CTGACATTGAGGGA**ttaaaa**CTAGAGTAGCG
AAACGGATTCGATACCCGTGTAGTTCTAGTAGTAAACTATGAATAC**aatt**
ATTTATAATATATATTATATATAAATAATAAATG**aaaa**TGAAAGTATTCC
ACCTGAAGAGTACGTTAGCAATAATGAAACTC**aaaa**CAATAGACGGTTAC
AGAC**ttaa**GCAGTGGAGCATGTTAT**ttaatt**CGATAATCCACGACTAACC
TTACCATA**tttt**GAATATTATAAT**aatt**ATTAT**aatt**ATTATATTACAGG
CGTTACATTGTTGTCTTTAGTTCGTGCTGCAAAG**tttt**AGA**ttaa**GTTCA
TAAACGAAC**aaaa**CTCCATATATAT**aattttaatt**ATATAT**aattttA**TA
TTATTTA**ttaa**TATAAAGAAAGG**aattaa**GACAAATCATAATGATCCTTA
TAATATGGGTAATAGACGTGCTATAAT**aaaa**TGATAATa**aatt**ATAT**aaa**
aTATAT**ttaatt**ATAT**ttaattaa**TAATAT**aaaa**CA**ttttaatttttaa**T
ATA**ttttttt**ATTATATA**ttaa**TATG**aatt**ATAATCTG**aatt**CGATTATA
TG**aaaaaa**G**aatt**GCTAGTAATACGT**aatt**AGTATGTTACGGTGAATATT
CTAACTGTTTCGCACTAATCACTCATCACGCGTTGAAACATATTATTATC
TTATTATTTATATAATA**ttttttaa**TAAATA**ttaa**T**aatt**A**ttaatt**TAT
ATTTATTTATATCAGAAATAATATG**aattaa**TGCGAAGTTGAAATACAGT
TACCGTAGGGGAACCTGCGGTGGGCTTATAAATATC**ttaa**ATATTCTTAC
ATAAATA**ttaa**TCTAAATA**ttaa**TATAAATA**ttaa**TA**ttaa**TAGTTCCGG
GGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGAAATA**ttaa**TATAAAT
ATAAATA**ttaa**TATAAATATAAATATAAATATAAATATA**ttttaa**TATAA
TATAATATAATATATAATATATTATATAAATATAATATATAAATAATATA
AT**aaaa**TA**ttttaa**TATATATATAATATAATAT**aatt**ATTATTAT**aattt**
aaTAT**aatt**ATTATTAT**aatttaa**TATAATAAATAAATAAAT**aatt**AT**aa**
ttAT**aatt**AT**aatt**ATAATCTCAATATATAAATGAT**aatt**ATTATAAATA
CAAAGGAAAT**aattG**A**ttttaaaa**TATAT**ttaa**T**aaaa**TATATAATAT**a**
attATAC**ttttttt**GTTATTATATAAT**aatt**ATA**ttaa**TATAT**ttaa**TAG
aattaaACTCCTTCGGCCGGACTATTATTCA**tttt**ATATA**ttaa**TGATAA
ATCA**ttaatt**ATTA**ttaa**T**aatt**tATTTATAATAT**ttaattttA**TATATT
ATTATTTATAAT**aaaaaaatt**ATATTATAAC**aatttaattttaattttt**A
ttttaattATa**aattaa**T**aatt**TATTTG**tttaaa**T**aaattt**ATAACTCC
TTCGGGGTTCGGCCGGACTA**ttaa**TATAAATAAATAATAAATATTTATAA
T**aaaa**TAATATACATCTTC**tttaaa**T**aaaaaaa**GGGGACATTATAAATAG
TATATAAATATATTATATC**tttttt**ATTATTATTA**ttaa**TAAATAATAAT

AAT**aatt**TATATATTTATAATATAT**ttaa**TAGTTCCGGGGCCCGGCCACG
GGAGCCGGAACCCCGAAAGGAGAATGTATTAT**aatt**ATTACATAT**aatt**A
TTATTATTCACTTCTTA**ttaaaaa**TAATACTCTATAT**aatt**TATAT**aatt**
TA**ttttaa**TATATATATATTTATATATAATATAATATATATATTTATTTA
TTATAATCA**ttttttttttaa**C**ttaaaa**T**aaaa**CTTATTAT**aatt**TATAT**a**
attTAT**aatttttt**ATAT**aaaaa**T**aatt**ATAT**aatttttt**ATTTATTTATAT
AATAATAATATTATTTGTTATATATTATATATTATATATATAATAAATAA
ATAAATAATAAATAATAATAATAAGGATATAGT**ttaa**TGGT**aaaa**CAGTT
GATTTCAAATCAATCATTAGGAGTTCGAATCTCTTTATCCTTGATAATAA
TAAT**aaaaa**TATGTATTTAT**ttaatt**A**ttttaa**TATTTCTCCTTTCGGGG
TTCCGGCTCCCGTGGCCGGGCCCCGGAACTA**ttaa**TATAATATAATATAA
TATAAATATTCATTTATC**tttttttttaa**TATTC**ttaattaattaattaat**
taaTATA**ttaatt**AT**aaaaaa**TATATTAT**aattttt**ATTA**ttaa**TAAGTAT
AAATATATTA**ttaa**TAAT**aatt**TA**ttaaaaa**TATATTATTATAATATA**tt**
aaTATATCAT**aatt**ATAATCAATATTATATTAT**ttaattttt**ATAATAC**tt**
aattA**ttaa**TATATTATTCATATATATAT**aattaattaattaatt**ATATT
GAATATATAAATATATATATATATAAATATAT**aaaaatt**ATAT**aatt**A**tt**
ttaaGT**aaaaa**TAATA**ttaa**T**aaaatt**ATACAATAATAATAATAAATATT
CATTATTAT**ttaattaa**TATCTCCTTTACTTC**ttttt**CCTCCGTTGAGGA
CTTATTA**ttaa**GTATATTATTATATACTAC**ttaa**GATTATATATATAATA
TATATATATATATTATATAT**aaaa**TATAAATATATAAATAATAT**aaaatt**
aaT**aaaa**TAAAT**aaaa**T**aatt**AGTCCGATCGAATCCCCTAT**ttaattaat**
taattaattaaGAAAGAGAT**aatt**tATAT**aaaa**TATTATTTAT**aattaat**
tAT**aattaatt**ATAATATAATATAATATAAATAATAATATAAT**aaaaa**T**a**
aaaaT**aaaa**TAATATTAGATTATATTATATAT**aatt**TATAT**aatttttttaa**T
AATAATAATAAATAAGTTTATTTAT**aatt**ATAAATATAAATATAAATATA
AATAAAGAAGGTATTATA**tttt**AT**aaaa**TATAATAATAATAC**aaatttA**T
A**tttaa**TAAATA**ttaa**TATAAG**tttaaa**GTTCCGGGGCCCGGCACGGGA
GCCGGAACCCCGAAAGGAGAAATAAATAATATATTTAT**aaaaattaa**ATA
AATAAATATTATCTA**tttaaaaa**TAAATATAATATAATATAATATAAT**aa**
ttCTAAATATAAATAATATTTATTAT**aatt**ATTATAAT**aatt**GTATTATT
TA**ttaa**TAATATATAT**aatt**ATA**ttaaaa**CTAATATTACATTA**tttt**GTA
TA**tttaaa**C**aattaatt**GATTATTCTTATTTGTAATCTTTATTTA**tttt**A
TTATATCTTA**ttaa**TGAT**aatt**AT**aatt**ATTA**ttaaaa**TAAT**aatt**TACT
TC**tttt**GATAT**aaaaa**T**aaaa**TAATATAGTTCCGGGGCCCGGCCACGGGA
GCCGGAACCCCGGAAGGAGATAAATATATTATA**ttttt**ATTCCTACCTA**t**
taaAGGTAAAGACTCGATTCTCAT**aattaatt**tATATCCTTCGGCCGGA**t**

taattTA**tttt**ATTTATATTTATATTTATAGTGAATACC**ttttttaa**TAT
TTA**ttttttaa**TATTTA**ttttttaa**TA**tttt**A**ttttttaa**T**aaaa**TATAATCT
TGTAAGTAAG**aaaa**G**aatt**TCGGTGATTGGAACCTTGAAAGGAT**aatt**tC
TTATTTATTATAATATTTATA**ttaa**TAGTTCCGGGGCCCGGCCACGGGAG
CCGGAACCCCGAAAGGAGTATTA**ttaa**ACAT**ttaa**TATATTATA**ttaa**TA
T**ttaatttaaa**TGA**ttaa**TATATTATTATAATAATATTTA**tttt**ATA**tta
aaa**TATTAT**aattaa**TATATATATATTTA**ttttaa**TAATATTATTATTAT
TATTA**ttaaatt**ATTA**ttttt**ATAAATATATATATATATATATATATT
A**ttttt**ATTCTTATAT**aatt**ATAT**aaaaaaaa**TATATATAATATAT**aatt
aattaa**TATATATTAT**ttaatt**ATATATTA**tttaaaa**TAC**ttttt**ATATT
ATATCTTCT**ttaattaaaa**TAT**aatt**ATTATTTATATTAT**aatt**ATTTAT
GAAATATTATTA**ttaaaa**T**aaaaaa**GAGGTTTAGACTATATATTTATTAT
TTATAAACTTATTATATTATTTATTA**ttaa**TAGTTCCGGGGCCCGGCCAC
GGGAGCCGGAACCCCGAAAGGAGAAATAAAT**aaaa**T**aaaaaa**TAATAAAT
A**ttaa**TATTA**ttaa**ATATTATTTATAATAAATA**ttaa**TATTA**ttaa**ATAT
TATTCATA**ttaa**T**aatttt**ATTATTATTTGTAATATA**ttaa**ATA**ttaa**TA
ATATATATATTATTTATTATAATG**aaaa**CCTATCCTATATTATCCTATCA
TATAATATCATATCATATTATATTATATCTTATTATATGATATATAAAGT
ATTCACTCTATATGAGGTTATGATTATTATATAAATCTTA**tttt**A**ttttt**
A**ttttt**ATTTGGACTAATAAT**aatt**ATAATAAT**aatt**ATTGATATGTTCT
AATA**ttaa**TAAATACATATTTATATTATAATATAAATATTCATTTCTTAC
T**aattaa**T**aaaaa**G**ttttt**ATATTCATTATAATATAAATATATAAATATA
TATAAATA**ttttaa**T**aatt**AT**aatt**ATA**ttaa**GATATTATAAATATATAT
TTA**tttttttt**AT**aaaa**TAAATAAATAAATAAAT**aattaa**TA**ttttt**AT
ATTATAACTTA**ttttt**ATAATAATAATAAGTA**tttt**A**tttttt**ATTATAT
TATTATTTATAT**aatt**ATATATATA**ttaatt**TC**aatttaattaattaatt
aatt**GGTATTTGGCATATAATATC**aattaatt**GT**aatt**CTTATAAG**aatt
aattaattaa**TATGC**ttttt**ATAT**aatt**TATAC**tttt**ATATTTCTCCTTC
CGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTATTATTATTA**ttttt**
ATTTATTTATTA**ttaaaa**TATAATAATAAATAGTCCGGCCCGCCCCGCGG
GGCGGACGCCGGAGGAG**aatt**ATA**ttttt**ATATAAT**aatt**TATATTTCTA
TATATATATATATATATTATATATAAATATTATTATATATA**ttttt**ATAT
ATATTAT**aatt**ATATTCA**ttaa**TA**tttt**ATTATAGTGGTGGGGTCCC**aat
t**ATTA**tttt**CAATAAT**aatt**TATCATGGGACCCGGATATCTTCTTG**tttt
t**ATTTATTA**tttt**A**ttaatt**tA**ttttaatt**ATTTATTTAT**aatt**TATATT
ATAC**aatt**TATTATTTCG**ttaa**TACCTTTATTTATATTATATAATATATT
ATATTATTATAATATATTTATTGATTATA**ttaa**TACAT**ttaa**CTAATGTG

TGCTCTATATTTATTGAATAGTTTGGTTCTTATCACCCACCCCCTCCCCC
TATTACGTCTCCGAGGTCCCGGTTTCGTAAGAAACCGGGACTTATATATt
taaTACTaaaaaTATAACTACATTACttttttaaTATATATAACAATATA
TATATATATATATATAttaattATATaaaaTATAATACTCTATAttaaATAT
TAttttATCAATATTTATTTATATATATAATAATAATAATAATAATCAA
TAttaattATTTATATATATAAGAttaaTATTATttaaTATATTATGAAT
aatttaattaaTAAATCtttaaaTATTATCATaaaaaTATaattaaATaa
ttTCTTATTTATAATAAAGAATAATAATATATATAAATATAATAAAGAAT
GTAAATAATATATATATAATATAATATAATATaaaaaaTATATATATATA
TAAATATATATATAATATATAGATAATAATAttttTATATaattTAtttt
ATTAttaaGTAATAAATAATaaaaaaaTCAATATAttaaATAATATATTT
ATATTAGTTCGGTTTAGTTGGTAttttGTAATGAGTaaaaaGTAATATAT
AATAttaaATAATAAGTATTGATATAAGTAATAGATATAATAATAATATT
AttaaTAttttATATAAATAATAttaaTAATATAGATTATGAAAGAGAGT
AttaaTATCAttaaATATATATATATGTTATATaatttaaaTGAttttaa
TATATATATATATATTATATTATAGATTATGATACATTTATATAAATAAT
ATATATATaaaattaattATACTATTACTTTATAATATAATAATATTTAT
TTATAAAGATATaaaaGaattGtttaaaGTTATAACTaaaaTATTATATA
GTATTCAttaaTaattaaTATTATaattCAACTATTGTTATATTTATAAA
TAGAATAATATATTATTATCCTttaaGATATAACAATaattATttaatta
attaattaatttaattaatttttttttttaaTGAATATAATAATAATAAT
ATTAttaaattaaTATATaaaaaaaaaGTaaaaaTGGTACAAAGATGATT
ATATTCAACAAATGCaaaaGATATTGCAGTATTATAttttATGTTAGCTA
tttttAGTGGTATGGCAGGAACAGCAATGTCTttaaTCATTAGATTAGaa
ttAGCTGCACCTGGTTCACAATATTTACATGGTaattCACaattATttaa
TGGTGCGCCTCTCAGTGCGTATATTTCGTTGATGCGTCTAGCATTAGTAT
TATGAATCATCAATAGATACttaaaaCATATGACTAACTCAGTAGGGGCT
AACTTTACGGGGACAATAGCATGTCATaaaaCACCTATGATTAGTGTAGG
TGGAGttaaGTGTTACATGGTTAGGttaaCGAACTTCTTACAAGTCTTTA
TCAGGATTACaattTCCTCTTATCATTTGGATATAGTaaaaCAAGTTTGA
TTAttttACGTTGAGGTAATCAGATTATGATTCATTGttttAGATAGCAC
AGGCAGTGTGaaaaaGATGAAGGACCTAAATAACACaaaaGGAAATACGa
aaaGTGAGGGATCAACTGAAAGAGGAAACTCTGGAGTTGACAGAGGTATA
GTAGTACCGAATACTCAAATaaaaaTGAGAttttaaaTCAAGTTAGATA
CTATTCAGTAAATAATaatttaaaaaTAGGGAAGGATACCAATATTGAGT
TATCaaaaGATACAAGTACTTCGGACTTGTTAGaattTGAGaattAGTAA
TAGATAATATAAATGAGGaaaaTATAAATAATaattTAttaaGTATTATa

aaaaaCGTAGATATA**ttaa**TATTAGCATATAATAG**aattaa**GAGTAAACC
TGGTAATATAACTCCAGGTACAACATTAGAAACATTAGATGGTATAAATA
TAATATA**tttaaa**T**aatt**ATCAAATG**aatt**AGGAACAGGT**aattC****aattt**
aaaCCCATGAGAATAG**ttaa**TATTCCTAAACCTAAAGGTGGTATAAGACC
T**ttaa**GTGTAGGTAATCCAAGAGATa**aatt**GTACAAGAAGTTATAAGAAT
aattttAGATAC**aatttttG**AT**aaaaa**GATATCAACACATTCACATGG**tt**
ttAGAAAGAATATAAGTTGTCAAACAGC**aatt**TGAGAAGTTAGAAATATA
TTTGGTGGAAGT**aatt**GATTTATTGAAGTAGAC**ttaaaaaaa**TG**tttt**GA
TAC**aatt**TCTCATGAT**ttaatt**A**ttaa**AG**aattaaaaa**GATATATTTCAG
ATAAAGG**tttt**ATTGATTTAGTATAT**aatt**A**ttaa**GAGCTGGTTATATTG
ATGAGAAAGGAACTTATCATAAACCTATATTAGGTTTACCTCAAGGATCA
ttaattAGTCCTATCTTATGTAATATTGTAATAACATTGGTAGAT**aattG**
ATTAGAAGATTATA**ttaatt**TATATAATAAAGGTAAAG**ttaaaaaa**CAAC
ATCCTACATAT**aaaaatt**ATCAAGAAT**aattG**C**aaaa**GCT**aaaa**TA**tttt**
CGACAAGA**ttaaatt**ACATAAAGAAAGAGCTAAAGGCCCACTATTTATTT
ATAATGATCCT**aatt**TCAAGAGAAT**aaaa**TACGTTAGATATGCAGATGAT
A**ttttaatt**GGGGTATTAGGTTC**aaaaaa**TGATTGT**aaaa**TAATC**aaaa**G
AGA**tttaaa**C**aatttttttaatt**CATTAGGT**ttaa**CTATAAATGAAG**aaaa**
aaCT**ttaatt**ACTTGTGCAACTGAACTACCAGCAAGA**ttttt**AGGTTATA
ATATTTC**aatt**ACACC**tttaaaaa**GAATACCTACAGTTACTAAACT**aatt**
AGAGGTAAACTTATTAGAAGTAGAAATACAACTAGACCTATTA**ttaa**TGC
ACC**aatt**AGAGATATTATCAAT**aatt**AGCTACTAATGGATATTGTAAGCA
TAAT**aaaaa**TGGTAGAATAGGAGTGCCTACAAGAGTAGGTAGATGACTAT
ATGAAGAACCTAGAAC**aatt**A**ttaa**T**aatt**ATAAAGCGTTAGGTAGAGGT
ATC**ttaatt**ATTAT**aatt**AGCTACT**aatt**AT**aaaa**GA**ttaa**GAGAAAGAA
TCTATTACGTATTATATTATTCATGTGTA**ttaa**CTTTAGCTAGTAAATAT
AGA**ttaaaaa**CAATAAGT**aaaa**CTA**ttaaaaatttG**GTTAT**aatttaaa**T
ATTATTG**aaaa**TGAT**aattaatt**GCC**aattttC**CAAGAAATAC**tttt**GAT
AATATC**aaaaaattGaaaa**TCATGGTATATTTATATATATATCAGAAGCT
AAAGTAACTGATCC**tttt**GAATATATCGATTC**aattaa**ATATATATTACC
TACAGCTAAAGCT**aattttaa**TAAACCTTGTAGTATTTGT**aatt**CAACTA
TTGATGTAGAAATACATCATG**ttaa**AC**aatt**ACATAGAGGTATA**ttaaaa**
GCAC**ttaa**AGATTATATTCTAGGTAGAAT**aatt**ACCATAAACAG**aaaaCa**
attCCATTATGTAAACAATGTCATA**ttaaaa**CACAT**aaaaa**T**aatttaaa**
aaTATAGGACCTGGTATAT**aaaa**TCTATTA**ttaa**TGATACTCAATATGGA
AAGCCGTATGATGGGAAACTATCACGTACGGTTTGGGAAAGGCTCT**ttaa**
CACGTGGCAACATAGG**ttaatt**TGCTATTTCA**ttttt**AGTAGTTGGTCAT

GCTGTA**ttaa**TGA**tttt**CTGTGCGCCGTTTCGC**ttaatt**TATCACTGTAT
TGAAGTG**ttaatt**GATAAACATATCTCTGTTTATTC**aattaa**TG**aaaa**CT
TTACCGTATCA**tttt**GGTTCTGATTATTAGTAGTAACATACATAGTATTT
AGATACGTAAACCATATGGCTTACCCAGTTGGGGCCAACTCAACGGGGAC
AATAGCATGCCAT**aaaa**GCGCTGGAGT**aaaa**CAGCCAGCGCAAGGTAAGA
ACTGTCCGATGGCTAGG**ttaa**CG**aatt**CCTGTAAAGAATGTTTAGGGTTC
TCA**ttaa**CTCCTTCCCACTTGGGGATTGTGATTCATGCTTATGTATTGGA
AGAAGAGGTACACGAG**ttaa**CC**aaaaa**TGAATCATTAGCT**ttaa**GT**aaaa**
GTTGACATTTGGAGGGCTGTACGAGTTCAAATGGa**aattaa**GAAATACGG
GATTGTCCGAAAGGGGAAACCCTGGGGATAACGGAGTCTTCATAGTACCC
aatttaatttaaaTAAAGTGAGATACTTTAGTACTTTATCT**aattaa**ATG
CAAGGAAGGAAGACAGTTTAGCGTAT**ttaa**CAAAGA**ttaa**TACTACGGA**t**
tttCCGAG**ttaa**AT**aattaa**TAG**aaaa**TAATCATAATAAACTTGAAACC
A**ttaa**TACTAG**aattttaaattaa**TGTCAGATATTAGAATGTTA**ttaatt**
GCTTATAATa**aattaaaa**GTAAGAAAGGTAATATATCTAAAGGTTCTAAT
AATATTACCTTAGATGGGA**ttaa**TATTTCATA**tttaaa**T**aatt**ATCTAAA
GATA**ttaa**CACTAATATGT**ttaattt**CTCCGGTTAGAAGAGTTG**aatt**C
CT**aaaa**CATCTGGAGGATTTAGACCT**ttaa**GTGTTGGAAATCCTAGAG**aa**
aaattGTACAAGAAAGTATGAGAATAATATTAG**aatt**ATCTATAATAATA
GTTTCTCTTATTATTCTCATGGATTTAGACCTAACTTATCTTGT**ttaa**CA
GCTATTATTCAATGT**aaaatt**ATATGCAATACTGT**aatt**GATTTA**ttaa**A
GTAGA**tttaaa**TAAATGCTTTGATAC**aatt**CCACATAATATG**ttaattaa**
TGTA**ttaa**ATGAGAGAATCAAAGATAAAGGTTTCATAGACTTATTATAT**a**
attA**ttaa**GAGCTGGATATGTTGAT**aaaaa**TAAT**aatt**ATCATAATACAA
CTTTAGG**aatt**CCTCAAGGTAGTGTTGTCAGTCCTA**tttt**ATGTAATA**tt**
tttttAGAT**aatt**AGATAAATATTTAG**aaaa**T**aatt**tGAGAATG**aatt**CA
ATACTGGAAATATGTCTAATAGAGGTAGAAATCC**aatt**TATAATAGTTTA
TCATCT**aaattt**ATAGATGT**aatt**ATTATCTG**aaaattaaatt**GATTAGA
ttaaGAGACCATTACCAAAGAAATATGGGATCTGAT**aaaa**G**ttttaaaa**G
AGCTTA**tttt**GTTAGATATGCTGATGATATTATCATTGGTGTAATGGGTT
CTCATAATGATTGT**aaaaa**TA**ttttaaa**CGATA**ttaa**TAACTTC**ttaaaa**
G**aaattt**AGGTATGTC**aattaa**TATAGATAAATCCGTTA**ttaa**ACATTCT
AAAGAAGGAGTTAG**tttttt**AGGGTATGATGT**aaaa**GTTACACCTTGAG**a**
aaaaaGACCTTATAGAATGA**ttaaaaaa**GGTGAT**aattttt**ATTAGGGTTA
GACATCATACTAGTTTAGTTG**ttaa**TGCCCCTATTAGAAGTATTGTAATa
aattaaATAAACATGGCTATTGTTCTCATGGTA**tttt**AGG**aaaa**CCCAGA
GGGGTTGGAAGA**ttaatt**CATGAAGAAATG**aaaa**CCA**ttttaa**TGCATTA

CTTAGCTGTTGGTAGAGGTATTATAAACTATTATAGATTAGCTACC**aatt**
ttACCACA**ttaa**GAGGTAG**aatt**ACATACA**ttttattttt**ATTCATGTTGT
ttaaCATTAGCAAG**aaatttaattaa**ATACTG**ttaa**GAAAGTTA**ttttaa**
attCGGTAAAGTATTAGTTGATCCTCATTC**aaaa**GTTAG**tttt**AGTATTG
ATGA**ttttaaatt**AGACAT**aaaa**TAAATATAACTGATTCT**aatt**ATACAC
CTGATG**aatttt**AGATAGATATAAATATATGTTACCTAGATCTTTATCAT
TATTTAGTGGTATTTGTC**aatt**tGTGGTTCTAAACATGATTTAGAAGTAC
ATCACGTAAGAACA**ttaa**ATAATGCTGCCAATa**aattaa**AGATGATTATT
TATTAGGTAGAATGA**ttaa**GATAAATAG**aaaa**C**aatt**ACTATCTGT**aaaa**
CATGTCA**ttttaaa**GTTCATCAAGGTAAATATAATGGTCCAGGTTTATAA
T**aatt**ATTATACTA**ttaa**ATATGCG**ttaa**ATGGAGAGCCGTATGATATGA
AAGTATCACGTACGGTTCGGAGAGGGCTC**tttt**ATATGAATGTTATTACA
TTCAGATAGGTTTGCTACTCTACTCTTAGTAATGCCTGCT**ttaatt**GGAG
G**tttt**GGTAACC**aaaaaa**GATATGAAAGTAATAATAATAATAATCAAGTA
ATAG**aaaa**TAAAGAATAT**aatttaaattaatt**ATGATAAGTTGGGACCTT
ATTTAGCTGGA**ttaatt**GAAGGTGATGGAACTATTCTAGTTCa**aatt**CAT
CTTCAAT**aaaaaaa**TCTAAATATAGACCG**ttaatt**GTTGTAGTAT**ttaat**
tAGAAGATTTAG**aatt**AGCT**aatt**ATTTATGT**aatttaa**CTAAATGTGG**a**
aaaGTGTAT**aaaaaattaa**TCGT**aatt**ATGTATTATGACTTATTCATGA**t**
ttaaaaGGTGTATATACATTA**ttaa**ATATTA**ttaa**TGGATATATGAGAAC
ACCTAAATATGAAGCATTTGTTAGAGGTGCTG**aatt**TATAAAT**aatt**ATA
ttaattCAACAAC**aatt**CTACATAAT**aattaaaaaa**TATAGATAATA**tta**
aattaaACCATTAGATACATCAGATATTGGTTCAAACGCTTGATTAGCTG
GTATGACAGATGCAGATGGT**aatttttt**CTA**ttaatttaa**TAAATGGT**aaa**
aaTCGTTCTAGTAGAGCAATGCCTTATTATTGTTTAG**aattaa**GACa**aat**
tATC**aaaaaatt**CTAATAATAATAATA**ttaattttt**CTTA**ttttt**ATATT
ATGTCTGC**aatt**GCACTATA**ttttaa**TG**ttaatt**TATATAGTAGAGAACG
T**aatttaatt**tATTAGTATCTC**ttaa**TAATACGTATAAACTATATTATAG
TTATAAAGTAATAGTGGCTAATCTATAT**aaaaa**TA**ttaa**AGTAATAGAAT
ACT**ttaa**TAAATATTCTTTATTATCATCTAAACACTTAGA**tttttt**AGAT
TGATCT**aatt**AGTTA**ttttaattaa**TAATGAGGGTCAAAGTAT**aaaa**C**tt**
aaTGGTAGTTGAG**aatt**AGGTAT**aatt**tACGTAAAGATTATAAT**aaaa**CT
AGAACTACGTTTACTTGATCTCA**tttaaaaaa**TACATATTTAG**aaaa**TAA
ATAAAT**aatt**ATTATTACTTTCTTCCCCTCCGAATCCGTAATATATTTAC
GGATATATAATCTCGTAGTGT**aaaa**GGTGTAACGAGATTA**ttaa**TAAGTT
GCCGTAATATATTGT**aaaa**TATATTATTATTACAACACTATATGCGGG**aa**
aaCCCTAAAGTCATAATATAATATTATCCCCACGAGGGCCACACATGTGT

GGCCCTCGCGGGGTATGGT**aatttaattaa**GTTATAAATGTACTATAGTA
ttaaaattATTATGAAT**aatt**TCCCCACCCCCATGCGAAGCATGGGGGGG
GGTATAAGTATGGACAATCCGCAGGAAACCAAATAAT**aattaa**TATCCTG
AAACAAAGTAAGTGAAGGAGATATC**ttaaaa**TATATATAATATATA**tttt**
AT**aatt**ATTATGTAGGATCCTCAGAGACTACACGTGTTGCACCCATTATA
TTATGTATAATGGGTTGAAGATATAGTCCAAATAT**aatt**GAAAGATTATA
AT**aaaa**TGAACTATTTATTACCA**ttaa**T**aatt**GGAGCTACAGATACAGCA
TTTCCAAG**aattaa**TAACATTGC**ttttt**GAGTATTACCTATGGGGTTAGT
ATGTTTAGTTACATCAACTTTAGTAGAATCAGGTGCTGGTACAGGGTGAA
CTGTCTATCCACCATTATCATCTATTCAGGCACATTCAGGACCTAGTGTA
GATTTAGC**aatttttt**GCATTACAT**ttaa**CATC**aatt**TCATCATTATTAGG
TGCTA**ttaatt**TCATTGTAACAACA**ttaa**ATATGAGAACAAATGGTATGA
CAATGCAT**aatt**ACCATTATTTGTATGATC**aattttt**CATTACAGCGTTCT
TATTATTATTATCATTACCTGTATTATCTGCTGGTATTACAATGTTATTA
TTAGATAGAAACTTCAATACTTCATTCTTTGAAGTATCAGGAGGTGGTGA
CCCAATCTTATACGAGCATTTA**tttt**GATTCTTTGGTCAAACAGTGGCCC
TTATTATTATA**ttaa**TAATATATAATGATATGCA**ttttt**CTAAATGCTGG
aattA**ttaaaaaaa**TG**aatt**ACAAATATTATAAGTCTATTA**tttaaa**GCC
TTATTTGT**aaaaa**TATTCATATCTTATAATAATCAGCAGGATAAGATAAT
AAATAATCTTATA**ttaaaaaaa**GATAATA**ttaaaa**GATCCTCAGAGACTA
CAAG**aaaaa**TA**ttaa**AT**aatt**CAATAAAT**aaaaatttaa**TCAATGATTAG
CTGGA**ttaatt**GATGGTGATGGATA**tttt**GGTATTGTAAGTAAGAAATAT
GTATCATTAG**aatt**CTAGTAGCATTAGAAGATGAAATAGC**tttaaaa**G**aa**
ttC**aaaa**T**aatt**tGGTGGTTCTA**ttaattaa**GATCAGGTGT**aaaa**GCTAT
TAGATATAGATTAC**ttaa**T**aaaa**CTGGTAT**aattaattaattaa**TGCAG**t**
taaTGGTAATATTAGAAATACT**aaaa**GATTAGTAC**aatttaa**TAAAGTTT
GTA**tttt**ATTAGGTATTGA**tttt**ATTTATCC**aattaattaa**CTAAAGATA
ATAGTTGATTTGTTGGA**tttttt**GATGCTGATGGTAC**aattaatt**ATTCA
tttaaaaaTAATCATCCTC**aattaa**C**aatt**TCTGTAACTAATAAATATTT
ACAAGATGTACAAGAATAT**aaaaa**TA**tttt**AGGTGGTAATATTTA**tttt**G
ATAAATCAC**aaaa**TGGTTATTATAAATGATCCATTCAATC**aaaa**GATATA
GTA**ttaattttt**A**ttaa**TGATTATA**ttaaaa**TAAATCCATCAAGAACACT**a**
aaaaTAAAT**aatt**ATAT**ttaa**GTAAAG**aattttA**T**aatttaaaa**G**aatta**
aaaGCTTATAATAAATCTTCTGATTCAATACAATATAAAGCATGA**ttaat**
tttG**aaaa**TAAATG**aaaaaa**TAAAT**aatt**AT**ttaa**TAAAGATATAGTCC**a**
attATATATATATAATATATATATATATAACAAGCACCCTGAAGTATATA
ttttaattATTCCTGGATTTGGTATTATTTCACATGTAGTATCAACATAT

TCT**aaaaaa**CCTGTATTTGGTG**aatt**tCAATGGTATATGCTATGGCTTC**a**
attGGATTATTAGGATTCTTAGTATGATCACATCATATGTATATTGTAGG
ATTAGATGCAGATCTTAGAGCATATTTCCTATCTGCACTAATGATTATTG
C**aatt**CCAACAGG**aattaaattttt**CTCATGA**ttaa**TAAATCCCTTTAGCA
AGGAT**aaaaa**T**aaaaa**T**aaaaa**T**aaaaa**GTTGATCAG**aatt**ATC**aaaaaa**
TAAATAATAATAATATAAT**aaaaa**CATA**tttaaa**TAATAATAATAT**aatt**
ATAATAAATATATATAAAGGT**aatt**TATATGATATTTATCCAAGATCAAA
TAG**aatt**ATATTCAACCAAATAATA**ttaa**TAAAG**aatt**AGTAGTATATGG
TTAT**aatt**TAGAATCTTGTGTTGGTATACCTCTATATACTAATATTGT**aa**
aaCATATAGTAGGTATTCCTAATAATA**tttt**ATATATTATAACAGGTA**tt**
ttA**ttaa**CAGATGGTTG**aatt**GATTATCTATCT**aaaaaa**GATTTAGAT**aa**
aaaaaC**aatt**ATAG**aattaatt**GTAGATTTAGA**ttaaaa**CAATCAAT**aat**
tCATAGTGAATAT**ttaa**TATATGTATTTATATTATTATCACATTATTGTA
TAAGTTATCCT**aaaa**T**aaaatt**GCTAAAG**ttaa**AGGTAAATCATATAATC
aattAG**aatttt**ATACTAGATCATTACCATG**tttt**ACTA**ttttaa**GATAT
ATA**tttt**ATAATGGTAGAGT**aaaatt**GTACCTAAT**aatt**TATATGATTTA
ttaattATGAATCTTTAGCTCATAT**aatt**ATATGTGATGGTTCATTTGT**a**
aaaGGTGGAGGTTTATAT**ttaatt**tACAATC**tttt**CTAACTAAAG**aatta**
atttttATTATAAATA**ttttaaaattaatttaatttaatt**GTCTATTACA
TAAATCTAGAAATAAATATCTTATTTATATAAGAGTAGAATCTG**ttaaaa**
GATTATTTCCTAT**aatt**TATAAATATA**tttt**ACCTTCTATAAGATAT**aat**
ttGATATTATATTATGAC**aaaaaaaa**TATAATATGA**ttaattaattaatt**
aattaattaattTATTTATTATTTAC**tttttt**GATATATATAGAGGCAAA
CTCGAGG**aaaa**CCATAT**aatt**AGAATAAGTAAT**aatt**ATATGACAACCGT
CGAACTAAATCATATTCAAG**aattaa**TATGT**aaaa**GCGTAGAGATTAGAC
GCCTCTGGTTATCTAAGTAATATATATATATATATTATATGATAACATAA
GGTATAATCCAATGAGATCAGTAATGA**ttttaaaa**CAAT**aattttG**t**ttt**
aaGTA**ttaa**TAATAATA**ttaa**TATTCGACCTC**ttaatt**GAGGATATTATA
ATCAT**aatttttt**ATATTATAATAT**aaatttaa**CTAGCTAGATAATATTA
TAT**aaaaaaaaaaaa**TAATATTATAT**aattaattaaaa**T**aatttttA**t**ta**
attGAAACTGAAATG**ttttaaa**G**ttaa**AT**aaaa**GAGCTCTAATCCATGGT
GGTTC**aatt**AGATTAGCACTACCTATGTTATATGC**aatt**GCATTCTTATT
CTTATTCACAATGGGTGGT**ttaa**CTGGTGTTGCCTTAGCTAACGCCTCAT
TAGATGTAGCATTCCACGATA**ttaatttaa**TAAGTGTCGTGC**ttaaatt**C
ACT**aaaa**TAATATATAAT**aatt**ATAATAAATATAT**aaaaaaaa**T**aaaaaa**
aaT**aaaaaaaattaa**TATCTTATGA**ttaatttt**ATATAAAT**aaaatttA**t
taaATATTATTGGTTATATATATATATATATA**ttaa**TAAT**aaaaaaa**TATAT

ATATATATATAGCTAACGGGGAAACTCTTAT**aatt**ATTATTTATATAATA
AATAAGACAATCCCGTGATAACT**ttaa**TATATATATATTATATA**ttaa**AG
TATTGTAGAGACTAAACGTGAATGA**ttttaa**TATTA**tttaaa**TA**ttaaat**
taaGAGATAGTCCAATCTTATATGTAAATATAAG**ttaa**TACC**aaaaaaaa**
aaTAATATTA**tttt**GACTTATTATATA**ttaa**TATTA**ttaa**TAAT**aatttt**
aaCTAATAATAAAG**ttttt**ATAGAAACTTTATATTATTAT**ttaa**TAT**tta**
attttC**aattaa**TATCTCC**tttt**GGGGTTCCGGTCCCTGGTCCGGCCCCC
GAAACTAAAGATA**ttaa**G**aatt**TATATGAATC**aatt**ATAAAT**aatt**ATA**t**
taaTA**ttttaaa**TAAATATCTTA**ttaa**TA**ttaa**TAAAGATAATA**ttaa**T**a**
attaaatttttAGAT**aatt**ATACTGAAGAAG**aaaaa**GGTTATTATTTATC
TGGATTATTTGAAGGAGATGGTAATATTTATACTAGATG**ttttt**C**aatt**A
C**ttttt**CTTTAGAAGATG**tttt**ATTAGCT**aatt**ATTTATGTCTTTA**tttt**
aaattGGTCATATTACAGCTAAATAT**aattttaa**TAAAG**aattaa**CAGCT
G**ttaa**ATGAAATATTAT**aaaaaaaaaa**GAACAAGAAGTATTTAT**aatt**AT
A**ttaa**TGGT**aatt**A**ttaa**CATAT**aaaa**GATATGATCAATA**ttttaaa**TAT
aattttaaTAATCG**tttaaa**TA**ttaatt**A**ttaaaa**CCTAAAG**aatt**TGAT
TTACTA**ttaa**ATCCTTGA**ttaa**CAGG**ttttaa**TGATGCTGATGGTTA**ttt**
ttATCTAGG**tttt**C**aaaaa**CAT**aaaaa**TAGTCAATGA**ttaaattt**CATTT
AG**aatt**ATCAC**aaaaa**GATAGTTATA**tttt**AGTCCGGCCCGCCCCCGCGG
GGCGGACCCCAAAGGAGATATTA**ttaaaaaa**TA**ttttaaa**CTTGGTGGTA
ttttaaaaaGAGATTATAAATCTGGTGCTACAGCTTATATTTATAAAGCT
CAATCATC**aaaa**GCTAT**aaaa**CC**tttt**ATTGAATA**ttttaa**T**aatt**ATCA
ACCA**ttaa**GTCTTAGAAGATATAAACAATATTTATTA**ttaa**ATATTGCTT
ACTTA**ttaaattaa**AT**aatt**ACATATATTAC**ttaatt**CTTTA**ttaa**TA**tt**
aaaaG**aattaa**TATTATTACAAAGTG**ttaaaaa**TATATCTTTAGAAAT**aa**
aaaaTG**aattaa**ATAATAGAG**ttaaatt**ATTA**ttaa**TAAACTTCATTATA
ACAATATCGAATAATGATAATA**ttaa**AGAGTa**aattC**t**taa**AGTG**ttaat**
taaATAATATTC**ttttttttttt**ATGACTTACTACGTGGTGGGACA**tttt**C
GTGCGGTCTGAAAGTTATCATAAATAATATTTACCATATAATAATGGAT**a**
attATA**ttttt**ATCAATATAAGTCT**aatt**ACAAGTGTA**ttaaaa**TGGTAA
CATAAATATGCTAAGCTGTAATGAC**aaaa**GTATCCATATTCTTGACAGTT
ATATTAT**aaaaaaa**GATGAAGGAACTTTGACTGATCTAATATGCTCAACG
AAAGTGAATCAAATGTTATa**aatt**ACTTACACCACT**aattG****aaaa**CCTGT
CTGATATTC**aatt**ATTATTTATTATTATAT**aatt**ATATAATAATAAAT**aa**
aaTGGTTGATGTTATGTATTGGAAATGAGCATACGATAAATCATATAACC
ATTAGTAATAT**aatt**TGAGAGCTAAGTTAGATATTTACGTATTTATGAT**a**
aaaCAGAATAAACCCTAT**aatt**ATTATTA**ttaa**TAAT**aaaaaa**TAATAAT

AATACCAATATATATATTAT**ttaatt**TATTATTATTATA**ttaa**T**aaattt**
aaTATATATTATAAAT**aatt**ATTGGA**ttaa**GAAATATAATA**tttt**ATAG**a**
attttCTTTATATTTAGAGGGT**aaaa**GATTGTAT**aaaaa**GCTAATGCCAT
ATTGTAATGATATGGATAAG**aatt**ATTATTCTAAAGATG**aaaa**TCTGCTA
ACTTATACTATAGGTGATATGCCTATCTTTATTTATATATATATTATTAT
TA**ttaa**TAAT**aaaaaaaaattaaaaaaaaa**GATAGGAGGTTTATATATAA
CTGATAAATATTTATTATATTA**ttttttttt**ATAATAAATA**ttaaaa**GAT
ATTGCGTGAGCCGTATGCGATGAAAGTCGCACGTACGGTTCTTACCGGGG
G**aaaa**CTTGTAAAGGTCTACCTATCGGGATACTATGTATTATCAATGGGT
GCTA**tttt**CTCTTTATTTGCAGGATACTATTATTGAAGTCCTC**aatttt**A
GG**tttaaa**CTATAATG**aaaatt**AGCTC**aatt**C**aatt**CTGA**ttaatttt**CA
TTGGGGCTAATGTTA**tttt**CTTCCCAATGCA**tttttt**AGGTA**ttaa**TGGT
ATGCCTAGAAG**aatt**CCTGATTATCCTGATGCTTTCGCAGGATG**aatt**AT
GTCGCTTCTATTGGTTCATTCATTGCACTATTATCATTATTCTTATTTAT
CTATA**tttt**ATATGATC**aatt**AG**ttaa**TGGA**ttaa**ACAATAAAG**ttaa**TA
ATAAATCAGTTATTTATAATAAAGCACCTGA**tttt**GTAGAATCTAATCTT
ATCT**ttaatttaaa**TACAG**ttaa**ATCTTCATCTATCG**aatt**CTTA**ttaa**C
TTCTCCACCAGCTGTACACTCAT**ttaa**TACACCAGCTGTACAATC**ttaa**G
TTAT**aaatttaatt**ATTTAC**ttaa**T**aattaaaaa**GTAAATATTATATCTA
AAC**ttaa**TAATATAATAATAATATTCTTAT**aaaaa**TATAT**aaaaaaaaa**T
ATAT**aaattt**A**ttaaaa**TATCTCCTTTCGGGAACTATAATATATTTATAT
AAATAAATACTAATATAATCCTATTATATATATATATATAT**aaaa**TAATA
TATATATAT**aattaa**TATAAATAATATTTATAAT**aattttttaa**TAATAT
ATAT**aatttaa**TATA**ttaa**TGAATATTATAT**aatt**A**ttaa**ATATATTATA
ATATTATTATTA**tttt**ATAAT**aaaaa**TA**tttttaa**TACT**aatt**ATTATTT
ATTATTTATAAATATATAAATAGTATGT**ttaa**TATTA**ttaa**TACT**aaaaa**
aaaTAT**aatt**AT**aatt**AGGATCTAACAATACATTTATCTGA**ttaa**TA**tta**
aTA**ttaa**TA**ttaa**TATTTATA**ttaa**TAAACGGA**ttaattaatt**GTATCC**a**
atttaattaattATAGATATATTATTTATAATA**ttaa**TATATTG**tttt**A**t**
taaaaaGGT**aaaaa**TAG**ttttt**A**tttt**ATATATAAATATAGGATATAAAT
AAATATATTATAGTGAACCCCGAAAGGAGAATATA**ttaa**GAATATATTTA
TA**tttt**ACATAT**aatt**ATTTATAATATAAATATCTCCGCAAAGCCGGA**tt**
aaTGT**aatt**AT**ttaa**T**aatttt**AT**ttaa**T**aatt**TA**ttaaaa**TAAATATTT
ACATTTGATAATATTTATATTATGTCAGTTA**tttt**ATA**ttaa**TGT**ttaa**T
CTATTATAATA**ttttttttt**ATAAATATATTATTTATTTATA**ttaatt**AT
ATATATATATTA**ttttt**ATAATATATATATA**ttttt**A**ttaa**ATATTTA**tt**
aaATATTTA**ttaatt**ATTATAATGTTGTTA**ttaa**TCTTA**ttaaaaaa**TAT

ATAT**aaaaa**TGCCAC**aatt**AGTTCCA**ttttAttttA**TGAATC**aattaa**CA
TATGGTTTCTTA**ttaa**TGATTCTATTA**ttaattttA**TTCTCAC**aattCtt
ttt**ACCTATGATC**ttaa**GATTATATGTATCTAGATTATTTATTTCT**aatt**
ATAATATATATTA**ttaatt**TATTTATTCATATAAATATTATTATTATATA
TAAATA**ttaa**TAATATTTATACTTAT**ttaa**TAATAAT**aaaaTaaaaaaTa
att**AT**aatttaa**TATAT**ttaa**TATATTTCCTTACGGACTATATATTTATA
TATATATA**ttaa**ATAC**aatttaatttaatttaatt**ATGTTATTTA**ttaa**A
TAAAGTTATATTATGATATAATAACAATATTATATATTATTATAT**aatt**A
TAATATA**ttttaa**TAT**aatt**ATC**aaaa**GAAATAAT**aaaaaaa**TA**ttaa**TA
AGAATAT**aatttaa**T**aatt**A**ttaaaaaaaatt**CTTATAGTCCGGCCCGCC
CCCCCCGCGGGGCGGACCCCAAAGGAGGAGTAAT**aaaatt**A**ttaa**ATACA
AATATTATATATATAT**aatt**CATTATATATATATATATATAAT**aattaa**T
CTTA**ttttttt**ATATATTTATTTATATATCTATTTATA**tttt**ATATATAT
TTATTTATATATCTAAGGGGTTCGGTCCCTCCCCCCGTAAGTATAATATA
CGGGGGTGGGTCCCTCACTATTTATA**ttttt**A**tttt**ATATA**tttt**ATATA
TTTATAAATAAAGTATAATAAGATAT**aatt**ATGA**ttaatt**ATTTATAAGT
TATAG**tttt**AT**aatt**tAT**aatt**ATTATGT**ttaatt**TA**ttaa**ATACATATA
TTACATCACCATTAGATC**aatt**TGAGATTAGACTATTATTTGGTTTACAA
TCATCATTTATTGAT**ttaa**GTTGT**ttaatttaa**CAACA**tttt**CATTATAT
ACTATTATTGTATTATTAGTTATTACAAGTTTATATCTA**ttaa**CTAATAA
TAATAATa**aatt**ATTGGTTCAAGATGA**ttaatt**TCACAAGAAGCTATTTA
TGATACTATTATAAATATGC**ttaa**AGGAC**aatt**GGAGGT**aaaatt**GAGGT
TTATATTTCCCTATGATCTTTACATTATTTATGTTTA**ttttt**ATTGCT**aa
tttaatt**AGTATGATTCCATACTCATTTGCATTATCAGCTCATTTAGTAT
TTATTATCTCT**ttaa**GTATTGTTATTTGATTAGGTAATACTA**tttt**AGGT
TTATATAAACATGGTTGAGTATTCTTCTCATTATTCGTACCTGCTGGTAC
ACCATTACCATTAGTACCTTTATTAGTTATTATTGAAACTTTATCTTATT
TCGCTAGAGCTATTTCATTAGGT**ttaa**GATTAGGTTCTAATATCTTAGCT
GGTCATTTA**ttaa**TGGTTA**tttt**AGCTGGTTTACTAT**ttaattttATG**tt
aattaatt**TATTTACTTTAGTATTCGG**tttt**GTACCTTTAGCTATGATCT
TAGCCATTATGATGTTAG**aatt**CGCTATTGGTATCATTCAGGGATATGTC
TGGGCTA**ttttaa**CAGCATCATA**tttaaaa**GATGCAGTATACTTACA**tta
att**AT**aaaa**Ta**aatt**AT**aaaa**T**aaaa**T**aatt**TACATATGGAGTA**ttaa**AC
TATAATAAATACAATATACCCCATCCCCCCC**ttttaa**TAATATTC**tttt**A
TCTAAT**aaaa**TATTTATTTA**ttaa**TATTATTATTATCTTCTTCAAGGACT
TAT**ttaa**TATAT**ttaa**TAACTTATTATACTTATTTATATTTAT**aattaa**T
ACAAATATATTA**ttaa**TCTTACTCCTTCGGAGTTCGGCCCCCCATAAGGG

GGGGACCTCACTCCTTCCCCACTGCACTGGATGCGGGGACTTA**ttttt**AT
TATTATTAT**ttaa**TCTTTATTTATa**aatt**ATATATTATATAT**aatt**ATTA
TAC**ttaa**T**aattaaaaaaaaa**CCTCT**aatt**ATTA**ttaa**TATTATATATAA
TATATATATTCTCA**ttaa**TGTTATATATAATATATATATTCTCA**ttaa**TA
TA**ttaa**TATAGTA**ttaaaaaaaa**T**aaaa**TAT**ttaa**TAAATATTATTA**tta**
aTAATATTTA**ttaaaaa**TAATATAACATAATAAATATAAGATTATTATAT
AATATATTTATTATATCATATAGTTCCGGGGCCCGGCCACGGGAGCCGGA
ACCCCGGAAGGAG**aatt**ATAACATA**ttttttaa**TAATATTCATATTTA**tt**
ttATATACAAATAAATATATTTATTTAGAATAAT**aaaaaaaaa**TAATAAA
TAAATATATTATTATCATTATTATACTTTATTCATTATTTATTATAAT**aa**
ttATATATAAC**aatt**ATAATATAT**aatt**ATA**tttt**ATATAATATTATA**tt**
aaTAT**ttaa**TATATTTATTATTATTATTACTTCTATGGAAACTTTATA**tt**
ttAGATA**ttttt**ATTATTATTA**ttaatt**TATAATGTTATA**ttttt**GATTT
ATAAATATATAAGTCCCGGTTTCTTACGAAACCGGGACCTCGGAGACGTA
ATAGGGGGAGGGGGTGGGTGATAATAACCAGAATATTCAATAAATACAGA
GCACACATTAGAT**aatttt**ATAATATAACCAATAT**aaaa**Ta**attaaaa**T
aattaaTATATATATATAAATATAAT**aatt**ATTATATATAAATATATAT**a**
atttttATAATAAATATTATAATATTATATAAATAAAT**aatt**ATAATATA
TAATAAATATATAATAATAAT**aaaaa**TA**ttaa**CAATATAAT**aaaattt**AT
AATATAAATAT**aatt**ATAAATAAG**ttaattaa**T**aaaa**TAATAAATGA**tta**
aCAAGAAGATATCTGGGGTCCCA**ttaa**T**aatt**ATTA**tttt**CAATAAT**aat**
tGGGACCCCCCACCATTATAATATCATA**ttaattaa**TATAATAATAATGT
ATAT**aaaa**TAGAAATAAT**aattaa**TATAATAATAATAATATATAT**aaaa**T
AGAAATAAT**aattaa**ATATATATATAAAT**aatt**ATTTATATAATATATTA
TAAATAATAATAATAATAAATATTTA**ttaattaa**TAATGATTATAAATA**t**
tttAT**ttaa**TAT**aatt**tATAACTA**tttt**ATTATATATATA**tttttt**ATTC
AT**aaaatt**CC**tttt**GAGGA**ttttt**A**tttt**ATATAAATATCTTCTAATATT
TATAATAAATAATAATATATTCATTATATTTAT**aatt**ATATATAATGTAA
TACGGGTAAACATTACCCGTTGTTCACGGGTAATGTTTACCCTA**tttt**AT
AT**aatt**C**ttaa**TAAATATATTTATA**ttttt**ATAT**aaaaaaatt**ATAAT**aa**
ttTA**ttaatt**CTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGAACTCCGGA
ACTAT**aaaaa**T**aattttaa**TAT**aatt**TATATA**tttt**ATGA**ttaa**TATAAT
ATATTA**ttaa**TGTAACTCCTTCGGGATTTGGTCCCCCTCGTAAGTATATA
GTATATAGTATATAGTATACGGGGGGTCCCTCACTCCTTCGGGGTTCGGT
CCTCCCTTACGGGTACGGATACGGATACGAATATGGGGAGTCCCTCACTC
CTTATCACTACGCTGAAGGTGG**aatt**TA**tttt**ATATTATTA**ttaa**ATCTT
TATTTAT**ttaatt**ATATAT**ttaa**TATATATATTATTATAAT**aaaa**CACCT

aattATTA**ttaa**TGTTATAT**ttaa**TATAATATATATATTC**ttaaaattt**A
TATAATATAAATAAAT**aaaaaaaaaa**GAAAGTACAT**aattaa**TATTATTA
TAAATAATATTA**ttaaaaa**GAATATAATAT**aattaa**TAGAAAGACG**tttt**
aaaaaT**aaaaa**T**aaaaa**T**aaaaa**T**aaaaa**T**aaaaa**T**aaaaa**T**aaaaa**T**aa
aaaT**aaaa**GAG**tttt**GGTTTACATATCAAGACCC**aattC****aatt**GAAACTA
TTTATTTA**ttaa**TCTCCTCCCCTCCCCCTCACTATTATTATAAGTAC**aat**
tAGGGCGCCAACCCCGCAGTGTTATTTACTGGGAAATGTTTATCCC**aatt**
aaTATAATAACGAGAGTTA**ttaatt**ATTATTTAT**aatt**CATATAATGTAA
TATAATGTAATGT**aattaa**TAGAACATTATTGTGTTATTCACCAGTG**tta**
aGATATA**ttaa**TCCC**aattttA**T**ttaa**TAGTGAAGATTATA**ttttA****ttaa**
ttATGAATCCATATTATTATTAT**ttaa**TATATTTATAATATTATATAT**aa**
ttAT**aatt**ATAAAT**aatt**TATAT**aaaaaaa**G**ttttA****ttaaaaaa**TATTA**t**
taaaaaTATAATA**ttaa**TAATAAAT**aaaaa**TAATATTATACTC**ttaa**TAG
aattTATAATGAT**aaaattaa**GATGAAGAC**tttttttt**AT**aatt**ATTAT**a**
atttATAT**aaaaa**TAATATATATATTTATATTTA**tttt**A**ttaa**TATAT
ATAATATATTTATGTATA**ttaaaaa**GATATA**tttaaa**TA**ttttt**A**ttttt**
tttttATAAGAT**aatttttt**GTAAATATATAAGTAAT**aattaa**G**tttt**ATA
GGGGGAGGGGGTGGGTGATTAGAAAC**ttaa**CTGAATAATATATATAAAGC
ATACATTAG**ttaa**TAT**ttaa**TAATATAATCAATATATAAT**aatt**AT**aaaa**
T**aattaatt**ATATAATAATAATAATGTATAAACAATATAAT**aattt**GTAT**a**
aaaT**aaaa**TATAAATCATAAATAAAGCT**aattaa**T**aaaa**TAATAAATGAT
AAACAAGAAGATATCCGGGTCCCAATAAT**aatt**ATTATTG**aaaa**TAAT**aa**
ttGGGACCCCATATAGAATATAAAT**aattaa**ATATATATATATAAATAAT
aattTATATAATATATTATAAATAAATAATAATAAATATTA**ttaa**TCTAT
AAT**aatt**ATAAATA**ttttA****ttaa**TAT**aatttaa**T**aatt**ATATATA**ttttt**
ATAATAACTCCGAAAGAGTAAGGAGATA**ttaatt**TCTTAT**aaaatttA****tt**
aaTAATAATAATATAT**aaaa**TATATAAATAATATATTATATAT**aaaa**T**aa**
aaT**aaaa**TAAATAATATA**ttaaaaa**TATTGAAAGTA**ttttaa**TAAATAAT
aatttaaattCATATTTATAATAATAAATAAATAAATAAATAAATAAGTA
AATATTTAGATTCTCA**ttaa**TA**ttaa**TATTTATATTTC**tttttttttt**ATA
ATAAT**aaaaa**TATCATATATAAATATAATATAATATAATATAAT**aatt**AT
TATATATAAATAATAAATA**ttaa**ATATAATATATAATAATATATAATCTT
AC**aatt**TAT**aatttaa**TAAAGAAGGAAATAAATAATAATAACTCC**tttt**G
GGGTTCCGGTGGGGTTCACACCTTTATAAATAATAAATAAAGATGTTTAC
TCCTCTTCGGGGTTCCGGTCCCC**ttttt**GGGTTCCGGAACT**aattaa**TA**t**
tttATATAATAATAATAATATA**ttaa**TAT**aatt**TCATTA**ttaa**TAAATAT
CTCCTGCGGGGTTCGGTTCCCCCCCGTAAGGGGGGGGTCCCTCACTCCTT

CGGAGCGTACTATTATTATAAAT**aatt**ATATATTATAATAT**aattaaaaa**
GTATTAT**aatt**GAAACGa**aatt**GT**aattttaaa**TGGAATAAT**aatt**ATTA
TATAT**ttaa**TATAT**ttaa**TAAAGTTATAATATCTCTTTCTACCGGACTA**t**
tttA**ttttttttttttttt**ATAAAG**aaaaa**TAGTAATAATATTAT
CTTCTCCTCCTTTCGGGGTTCCGGTTCCCGTGCCGGGCCCCGGAACTA**tt**
aattATATAATATAATATAATATAATATAATATAATATGATACGGATCAA
ACATTACCCGTTGTTCACTGGCAATGT**ttaa**TCCTATTGTATATAAATAT
AAT**aaaa**T**aatt**ATCCCTCTCGTAATACATATAT**aaaa**TAT**aaaa**TAT**aa**
aaT**aaaaa**TATTATGATTATTATAATATATATATATATATATATAAATAT
ATATATAT**aatt**TAT**aatt**TATATGA**ttaa**TATATTATATATAT**aaaaaa**
TATA**ttaatt**tAC**ttttt**ATAGAAAGGAGTGAGGGACCCCCCCCCCTTAC
GGGGGGGAACCGAACCCCGCAGGAGATATTTA**ttttaa**TACTTATATAGT
ATTTA**ttaa**TAATATAAT**aatt**GTTATTATAAATA**ttaa**TAATAATAT**aa**
aaaTAGGGTAAATAATATAAATAATATGAATAAATAT**aaaaa**CATA**ttaa**
ATAT**aaaa**TATATCAT**aatttaa**TAAATATTATAAT**aatt**TATAAATGAT
AGATATCTGGGGTCCTATAAATAAT**aatt**A**tttt**CAAT**aatt**ATAGGGAC
CCCCACCTATTATATAAATATAAATATAAATATAAATATAAATACAAATA
TAAATATATAAATATATAAATATAATATAAATACAAATATAATATATAAA
TATAAATATAAATATATAAATATAAGTCCCCGCCCCGGCGGGGACCCCGA
AGGAGTGAGGGACCCCTCCCTATACTAATGGGAGGGGGACCGAACCCCGA
AGGAGTATAAAT**aaaattaa**TAATATATATAT**aatt**ATAATAGTTCCGGG
GCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGAAATAATAATATAATAT
ATAAT**aaaa**TATAACTTA**ttaa**TATAATA**ttaaaaa**TAT**aattaa**CAAGA
ATAAATAGTCCGTGGGATCGAACCCCC**tttttt**AT**ttaa**TAT**ttaa**TA**tt**
taaaGAAGG**aatt**GTTTATATATA**ttaa**TATCTTATTTGGGGA**ttaa**TAT
AATATATAAG**tttt**GGATACCAGGCCAAAGACCGGAATCCC**aaaa**GGAGA
TTATATAAATATTATTTATCTCCC**ttttttaa**TATTATAAT**aattttA**tt
aaaaaT**aaaa**TAATAATAAT**aatt**AT**aatt**TATAATAAC**aatt**ATAAT**aa**
tttaattaattaattaattaattaattaattaattaattaattaaTAATA
AATATAAATAT**aaaaa**GAATAT**aatt**TATAATAAAT**aatt**tATATATATA
TATATATA**ttaa**AT**aaaa**TATTTACTTCA**ttaa**TAT**aaaa**TATAAATATA
T**ttaattaa**TAAGTATATATATATAATAATATATAATAACCTATTTATAT
ATATAATC**ttaa**TAT**aatt**ATAAGAAATATTATATAAGTAATATAT**aaaa**
aTAATAT**aaaa**T**aatt**AT**aatt**C**aatt**TATATA**ttaa**TAGTTCCGGGGCC
CGGCCACGGGAGCCGGAACCCCGAAAGGAGGAATAAGATAAATATAT**aat**
tATA**ttaa**TAAATAT**aattttaaa**TG**aattaa**Ta**aattaa**TATATATATG
TATATATATATATATA**ttaaaaa**TAT**ttaatt**A**ttttt**AGGAAGGAGTGA

TAGATCCCTTTGGGGGACCGAACCCCTAT**ttaa**GAAGGAGTGCGGGACCC
CGTGGGAACCGAACCCC**tttttt**A**tttaaa**GAAGAAG**tttt**A**tttt**A**ttt**
tA**tttt**A**tttt**A**tttt**A**tttt**A**tttt**A**tttt**A**tttt**A**tttt**A**tttt**AT**tt**
aatttaattttaattAGG**ttaa**TAAATAGTAATAATAAAC**ttaa**TAATAA
TAATAAT**aattttA**t**tttt**AT**aatt**TA**ttaa**TAATAATAAT**aatt**ATATA
TATATATATTA**ttaa**TAAATATAGACCTTATCGTCTAATGGTTACGACAT
CACCTCTTCATGTTGATAATATCGGTTCGATTCCGA**ttaa**GGTTATTCAT
AATAATAAATATTTGT**aaaaaaa**GTATATAT**aattaa**ACATATTCTTTAT
A**ttaattaa**T**aatt**A**ttaa**TAATATACA**tttt**ATATAATAC**aatt**ATATA
TATATATATA**ttttttttttaa**TACAAATAATATATTCATAATAATAAATA
CCGATTGTTATTATACTATAAT**aaaa**TATATAATATA**ttttt**CATTATAA
TA**ttttttaaa**TAAATATTATAAT**aatt**ATATAAATAATATTTATGTATAA
TAATAATAATAAT**aattG**TTA**ttaattaatt**CTAT**aatt**ATTATATAT**tt**
aatttttttttttaaTATAATATATAATAATAT**aatt**TA**tttt**A**tttttt**
tttATAGTTCCGGGGCCCGGTCACGGGAGCCGGAACCCCGAAAGGAGAAT
AT**aattaa**TAATAATATAAATAACATA**ttaa**CAAT**aatt**ATTG**ttaa**TAT
AATAATAATAATAACAATA**ttaa**TAAATAATAT**aaaatt**A**ttaa**TATTAT
ATTTATATAATA**ttaa**TAT**aaaaa**TCTTTCATAATA**ttaatt**ATTA**ttaa**
ATAATAATGATATCA**ttaa**TA**ttaa**TATAATCGTCAATATTATTTATTTA
TTTATTTATTTATTTATTTATTTATTTATTTATTTATTA**ttaa**ATAAATA**tttt**
taaaTATTATATTATATTA**ttaa**C**ttttt**A**ttaaaaaattaa**TAATGATA
TAATAT**aattaa**TATTATCCACGGGACCAATGACCAACCCAGTAGTTGAC
CGGATTGGCGCCCGCGAGGTTTATAT**ttaa**TAAATAATAATAATAATA**tt**
aaT**aaaa**TCTA**ttaa**C**tttttttttaa**TGGATTATA**ttaa**TG**aaaaaaa**
aaTGAGAAATATC**tttttttttaa**T**aatt**AT**aatt**TATATATAAT**aaaa**
TATGTATATATAAT**aaaaaaa**TAG**tttttaa**TATTATAATAT**aatt**ATAT
ATAT**aatt**ATAAATATATATATATATAATAAGTA**ttaattaa**TAATATAT
ATTTATATA**tttttt**A**ttaattaa**TATATAT**aaaa**TATTAGTAATAAATA
ATATTA**ttaa**TA**tttt**ATAAATAAATAATAATAATATGGCATTTAG**aaaa**
TCAAATGTGTAT**ttaa**GTTTAGTGAATAGTTATATTATTGATTCACCACA
ACCATCATC**aattaatt**ATTGATGAAATATGGGTTCATTATTAGGTTTAT
GTTTAGTTATTC**aatt**GTAACAGGTA**ttttt**ATGGCTATGCATTATTCAT
CTAATATTG**aatt**AGC**ttttt**CATCTGTTGAACATATTATAAGAGATGTG
CATAATGGTTATA**ttttaa**GATATTTACATGCAAATGGTGCATCATTC**tt**
ttttATGGTAATGTTTATGCATATGGCTAAAGGTTTATATTATGGTTCAT
ATAGATCACCAAGAGTACTATTATGAAATGTAGGTGTTATTA**tttt**CA**tt**
ttaaCTATTGCTACAGC**tttttt**AGGTTATTGTTGTGTTTATGGACAGAG

TGAGACAAGTATAAGTATATTATTATAATATCATACCA**ttaa**AT**aatt**A**t**
tttaaTGAAATGATTATGTTTATATATAACATATACCT**aatt**AGACATGC
ATTATTAGTAAT**aattttG**TATGAAACTCTAATAATAAT**aatt**ATTA**tta**
attA**ttaa**GGTAAGATTCATATGGATAGCGTAAGTCAATCTAATATTAT**a**
aaaTATCGTAACATAAACAATA**ttttttt**CTATTA**ttaattaa**TAAATAA
TAATAAAT**aaaaa**T**aatt**ATATGAGAAGTAAGATATTC**aatt**CTGTCTAG
AATACATATATATACG**ttaa**TACTCATCGGTATa**aatt**AGAATCCTAAGT
G**aatt**ATTGAAAGTATAATAATATAAACTTGGTAAGCCC**aatt**ATTTCCA
TATAATA**ttaa**TATAAATATTATATGGTAGTTATATATAATATTA**ttaa**A
TAAATAATAATAG**aatt**ATAATATAGATAAGTGGGT**aaaa**GACTATTG**aa**
aaaGCTAAAGATTATATGTAATGTATAATATAGATC**aatt**ATTTATATA**t**
tttaaT**aaaaa**TATA**ttaa**TAATGG**ttaa**TATTATTA**ttaattaattaat**
taattaattaaTAATAATAACGAATAAATGA**ttaa**TGTGAAAGCATGCTA
ACTTCAATATAGGATGATTTATATAGTATAT**aatt**GTTTGAGCTGTATAC
TATGAAAGTAGTACGTACAGTTCTGAGTGGGGG**aaattt**GTAAAGATCTA
CCTATCAC**aatt**GTCACATTGAGGTAATATAAATATCGCCTCAAATATAT
ttaaTATAAT**aaaa**CT**aatt**TATATAATAATG**ttaa**TATTA**ttaatt**TAT
A**tttttt**ATACGATTATAATAAGACAAATAAT**aaaaa**CTAAAGAATATCT
TATA**ttaattaa**GAGTATAGATTATA**ttaa**T**aaaaa**TAAATATAT**aatta**
atttaaaTATAACAAATAAGAAAGATATAAATAATAATATTGGTCCA**tta**
aATATAAACA**tttt**ATC**aatt**ATTTATGGTTCAATATTAGGAGATGGTCA
TGCTG**aaaaaa**G**aaaa**GGTGGTAAAGGAACAAG**aatt**GTATTTCAACAAG
AATATTGTAATA**ttaatt**ATTTATATTATTTACATAGTTTATTAGCT**aat**
tTAGGTTATTGTAATACT**aatt**TACCT**ttaattaaaa**CTAGATTAGGT**aa**
aaaaGGTa**aatt**AGACAATA**tttaaatttaa**TACATGAACTTATGATTCA
T**ttaa**TATGATTTATTCAGAATGGTATA**ttaaaaa**TATATCTGG**aaaa**GG
TAATA**ttaa**AGTTATTCCTAAATCTTTAGAC**aatt**AT**ttaa**CTCCTTTAG
CTTTAGCTATTTG**aatt**ATAGATGATGGATGT**aatt**AGGTAAAGG**tttaa**
attCACAACT**aattGtttt**AGTTATAAAGATGTTCAATATTTACTTTATT
TATTACATAATAAATATAATA**ttaa**ATCTACTATTC**ttaa**AGGCAATAAA
G**aaaa**TACAC**aatt**TGTTATTTATGTATG**aaaa**GAATCTATACCTA**tttt**
aaCTa**aatt**GTATCTCCTTATATTATTCCTAGTAT**aaaa**TAT**aatt**AGGT
aattATTTATAAT**aaaa**TATATAGTATTATA**ttaatt**ATTATATTATTAT
AATGCGATATTATTG**aaaa**CATGTCa**aatt**ATATTA**ttaa**GTAACAAGAC
AGTGGGTTATAT**aatt**ATATGATCCCAACAGAATACACCAATAATAGGTA
TTATTAT**aaaaaaaa**TAATAATAT**ttaa**TGTTTATTCGAAG**aaattt**ATA
ATATTATTATTATAACACAAGGT**ttaa**TAATCTATATATATATATTATAT

ATATAACTACTGTTATTATTCCATTTACCT**aattaa**TATATAAATAATG**a**
attAT**aatt**ATTATGA**ttaa**TA**ttttt**ATAATAATAACCCCATCATAACA
TTTATATATAACATTTATATATAACATTTATATATAATATTTATATTATG
GTATTATTAGGTATAAATATTTATTCATAAGAG**aaaa**TAGTGA**ttaa**ATG
G**aatt**AT**aaaaa**GGGTAGATATTA**ttaa**ATACAGGGTATTATTTATA**tta**
aTAAATCAATAAATATTGAGATTATTATTA**ttaaaaaa**TAATAAT**aatt**T
ATAAATAATATTA**tttt**CTTGGCACTAGTTATTACT**aatt**TATTCTCAGC
aattCCATTTGTAGGTAACGATATTGTATCTTGATTATGAGGTGGGT**tta**
aTATAGAGGATCCATATTATAGTAATATAATA**ttaa**ATAAATCTG**tttt**A
TGCTGAAATATCTTCATTTGAATAAT**aatt**ACTATATTATTC**aattaatt**
ATTTATAATAATAT**aatt**TGAAAT**aaaaa**TAATATAG**ttaaaa**TATTTAT
TATAAGAAGa**aatt**AGCAGT**aattaa**TATATATATATATATa**aattaatt**
ATTCAGAGACTTTATAGTTATTATATAAATAATACTATTATTTATGAT**aa**
aaaTCAT**aattaa**ACACAGATAATCCTATTTATGCATATATTGGTGGTTT
ATTTGAAGGAGATGGTTG**aatt**ACTATTTC**aaaaaaa**GGTAAATATTTAT
TATATG**aatt**AGGTATTGAAATACATATTAGAGATATTC**aatt**ATTATAT
a**aattaaaaa**TA**tttt**AGGTATTGGTAAAGTAAC**aattaaaaattaaaaa**
T**aaaa**GATGGTACTA**ttaa**AGAAATATGT**aatttaa**TGTAAGAAAT**aaaa**
aTCA**tttaaa**GAATATTATTATTCCTA**ttttt**GATAAATATCCTATA**tta**
aCTAATAAACATTATGATTATTTATA**ttttaaa**GAT**aatt**TA**ttaaaa**GA
TA**ttaa**ATATTATAATGATTTATCTTATTATTTACGTCCTA**ttaa**ACCAT
ttaaTACTCTTGAAGATA**ttttaaa**T**aaaatt**A**ttttt**CTTCATGA**ttaa**
ttGG**ttttttt**GAAGCTGAAAGTTG**tttt**AGTATTTATAAACCTATAAAT
aaaaaaaT**aaaa**CTTGCTAG**tttt**GAAGTATCTC**aaaa**TAATAGTATAGA
AGTTATATTAGCTA**ttaa**ATCATA**tttaaaatt**ACTC**aaaa**TATTTATAC
AGAT**aatttaa**T**aatt**CAAGAATAACAC**ttaaaa**GTA**ttaa**TGGTA**ttaa**
aaaTGTTGTAATATTTA**ttaa**TAATAACCCTA**ttaatt**ATTAGGTTAT**aa**
aaattACAATATTTATTATTC**ttaaaa**GATTTACGTCTTATTC**ttaa**ATA
TAAT**aatt**A**ttttaaatt**CCTCCTAAATA**ttaa**TCTTATAT**aaaaa**TATA
ATAATAATATATTTATATATTATAT**aatt**ATATAAAC**aaaa**TAT**aatt**TA
TATAT**aatt**ATTTATTATAAATATAGTCCGGCCCGCCCCGCGGGGCGGAC
CCCGGAGGAGTGAGGGACCCCTCCCTATTCTAACGGGAGGGGGACCGAAC
CCCGAAGGAGT**ttaatt**ATATA**ttaa**ATATATTATTATCAATAAAT**aatt**
CCTTTGAACTATTTATTA**tttt**ATTATATTTA**tttt**CTCCTTCATTA**tta**
atttttA**ttaa**T**aattaaaa**TCTTATCA**tttt**ATGGTA**ttttt**ATTTCTA
ttttAGGATATCGAAACTAT**aattaaaaa**GTAT**aattttA**ttaatt**ATaa**
ttTATGA**ttaa**TAAATAAGAAAT**aaaaa**CTTTAGAAGTAATATTTATC**tt**

tttttttATAAATAAATATTATGA**ttaa**TATATAATCATTTATAAATAT
TTATATAT**aatt**ATATATATACATAAATAGGA**ttaa**GATATAGTCCGAAC
AATATAGTGATATATTGATAATAG**tttt**CAAATATGTAACTA**tttaaa**CA
ttaaaaGCTCAGTATCTAACCCTCTAATCCAGAGATTCTTTGCGTTACAT
TATTTAGTACC**tttt**ATCATTGCTGCAATGGTTATTATGCAT**ttaa**TGGC
ATTACATATTCATGGTTCATCTAATCCATTAGGTATTACAGGT**aatt**TAG
ATAG**aatt**CCAATGCATTCATACTTTA**ttttaaa**GATTTAGTAACTG**tt**
ttCTTATTTATG**ttaatttt**AGCATTATTTGTATTCTATTCACCTAATAC
TTTAGGTC**aaaa**TATGGCCTTATTA**ttaatt**ACATATGT**aattaa**TA**ttt**
tATGTGCTGTATGCTGGAAATCTTTATTTA**ttaa**ATATCAATG**aaaattt**
ATAAT**aaaa**CTCTATATTA**tttt**ATTATTC**aaaa**TA**ttttaaa**TAC**aaaa**
C**aattaa**AT**aatt**TCGTA**ttaaatttaatt**GAACAAAGCAATATAAT**aaa**
aTAAATATTGTAAGTGATTTAT**ttaa**TCCCAATAGAGT**aaaa**TATTATTA
TAAAGAAGATAATCAGCAGGTAACCAATAT**aatt**CTTCTAATACTCAC**tt**
aaCGAGTAAT**aaaaa**G**aatt**TATTAGTAGATACTTCAGAGACTACACGCA
CACT**aaaaaa**T**aatttaatt**ATTTA**ttaa**ATA**ttttttaa**TAT**aaaaaaaa**
TAAATC**aatt**ATTC**ttaaaa**GACATTATAGTATTTATAAAGATAGTAATA
TTAGAT**ttaa**CCAATGATTGGCCGGT**ttaatt**GACGGAGATGGTTA**tttt**
tGTATTACT**aaaaa**TAAATATGCATCTTGTG**aatt**CTTGTAG**aattaaaa**
GATG**aaaaaa**TG**ttaa**GACAAATCCAAGAT**aatt**tGGTGGTTCTGTa**aat**
taaGATCAGGTG**ttaa**GGCTATTAGATATAGATTAC**aaaa**TAAAGAAGGT
AT**aattaattaattaa**TGCCG**ttaa**TGGTAATATTCGTAATAGT**aaaa**GA
TTAGTAC**aatttaa**TAAAGTATGTA**tttt**A**ttaa**ATATCGA**ttttaaa**GA
ACCTA**ttaattaa**CTAAAGATAATGCTTGATTTATAGGGTTCTTTGATGC
TGATGGTACTA**ttaatt**ATTATTATTCCGGT**aattaaaatt**AGACCTC**aa**
ttaaCTATTAGCGTTACAAATAAATATTTACATGATGTTGAATACTATAG
AGAAGTATTTGGTGGTAATATTTA**tttt**GATAAAGCT**aaaaa**TGGTTA**tt**
ttaaaTGATCTA**ttaa**TAATAAAG**aatt**ACATAATA**tttttt**ATCTTTAT
AAT**aaaa**GTTGTCCTTCTAAATCTAATAAAGGTAAACGTTTA**tttttaat**
tGAT**aattttt**ATTATTTATATGATTTATTAGC**ttttaaa**GCACCTCATAA
TACTGCTTTATATAAAGCTTGA**ttaaatttaa**TG**aaaaa**TGAAATAAT**aa**
ttaatttttCTCCGTATTCATTATTATATTATCT**aatt**TAT**aaaa**TA**ttta**
aaGATTCCTTATAATAATATAACATCTTTGT**aatt**ATTG**ttaa**AGATAAT
AT**aatt**ATTATGAATCGGTAGATTATA**ttttt**ACAATCTTA**ttaa**ATa**aa**
ttCTGATCA**ttaa**ACATGATTGAAGAAATAATAATAGTTTATGAAATAAG
ATAGTGTAATAT**aatttttt**ATGAAGATATAGTCCA**tttt**ATATTTATTAT
aaaaGCATCCTGATAACTATATTCCTGGTAATCCTTTAGTAACACCAGCA

TCTATTGATA**ttaaaaa**TA**ttaa**Ta**aatt**ATTATTAT**ttaa**TCTTATTTA
ttttATAT**aaaaaaaa**TAAATAAT**aatt**A**ttaa**T**aaaaa**TATATTATTTA
TTTCTCCTTTCGGGGTTATTTATATATATTCCTTTAT**aatt**TATAT**ttaa**
TATATTATA**ttaa**ATATATG**aaaatt**ATAATAAAT**aattaattaattaa**T
AATAAATAATAAT**aaaaa**GTACAGTAGCA**ttaa**ATATTC**ttaa**GTTTCCG
CTTTGTGGGAACTCCCATAAGGAGT**ttaa**TGA**ttaaatt**GG**ttaatt**GTC
AAG**aaaa**TCTAAGGTA**ttaa**TAAATAAATAATACTATGACAACTTGCAGC
GAAGTTTATATCATCTCTATATTATATA**ttaa**TATATATATATAATAATA
ATAATAATA**ttaa**TATAATATAAGATAT**aaaaa**CGTTCAACGACTAGAAA
GTGAACTGAGATAGTAATACCTTTCCACG**aaaa**CC**aattaatt**TAT**aatt**
A**tttttaaa**TAAAGAATAGATTA**ttaatttttttt**ATATAGTTCCGGGCC
CCGGCCACGGGAGCCGGAACCCCGGAAGGAGTAATATATATTATATAT**aa**
aaT**aaaaaa**TATATATATATATATTAT**aaaa**TATC**aaaa**G**ttttaa**TC**tt**
ttATTAT**aattaa**TGACATAGTCTGAACAATAATGa**aatt**ATTGAGATAA
GATA**ttaa**ATAATCTTATG**ttaa**CATATAT**aatt**GTGTACCTGAATGATA
CTTATTACCATTCTATGCTA**ttttaa**GATCTATTCCTGAT**aatt**ATTAGG
AGTTATTCTAATGTTTGCAGCTA**tttt**AGTATTATTAG**tttt**ACCATTTA
CTGATAGAAGTGTAGTAAGAGGTAATAC**ttttaaa**GTATTATCT**aatt**CT
TCTTCTTTATCTTTGTATTC**aatt**TCGTATTATTAGGAC**aatt**GGAGCAT
GCCATGTAGAAGTACCTTATGTC**ttaa**TGGGACAAATCGCTACATTTATC
TACTTCGCTTATTTC**ttaatt**ATTGTACCTGTTATCTCTACTATTG**aaaa**
TG**tttt**ATTCTATATCGGTAGAG**ttaa**TAAATAATATAT**aattaattaa**T
ACATAGATATAATATATATATTATTATTA**ttaa**TAATATAAT**aaaaa**T**aa**
aaaTa**aatt**A**ttaa**TAATAATAATACT**ttaa**TAATATTC**ttaaaaa**TAAT
ATATCTCT**aatt**TAT**aaaattaa**ATAATAATAAT**aaaaaaaaaaa**TATTAT
aaaaTAT**aattaattaa**TAATG**aaaa**TAATATACTTA**ttaattaa**TATAA
ATAAATGAATAATATAATATAACTATATTG**aatt**ATAATCTATCTATC**tt**
tttttttCATAT**aatt**ATAATATATATA**ttaa**TATATAT**aatt**ATTA**ttt**
tATATATTATAGTTCCGGGGCCCGGTCACGGAAGCCGGAACCCCGCAAGG
AGATTTA**ttaatt**ATTATTATCATTATTA**tttttt**AT**ttaa**TCTTATTTA
TTAT**aaaa**T**aattaatt**ATCATAAAGCAT**aatt**ATTATAGAATCTTATTA
ttttCTTTAT**ttaatt**tAT**aaaaa**TATAAAGTCCCCGCCCCC**ttttt**A**tt**
ttAT**ttaattaa**GAAGGTA**ttttaaaaaa**GGAGTGAGGGACCCCCTCCCG
TTAGGGAGGGGGACCGAACCCCGAAGGAGTACTCAT**ttaa**TATAAATA**tt**
aaAT**aaaatt**A**tttt**ATATATA**ttaa**TGATTA**ttaa**TATTGATAATAT**aa**
ttA**tttt**AT**aattaatt**ATTATAAATATATAACTA**ttaa**T**aattaatttt**
taaTCTAGGGGTTTCCCCCACTTACATAAACTTACGTATACTTACATATA

CTTATGTATACTTACATATACTTACGTATACTTATATATACTTATGTATA
CTTACGTATACTTACATATATGGGGGATCCCTCACTCCTCCGGCGTCCTA
CTCACCCTATTTA**ttaa**TCA**ttaa**TAAG**aatt**ATTA**ttaaaaaatt**AT**aa**
ttTACTCAAAG**ttaatt**ATAAATATA**ttttttaaa**TATCTA**ttttt**A**ttaa**T
C**ttttt**AT**aaatttaattaatt**GT**aattaattaa**TATTATAAT**aatt**ATTC
TTAGGAAGGATATTTATTTA**ttttaatt**ATG**aatt**CCTGACATAGAGAC**a**
attaattAGAACTTCTTATTATTATTATAGTAATAAT**aaaaa**TATTCTAA
ATATATTATATATATTATTA**tttttttt**ATTA**ttaa**T**aaaa**TATTATAAT
aatttaaaTAAGTTTAT**aatttttt**GATAAGTATTGTTATA**tttttt**ATTT
CCAAATATATAAGTCCCGGTTTCTTACGAAACCGGGACCTCGGAGACGTA
ATAGGGGGAGGGGGTGGGTGATAAGAACCAAACTATTCAATAAATATAGA
GCACACATTAG**ttaa**TAT**ttaa**TAATATAACTAATATATAAT**aatt**AT**aa**
aaT**aattaatt**ATATAATATAATATAAAGTCCCCGCCCCGGCGGGGACCC
CAAAGGAGTA**ttaa**CAATATAATATATTGTAT**aaaa**T**aatt**ATAAATA**tt**
aaAT**aaaaa**CCAAATAAATAATATAATAAATGATAAACAAGAAGATATCC
GGGTCCCAATAAT**aatt**ATTATTG**aaaa**TAAT**aatt**GGGACCCCCATCT**a**
aaaTATATATATAACTAATAATATATTATATATA**ttaa**TATATAATAATA
TTA**ttaaaa**TATAATATTA**ttaaaaaaaaaa**GTATATAT**aaaa**TAAGATAT
ATATATATAAATATATATATTC**ttaa**TAAATATTATATATAATAATAAT**a**
attATTTCATAAT**aatt**ATTTC**ttttt**A**ttaa**T**aaaatt**ACTTATCTCCT
TCGACCGGACTA**ttaa**ATA**ttaa**ATAT**ttaa**TAT**ttaa**TAT**ttaa**TA**ttt**
tATTCTATAGATATTCATATG**aaaaa**TAATAAGTATAT**aatt**ATGATAAT
GAATATA**ttttt**ATTTAT**aatt**TATTATTAT**aaaaa**TA**ttttaatttaa**T
AATAATAATAAATCATTATA**ttaatt**C**ttttaa**G**aatt**TAT**aatt**GTCAT
TATTTATTATATACTCCTTA**ttaaaa**GGGATTCGGTTTCCCTCATCCTCA
TGGGTATCCCTCACTCCTTCTGAT**aattaatttt**ATAATAATAAT**aaaa**T
AAAC**ttaattaa**ATATTATATATTTATTTAC**aatt**ATATATATATATTAC
TCAT**aattaattaattaa**GATGC**aatt**CAATACGGTTGTATTATATTATT
CATCAAATATTG**ttaa**TATTGATACCTACAGAGATAT**ttaa**TA**ttttt**AT
TATTATTATCCATTAC**ttttttt**ATTATA**ttttaatt**ATTTATTTATTTA
TTTATTTATAATAATAATATTTCATATTATC**aatt**ATTA**ttttttttttt**
tATAATATAT**aattaatt**ATTTATATAGTTCCCCGAAAGGAGAATAAAT**a**
aaaTATTATATAAATATTTATATCTTTA**ttaa**TA**ttaa**TATAAGTAATAT
ATATAGTTTATGATAT**ttaattttt**ATCATAATATAATAAT**aatt**ATATAA
ATCTTATACACATTTATATAAGTATATATATATATTA**ttaa**TATAATGAA
CATCTA**ttaa**AT**aaaa**T**aatt**GTAAATCTCAAGT**aatt**ATTATTA**tttt**A
ttttaaTAAT**aatt**TATGATTTAT**aattaa**TAAAT**aaaa**GAGT**aatt**AT

ATGAT**aaaaaa**GGTAATAAAT**aaattt**ATAGTTCCGGGGCCCGGCCACGG
GAGCCGGAACCCCGAAAGGAGTTTATTTATATATATATATATATG**aatta**
aTAT**ttaa**TAATAAATAATAATAT**aattaa**TAATATTATTATTATTAT**aa**
ttttttATTTATAATA**ttaa**T**aaaa**TATTATTATATATATATTATAATA
ATA**ttaa**TAAGATATATAAATAAGTCCC**ttttttttt**A**tttaaaa**TAAAG
AAAGAAT**aattaa**ATAATA**ttttaa**T**aatttaattaa**ATAGTGTA**ttaaa**
aGATAAT**aaaaa**GTAATA**ttaa**TATG**ttaatt**ATATATAATATATTTATA
TAT**aatt**ATATATATATATATAAATAATAATAAATATATATATAATAT**aa**
aaaTAAGAATAGA**ttaa**ATAT**ttaa**TAAATAAATATTATGC**aatt**AGTAT
TAGCAGCTAAATATATTGGAGCAGGTATCTCAAC**aatt**GGTTTATTAGGA
GCAGGTATTGGTATTGCTATCGTATTCGCAGCT**ttaattaa**TGGTGTATC
AAGAAACCCATC**aattaa**AGACCTAGTATTCCCTATGGCTA**tttt**AGGTT
TCGCCTTATCAGAAGCTACAGGTTTATTCTGT**ttaa**TGGTTTCATTCTTA
TTATTATTCGGTGTATAATATATATAATATATTATAAATAAAT**aaaaaa**T
AATG**aattaa**T**aaaaaaa**T**aaaa**T**aaaa**T**aaaa**TCTCATTTGA**ttaatta**
aTAACATTCTTAT**aatt**ATAT**aatt**ATTAT**aaaa**TATATAAATATTATAA
TAATAATAATATATAT**aatt**ATAAT**aaaaaa**TAATAATAATATATAATAT
ACC**tttttttttaa**TATA**ttaa**TATATAAATAAATAAATAATGGATAATAT
AT**aatt**AC**ttttttt**ATATTA**ttaa**TAATAAT**aatt**TATAAATATTGTTA
TAATAAACATTTATATAAATAAATAT**aatt**ACCATAATAAGATATATTAT
TTA**ttaa**TAAT**aaaaa**TATTTA**ttaa**TAAATAAGAAATATATATATTATG
ATAATATTTA**ttaa**TAAATAAT**aatt**CTTTATATATAAATATA**ttaa**ATA
TAT**ttaatt**GAACACAATAT**aatttttt**ATTGTATTATTCAT**ttaa**TAATA
ttaaTA**ttaa**TA**ttaa**TATAATATTAGTGAACATCTCCTTTCGGGGTTCC
GGCTCCCGTGGCCGGGCCCCGGAACTA**ttaa**TAT**ttaa**T**aaaa**TATATAT
aattTAT**aatttt**CATAT**aattaa**TATAAT**aatt**AGGTTTATAAAT**aatt**
ATAATATATTATAACAATATAAT**aaaa**TATATTATAAATCTATCTATCTA
TCTATATAATATAT**aatt**tATATATACA**ttaa**TAATAT**ttaatt**AT**aatt**
A**tttaaa**TAT**ttaatt**TA**ttaa**TATTCCCCGCGGGCGCCAATCCGGTTGT
TCACCGGATTGGTCCCGCGGGGTTTATATTA**tttaaa**TA**ttaa**ATA**ttaa**
ATAAT**aatt**TATATTATA**ttaa**TAAATATAAT**aattaaaaa**TATATGA**tt**
aattATATAATAATAATAAT**aatt**A**ttttaa**TATTAT**aatt**TATa**aatta**
attATA**ttaatt**ATA**ttaatt**CTTATTATATAAT**aatt**A**ttaa**TAAT**aat**
tTA**ttttaa**GAAAGGAGTGAGGGACCCCCTCCCGTTAGGGAGGGGGACCG
AACCCCGAAGGAG**aaaa**T**aattaa**T**aaaa**G**tttaaaa**GTTCTTATA**ttaa**
T**aatt**ATATAATATTATA**ttaa**AGA**ttttt**ATAATATATATATATAATAT
ATTTATAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGTT

TAT**ttaa**TATTTATATTTATA**ttaa**TATTTATATTTATATTTATATTCCT
C**ttaa**GGATGGTTGACTGAGTGG**tttaaa**GTGTGATATTTGAGCTATCAT
TAGTCTTTATTGGCTACGTAGGTTCAAATCCTACATCATCCGTAATAATA
CATATATATAATAAT**aattttaa**TATTATTCCTAT**aaaaa**T**aaaa**TAAAT
AAATAAATAATAAT**aattaattaattaattttaa**TAAATAT**aaaa**T
ATAT**aaaaa**TAATAATAATAAT**aatt**ATTA**ttttaa**TAATATTATTTATA
TAATAGTCCGGTCCGACCC**ttttt**ATTC**ttaa**GAAGGGA**tttt**A**tttt**A**t**
taattaaTAATAATATA**ttaaaatt**ATAAAT**aattaa**T**aatt**CTTTATAT
TTATATATATATATATATATTTATATATTTATATATATA**ttttaa**TAATA
TTATGATATA**tttt**A**ttttaa**TAATA**ttttt**A**ttttt**ATATATa**aatt**AT
AATA**tttt**A**tttt**AT**aatt**ATTTATATAT**aatt**A**ttaa**TAAT**aatt**A**ttt**
ttttttATTTGGGATTTATATTATTATTATAAAGAATATAATGTTA**ttaa**
TAACTGC**aaaaaa**TATCTAATATATTATTATTTATAATAATAAATAATAT
TATAATAAGGATGCATATTATATATATATATATATTTCTATTTATA**ttaa**
TA**ttaa**TA**ttaa**TATGTATATATAATAGAT**aaaaa**GT**aaaaa**T**aaaaaa**T
AATG**aattaaatt**A**ttaa**ATAT**aattttA**TCAATAATAAATAAAC**ttaa**T
AATAATAATAATATTATTA**ttaa**TAATCTATTAGATTCA**ttaa**TAAATAA
G**aatt**ATTA**ttaa**AGAATATATTATTAGATATAAATAAT**aaaaaaa**TAAA
TAATAT**aaaaa**GAATA**ttaa**ATAATAATAATATAAACCCCGCGGGCGCCA
ATCCGGTTGTTCACCGGATTGGTCCCGCGGGGAATA**ttaa**TAAT**aatt**AC
AACA**tttaaa**TAATATAAAT**aattG**AAATCTAC**aatt**tAT**aatt**ATAAT**a**
aaaaTATAG**aatt**ATAAATACTATAAATGAT**aattaattaa**T**aatt**ATTA
TAT**aaaa**TAATAAC**tttaaattaa**ATAATATAAATA**ttaa**Ta**aatt**ATTA
TAAGTAAACTTA**ttaa**TCAACATAG**tttaaa**T**aattaa**ATA**ttaattttA**
TTATTATAATAATGATA**ttaa**TAATAATAATAATAATAATAATAAT**aatt**
ATTATATAAATATAATAAAT**aattaa**TAAATATTATAAATAATAATATAA
ATAAT**aatt**TATGTAATA**ttttaa**GTTATTATTAT**aaaaaaaaa**GTAACT
ATTGAACCTA**ttaatt**ATCATATATTTA**tttaaa**TAGTGATA**ttttt**AGT
AAATATATTAG**tttaaa**TGATATAGATAAATATAATAATGGTATC**ttaa**C
T**aatt**ATCAACGTATA**ttaa**ATAATATTATGCCT**aattaa**ATGATCATAA
TATTTCTAT**aatt**ATA**ttaa**TAATA**ttaa**TAATA**ttaa**TAATAATAAATA
TAATAATAT**aattaatt**TA**ttaa**ATAATAATAATAATA**ttaa**TAATAATA
AT**aatt**ATAATAATAATAATAAT**aatt**ATATTGGTAATA**ttaa**TAATATT
TATAATAATATAACTATTGATAATATTCCTATAGATA**ttttaa**TATATAA
ATATTTAGTTGGTTGATCTA**ttaatttaaa**GGTAGA**ttaa**GTAATAATAA
TGGTAGAACTAGTACAC**ttaatt**TA**ttaa**ATGGTAC**ttttaa**TAAT**aaaa**
aaTATTTATGAAGTAATA**ttaa**TAAT**aatt**AT**aattaatt**ATATCCCTTC

TAATCAT**aatt**TATATAAT**aatt**CTAATA**ttaa**T**aaaaa**TGGTAAATATA
ATA**ttaa**AG**ttaattaa**ACTTTAT**ttaa**TATATATA**ttaa**TAGTTCCGGG
GCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGAAAT**aaaa**TAAATATAA
TAAAT**aaaa**TAAATAAATAAATAATATATATATATATATAAATATAT**aaa**
aTAATATTTAC**ttttt**ATATATATAT**aatt**ATATATAAAT**aaaa**TATAAT
ATAATATCATAT**aatt**ATAT**aaaaa**Ta**aatt**AT**aatt**TATTTATA**ttaaa**
aaTA**ttaattaattaatttttt**ATAT**aatt**ATTATAATAAT**aatttaat**
taaaaaTAAATATCAAATa**aatt**AT**aattaa**TCCTAC**tttt**GGATCCTAT
TTATA**tttt**ATTATTATAAAT**aatt**ATTATTGATAG**ttaattaa**AT**aaaa**
aTATATATATATATTACTCCTTCGGGGTCCGCCCCGCAGGGGGCGGGCCG
GACTATTAT**aatt**ATTA**ttaa**TATA**ttaatt**A**ttaatt**ATATAAACCGCC
CCCGCGGGGGCGGTTAGTTATTTATA**ttaa**TATA**tttt**ATA**ttaa**TATAT
AATACTC**tttttt**CTATTATA**ttttaa**TATATAATA**ttaaaaaaaa**TAAA
T**aaaa**TAATATTC**ttaatttt**ATTCTTTATCTTCT**ttaa**CCAAACTCCT
TCGGGGTTCGGTCCCCCTCCCATTAGGTTAGGGAGGGGGTCCCTCACTCC
TTCGGGGTCCGCCCCCCCCCGCGGGGGCGGGCCGGACTA**ttttaatttta**
atttaattttATAAATATAATAT**ttaatt**AT**aatttaa**TAATAATATAT**a**
aaaaaTATATATATGG**ttaa**TATATATAAAGATTATAATC**ttttt**A**ttaa**
ATAAAGG**aaattt**ATTATAT**aatttttt**CTCTATAGTTATATA**tttaaaa**C
TTA**ttttttttttttt**ATAAATAAT**aatt**ATAATAAATAATA**ttaatt**AT
TTATTATAT**aattaatt**GGCCCCCATGCTGGGTTCCGGAACTCCTCCTTC
TCGCGAGG**ttaa**CACCTATTATATAACTATAACTATAACTATAACTAT**aa**
ttAT**aatt**ATAACTATAACTATAAATATTCA**ttttaa**TAATAATAATAAT
AATAATA**ttaa**TATAAATAGTCGAAGAATATATTTATTTA**ttttaa**TATA
AAT**aaaaa**GTTTC**aattaatt**TG**aatt**TGG**aattaatt**ATTACTTCATAT
GGGGTTATGGATTTCGTTCGGAACTCCTCCCTCCTACCTCTATTTA**ttaa**
TCATAAATCAT**aatt**ATTA**ttaattaa**TAATAAT**aatt**TACTCGAGGTTC
ATACCTA**ttttaa**TA**ttaa**TA**ttaa**TATTGAT**aaaa**TATATATTCACT**aa**
aaaGTATAT**aatt**TACTC**aatt**TATACTAT**aattttA**TA**ttttttt**ATTA
T**aatttaatt**ATTTCAAATAAAGT**aatt**ATAATAATATATATCCTTTA**tt**
aaATATATA**ttaattaa**TATATATAT**aaaaa**GTAAATATTA**ttaatt**GTA
TAT**aatt**ATAAAT**aattaa**TATTTA**ttaaaa**TATATAT**aatt**TATAATCC
TCATAT**aattaa**TATAATAAATAATATAACACAATGT**aatttaatttaat**
tACATAAT**aatt**tATTATTATTAT**aatt**ATTATTTATTTATTTATTTATT
ATTAT**aatt**ATAAATATTATTAT**aattaaaa**TC**aatt**A**ttaatt**A**ttaa**T
GATAAAT**aattaa**TGAT**aatt**ATCAATAACC**aatt**AGATTATTTATCGAT
AT**ttaatt**ATATTATATTATATTATATTATATATATATATATATATTATA

TTAT**aaattt**ATTTATAAATATTTGTTTATTTATTTATTTATTGAATAAC
AATAG**aattaa**ATATTGTCAATAAATAATAAATAATGT**ttaa**TATATATT
ATATTATA**ttaa**TA**ttaa**TATTATTATTA**tttttttt**ATTATA**ttaa**TAT
aattTAT**aaaaa**TATa**aatt**ATTA**ttttt**ATTAT**aatt**TATATATATATA
ATATATATATTTA**ttaaaa**TA**ttttaa**GAAAGGAG**aaaaa**T**aattaatta**
attaattaattATTTATTATTATTATTATTATTTATATAATAATATATTA
tttaaaTATTTATATATTA**ttttt**ATA**ttaa**TATTTATAGATGGGGGGTC
CCTATTATTATTG**aaaa**TAAT**aatt**A**ttaa**TGGACCCCAGATAGCTTCTT
GTTTATCATTTATATATATATATATATTA**ttaatt**A**tttt**ATTCTCCTTT
CGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTTTATAATATTATTA**t**
taattAT**ttaattaa**TATTATAATCATAT**aatttaa**TA**tttt**AT**ttaatt**
ttA**ttaaatttaa**TATATATA**ttttt**ATTATTAT**ttaattaatt**TATAAA
TAT**aaaa**TATTC**ttaa**TA**ttaaaaa**TAAATAAATAATAAAGTTTATAAAT
CATATATTAT**aatt**ATTTATTA**ttttt**ATATTATA**ttaa**T**aaaa**TATTAT
TATTAT**aaaaaaaaa**TAG**aatttt**ATAATA**ttttt**ATATA**ttttttaatt**A
TTATTA**ttaa**TATTTA**ttaa**AGGAAATAT**aaaaa**CCGAAGGAATATTAT**a**
attAT**aatt**AT**aatt**ATTATTATATT**taatt**TATTATTATAATAAT**aatt**
ATAGTCTGCCCCCTCTTTATCTTTA**ttttaaa**GTTCCGGGGCCCGGCTAC
GGGAGCCGGAACCCCGAAAGGAGAAGGATAT**ttaa**T**aatt**TATAATAT**tt**
aattCATATATATATATATATA**tttt**A**tttttt**ATATATATAT**ttaa**TAT
ATTATATTTATATTTATATTATTATTATATTTATATTATATTAT**ttaatt**
A**ttttttaa**TAATATATTA**ttaa**TA**tttt**ACC**tttt**GATAAAT**aaaattt**
A**ttaaaatttt**ATAATAAGTA**ttaaaa**TATCAT**aaaa**GTATAATATTTAT
AT**aaaa**TGTAT**aatt**tATAATCTTCT**aattaattaattaa**ATAAAT**aaaa**
T**aaaa**T**aattaa**ACTCC**tttt**GAGATTCACACCTA**tttt**A**ttaaaaa**TAG
GTATTCAC**ttaattaattaattaattaattaatt**ATGGAT**aatt**TAT**tta**
aTAAATATATATA**ttaatt**AT**aaaa**TAATAGTCCGGCCCGCCCCGCGGGG
CGGACCCCGAAAGAGTCTGCCCC**ttttt**AT**ttaa**TAT**ttaa**TAT**ttaa**TA
T**ttaa**TAT**ttaa**TAT**ttaa**TA**atttaaa**GAAGGATATATTTAT**aatt**TATC
ATAATATTAT**ttaa**TAAG**aatt**A**ttaattaattaattaattaatt**TATTT
ATTGTTTATATTTA**ttaa**TA**ttaa**TATAAT**aaaaa**TGT**aaaa**TAC**ttaa**T
ATTA**ttaa**TATTATTATATATAATATATATAATAATATATTATATTTATA
TCTCCTTTATTCC**ttttt**CCCCCGATGGGGACTTATTATATTATATTATT
ATATATTTCTTCGATAACTTTATATATA**tttt**A**ttttt**AT**aaaaaaa**TAT
TTATATATTATTATTTACAATAAT**aatt**A**ttaa**TAGTCCGGCCCGTCCCG
CGGGGGGGAACCGAAGGAGTGCGGGACCCCGTGGGAACCGCATCCC**tttt**
tA**ttttttaattaa**GAAGGAGTGAGGGACCCCGTGGGGACCGAACCCCGAA

GGAGTC**tttttt**CTATTTA**ttaa**TAATAACTAT**aatt**ATA**tttaaaa**TAA
T**aatt**TACTTGTTATAATC**ttaa**TGTTCCGGGGCCCGGCCACGGGAGCCG
GAACCCCGAAAGGAGAAGTATATAAATATTTACTTGTTAT**aatt**TATTAT
ATATTTATAACCTCCTTC**ttaatt**ATCTTTACTTTATAATa**aattaa**TAT
AATATAATCTGATAATAATCG**aattttA**TTATAT**ttaatttaattaa**TAA
TAGAC**aatt**ATTATTATTA**tttt**ACTTA**ttaa**TA**ttaatt**TAGATTTATA
TATATAAATATTA**ttaattttA**TA**ttaatttttt**A**ttaatt**ATTTA**tttt**
tATATTCATA**tttttt**A**ttaa**TATTA**ttttt**A**ttaa**TAAC**tttttaaa**T**a**
attATAAACTATATATTATTTATATTTATATTTATAATAAATGAAAC**aat**
tATAAT**aaaatt**AC**aatt**AC**aatt**ATATTAT**aatt**ATGATTACAATAGGG
ttaaACATTACCTGTGAACAACTGGTAATGT**ttaa**CCCGTATTATTATTT
ATTATATTATATATATA**ttaaaa**TA**ttaa**TA**ttaa**TA**ttaa**TATTATATT
ATATTATATTATATTATATTATATTATATTATATTATATTTAT**aatt**ATA
TTATATTATAT**aatt**TATATAC**tttt**AT**aatt**CTTATTATTATTTATTTA
TTATTTATTTATTATTA**tttaaa**TATATTATTATTATATA**ttaa**TAATAT
ATATATTA**tttt**ATATA**tttt**AT**ttaa**TAT**aatt**ATTTATA**ttttt**ATA**t**
tttATTATGAGGGGGGGTCCC**aatt**ATTA**tttt**CAATAAT**aatt**TATCAT
GGGACCCGGATATCTTCTTGTTTATCATTTATTATTCTTATTATTTGG**tt**
tttAT**ttaa**TATTTAT**aatt**TA**tttt**ATAC**aatt**TATTATATTGTTTATA
CCTTATTATTATTATATAATATATTATATTATTATAAT**aatttaattaat**
tATATTA**ttaa**ATA**ttaa**CTAATGTGTGCTCTATATATATTATTCATTCT
AGTTTCTAATCACCCACCCCCTCCCCCTATTACTTATATATCTAGAAAT**a**
aaaaTACATAACATATA**ttttaaa**TATATATATAT**aatt**ATATAAT**aatt**
ATTATATAT**aaaa**TATATATATATATAATATATATTTAT**aaaa**TAATAAT
AATAAATATTATTACTCCATTAGAGG**tttt**GGTCCCATATCAGGAACCGA
AACTATAATAATATATAATATTATAATAAAGATATTCTTATTTATAATAT
ATTA**ttaa**AT**aattaa**TAAT**aatt**ATAATATATATATATAATATATTATA
ATATATTTATTCGAGAAC**ttttt**ATTTATTAT**aaaa**T**aaaa**TA**tttt**ATT
TATTATTTAG**tttttttttA**ttaa**ACA**tttt**AT**aaaaa**TATAAATG**ttaa**
TAATATTATGA**ttaa**TAAGTAATAAT**aatt**tATTTA**ttttt**A**ttaatt**AC
TTCTTCGAGGTATTAGTATCAGTATCAGTATCAGTATCGT**aaaaaa**CGGG
TGACT**aaaa**TATATATATATATa**aatt**ATAAAT**aaaaa**TATTATAAT**aat**
tttaaaTAAATAAATATCAATATATTATTATTATTTATATTATAATAAAT
ATTATCTAATAATAGTCCGGCCCGCCCCCGCGGGGCGGACCCCGAAGGAG
TCCGAACCCC**tttttt**AT**ttaattttA**t**ttaaa**GAAGGAGTGAGGGACCC
CTCCCGTTAGGGAGGGGGACCGAACCCCGAAGGAGAT**aatt**AGATAT**aat**
tATA**tttt**A**tttt**ATAT**aatt**ATATAATATTATATAATAAT**aatt**ATATA

ATAAG**ttaa**TAAT**aatt**ATATAATAAG**ttaa**TAATAATCATATCTCCTTT
ATAAATGAAC**ttttA**t**taa**ATATA**ttttA**t**taa**ATA**ttaa**ATATA**ttttt**
tATAATA**ttaa**ATATA**ttttA**t**taaaa**TAT**ttaa**TATA**ttttA**t**taa**ATA
ttaaATATA**ttttA**t**taa**ATA**ttaa**ATATAAATAAAGGTTTATATTAT**aa**
ttCATTATTTATATCTTCTTTAT**aattaa**TATTCGTATTAGATCCTTAT**t**
taattTATAATCC**tttaaaaaa**C**ttttaa**TAAATATAATATAATATATAT
AT**aatttttt**ATTA**ttttt**ATATTA**ttttt**ATTA**tttaaa**TATATTATATA
TTTCATTATAAT**aatt**A**tttaaaaa**GTTAT**ttaa**TAAATAATCTGATATT
ATA**ttttt**AT**aattaattttt**ATTTA**ttttt**ATTTATTATATATATTATTATA
TAT**aattaaatt**AT**aatt**AC**aatt**ATAACTAT**aattaattaattaattaa**
ttGGA**ttaattaattaatt**GGGCGCCAAGCCGGTTGTTCACCGACTTGGT
CCCAATATAATATGAGATAATATAATATACTATATGATATAACATAAATA
TAATATATTATATGATATAACATAAATATAATATACTCCTTCGGGGTCCG
CCCCCGCGTGGGCGGACCGGACTATATGAATATATTATTATTAT**aatt**AT
a**aatt**ATAATAAATAAAT**aaattt**CT**ttaa**T**aatt**A**ttaattaa**TATTA**t**
taattTATTTACAAATA**ttttA**t**taattttt**A**ttttt**A**ttaa**ATATAAAT
ATATAAATATATATATATTTATTTATAATATTATTTATATTTATTATATA
TTATTA**ttaa**ATATA**ttttt**ATTATATATCA**ttaa**ATA**ttaa**TATGTTAT
TATAGTGGTGGGGGTCCC**aatt**ATTA**tttt**CAATAAT**aatt**ATTATTGGG
ACCCCGGATATCTTCTTG**ttaa**TC**aatt**ATTATATTAT**ttaatt**TA**tttt**
ATTTCTTATTTAT**aatt**TATATTATAT**aatt**TATTATATTG**ttaa**TACTC
CTTCGGGGTCCCCGCCGGGGCGGGGAC**tttt**ATTTATATTA**ttaatt**ATA
TTATATTATTATAATATAT**ttaatt**GATTATATTATa**aatt**ATAACTAAT
GTATGCTTTGTATTTATTGAATAGTTTGGTTCTTATCACCCACCCCCTCC
CCCTATTACTTCTCCGAGGTCCCGGTTTCGTAAGAAACCGGGACTTATAT
ATTTGGT**aattaaaaa**TATAACTTATATAAATAT**ttaa**TAAATATATA**tt**
aaATATATTATTA**ttaa**T**aatt**TATTATTATAT**aaaaaaa**TAATAAATAT
TA**ttaa**TGAT**ttaatt**ATATAAATA**ttaatt**A**ttaa**ATAAAT**aatt**ATAC
TTTCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAAC**tttaaa**
aTAATATATATATATAT**aaaa**GTA**ttttA**TAAT**aatt**AGT**ttaatt**ATTA
TTC**tttttttt**A**ttaa**ATAT**aaaa**TCA**ttttA**GGTTA**ttaattttt**ATT
TA**ttaaaaa**T**aatttt**AT**aattaa**TATTTCTCCTTTC**ttaaaa**TAAATAA
TATTATTATTAT**aatt**A**ttaattaa**TGAATACTCTTCTC**tttt**GGGGTTC
GGTCCACCCTCCCGTATACTTACGGGAGGGGGGTCCCTCACTCC**tttt**GA
GACT**ttaattttt**ATAAATATAAATATAAATATAATAAGATG**ttaa**CTC**tt**
ttATAAATAAATAATAAATAT**aatt**CTA**ttttttaa**TAATAATATATAATA
tttttATAAT**aaaa**TATATAAATAATAATATTTATATATATATATATAC**t**

tttttttATATAAGAATAATATATATAGTTCACATTGGAGGCGAGT**aaaa**
GGAGATAAGAAATATAATATAATATAATAAT**aaaaa**TATAATGAATAATA
ATAAT**aaaattt**ATATAATAAC**aaaa**TAGTCCGACCGAAGGAGATGAGAT
TA**ttaa**TATTA**ttaa**ATAAT**aaaa**TGTA**ttaatt**AT**aaaa**TAT**aaaa**CCT
ATAAAT**aatt**TATAATAT**aatt**TATATTATGATAATAATAATATATATAT
TATAATA**tttt**ATATATATATTTATTATATTTATATTTATAT**aaaaaa**GT
GATATTGA**ttaattaattaatt**TAT**aattaa**T**aatt**A**ttaa**TATAGTCCG
GCCCGCCCCCGCGGGGCGGACCCCGAAGGAGTCCGGCCGAAGGAGTTTAT
TATATTATA**ttaa**ATAAGATTTATAATAT**aattaa**TATATA**ttttaa**TAA
ATAT**aaaa**GATTATATTATATTAT**aaaaa**GTATA**tttt**ATATTTATA**ttt**
tATTTATTATTATTATTATATATATAAGTAGT**aaaaa**GTAGAATAATAGA
TTTGAAATATTTATTATATAGA**tttaaa**GAGATAATCATGGAGTATAAT**a**
attaatttaaT**aatttaa**TATAACTA**ttaa**TAG**aatt**AGGTTACTAAT**aa**
ttaaTAAC**aattaattttaaaa**CCTAAAGGTAAACCTTTATA**ttaa**TAAT
GTTA**tttttt**ATTA**ttttt**ATAATAAGAAT**aatt**A**ttaa**TAATAATAAAC
TAAGTGAACTGAAACATCTAAGTAAC**ttaa**GGATAAGAAATCAACAGAGA
TATTATGAGTATTGGTGAGAG**aaaa**TAATAAAGGTCTAATAAGTATTATG
TG**aaaaaaa**TGTAAG**aaaa**TAGGATAAC**aatt**CTAAGACTAAATACTA**tt**
aaTAAGTATAGTAAGTACCGTAAGGGAAAGTATG**aaaa**TGATTA**tttt**AT
AAGCAATCATGAATATATTATATTATA**ttaa**TGATGTACC**tttt**GTATAA
TGGGTCAGCAAGT**aattaa**TATTAGT**aaaa**CAATAAGTTATAAATAAATA
GAATAATATATATATAT**aaaaaaa**TATA**ttaaaa**TAT**ttaattaa**TA**tta**
attGACCCGAAAGCAAACGATCTAACTATGATAAGATGGATAAACGATCG
AACAGGTTGATGTTGCAATATCATCTGA**ttaatt**GTGGTTAGTAGTGAAA
GACAAATCTGGTTTGCAGATAGCTGG**tttt**CTATGAAATATATGTAAGTA
TAGCCTTTATAAATAAT**aatt**ATTATATAATATTATA**ttaa**TATTATATA
AAGAATGGTACAGC**aattaa**TATATATTAGGGAACTA**ttaa**AG**tttt**A**tt**
aaTAATA**ttaa**ATCTCGAAATAT**ttaatt**ATATATAATAAAGAGTCAGAT
TATGTGCGATAAGGTAAATAATCTAAAGGGAAACAGCCCAGA**ttaa**GATA
TAAAGTTCCTAATAAATAATAAGTGAAATAAATA**ttaaaa**TATTATAATA
TAATCAG**ttaa**TGGGTTTGACAATAACCA**ttttttaa**TGAACATGTAACA
ATGCACTGATTTATAATAAAT**aaaaaaaaa**TAATA**tttaaaa**TCAAATAT
ATATATATTTG**ttaa**TAGATAATATACGGATC**ttaa**TAATAAG**aatt**AT**t**
taattCCTAATATGGAATATTATA**ttttt**ATAAT**aaaaa**TATAAATACTG
AATATCTAAATATTATTATTAC**tttttttttaa**TAATAATAATATGGTAA
TAGAACAT**ttaa**TGATAATATATATTAGTTA**ttaattaa**TATATGTA**tta**
attaaATAGAGAATGCTGACATGAGTAACG**aaaaaaa**GGTATAAACC**ttt**

tCACCT**aaaa**CATAAGGT**ttaa**CTAT**aaaa**GTACGGCCCCT**aattaatta**
aTAAGAATATAAATATAT**ttaa**GATGGGATAATCTATA**ttaa**T**aaaattt**
ATC**ttaaaa**TATATATATTA**ttaa**T**aatt**ATA**ttaattaattaa**TAATAT
ATAT**aatt**ATATTATATATTATATA**tttttt**ATATAATATAAACTAATAA
AGATCAGGAAAT**aattaa**TGTATACCGTAATGTAGACCGACTCAGGTATG
TAAGTAGAGAATATGAAGGTG**aatt**AGAT**aattaa**AGGGAAGGAACTCGG
CAAAGATAGCTCATAAGTTAGTCAATAAAGAGTAATAAGAACAAAGTTGT
ACAACTGTTTACT**aaaaa**CACCGCACTTTGCAGAAACGATAAGT**ttaa**GT
ATAAGGTGTGAACTCTGCTCCATGC**ttaa**TATATAAATa**aatt**AT**ttaa**C
GAT**aatttaattaatt**tAGGTAAATAGCAGCCTTATTATGAGGGTTATAA
TGTAGCG**aatt**CCTTGGCCTAT**aatt**GAGGTCCCGCATGAATGACGTAAT
GATACAACAACTGTCTCCCCT**ttaa**GCTAAGTG**aatt**GAAATCGTAGTGA
AGATGCTATGTACCTTCAGCAAGACGGAAAGACCCTATGCAGCTTTACTG
T**aatt**AGATAGATCG**aatt**ATTGTTTATTATATTCAGCATA**ttaa**GTAAT
CCTATTATTAGGTAATCGTTTAGATA**ttaa**TGAGATACTTATTATAATAT
AATGAT**aatt**CTAATCTTATAAAT**aatt**ATTATTATTATTA**ttaa**TAATA
ATAATATGCTTTCAAGCATAGTGAT**aaaa**CATATTTATATGATAATCACT
TTAC**ttaa**TAGATAT**aatt**C**ttaa**GTAATATATAATATATA**tttt**ATATA
TATTATATATAATATAAGAGACAATCTCT**aatt**GGTAG**tttt**GATGGGGC
GTCATTATCAGC**aaaa**GTATCTGAATAAGTCCATAAATAAATATATa**aat**
tATTGAAT**aaaaaaaaaa**TAATATATATTATATATA**ttaatt**AT**aatt**GA
AATATGTTTATAT**aatt**tATATTTATTGAATATA**tttt**AGTAATAGAT**aa**
aaaTATGTACAGTa**aatt**GTAAGG**aaaa**CAATAATAACTTTCTCCTCTCT
CGGTGGGGGTTCACACCTA**ttttaa**TAGGTGTGAACCCCTCTTCGGGGT
TCCGGTTCCCTTTCGGGTCCCGGAAC**ttaa**AT**aaaaa**TGGAAAG**aattaa**
ttaaTATAATGGTATAACTGTGCGAT**aatt**GTAACACAAACGAGTGAAAC
AAGTACGTAAGTATGGCATAATGAACAAATAACACTGATTGTAAAGGTTA
TTGATAACGAAT**aaaa**GTTACGCTAGGGAT**aatt**TACCCCCTTGTCCCAT
TATATTG**aaaaa**TAT**aatt**ATTC**aattaatt**AT**ttaatt**GAAGT**aatt**GG
GTG**aatt**GCTTAGATATCCATATAGAT**aaaaa**TAATGGACAATAAGCAGC
GAAGCTTATAACAACTTTCATATATGTATATATACGGTTATAAGAACGTT
CAACGACTAGATGATGAGTGGAG**ttaa**CAAT**aatt**CATCCACGAGCGCCC
AATGTCGAATAAAT**aaaa**TA**ttaa**ATAAATATCAAAGGATATATAAAGA**t**
ttttaaTAAATC**aaaaaa**T**aaaa**T**aaaa**TG**aaaaa**TA**ttaaaaaaaa**TCA
AGTAAT**aatt**tAGGACCT**aatt**CT**aatt**A**ttaaaa**GAATATAAATCAC**aa**
ttaattG**aattaa**ATATTGAAC**aatt**TGAAGCAGGTATTGGT**ttaatttt**
AGGAGATGCTTATATTCGTAGTCGTGATGAAGGTAAACTATATTGTATGC

aattTGAGTG**aaaaaa**TAAGGCATACATGGATCATGTATGTTTATTATAT
GATCAATGAGTATTATCACCTCCTCAT**aaaaaa**GAAAGAG**ttaa**TCATTT
AGGT**aatt**TAGT**aatt**ACCTGAGGAGCTCAAAC**ttttaaa**CATCAAGC**tt**
ttaaT**aatt**AGCTAACTTATTTATTGTAAATAAT**aaaaaa**CTTATTCCTA
AT**aatt**TAGTTGa**aatt**AT**ttaa**CACCTATAAGTTTAGCATATTGATTTA
TAGATGATGGAGGTAAATGAGATTATAAT**aaaatt**CTC**ttaa**T**aaaa**GTA
TTGTA**ttaa**ATACACAAAG**tttt**AC**tttt**GAAGAAGTAGAATATTTAG**tt**
aaAGGT**ttaa**GAAAT**aatt**tC**aattaatt**GTTATG**ttaaattaa**T**aaaaa**
TAAACC**aatt**ATTTATATTGATTCTATAAGTTAT**ttaatttttt**AT**aatt**
taattaaACCTTAT**ttaatt**CCTCAAATGATATAT**aatt**ACCTAATACTA
TTTCATCCGAAAC**ttttttaaaa**TAATATTCTTA**ttttt**A**tttt**ATGATA
TATTTCATAAATATTTATTTATA**ttaatttt**ATTTGATAATGATATAGTC
TGAACAATATAGTAATATATTGAAGT**aatt**A**tttaaa**TGT**aatt**ACGATA
AC**aaaaattt**GAACAGGGTAATATAGCGAAAGAGTAGATATTGTAAGCTA
TGTTTGCCACCTCGATGTCGACTCAACATTTCCTCTTGGTTGT**aaaa**GCT
AAGAAGGGTTTGACTGTTCGTC**aattaaaa**TGTTACGTGAGTTGGG**ttaa**
ATACGATGTGAATCAGTATGGTTCCTATCTGCTGAAGGAAATATTATC**aa**
ttaaATCTCATTATTAGTACGCAAGGACCATAATGAATCAACCCATGGTG
TATCTATTGATAATAATATAATATAT**ttaa**T**aaaaa**TAATACTTTA**ttaa**
TATATTATCTATATTAGTTTATA**ttttaatt**ATATATTATCATAGTAGAT
AAGCTAAGTTGATAATAAATAAATATTGAATACATA**ttaa**ATATGAAGTT
G**ttttaa**TAAGAT**aattaa**TCTGAT**aattttA**TACTa**aattaa**T**aatt**AT
AGG**tttt**ATATATTATTTATAAATAAATATATTATAATAATAAT**aatt**AT
TATTA**ttaa**T**aaaaaa**TA**ttaatt**ATAATA**ttaa**T**aaaa**TACT**aatt**TAT
CAGTTATCTATATAATATCTAATCTATTATTCTATATACTTATTACTCC**t**
ttttaattaattaaaaaGGGGTTCGGTTCCCCCCCCCCATAAGTATGATT
AT**aatt**AT**aatt**ATAATATAAGGGAGGGGTCCCTCACTCCTTATGGGGTC
CCGGTTGGACCGAGACTCCTCCCTTGCGGGATTGGTTCACACCTTTATAA
ATAAATAATAAATAATAAATAAAGGTGTTCACTAATAAATATATATATAT
ATATATATATATTATATTATAATATTAT**ttaa**TAC**ttaa**TATATTATATA
ttttATAT**ttaa**TAAAT**aaaaaaaa**TA**ttaa**TAAATAATAATA**ttaa**TAA
TAAAG**aatt**AT**aattaa**TACCCTCTATATAT**aatt**CT**aattaattaatta**
aATATTTATATATAATAATCAATATATTA**ttaatttaa**T**aatt**ATTATAA
TAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGTTTAT**aa**
aaGATATA**ttttt**ATATTATATTATATTATATT**ttaa**TAAATATTACC**ttt**
ttttATTATTA**ttttt**ATATATTATATAATATTA**ttaatttttt**ATTATAA
TATTATTTAC**tttttt**ATTGGATTATTTATTTATTTATTTATTTA**ttaat**

taattaattaaATATTTA**ttaattaa**TATATATA**ttaa**ATA**ttaa**TATTT
CA**ttaaaaaaaa**GAGATATATGAATAATATATTATGTTATATTATATTAT
AT**aatt**ATATTA**ttttt**ATAATA**ttaa**T**aattaaaaa**TAAGAACTTA**ttt**
aaaattAT**aatt**ATGATAAT**aattaa**TAC**ttttttaatt**TAT**aaaaa**TAT**a**
atttCTTTACATATATATATATATATTATTATTATTTATA**ttaa**TCAT
aattttaaTATTTATAAT**aatt**tATAT**aaaa**TC**aatt**ATAATATTATATA
C**ttttt**ATATACTTTATAATCTTTATATCTTCACCCCCCC**tttttttaa**TA
ATATATTATA**ttaaaaa**TATAAT**aatt**TATATGATTTA**ttaa**TAC**ttttt**
ATAT**aatt**ATATTATTA**tttttttt**ATAGATGTTATATTA**tttttt**ATA
AT**aatttttttt**A**tttaaa**T**aaattt**ATAACTCCTTC**ttaattaa**AGAT
aaaaGGGGTTCCCCCC**ttaa**GTATAAGTATAAGTATAAGTATAAGTATAA
GTATAAGTATAAGTATAAGTATAAGTATAAGTATAGTATACGGGGGGGGG
GTCCCTCACTCCTTCG**ttaatt**TATATATATTA**ttaa**T**aatt**AT**ttaatt**
tttATTATTTATTTATATAT**aaaaa**TATTCTa**aatt**A**ttaa**TATTTATAA
TAGAATAAATATTATAAAGTAT**aatt**ATAAAT**aattaatt**A**tttaaa**TAA
TAATAATATATTTATTATTATATAATAAATATATTATAAATAATAGTTAT
ATTAGC**ttaatt**GGTAGAGCATTCG**tttt**GTAATCGAAAGGTTTGGGGTT
CAAATCCCTAATATAACAATAATAATAAT**aaaa**TA**ttaaaa**TAAATATAA
TATTTAT**aaaaattt**A**ttaatt**TATAT**aaaaaa**TATATATATAAATAAT**a**
attATAAT**aaaa**CA**tttt**ATAATCAAT**aatttaa**TAAATAATCTTCTTAT
TATAATATTATG**tttaaa**TATTACTCTTTATGAGGTCCAACAAACTAATA
AGATATAAATATATATATATTATATAATAATAATAATAATAATATATTAT
ttaaTATATTATCAAGAAGATAAATATAAATAATA**ttttaa**T**aattttaa**
aTAAATCT**aatt**TATATA**ttaa**T**aatttaa**TAATC**ttaa**TATTTATTATC
ATTATTTCATATTTATATTATATAAATATTTA**tttaaa**T**aaaaaa**TA**tta**
aAGAGTTTA**tttt**ATTTATTAT**aatt**AT**ttaa**T**aaaa**TATATATAATAAT
ATATAGAATAAAGATATAAAT**aatt**ATAAGTATATAAAGTAATAAAGGAG
ATGTTG**ttttaa**GG**ttaa**ACTATTAGATTGCAAATCTACTTA**ttaa**GAGT
TCGATTCTCTTCATCTC**ttaa**ATAAATAATATAATAAT**aaaa**TATTATAG
TTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGGAAGGAGATAAATATAT
ATATATTTATAAT**aatt**ATATAATAAAGGTGAATATATTTCAATGGTAG**a**
aaaTACGCTTGTGGTGCG**ttaa**ATCTGAGTTCGATTCTCAGTATTCACCC
TATAAATAATAATAATAATATA**tttt**ATTATTC**ttaatttttt**ATTCTTT
ATATTATATATATAATA**ttaa**TATTATTAC**tttttaa**TAAC**aaaa**TATTA
T**aattaatt**GATATATATATATACCAAATAT**aattaatt**G**aattaa**ATAA
TAAAT**aaaa**TATTTACTTCTTTA**ttaaatt**CT**aattaatt**GATTC**ttttt**
ATTGAATA**ttaatt**CTATTATAACTTA**ttaattaattaattaattaatt**A

TAATAATAATAATATTTATTA**ttaatt**A**ttaa**ATATTTATTATTATATAT
AAGAT**ttaattttaaa**TA**ttaa**T**aaaaaaa**GAAT**aaaa**T**aaaa**T**aaaa**TG
AATAATATTTCTTTATCTCTTTCGATCGGACTCCTTCGGCCGGACTCCTT
CGGGGTCCGCCCCGCGGGGCGGGCCGGACTATTTATTATTATAATATAAT
AT**ttaa**TCAATAGATTTAT**aatt**TAT**ttaa**TGAATA**tttt**ATAAATATAT
aaaaC**aatt**CC**ttttt**ATTATTAT**aatttttt**CATTATTTATTTATTTATT
TATTTATTTATTCAATATAT**aaaaa**T**aatt**AT**aaaaa**GATTA**ttaaaaa**T
AAT**aatttaa**TGATAAATATATATTATATATA**ttaa**TAT**aaaaa**TAATAA
ATATAAATATATTATGTAAATATTATAT**aatt**tGTATATGTATATATTAT
AATAATGTTATATAAGTAATAATATAAT**aaaa**TA**tttt**ATGT**aatt**TATA
TATATTTAT**aatt**AT**aaaa**T**aaaaa**TATTATAAATAATa**aattaa**TAATA
ATAAT**aattttaa**T**aaaa**T**aatt**ATATAT**ttaattttt**ATTATGAAGTTTA
TAC**ttaa**TAT**aatt**ATATTTCCTTTAT**aatt**A**ttaa**TATATCC**ttttttaa**
ttaaAT**aaaa**T**aaaaa**TATTATAAATA**ttaa**T**aattaatttttt**ATTTAT
ATTTATATATATA**ttaa**AGA**ttaa**ATATATTA**ttaa**TACT**aatt**TAT**aat**
tTATTA**ttaa**TAAATAGTCCGGCCCGCCCCCTGCGGGGCGGACCCCGAAG
GAGTTCGAC**ttaatt**AT**aatttaa**T**aatttttt**ATTTA**ttaa**TAGTTTCGG
GGCCCGGCCACGGGAGTCGGAACCCCGAAAGGAG**tttt**ATTA**ttaa**TAT**a**
aaaaGAGTAAGGATAATAAT**aattC**t**ttttaatt**TA**ttttttaa**T**aaaa**TAT
aattttaaaaTAG**ttttt**ATAGTCCGGCCCGCCCCGCGGGGGGGGGCGGA
CCCCGAAGGAGTTCGGTCTGGCA**ttaatt**ATAAT**aatt**ATA**ttaa**TATTA
TTATTATTTATTATATTATAATATATTTATTTATA**tttt**ATAATA**ttaa**T
aattA**tttt**ATAT**ttaa**TAAATATAATATATATATTA**tttttttttaa**TAA
CTATCT**aattaa**TAGCTA**tttt**GGTGG**aatt**GGTAGACACGATACTC**ttta**
aGATGTATTACTTTACAGTATGAAGGTTCAAGTCC**tttaaa**TAGCAATAA
ATATATATAATATATATAATATATATAAATGAGTCGTAGACTAATAGGTA
AGTTACC**aaattt**GAGTTTGGAGTTTGTTTGTTCGAATCAAACCGATTCA
ATATTATAATATATATATTATTTATATATAAATATAT**aatt**ATACTCCTA
tttttATA**ttaattaattaa**TAATATATGATAATAT**aaaatt**ATTG**aatt**
A**ttaa**CTCTTA**ttaa**TAATAATAATAATCATAATAATAATATATATATAT
ATAGTATATATAT**aaaa**G**tttt**ATTATATTATATTATATTATATATTTAT
TTATATAT**aatt**CTTA**ttaattG**a**aaaaaa**GAAT**aattaa**TAATCTTA**tta**
aaaaaaTAAATACTTTCA**ttttAttttAttttA**T**ttaatttaatt**ATAAT
ATATAAATA**ttaaaaaaa**GGATATAAG**tttttt**ATAAGATATAATATATA
TATATA**ttaa**ATATAAAGAAG**ttaa**TATTTATA**ttttaatt**AT**aaaa**TG**t**
taaTACTCCTTTGGGGACTTA**ttaattaatt**A**ttaattaa**TAAT**aatt**TA
TGATTTATAAATAATAAATAAAGGAATAAGTATC**aattaattaa**TATATT

ATAT**ttaa**TA**tttt**ATAT**ttaa**TAT**ttaa**TAT**ttaa**TA**ttttaa**GTTCCG
GGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGTAGTA**ttaatt**ATGG
ATAGTGAGGGTGGAT**ttaa**TCC**tttt**GTTATGTTA**ttaattaattaatta**
attTATATATAT**aaaa**TA**ttttaattaatttt**ATATAAATATATATATA
TATATATA**ttaa**TAATAGTCCGGCCCGCCCCGTGGGGCGGACCCCAAAGG
AGTAATATATATTATGTATAAACAATAGAGAATATTGT**ttaa**TGGT**aaaa**
CAGTTGTC**ttttaa**GCAACCCATGCTTGGTTCAACTCCAGCTATTCTCAT
AATATTATATATATATATTTCCCTTTCT**aaaaa**TAATAAT**aatt**ATATAT
AATAATAATAT**aatt**ATATATATATATATTATAATAATAATAATAATAAT
AATAATAAATAATAAT**aatt**A**ttttt**A**ttaa**TAATA**ttaa**TATATTAT**aa**
ttA**ttaa**TAAATA**ttaa**T**aaaaa**TAGCTCTCTTAGC**ttaa**TGG**ttaa**AGC
ATAATACTTCTAATA**ttaa**TATTCCATGTTCAAATCATGGAGAGAGT**aat**
tATATTATA**ttaa**TAATCCCCCCCCCA**tttttaattaattaa**GAAGT**tta**
attTACTAT**ttaa**TAATAAATGAAATAATAATAATAGATATAAG**ttaatt**
GGTAAACTGGATGTCTTCCAAACATTGAATGCGAGTTCGATTCTCGCTAT
CTAT**aattaa**TA**ttaa**TAT**aattaa**TATCCTAT**aattaattaa**ATACa**aa**
ttATA**ttaaaa**CTTATATTATATTATATTATAATATTATATTATTATTAT
AT**aaaaa**TATAATAATAATAATAT**ttaattttA**T**ttaa**TAATAATA**tttt**
ATATAAT**aaaa**TAATCATATTTATAATAT**ttaa**TA**ttaa**TAAT**aatt**TAT
TATAAT**aatt**CT**ttaa**TATACTTATTTATTATTA**ttttaa**TAAATAAATA
T**aatt**CTTATAAATATATTATAAC**aaaa**TATATTATA**ttttaattaa**ATA
CAATATTATAAATATATATATATATATAAATATTTATAT**aaaaaaaaaa**T
aaaaaTA**ttttaa**T**aatt**ATTCTTTATAAATAAATAATGATAATAAT**aat**
tTATAATAATCTCCTTGTGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAA
CTATAATATA**ttttaa**TATA**tttttt**ATTACTCCTCCTTTGGGGTCCGCC
CCGCGGGGGCGGGGCGGACTATAAT**aatttttt**ATTGAT**aaaaaa**GTATA
TATAATAT**aattaa**TATATTTC**ttttt**ATAT**aatt**ATAAATATTA**tttt**A
TAAT**aaaaaaa**GTATATATAATATTATATAT**ttaa**TAAATAATATAATAA
TAATATAAATAAATATATATATATTA**ttaa**TATA**ttaattttA**TAATAAT
aattATAATAATAGTAGTAGGTAT**aattttaa**TAAAGAG**tttt**ATTCCAA
TGGAGTAATAATAATAATAATAAT**aaaa**TAAAGGATCTGTAGC**ttaa**TAG
TAAAGTACCA**tttt**GTCATAATGGAGGATGTCAGTGCAAATCTGATTAGA
TTCGTATATTTATAC**ttaa**TAT**aaaaaaaa**TAAATAATAATC**tttttt**AT
TATTATATTTA**ttaa**TAAT**aatt**A**tttt**GTTATTATTA**ttaatt**TATA**tt**
aaTA**tttt**ATAT**aatt**ATTTAT**ttaa**TCTTTCATTATATAT**ttaa**TATAT
TA**ttaa**TA**ttaattaa**TA**tttt**ATAATAAATAAAT**aaaa**T**aaaa**TAAATA
ttttaaTATAATACTCCTTCGGGGTTCGGTCCCCCTCCCATTAGTATAGT

ATAGGGAGGGGTCCCTCACTCCTTCGGGGTCCCCGCCGGGGCGGGGACTT
A**ttttt**ATATTTA**ttaa**TAAT**aattaatttttt**ATAT**aatt**tATTATTTCT
TACAATATATTTATTACTATTA**ttttttaa**TAATCTTATATATAATATAT
aaaaTATATATATATTATATATATATATAAATATAATATATATTATTATA
AATATTTATAATCTTA**ttaattaatt**AGATTATATTATATTATATTAGAT
CATATTATATTATATTATATTATATTATATTATTATTA**ttaa**TA**ttttt**A
tttttA**tttt**ATAT**ttaa**TAGT**aaaaaa**TCAT**aattttA**T**aatt**TA**ttaa**
ttATTATAT**aatt**TCA**ttaa**TATATTTCTTC**ttttt**ATTTATTTATTTAT
TACTTA**ttaa**TAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAG
G**aaaa**TAATAT**aaaaaa**T**aatt**AT**aatt**TATTAT**aatt**TA**ttaatt**TA**tt**
aattTA**ttaatt**TATTTA**ttaatt**TA**ttaatt**TATTTATTATTATA**tttt**
ttttaaTAAAGGa**aattaa**CTATAGGTAAAGTGGATTATTTGCTAAGT**aa**
ttG**aatt**GT**aatt**CTTATGAGTTCGAATCTCATA**ttttC**CGTATATATCT
ttaatttaaTGGT**aaaa**TATTAGAATACGAATCT**aatt**ATATAGGTTCAA
ATCCTATAAGATATTATATTATATTATATAATATTATATA**ttaa**TAAATA
TTA**ttaattaatt**TATTTATTTATTTATTA**ttaa**AT**aaaaa**TAT**ttaa**TA
GTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGAATAATAT**a**
aaaTATTAT**aatt**ATTTATATA**ttaatt**A**ttaatt**ATTTATTATTTATTA
TAT**aaaaa**GTATAT**aattttA**TA**ttttaa**TATAGGG**ttaattaattaatt**
A**ttaatttttt**ATAATAAGATAATAATATA**ttaaaaa**CTTATTAT**aattt**
AT**aaaa**TAATATTTATTTACTTTGATATTA**ttttttaa**TCTTTCA**ttaa**TA
TATA**tttt**ATTATAAGTAATAATATAGT**ttaatttaattaa**TATAAAT**aa**
ttACATAAGAATAATATTATAATAATATTATATATTATATAAAGAAATAA
T**aatt**TATA**ttttt**A**tttttttt**ATAAATAATATAAATATAAATATAATG
GGGTTATAG**ttaatt**tGGTAGAACGACTGCGTTGCATGCAT**ttaa**TATGA
GTTCAAGTCTCA**ttaa**CTCCAAT**aatt**ATATTATATAATATATATA**ttaa**
T**aatt**ATATATATATATATATATATAAATA**ttaa**ATAAATATTATA**ttaa**
TAAATAATAT**aatt**ATCTAATCGAAGGAGATATTTATAATATAATATAAA
TA**ttttaa**T**aattaa**TAAATATTATA**ttaa**TAAAT**aattaa**TAAATATAT
aattATAAT**aattttaa**TATTATTATAT**aattaattaa**ATATAAT**aatta**
aTGAAATAGAAACTAT**aattC**a**attG**GTTAGAATAGTA**ttttG**ATAAGGT
ACAAATATAGGTTCAATCCCTGTTAGTTTCATATTATATATCA**ttaa**TAT
AT**aaaa**TATAAATATATATATTATAATAATAATAATAATAAATATAAATA
T**aatt**ATATATATATATATATATAAATAAAT**aattA**T**ttaatt**TATAATA
AATATATATAGTTCCCGCGAAGCGGGAACCCCATAAGGAG**tttt**ATTA**tt**
aattATAT**ttaa**TAAATA**ttaatt**A**ttaatttt**ATATTTATAAAT**aattt**
ATTACTCCTTC**ttaattaa**GAAT**aaaaa**GGGATGCGGTTCCCATGGGGTC

CCGCACTCCTTCGGGGTCCGCCCCCTCCCCTGCGGGAGGGGAGCGGACTA
ttttA**ttaaaaa**TATTAT**aattaa**ATAATAATATAAAT**aatt**TATAATAT
AATAATATATACTTATAAATAATA**tttaaa**TCTTATTA**ttaatt**TATAAA
TCAT**aatt**ATTA**ttaa**TAAATATCTC**tttt**AGATAAGAT**aatt**GAACTTA
TATTTATATTATATATATATAGATATAAATC**ttaa**ATAGAGTAAATATAT
TATAAT**aatt**ATATAAATATATATATATTATA**ttaa**GATAATAATATATA
TATATA**ttaa**TATATAAGGAGGGA**tttt**CAATGTTGGTAGTTGGAGTTGA
GCTGTAAACTCAATGACTTAGGTCTTCATAGGTTC**aatt**CCTATTCCCTT
CAT**aatt**TATTA**ttaatt**ATATATTATTATAAATCAAATCCATTG**aatta**
aTATAATCCAATGAAT**aattaatttaa**TACAT**aatttaa**TATATa**aatt**A
TATATATATATACTTTAT**aaaaaaaaaatt**ATATAAT**aatt**ATA**ttaa**TA
TATTTATATATATAAATAAATAAATAAATAATAAT**aatt**AT**aatt**AT**aat**
tAT**aattaattaattaa**TAAATAAATAAT**aatt**TATATTATCTTTATAAT
ATATATATATAC**tttt**AT**aaaaaaaa**TATATAAAT**aatt**CT**aaaa**TGTAT
ATTTCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAACTA**tta**
aTa**attaa**TAAT**aaaa**T**aatt**ATTATCTGTAT**ttaa**T**aatttaatt**ATA
GAGTTATATTTCTATATATTTATATATTTATTTATTTATTCTCCTTCCGG
AACTAAT**aaaa**TATAT**aaaa**TAAGGG**ttttt**ATTTAT**ttaattaa**TATAT
ATTTATTC**tttt**ATATAATATGTCCTTATAGCTTATCGG**ttaa**AGCATCT
CACTG**ttaa**TGAGAATAGATGGGTTC**aatt**CCTA**ttaa**GGACGATAATAA
TATATATATA**ttttaatt**TATATATCATATATATATATATA**ttaa**AG**aaa**
aTAATAT**aaaaa**GTATGTA**ttaa**TAATAATAATAAATAAATAATAATAAA
T**aatttt**ATTATATTATATTATATTATATTTATTGATATATTTATTGATA
TTTA**ttaatttaa**GATTATTCA**ttaa**ATATAT**aatt**A**ttaa**T**aatttaa**T
ATA**tttt**AT**aatttttt**ATTATA**tttt**ATGTAAGAAGAAACTA**tttt**ATAT
ATTATATATATATATAT**aatttttt**AT**aaaa**TGAT**aatttt**ATATTATAAA
TATTA**ttaaaa**TA**ttttt**ATAAATAT**ttaatt**ATTTAT**aaaaa**GGTATAT
AATAAT**aatt**A**ttaa**TATTATATTATATTATATATTTATTTATATTATAT
ATAATAATATATTTATATATATA**ttaattaa**T**aattaa**ATAAGTATCTAT
A**tttt**ATATTATATTATATTA**tttt**A**tttt**A**ttaatt**CCGGAAGGAGAAT
aaaaaGTATTCTAAAG**aatt**ATATATTTATTA**ttttt**A**ttaa**TATGTTAT
aattaaT**aaaaaa**TAAATATGTATATAT**aatt**ATATTTATTATGT**ttaat**
tATTTAT**aatt**TATTATAATATATAGTATAAGATATCTTATTTATATTTA
TATATAATAAAGAATATTA**ttaa**ACTAACACCTATATTATATATATTATA
TTATATAATATTATATATATA**ttaatt**ACTAAGAAT**aatt**tAT**aatt**AGA
TAATATTTATATTTATTTATTTAT**ttaattaa**CAAATATA**ttaa**TA**tttt**
taattaattaaTAATACCTTTATATATATATATATATATATA**ttaatttt**

aattATAT**aatt**ATC**tttttt**A**ttaa**T**aatt**ATAAATATATTATATA**ttt**
tATATAATAAGATTAT**aattttA**T**aatt**A**tttt**A**tttttt**A**ttaaaatt**A
TTATTATTAT**aatt**ATTATATTAT**aatt**AT**aatt**A**ttaa**AGAATATATTT
A**ttaa**TA**ttttaa**T**aattaa**TATC**tttt**ATTTATATTTAT**aaaa**TAAGGT
ATAAATATTGATAATAAAGAGTAAATATTGTA**ttaatt**ATAATAAT**aatt**
AT**aattaa**GGAGCTTGTATAGT**ttaatt**GG**ttaaaa**CATTTGTCTCATAA
ATAAATAATGTAAGGTTC**aatt**CCTTCTACAAGTAATAATGATTATAATA
TTTATATATA**ttaaaa**TAATA**ttaa**TAAAT**aatt**ACTCCTCCTAGCAGGA
TTCACATCTCCTTCGGCCGGACTCCTTCGGGGTCCGCCCCGCGGGGGCGG
GCCGGACTA**tttt**ATTATTA**ttaa**ATAGATGTTCA**ttaa**AT**aatt**ATAAA
TAT**aatt**TATC**ttttaaa**TATATATATATAATATAATA**tttaaa**TATATA
TTATAAATAAATAAATAAAT**aattaattaa**T**aaaaa**CATATAATGTATAT
TTATCTAT**aaaaaa**TA**ttaattaattaa**TATATTATTACAGTTCCGGGGG
CCGGCCACGGGAGCCGGAACCCCGAAGGAGATAAATAAATAAATAAATAT
AAAT**aatt**CTTCTTC**tttaaattaa**AT**aaaa**T**aaaa**T**aaaaa**GGGGGGCG
GACTCCTTCGGGGTCCCGCCCCCCTCCGCGGGGCGGACTA**tttt**A**ttttt**
aaaTATATATTATA**ttaa**TAATATAAATATAAGTCCCCGCCCCGGCGGGG
ACCCCGAAGGAGTATAAAT**aaaattaa**TAATATATTATATATATATTATA
ttaaTAATAATAATAATAATAATAATAATAAATAATAACTCCTTGCTTCA
TACCTTTATAAATAAGGTAATCACTAATATATTATAATAAT**aaaatt**ATA
TATATTATATATAATCTAAATATTATATA**ttttaa**TAAATA**ttaa**TATAT
ATGATATGAATATTATTAG**ttttt**GGGAAGCGGGAATCCCGTAAGGAGTG
AGGGACCCCTCCCTAACGGGAGGAGGACCGAAGGAG**tttt**AGTA**tttttt**
ttttttaaT**aaaa**TATATATTTATATGA**ttaa**TAATATTATATATATTA
TTTAT**aaaaa**TAATATAT**aattttaatt**A**ttttta**T**aaaaaaa**GGTGGG
GTTGATAATATAATATAATA**tttttt**A**tttttaatt**TATAATATATAATAA
T**aatt**ATAAAT**aattttaattaaaa**GTAGTA**ttaa**CATATTATAAATAGA
C**aaaa**GAGTCTAAAGG**ttaa**GATTTA**ttaaaa**TGTTAGATTTA**ttaa**GAT
TAC**aattaa**CAACATTCATTATGAATGATGTACCAACACCTTATGCATGT
TA**tttt**CAGGATTCAGCAACACCAAATCAAGAAGGTA**tttt**AG**aatt**ACA
TGATAATATTATG**tttt**ATTTATTAGTTA**tttt**AGGTTTAGTATCTTGAA
TGTTATATAC**aatt**GTTATAACATATTC**aaaaaa**TCCTATTGCATATAAA
TATA**ttaa**ACATGGACAAACTATTGAAGTTATTTGAAC**aatttttt**CCAGC
TGT**aattttA**ttaatt**ATTGC**tttt**CCTTCATTTA**tttt**ATTATATTTAT
GTGATGAAGTTATTTCACCAGCTATAACTA**ttaa**AGCTATTGGATATCAA
TGATATTG**aaaa**TATGAATATTCAGA**ttttA**ttaa**TGATAGTGGTGAAAC
TGTTG**aatt**TGAATCATATGTTATTCCTGATG**aatt**ATTAGAAGAAGGTC

aattaaGATTATTAGATACTGATACTTCTATAGTTGTACCTGTAGATACA
CATATTAGATTCGTTGTAACAGCTGCTGATGTTATTCATGA**tttt**GCTAT
TCCAAGTTTAGGTA**ttaa**AGTTGATGCTACTCCTGGTAGA**ttaa**ATCAAG
TTTCTGCT**ttaatt**CAAAGAGAAGGTGTCTTCTATGGAGCATGTTCTGAG
TTGTGTGGGACAGGTCATGCAAATATGCC**aattaa**GATCGAAGCAGTATC
ATTACCT**aatttttt**GGAATGA**ttaa**ATGAACAAT**aattaa**TATTTACTTA
TTA**ttaa**TA**ttttttaatt**A**ttaaaaa**TAATAATAATAATAAT**aatt**ATAA
TAATATTC**ttaa**ATATAATAAAGATATAGATTTATATTCTATTCAATCAC
CTTATA**ttaaaaa**TATAAATATTA**ttaaaa**GAGGTTATCATACTTC**ttta**
aaTAAT**aattaatt**ATTGTTC**aaaaa**GATAAT**aaaaa**TAATAATAAGAAT
aattTAGAAATAGAT**aatttttt**ATAAATGATTAGTAGGATTTACAGATGG
AGATGGTAG**ttttt**ATA**ttaattaa**ATGAT**aaaaaa**TAT**ttaa**GA**ttttt**
ttATGG**tttt**AGAATACATATTGATGATAAAGCATGTTTAG**aaaa**GATTA
GAAATATA**ttaa**ATATACCTTCT**aattttt**GAAGAACTAC**ttaaaa**C**aatt**
ATATTAGT**aatt**CAC**aaaa**GAAATGGTTATATTCTAATATTGTAACTA**tt**
tttGATAAGTATCCTTGT**ttaa**C**aattaa**ATATTATAGTTATTATAAATG
aaaaaTAGCTAT**aattaa**T**aatttaaa**TGGTATATCTTATAATAATAAAG
ATTTA**ttaa**ATA**ttaaaaa**TAC**aattaa**T**aatt**ATGAAGTTATACCT**aat**
ttaaaattCCATATGAT**aaaa**TAAATGATTATTG**aattttt**AGG**tttt**ATT
GAAGCTGAAGGTTCATTTGATCTATCTCC**aaaa**CGTAATATTTGTGG**ttt**
taaTGTTTCACAACATAAACGTAGTA**ttaa**TACA**ttaaaa**GCTA**ttaa**AT
CTTATGTA**ttaa**AT**aattG**aaaa**CC**aatt**GATAATACACCATTA**ttaatt**
aaaaaT**aatt**A**ttaaaa**GATTGAGATTCATCTA**ttaattaa**CTAAACCTG
AT**aaaaa**TGGAGTTA**ttaatt**AG**aatttaa**TAGAATAGA**tttttt**ATATT
ATGTTA**tttt**ACCT**aatt**ATATTCA**ttaaaa**TGATATAGTCGTAAAG**aat**
tGATTTCC**aatt**ATG**aaaaa**CACTTATAGAAATCTATAT**aaaa**GGTTTAC
ATAATACAC**ttaa**AGGTTCT**aatt**TA**ttaaattaattaa**TAATAATA**tta**
aT**aaaaaaa**GATATTATTCT**aatt**ATAATATTTCTCCTTTCGGGGTTCCG
GCTCCCGTGGCCGGGCCCCGGAACT**aaaaa**TATTATTGATGATGTA**ttaa**
ATATAAATCTTATCTAT**aatt**AT**aatt**ACCATATCGTATAAATAGTGATA
TTCAACGT**ttaatt**CTATAAATAATAATAATACT**aatt**tA**ttaa**TGTTGG
AGTATTTGTTTATGA**tttaaa**TAATACA**ttaatt**ATAACATTTACTGGTT
ATAGACCAGCAGCTCTTTACT**ttaatt**GTTCTCC**tttt**CGGGGTCCCGAC
TGGGGCCGGGACTAAACATG**aatt**GCTAAATATA**ttaaaaa**TGGTAATGT
ATTTATAAATAAATATA**ttttaaaaaa**TA**tttt**ATTAGA**ttaatt**ATTA**t**
ttttACTTCTTC**ttaaattaaaaaa**GGAGAC**tttttt**ATATTTATAT**aat**
tATATAT**aatt**ATTC**tttt**ATTATAAATATATa**aatt**A**tttt**C**tttaat**

tA**ttttt**AT**aattaattaatt**CTTCATGGCTATAGCCATAAC**ttttaa**TA
ATATTC**tttt**ATTCTTTATTATTATATATATATATATTTATTATTTATTA
TTATAG**aatt**TATATTTAT**aaaaa**TA**ttaa**TA**ttttatttaaaa**TAAATA
ATGA**ttaatt**TAT**aaaa**TATATA**ttaattaa**GTTTCGGGTCCCGGCTACG
GGACCCGGAACCCCCGAGAGGAGTTATTATATTTAT**aattaa**ATC**tttaa**
aTAATATATC**ttaatt**ATTATATTGATA**ttaa**TATTATATTGATA**ttaa**T
A**ttaa**ATATATAT**ttaa**TATTTAGCTTATTA**tttt**ATa**aatt**ATATTTAT
ATATTATAATAT**aattaa**ATATATTAT**aatttaa**T**aatttaa**T**aaaaa**TA
TTC**ttttt**AT**aatt**ATTATAAT**aattaa**ATAAATAATAATAATAAGAAT**a**
attaaTGTAT**aatttttttt**ATAAATATTATATA**ttttt**ATA**ttaa**TAGTT
CCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGAAATA**ttaa**T**aa**
aaT**aaaa**Ta**aatt**ATAATAT**aattaatt**ATAAG**aatt**ATATTTACTCC**tt**
ttAT**aatt**TATATTTATAATATAATATAATAT**aaaa**TAAATATAATATAA
TAT**aaaa**TAAATATAATGTAATAGGTATTCACTCCTCTTTGGGGTTCCGA
TCCCCCATACGGATACGGATACGGATACGAATACGGATACGGATACGGAT
ACGGGGGGCCGTCCCCCAGAAC**ttaa**TATTATATC**ttaa**AT**aattaa**TAT
AAATATAATATATTAT**ttaa**TAATAATAATAAATAAATAAATAAATAAAT
AAAT**aattaa**ATAAATAATAATATTATTAT**aatt**AC**tttttaa**TAAATAA
TA**ttaa**TATAATATTATATTAGTATTATAAATAGAC**ttttt**ATTA**tttt**A
TATATAATATAGTCCGGCCCGCCCCCGCGGGGCGGACCCCGAAGGAGTAA
TATATTATAT**aatt**ATTA**ttttttaatt**ATAAAT**aaaa**TAT**aatt**ATTATT
TATTATAT**aatt**TATATAAATATATATATATATTTATTATATATATAAAT
ATAAATATAAATATAAT**aattaa**TAATA**ttaa**AG**tttt**ATATATA**ttaa**T
ATATTAT**aaaa**GGTTTATATATATATATAATAAGATAAGTAAT**aattaat**
taattaaTAATAT**aaaaa**TATATATTATATATTATG**tttt**ATTTATATAT
ATATATATATTATGTATTATTATATAAATATATATATATATTATATTATA
AGTAATAATAAGTATTATATTATATATAGC**tttt**ATAGCTTAGTGGTAAA
GCGAT**aatt**GAAGATTTATTTACATGTAGTTCGATTCTCA**ttaa**GGGCAA
TAATAATAATATA**ttaattaa**T**aattaatt**TATAATAAATATATTATAAT
aattaaTATATATATATATATAATATAT**ttaa**TACAAAG**aaaa**TATATATTA
TATCTCTTATTTATTTATTTA**ttaa**TA**ttttaa**TAAATATAATATTAT**aa**
aaaaaaGTTTATATATTTAGTTCCGGGGCCCGGCCACGGGAGCCGGAACC
CCGGTAGGAGAAATATAAATATAAATATAAATATAATATAAGTTTGGTAT
TCAT**ttaatt**ATATTAT**ttaattaaaaa**TATTCTAAATAAGAATAAATAT
CAATAAAGGAGTTATAAATATATATATATATA**ttaa**TATATATAT**aaaaa**TA
TATATTATTATTAGTTCCCGCTTTGCGGGAACCCCGTAAGGAGTGAGGGA
CCCCATGGGAACCGAACCCCTAT**ttaa**GAAGGAG**tttt**ATTATAAT**aaat**

ttATATATAT**ttaa**TATAT**aatt**AT**aaaaa**TATTATATAATAAATAATAA
AT**aatt**A**ttaa**TAATAAATAAATATAATAATAATATTATAAT**aatt**tATA
AATGATTATAAT**aatt**tATA**ttaatttttt**A**tttt**GTAAATACTAAGATT
TGAACTTAGATAATATGCACCT**aaaaa**CATACA**tttt**ACCA**ttaatt**ATA
TTTACCTTA**ttaatt**ATAT**aaattt**A**ttaa**ATATATAATATA**ttaatt**AT
AT**aaaatt**A**ttaa**ATAAATATATAATATATTATATAT**aatt**TATAATATA
TATATTATAAATATTATTATATAT**aaaa**TATAATATACTACTTAT**aaaaa**
TATATATATATATATAAATATATATATAAATAAATA**tttt**ATATA**ttaat**
taaAT**aatt**A**ttaa**T**aatttaatt**ATAAAGTAT**aattttt**CAATAGGAATA
TTTATAAGATTATAAT**aatt**ATATG**aatt**ATTAT**aatt**ATATATATATAA
ATAAAT**aaaa**TAAT**aatt**ATAAT**aattaa**TAAGAG**tttt**GGATATATATC
TGTGGAGTATATA**tttt**ATAAAGGAGATTAGC**ttaatt**GGTATAGCATTC
G**tttt**ACACACGAAAGATTATAGGTTCGAACCCTATATTTCCTAAATCTA
GATATAATATTATATCTATC**ttaa**TATAATAATATTTATTTA**ttaa**AT**aa**
aaaaaaaaTAAATAATA**ttaattaa**TATAAGATTC**ttttttaatt**ATAAT
AATAAATAAAT**aaaaa**GAAGATATTATCAATGATTTATA**ttaa**TAATAAA
TATAAATAAT**aaaaaa**TATATATAATATAATATAATAAATATATTTCCT**t**
taaTA**ttaa**T**aattaa**TAATAATAATAATAATAATAAT**aaaa**TA**tttaaa**
T**aatt**ATATTCAATAC**aattaatt**ATTTATATTA**ttaa**T**aatt**GAATAAA
TAATCCGGTCGAAAGAGATA**ttaatt**CGATTATATTATTTAT**ttaatt**AT
AT**ttaatttaaa**TATAT**aattaa**TATATATATATTG**aatt**ATATAT**aatt**
tA**tttt**AT**aattttt**ATAAATAATATATTATTATAAATAT**ttaa**TAT**aatt**
TATATTATTA**ttaa**AT**aaaa**GATTTA**ttaattaa**TATTATTAT**ttaattt**
tATTATATAGT**ttaa**GGGATAATA**tttt**A**ttaa**TA**tttttttt**ATTTATT
TAT**ttaatt**ATATTATATATATAATATATATATAACAAT**aatt**tATGACA
CATTTAGAAAGAAGTAGACATCAACAACATCCATTTCATATGGTTATGCC
TTCACCATGACCTATTGTAGTATCATTTGCATTATTATCATTAGCATTAT
CACTAGCA**ttaa**CAATGCATGGTTATATTGGTAATATGAATATGGTATAT
TTAGCATTATTTGTATTA**ttaa**CAAGTTCTA**tttt**ATGATTTAGAGATAT
TGTAGCTGAAGCTACATATTTAGGTGATCATACTATAGCAGTAAG**aaaa**G
GTA**ttaatt**TAGGTTTC**ttaa**TGTTTGTATTATCTGAAGTA**ttaa**TCTTT
GCTGGTTTATTCTGAGCTTATTTCCATTCAGCTATGAGTCCTGATGTACT
ATTAGGTGCATGTTGACCACCCGTAGGTATTGAAGCTGTACAACCTACCG
aattACCTTTA**ttaa**ATACTATTATCTTATTATCTTCTGGTGCTACTGTA
ACTTATAGTCATCATGCC**ttaa**TCGCAGGTAATAGAAATAAAGCCTTATC
AGGTTTA**ttaatt**ACATTCTGA**ttaatt**GTTA**ttttt**GTTACTTGTCAAT
ATATTGAATATACTAATGCTGCATTCACTATCTCTGATGGTGTTTATGGT

TCAGTATTCTATGCTGGTACAGGATTACATTTCTTACATATGGTAATGTT
AGCAGCTATGTTAGGTG**ttaatt**ATTGAAGAATGAG**aatt**ATCAT**ttaa**C
AGCTGGACATCATGTTGGATATGAAACAACTATTATTTATCTACATG**ttt**
tAGATGTTATCTGATTA**ttttt**ATACGTAGTCTTCTACTGATGAGGAGTC
TAAGGCTATAG**aatt**ATATATCTAAATGA**ttaa**TATATATATTA**ttaa**T**a**
attaaCAAT**aattaa**TATATTAT**aatt**TATATATATATA**tttt**ATATTAT
TATAATAATATTCTTACAAATAT**aatt**ATTATATATTATTCCTTC**aaaa**C
TCCTAACGGGGTTCCCGCGAAGCGGGAACTAATAATAATATAATCATTAT
ACTC**tttttt**CATTTACC**tttt**ATAAAGAT**aattaa**T**aatt**tAT**ttaa**TA
TTTAT**aaaaaaaaaaa**TATAATA**ttaa**TATAATATAATATAATAATGT**aa**
ttATTTATA**ttttt**ATATTCCTTCGAGGTCACCGCCTCACCTCCAGCGGG
AC**ttttttaa**TATGATATAATATAATATAAATATTA**ttaatttaa**CTAAT
ATAT**aatt**CATATATATATATATATTA**ttaa**TATTA**tttt**AT**aaaaaa**TA
ttttttATTTGATTATTA**ttaa**ATATTATATAGTTCCGGGGCCCGGCCAC
GGGAGCCGGAACCCCGAAAGGAGAAATA**ttaa**TATATTATAAATATACTA
TTTATGT**aatt**A**tttttt**GAAGTGAGCACCTA**tttt**ATATATA**tttt**ATA
TATA**tttt**ATTATA**tttt**A**ttaaaaa**TAGGTGTGAACCTCCATGAGAGAG
GAATGAATACCTA**tttt**ATAAAGTATA**tttt**ATATTCTATATATTATAAA
TATGAACC**aaaaaaa**GGAG**tttaaatttaattaatttaattaattG**a**att**
TCTTTATTATTATTATCAT**aatt**A**ttaa**ACCCTTTA**ttaa**TATAATAATA
TATTATTTATTATC**aaaa**TACCTACCC**ttttt**AT**aatt**TATATCT**ttaa**T
AATAT**aattaa**ATAT**aaaa**TGTTTA**ttaa**ATATTATAT**aaaaa**T**aaaaa**T
aaaaaTATATATATATATATAAATGATAAATAATAAGG**aatt**CACACTTA
TAT**aatttaaa**TATAAAGTCCC**aaaa**GAAGTATTCA**ttaa**AT**aatt**ATCA
ttaattaattATAATAAACTTAT**ttaa**TATTA**ttaa**AGA**ttaatt**TATAA
TAAT**aatt**ATTATTATTATTA**ttaa**TA**ttaa**T**aaaa**TATATAAAT**aatta**
aATAGTTCATATA**ttaaaaa**G**aatt**AG**aattaa**ACT**ttaa**TAAGTGTAT**t**
taaTATATAGAATA**ttaa**TAGAATATTTATTCTATTTATATATATATTTA
TATATATATATA**ttaa**ATAATATTATTTATATTATA**tttt**ATATATATAT
TA**ttaa**TAT**aaaaa**GTATATTATATGTATTATATATATTATATATTATAT
AT**ttaa**TAATATATTACTCCTTTGGGGTGGGTCCGCCCCACGGGGCGGGC
CGGACTATTAT**aattaa**T**aatttt**ATAAAGTTCCGGGGCCCGGCCACGGG
AGCCGGAACCCCGAAAGGAGAATAAAT**aatt**ATATATCTTCTTC**ttaatt**
aattaattaattaattaattaattaattaattaaaaaGGGGTTCGGTCCC
CCTCCCTAACGGGAGGGGGTCCCTCACTCATTCAAACTAT**aatttaa**TAT
ATTATGATATTATTTAT**aatt**TATAATATAATGTATAATATTATATTATA
AATATTATAT**aaaaa**T**aaaa**TGATATATATAATAATAATAATAATAATAA

T**aaaaaaa**TAG**aaaa**GAAT**aatttttt**ATTA**tttt**AGTATATATAAG**aatt**
taaTAAGTTATATTATTGCGGACACCGTTACGCGGAGTGGGGACTATTAT
A**tttt**ACCTATATATA**ttaa**TATTATTAT**aatt**TCCTTC**tttaaaa**G**aaa**
aaaGG**aatt**CGAGAACTTATTATTATA**ttaa**TATA**ttaa**TAATAAATAAT
AATAAATAAT**aaaaaa**GTAAAT**aatt**AT**aatt**ATAT**aaaaa**TAT**aatttt**
ATTA**ttaa**GAAAGGAG**tttaaa**TAT**aaaa**TATAATATTATCA**ttaa**GTTC
TAATAAAGGTATATAATGAAGATCTATTAGAACCT**aaaaa**GAATA**ttaa**T
ATATCTATTAT**aaaa**TAATAATAATAAATATAAATAT**aaaaa**T**aatt**GTA
ATATTTATAAATAATAAT**aaaaaa**TAAATAAGGAATATA**ttaatt**A**ttaa**
TAATAAAT**aatt**ATA**ttaaaa**TATAATATTATTA**ttaattaa**AG**aatt**AT
A**ttaa**ATATATTTA**ttaaattt**ATAAATAAG**ttaa**TA**tttt**A**ttaa**ATA
ATATTTATAAATAAT**aaaaaaaaa**TAAGTATAT**aatt**A**ttaa**TATA**ttaa**
ttTATTATGTTATATATTTATATATTTCAAATATATAAGTAATAGGGGGA
GGGGGTGGGTGATAATAACCAGAATA**ttaa**ATAAATACAGAGCACACATT
TG**ttaa**TAT**ttaa**TAATATAATCAATAAATATATTATAATAATATAATAT
aattaaTAATAGATATAAAGTATAAACAATATAAT**aatt**ATAT**aaaa**TAA
ATAT**aattaaaaa**TAATAACCAAAT**aattaa**TATAATAAATGATAAACAA
GAAGATATCCGGGTCCCAATAAT**aatt**ATTATTG**aaaa**TAAT**aatt**GGGA
CCCCCACAATAGAAT**aaaaaa**T**aaaaa**G**aattaa**TAATATATAAATAATA
T**aaaa**TATATTATATATATATATAATATATATATATATATAAT**aaaaaaa**
aaTATATAATATAATATATATATAT**aaaa**TAAT**aatt**ATATATATATATA
T**aaaa**TAAT**aaaaaa**TAATAATCATATG**aatttt**ATAAATAT**aatt**ATTA
ttaaTAATAATAATAATAATAATAAAGTCCGGTCCGCCCCGCGGAGGGGG
CGGACCCCCGAAGGAGTGCGGGACCCCGTGGGAACCGCATCCC**ttttt**AT
TC**ttaattaa**GAAGGAGATAAT**aatt**TAT**aaaattaa**TATTTA**tttt**ATG
TAATA**ttaa**TA**ttaa**TA**ttaa**TATAATATAATATAATATAATACGGA**tta**
aATATTACCAGTTGTTCACAGGTAATAT**aaaa**TCCTATTGTTTCACCTAT
TA**ttaattaa**TAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAG
GAGAATAAGTATATATAAT**aaatttaa**T**aaaaaaaaa**T**aatt**ATATAATA
AATATATATATTATAATATTATATAAATAT**aaaa**TAT**aatt**GATA**ttaa**C
ATTATAT**aattaa**TAATATAATCAAATAATATAAATATAATAT**aaaaa**G**t**
tttaattA**ttaaatt**ATATAAATATTAT**ttaa**T**aaaaa**T**aaaaa**TAATAA
TAATAATAATAATAATAAAGTCCGGTCCGCCCCCTCCGCGGAGGGGGCGG
ACCCCGAAAGAGTGAGGGACCCCCCCGTATACTTACGGGGGGAGAACCGA
ACCCC**ttttttt**A**tttaaa**GAAGGAGATAAATATTTATATCTTTATTTAT
aattATATATAAAT**aaaa**GTTTA**ttaaattt**ATAATAATAATAT**aaaaaa**
GTATATAAT**aatt**tATTATAAATAAATAAATATTTAGTAATAATAT**ttaa**

Ta**aatt**ATAAATATTATAAAT**aaaa**TA**ttaa**TAATAAATAATAAATATAT
AATATAATATAATATAAT**aattaa**TAACAATAAGATATCCGGGTCCCCTA
AAT**aatt**ATTATAT**aaaa**TAAT**aatt**GGGACCCATACATATAAATAT**aaa**
aTA**ttttaa**TATTTATATATAAATAATAATAATATATATTTATATTATAT
TATAATATAACCCTTTCC**aattaa**TA**ttaa**TA**ttaa**TA**ttaatt**ACTTCC
ttaaaaaaaTAAT**aattaattaatt**GA**ttttt**ATA**ttaa**TAT**aaaaaa**G**t**
taaTATATATATTTATATATAAATAATAT**aattaa**TATAAAGATAATAAG
TCCCCGCTTTCAGCGCAGTGAGGGACCCCCTCCCGTAAATATACGGGAGG
GGAGACCGAACCCCAAAGGAATAATAAATAATAGTATGTA**tttaaa**TAAA
TAT**ttaa**TATACTA**tttttttttt**ATTA**ttttt**ATAATATATTTATAATAA
TATAT**ttaatt**AT**aatt**TAT**aaaaaa**GAGATATAATA**tttt**ATTATATAT
AATA**ttaa**TATAATAC**aattaa**CATTAT**ttaatt**ATTA**ttaa**TAATAT**tt**
aaCTTTATTATTATCTTCTACGGTTGGACTCCTTC**ttaaaaa**GGGGTTCG
GTCCCCCTCCCATTAGGGAGGGGTCCCTCACTCCTTCGGGGTCCGCGCCC
CCCGCGGGGGGGGGCGGACCGGACTATTATTACTATTTATTTA**ttaa**TAA
TAAATAAT**aatt**ATAAAGTCACTGAAAGAGTGAGG**aattttCCtttt**CCC
AAGGG**aaaa**CCCCAAAGGATAATATAAATATTAT**aaatttttA**ttaaATA
ATATa**aattC**AAT**aaaa**T**aattttaattaattaattaattaattaa**TAT**a**
aaaaTAAATA**ttttaattaa**TA**ttaa**TA**ttaa**TAGTTCCGGGGCCCGGC
CACGGGAGCCGGAACCCCGGAAGGAGAAATATAAATATAATAGTATAGTA
TATAGGAAG**ttaa**TAATAATATAAATATTATATAATATATATATGTATAT
ATATTATATTATAT**aattaattttCTCCtttt**GTATTTACATC**ttaaTaa**
aaTAT**aaaa**TAT**aaaa**TGTTA**ttaa**CAAT**aaaattAttaa**TCTTTATAAT
A**ttaa**TAATAGT**aatt**tATTTATATATCTCCTTTAGGATGGACTCCTTCG
GCCGGACTCCTTCGGGGTCCGCCCCGCGGGGGCGGGCCGGACTA**ttttt**A
ttttttttttaaaaaaTA**ttaa**ATATTATAAATATATTATAAATATATTA
TAAATATGTTATAAATATATTATAAATAGAATATAATATAATATTATATA
TTATAATGATAAAGATTATATATA**tttt**C**ttttttttttt**ATTTATTA**tt**
tttaaTAAGT**aaaatt**ATATTATATATATATATATATTAGA**tttt**ATAAG
TAATATAATATAAGTA**ttaa**TATATAAATGCAATATGATGT**aatt**GG**tta**
aCA**tttt**AGGGTCATGACCT**aatt**ATATACGTTCAAATCGTATTATTGCT
AAT**aattaa**TATATAATATTTAT**aaaaaa**GTATAAT**aaaa**TATATTATAA
GAAGAATATATTATATAAT**aatt**ATA**ttaa**TAATA**ttaa**TAAATAATATA
TAAAT**aatt**AT**aaaaaa**GTATATAATA**ttaa**TC**aattaattaattaa**TAA
ATATAAATAATATA**ttaatttttaattaatt**TGAATAAGATATTTATATT
A**ttaa**TAGGAAAGTCATAAATATAT**aatt**ATATTATAT**aattaa**TATAAT
AAT**aaaa**T**aatt**ATATA**tttt**ATTTATAATATTATTTCTTTATAAGAT**aa**

aaTATTATCTGATGAAT**aatt**AGATTGAATAATATTTATAAAGAAATATA
TAT**aaaaa**GTCATTATAT**aatttaatt**AT**aatttaaa**T**aattttA**TAT**aa**
ttaaTATAATA**ttaa**TAAAGT**aatt**AGTATAAATAAATAATATG**aaaa**T**a**
aaaC**ttaa**TAAATATATAAATATAGTCCGGCCCGCCCCCCCGCGGCGGGC
GGACCCCGAAGGAGTGAGGGACCCCTCCCTAATGGGAGGGGGACCGAACC
CC**ttttttaa**GAAGGAGTCCATATATATATA**ttaa**T**aaaaaaaa**GTAATAT
ATATATATATATTGGAATAGTTATATTATTATACAGAAATATGC**ttaatt**
ATAATATAATATCCATA

References

1. Ernster. Mitochondria: a historical review. *The Journal of Cell Biology* **91,** 227s–255 (1981).
2. Nunnari, J. & Suomalainen, A. Mitochondria: in sickness and in health. *Cell* **148,** 1145–59 (2012).
3. Ballmoos, C. von, Wiedenmann, A. & Dimroth, P. Essentials for ATP Synthesis by F1F0 ATP Synthases. *Biochemistry-us* **78,** 649–672 (2009).
4. Sibson, N. R. *et al.* Functional energy metabolism: in vivo 13C-NMR spectroscopy evidence for coupling of cerebral glucose consumption and glutamatergic neuronalactivity. *Dev. Neurosci.* **20,** 321–30 (1998).
5. Sibson, N. R. *et al.* Stoichiometric coupling of brain glucose metabolism and glutamatergic neuronal activity. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 316–21 (1998).
6. Baughman, J. M. *et al.* Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476,** 341–5 (2011).
7. Li, X. *et al.* Targeting mitochondrial reactive oxygen species as novel therapy for inflammatory diseases and cancers. *Journal of Hematology & Oncology* **6,** 19 (2013).
8. Green, D. Apoptotic Pathways. *Cell* **94,** 695–698 (1998).
9. ROSSIER. T channels and steroid biosynthesis: in search of a link with mitochondria. *Cell Calcium* **40,** 155–164 (2006).
10. Klinge, C. Estrogenic control of mitochondrial function and biogenesis. *Journal of Cellular Biochemistry* **105,** 1342–1351 (2008).
11. Taylor, R. & Turnbull, D. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics* **6,** 389–402 (2005).
12. Sherer, Betarbet & Greenamyre. Environment, Mitochondria, and Parkinson's Disease. *The Neuroscientist* **8,** 192–197 (2002).
13. Lim, Y.-A. *et al.* Aβ and human amylin share a common toxicity pathway via mitochondrial dysfunction. *PROTEOMICS* **10,** 1621–1633 (2010).
14. Schapira, A. Mitochondrial disease. *The Lancet* **368,** 70–82 (2006).
15. López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes. *Trends Ecol. Evol. (Amst.)* **30,** 697–708 (2015).
16. Martijn, J. & Ettema, T. J. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochemical Society Transactions* **41,** 451–457 (2013).
17. López-García, P & Moreira, D. Selective forces for the origin of the eukaryotic nucleus. *Bioessays* (2006). doi:10.1002/bies.20413
18. Zamaroczy, M. de & Bernardi, G. The primary structure of the mitochondrial genome of Saccharomyces cerevisiae--a review. *Gene* **47,** 155–77 (1986).
19. Foury, F., Roganti, T., Lecrenier, N. & Purnelle, B. The complete sequence of the mitochondrial genome of Saccharomyces cerevisiae. *Febs Lett* **440,** 325–31 (1998).
20. Taanman, J.-W. The mitochondrial genome: structure, transcription, translation and replication. *Biochimica Et Biophysica Acta Bba - Bioenergetics* **1410,** 103–123 (1999).
21. Herrmann, J. & Neupert, W. Protein transport into mitochondria. *Current Opinion in Microbiology* **3,** 210–214 (2000).
22. Rao, S. *et al.* Biogenesis of the preprotein translocase of the outer mitochondrial membrane: protein kinase A phosphorylates the precursor of Tom40 and impairs its import. *Mol. Biol. Cell* **23,** 1618–27 (2012).
23. Chipuk, Bouchier-Hayes & Green. Mitochondrial outer membrane permeabilization during apoptosis: the innocent bystander scenario. *Cell Death and Differentiation* **13,** 1396–1402 (2006).
24. Haanen, C. & Vermes, I. Apoptosis: programmed cell death in fetal development. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **64,** 129–33 (1996).
25. Hayashi, T., Rizzuto, R., Hajnoczky, G. & Su, T.-P. MAM: more than just a housekeeper. *Trends in Cell Biology* **19,** 81–88 (2009).
26. Westermann, B. Mitochondrial fusion and fission in cell life and death. *Nature Reviews Molecular Cell Biology* **11,** 872–884 (2010).
27. Herrmann, J. & Neupert, W. Protein transport into mitochondria. *Current Opinion in Microbiology* **3,** 210–214 (2000).
28. Scheffler, I. *Metabolic Pathways Inside Mitochondria*. 246–272 (1999). doi:10.1002/0471223891.ch6
29. Chen, X. J. & Butow, R. A. The organization and inheritance of the mitochondrial genome. *Nat. Rev. Genet.* **6,** 815–25 (2005).

30. Tzagoloff, A., Wu, M. A. & Crivellone, M. Assembly of the mitochondrial membrane system. Characterization of COR1, the structural gene for the 44-kilodalton core protein of yeast coenzyme QH2-cytochrome c reductase. *J. Biol. Chem.* **261,** 17163–9 (1986).
31. Otterstedt, K. *et al.* Switching the mode of metabolism in the yeast Saccharomyces cerevisiae. *EMBO reports* **5,** 532–537 (2004).
32. Williamson, D. The curious history of yeast mitochondrial DNA. *Nat. Rev. Genet.* **3,** 475–81 (2002).
33. Fangman, W. L., Henly, J. W., Churchill, G. & Brewer, B. J. Stable maintenance of a 35-base-pair yeast mitochondrial genome. *Mol. Cell. Biol.* **9,** 1917–21 (1989).
34. Maleszka, R., Skelly, P. J. & Clark-Walker, G. D. Rolling circle replication of DNA in yeast mitochondria. *EMBO J.* **10,** 3923–9 (1991).
35. Blanc, H. Two modules from the hypersuppressive rho- mitochondrial DNA are required for plasmid replication in yeast. *Gene* **30,** 47–61 (1984).
36. Zweifel, S. G. & Fangman, W. L. Creation of ARS activity in yeast through iteration of non-functional sequences. *Yeast* **6,** 179–86 (1990).
37. Shadel, G. Yeast as a Model for Human mtDNA Replication. *Am J Hum Genetics* **65,** 1230–1237 (1999).
38. Lecrenier, N. & Foury, F. New features of mitochondrial DNA replication system in yeast and man. *Gene* **246,** 37–48 (2000).
39. Williamson, D. H. & Fennell, D. J. Apparent dispersive replication of yeast mitochondrial DNA as revealed by density labelling experiments. *Mol. Gen. Genet.* **131,** 193–207 (1974).
40. Arbuckle, J. & Medveczky, P. The molecular biology of human herpesvirus-6 latency and telomere integration. *Microbes and Infection* **13,** 731–741 (2011).
41. Gerhold, J. M., Aun, A., Sedman, T., Jõers, P. & Sedman, J. Strand invasion structures in the inverted repeat of Candida albicans mitochondrial DNA reveal a role for homologous recombination in replication. *Mol. Cell* **39,** 851–61 (2010).
42. Bendich, A. J. The end of the circle for yeast mitochondrial DNA. *Mol. Cell* **39,** 831–2 (2010).
43. Gao, L., Laude, K. & Cai, H. Mitochondrial Pathophysiology, Reactive Oxygen Species, and Cardiovascular Diseases. *Veterinary Clinics of North America: Small Animal Practice* **38,** 137–155 (2008).
44. Thomas, C., Mackey, M., Diaz, A. & Cox, D. Hydroxyl radical is produced via the Fenton reaction in submitochondrial particles under oxidative stress: implications for diseases associated with iron accumulation. *Redox Report* **14,** 102–108 (2013).
45. Ljungman, M. & Hanawalt, P. C. Efficient protection against oxidative DNA damage in chromatin. *Mol. Carcinog.* **5,** 264–9 (1992).
46. Gilkerson, R. *et al.* The mitochondrial nucleoid: integrating mitochondrial DNA into cellular homeostasis. *Cold Spring Harb Perspect Biol* **5,** a011080 (2013).
47. Guliaeva, N. A., Kuznetsova, E. A. & Gaziev, A. I. [Proteins associated with mitochondrial DNA protect it against the action of X-rays and hydrogen peroxide]. *Biofizika* **51,** 692–7 (2006).
48. Tuyle, G. C. Van & McPherson, M. L. A compact form of rat liver mitochondrial DNA stabilized by bound proteins. *J. Biol. Chem.* **254,** 6044–53 (1979).
49. Miyakawa, I., Sando, N., Kawano, S., Nakamura, S. & Kuroiwa, T. Isolation of morphologically intact mitochondrial nucleoids from the yeast, Saccharomyces cerevisiae. *J. Cell. Sci.* **88 (Pt 4),** 431–9 (1987).
50. Kukat *et al.* Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. *Proceedings of the National Academy of Sciences* **108,** 13534–13539 (2011).
51. Garrido, N. *et al.* Composition and dynamics of human mitochondrial nucleoids. *Mol. Biol. Cell* **14,** 1583–96 (2003).
52. Rubio-Cosials, A. *et al.* Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. *Nat. Struct. Mol. Biol.* **18,** 1281–9 (2011).
53. Rubio-Cosials, A. & Solà, M. U-turn DNA bending by human mitochondrial transcription factor A. *Curr. Opin. Struct. Biol.* **23,** 116–24 (2013).
54. Ngo, H. B., Lovely, G. A., Phillips, R. & Chan, D. C. Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation. *Nat Commun* **5,** 3077 (2014).
55. Kukat, C. *et al.* Cross-strand binding of TFAM to a single mtDNA molecule forms the mitochondrial nucleoid. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 11288–93 (2015).
56. Malarkey, C. S. & Churchill, M. E. The high mobility group box: the ultimate utility player of a cell. *Trends Biochem. Sci.* **37,** 553–62 (2012).

57. Bustin, M. Regulation of DNA-Dependent Activities by the Functional Motifs of the High-Mobility-Group Chromosomal Proteins. *Molecular and Cellular Biology* **19,** 5237–5246 (1999).
58. Ferrari, Finelli, Rocchi & Bianchi, M. E. The Active Gene That Encodes Human High Mobility Group 1 Protein (HMG1) Contains Introns and Maps to Chromosome 13. *Genomics* **35,** 367–371 (1996).
59. Shirakawa, H. & Yoshida, M. Structure of a gene coding for human HMG2 protein. *J. Biol. Chem.* **267,** 6641–5 (1992).
60. Kuhn, A. *et al.* Functional differences between the two splice variants of the nucleolar transcription factor UBF: the second HMG box determines specificity of DNA binding and transcriptional activity. *EMBO J.* **13,** 416–24 (1994).
61. Kukat, C. & Larsson, N.-G. mtDNA makes a U-turn for the mitochondrial nucleoid. *Trends Cell Biol* **23,** 457–63 (2013).
62. Morozov, Y. I. *et al.* A model for transcription initiation in human mitochondria. *Nucleic Acids Res.* **43,** 3726–35 (2015).
63. Brewer, L. *et al.* Packaging of Single DNA Molecules by the Yeast Mitochondrial Protein Abf2p. *Biophys J* **85,** 2519–2524 (2003).
64. Friddle, R. W. *et al.* Mechanism of DNA compaction by yeast mitochondrial protein Abf2p. *Biophys. J.* **86,** 1632–9 (2004).
65. MacAlpine, D., Perlman, P. & Butow, R. The high mobility group protein Abf2p influences the level of yeast mitochondrial DNA recombination intermediates in vivo. *Proceedings of the National Academy of Sciences* **95,** 6739–6743 (1998).
66. Zelenaya-Troitskaya, O., Newman, S. M., Okamoto, K., Perlman, P. S. & Butow, R. A. Functions of the high mobility group protein, Abf2p, in mitochondrial DNA segregation, recombination and copy number in Saccharomyces cerevisiae. *Genetics* **148,** 1763–76 (1998).
67. Stigter, D. Packaging of single DNA molecules by the yeast mitochondrial protein Abf2p: reinterpretation of recent single molecule experiments. *Biophys. Chem.* **110,** 171–8 (2004).
68. Bakkaiova, J. *et al.* Yeast mitochondrial HMG proteins: DNA-binding properties of the most evolutionarily divergent component of mitochondrial nucleoids. *Biosci. Rep.* **36,** (2015).
69. Kucej, M., Kucejova, B., Subramanian, R., Chen, X. J. & Butow, R. A. Mitochondrial nucleoids undergo remodeling in response to metabolic cues. *J. Cell. Sci.* **121,** 1861–8 (2008).
70. Diffley, J. F. & Stillman, B. A close relative of the nuclear, chromosomal high-mobility group protein HMG1 in yeast mitochondria. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 7864–8 (1991).
71. Diffley, J. F. & Stillman, B. DNA binding properties of an HMG1-related protein from yeast mitochondria. *J. Biol. Chem.* **267,** 3368–74 (1992).
72. Kao, L. R., Megraw, T. L. & Chae, C. B. Essential role of the HMG domain in the function of yeast mitochondrial histone HM: functional complementation of HM by the nuclear nonhistone protein NHP6A. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 5598–602 (1993).
73. Bustin, M. Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Mol. Cell. Biol.* **19,** 5237–46 (1999).
74. Bowles, J., Schepers, G. & Koopman, P. Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators. *Developmental Biology* **227,** 239–255 (2000).
75. Milatovich, A., Travis, A., Grosschedl, R. & Francke, U. Gene for lymphoid enhancer-binding factor 1 (LEF1) mapped to human chromosome 4 (q23–q25) and mouse chromosome 3 near Egf. *Genomics* **11,** 1040–1048 (1991).
76. Chen, J., Wang, H. & Wang, Y.-F. F. Overexpression of HmgD causes the failure of pupariation in Drosophila by affecting ecdysone receptor pathway. *Arch. Insect Biochem. Physiol.* **68,** 123–33 (2008).
77. Wong *et al.* Biophysical characterizations of human mitochondrial transcription factor A and its binding to tumor suppressor p53. *Nucleic Acids Research* **37,** 6765–6783 (2009).
78. McCulloch, V. & Shadel, G. S. Human mitochondrial transcription factor B1 interacts with the C-terminal activation region of h-mtTFA and stimulates transcription independently of its RNA methyltransferase activity. *Mol. Cell. Biol.* **23,** 5816–24 (2003).
79. Newman, S. M., Zelenaya-Troitskaya, O., Perlman, P. S. & Butow, R. A. Analysis of mitochondrial DNA nucleoids in wild-type and a mutant strain of Saccharomyces cerevisiae that lacks the mitochondrial HMG box protein Abf2p. *Nucleic Acids Res.* **24,** 386–93 (1996).

80. MacAlpine, D., Perlman, P. & Butow, R. The numbers of individual mitochondrial DNA molecules and mitochondrial DNA nucleoids in yeast are co-regulated by the general amino acid control pathway. *Embo J* **19,** 767–775 (2000).
81. Chen, X. J., Wang, X. & Butow, R. A. Yeast aconitase binds and provides metabolically coupled protection to mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 13738–43 (2007).
82. Chen, X. J., Wang, X., Kaufman, B. A. & Butow, R. A. Aconitase couples metabolic regulation to mitochondrial DNA maintenance. *Science* **307,** 714–7 (2005).
83. Cheng, X. & Ivessa, A. Association of the yeast DNA helicase Pif1p with mitochondrial membranes and mitochondrial DNA. *European Journal of Cell Biology* **89,** 742–747 (2010).
84. Petersen, J. G. & Holmberg, S. The ILV5 gene of Saccharomyces cerevisiae is highly expressed. *Nucleic Acids Res.* **14,** 9631–51 (1986).
85. Bateman, J. M., Perlman, P. S. & Butow, R. A. Mutational bisection of the mitochondrial DNA stability and amino acid biosynthetic functions of ilv5p of budding yeast. *Genetics* **161,** 1043–52 (2002).
86. Kaufman, B. A., Kolesar, J. E., Perlman, P. S. & Butow, R. A. A function for the mitochondrial chaperonin Hsp60 in the structure and transmission of mitochondrial DNA nucleoids in Saccharomyces cerevisiae. *J. Cell Biol.* **163,** 457–61 (2003).
87. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41,** W349–57 (2013).
88. Dairaghi, D., Shadel, G. & Clayton, D. Addition of a 29 Residue Carboxyl-terminal Tail Converts a Simple HMG Box-containing Protein into a Transcriptional Activator. *J Mol Biol* **249,** 11–28 (1995).
89. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
90. Chong, Y. *et al.* Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161,** 1413–1424 (2015).
91. Clyne, R. K. & Kelly, T. J. Identification of autonomously replicating sequence (ARS) elements in eukaryotic cells. *Methods* **13,** 221–33 (1997).
92. Raychaudhuri, S., Byers, R., Upton, T. & Eisenberg, S. Functional analysis of a replication origin from Saccharomyces cerevisiae: identification of a new replication enhancer. *Nucleic Acids Res.* **25,** 5057–64 (1997).
93. Poloumienko, A., Dershowitz, A., De, J. & Newlon, C. S. Completion of replication map of Saccharomyces cerevisiae chromosome III. *Mol. Biol. Cell* **12,** 3317–27 (2001).
94. Nieduszynski, C. A. & Donaldson, A. D. Detection of replication origins using comparative genomics and recombinational ARS assay. *Methods Mol. Biol.* **521,** 295–313 (2009).
95. Goulas, T. *et al.* The pCri System: a vector collection for recombinant protein expression and purification. *PLoS ONE* **9,** e112643 (2014).
96. Munteanu, B., Braun, M. & Boonrod, K. Improvement of PCR reaction conditions for site-directed mutagenesis of big plasmids. *Journal of Zhejiang University. Science. B* **13,** 244–7 (2012).
97. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols* **2,** 1849–1861 (2007).
98. McPherson, A. Crystallization of biological macromolecules. (1999). at <http://agris.fao.org/agris-search/search.do?recordID=US201300029639>
99. Grimmer, H. The Basics of Crystallography and Diffraction. Fourth Edition. By Christopher Hammond. IUCr Texts on Crystallography, No. 21. IUCr/Oxford Science Publications, 2015. Paperback, Pp. 544. Price GBP 34.99. ISBN 978-0-19-873868-8. *Acta Crystallogr A Found Adv* **72,** 173–5 (2016).
100. Dauter, Z. & Wlodawer, A. On the accuracy of unit-cell parameters in protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **71,** 2217–26 (2015).
101. Bragg & Bragg. The Reflection of X-rays by Crystals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **88,** 428–438 (1913).
102. Liljas, A. Background to the Nobel Prize to the Braggs. *Acta Crystallogr., A, Found. Crystallogr.* **69,** 10–5 (2013).
103. Wang, B. C. Resolution of phase ambiguity in macromolecular crystallography. *Meth. Enzymol.* **115,** 90–112 (1985).
104. Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 282–92 (2011).
105. Powell, H. R., Johnson, O. & Leslie, A. G. Autoindexing diffraction images with iMosflm. *Acta Crystallogr. D Biol. Crystallogr.* **69,** 1195–203 (2013).
106. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 125–32 (2010).

107. Diederichs, K. & Karplus, P. A. Better models by discarding data? *Acta Crystallogr. D Biol. Crystallogr.* **69,** 1215–22 (2013).
108. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **34,** 60–8 (2015).
109. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* **19,** 1395–412 (2011).
110. Hendrickson, W. A. Anomalous diffraction in crystallographic phase evaluation. *Q. Rev. Biophys.* **47,** 49–93 (2014).
111. Evangelista, L. R., Lenzi, E. K. & Barbero, G. The Kramers-Kronig relations for usual and anomalous Poisson-Nernst-Planck models. *J Phys Condens Matter* **25,** 465104 (2013).
112. Sheng, J. & Huang, Z. Selenium derivatization of nucleic acids for X-ray crystal-structure and function studies.*Chem. Biodivers.* **7,** 753–85 (2010).
113. Sun, H., Jiang, S. & Huang, Z. Nucleic Acid Crystallography via Direct Selenium Derivatization: RNAs Modified with Se-Nucleobases. *Methods Mol. Biol.* **1320,** 193–204 (2016).
114. Terwilliger, T. C. *et al.* Can I solve my structure by SAD phasing? Planning an experiment, scaling data and evaluating the useful anomalous correlation and anomalous signal. *Acta Crystallogr D Struct Biol* **72,** 359–74 (2016).
115. Isaacs, N. A history of experimental phasing in macromolecular crystallography. *Acta Crystallogr D Struct Biol* **72,** 293–5 (2016).
116. McCoy, A. J. & Schneider, T. Advances in experimental phasing. *Acta Crystallogr D Struct Biol* **72,** 291–2 (2016).
117. Giacovazzo, C. Solution of the phase problem at non-atomic resolution by the phantom derivative method. *Acta Crystallogr A Found Adv* **71,** 483–512 (2015).
118. Scapin, G. Molecular replacement then and now. *Acta Crystallogr. D Biol. Crystallogr.* **69,** 2266–75 (2013).
119. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–21 (2010).
120. Aldrich, J. R.A. Fisher and the making of maximum likelihood 1912-1922. *Stat Sci***12,** 162–176 (1997).
121. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53,** 240–55 (1997).
122. Bricogne G., Blanc E., Brandl M., Flensburg C., Keller P., Paciorek W., Roversi P, Sharff A., Smart O.S., Vonrhein C., Womack T.O. (2016). BUSTER. Cambridge, United Kingdom: Global Phasing Ltd.
123. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 12–21 (2010).
124. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D Biol Crystallogr* **66,** (2010).
125. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 235–42 (2011).
126. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).
127. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69,** 1204–14 (2013).
128. Thorn, A. & Sheldrick, G. M. ANODE: anomalous and heavy-atom density calculation. *J Appl Crystallogr* **44,** (2011).
129. Svergun, D. & Koch, M. Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* **66,** 1735 (2003).
130. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589,** 2570–7 (2015).
131. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45,** (2012).
132. Bernadó, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* **8,** 151–67 (2012).
133. Tria, G., Mertens, H. D., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2,** 207–17 (2015).
134. Paquet, E. & Viktor, H. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review. *BioMed Research International* **2015,** 1–18 (2015).

135. Moore, C. Ergodic theorem, ergodic theory, and statistical mechanics. *Proc Natl Acad Sci* **112,** 1907–1911 (2015).
136. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **13,** 55–8 (2016).
137. D.A. Case, R.M. Betz, W. Botello-Smith, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2016), AMBER 2016, University of California, San Francisco.
138. Berendsen, Postma, van Gunsteren, DiNola & Haak. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81,**3684 (1984).
139. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9,** 3084–95 (2013).
140. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14,** 33–8, 27–8 (1996).
141. Lavery, Moakher, Maddocks, Petkeviciute & Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* **37,** 5917–5929 (2009).
142. Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.* **39,** W68–W73 (2011).
143. Dršata, T. *et al.* Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res.* **42,** 7383–94 (2014).
144. Portella, G., Battistini, F. & Orozco, M. Understanding the Connection between Epigenetic DNA Methylation and Nucleosome Positioning from Computer Simulations. *Plos Comput Biol* **9,** e1003354 (2013).
145. Scarlett, G., Siligardi, G. & Kneale, G. G. Circular Dichroism for the Analysis of Protein-DNA Interactions. *Methods Mol. Biol.***1334,** 299–312 (2015).
146. Velazquez-Campoy, A., Leavitt, S. A. & Freire, E. Characterization of protein-protein interactions by isothermal titration calorimetry. *Methods Mol. Biol.* **1278,** 183–204 (2015)
147. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122,** 19–27 (1989).
148. Huxley, C., Green, E. D. & Dunham, I. Rapid assessment of S. cerevisiae mating type by PCR. *Trends Genet.* **6,** 236 (1990).
149. Lundblad, V. & Zhou, H. *Current Protocols in Molecular Biology*. 13.9.1–13.9.6 (wiley, 2001). doi:10.1002/0471142727.mb1309s39
150. Boeke, J., Trueheart, J., Natsoulis, G. & Fink, G. *[10] 5-Fluoroorotic acid as a selective agent in yeast molecular genetics*. *Method Enzymol* **154,** 164–175 (sciencedirect, 1987).
151. Pike, A. C. W., Garman, E. F., Krojer, T., Delft, F. & Carpenter, E. P. An overview of heavy-atom derivatization of protein crystals. *Acta Crystallogr Sect D Struct Biology* **72,** 303–318 (2016).
152. Banumathi, S., Dauter, M. & Dauter, Z. Phasing at high resolution using Ta6Br12 cluster. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 492–8 (2003).
153. Doublié, S. Production of selenomethionyl proteins in prokaryotic and eukaryotic expression systems. *Methods Mol. Biol.* **363,** 91–108 (2007).
154. Haran, T. E. & Mohanty, U. The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **42,** 41–81 (2009).
155. Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D. & Shakked, Z. DNA bending by an adenine--thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 8490–5 (2001).
156. Deniz, O. *et al.* Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics* **12,** 489 (2011).
157. Rhodes, D. Nucleosome cores reconstituted from poly (dA-dT) and the octamer of histones. *Nucleic Acids Res.* **6,** 1805–16 (1979).
158. Kunkel, G. R. & Martinson, H. G. Nucleosomes will not form on double-stranded RNa or over poly(dA).poly(dT) tracts in recombinant DNA. *Nucleic Acids Res.* **9,** 6869–88 (1981).
159. Prunell, A. Nucleosome reconstitution on plasmid-inserted poly(dA) . poly(dT). *EMBO J.* **1,** 173–9 (1982).
160. Puhl, H. L., Gudibande, S. R. & Behe, M. J. Poly[d(A.T)] and other synthetic polydeoxynucleotides containing oligoadenosine tracts form nucleosomes easily. *J. Mol. Biol.* **222,** 1149–60 (1991).

161. Ehrlich, S., Thiery, J.-P. & Bernardi, G. The mitochondrial genome of wild-type yeast cells III. The pyrimidine tracts of mitochondrial DNA. *J Mol Biol* **65,** 207–212 (1972).
162. Farge, G. *et al.* In vitro-reconstituted nucleoids can block mitochondrial DNA replication and transcription. *Cell Rep* **8,** 66–74 (2014).
163. Lavigne, M. & Buc, H. Compression of the DNA minor groove is responsible for termination of DNA synthesis by HIV-1 reverse transcriptase. *J. Mol. Biol.* **285,** 977–95 (1999).
164. Lavigne, M., Roux, P., Buc, H. & Schaeffer, F. DNA curvature controls termination of plus strand DNA synthesis at the centre of HIV-1 genome. *J. Mol. Biol.* **266,** 507–24 (1997).
165. Falconi, M., Colonna, B., Prosseda, G., Micheli, G. & Gualerzi, C. O. Thermoregulation of Shigella and Escherichia coli EIEC pathogenicity. A temperature-dependent structural transition of DNA modulates accessibility of virF promoter to transcriptional repressor H-NS. *EMBO J.* **17,** 7033–43 (1998).
166. Prosseda, G. *et al.* The virF promoter in Shigella: more than just a curved DNA stretch. *Mol. Microbiol.* **51,** 523–37 (2004).
167. Hagerman, P. J. Sequence-directed curvature of DNA. *Annu. Rev. Biochem.* **59,** 755–81 (1990).