High Throughput Computational Studies of Macromolecular Structure Flexibility

Adam Hospital Gasch

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR

High Throughput Computational Studies of Macromolecular Structure Flexibility

Adam Hospital Gasch

2014

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR

DOCTORAT EN BIOTECNOLOGIA

High Throughput Computational Studies of Macromolecular Structure Flexibility

Memòria presentada per Adam Hospital Gasch per optar al títol de doctor per la Universitat de Barcelona

Dirigida per:

Josep Lluís Gelpí Buchaca

Modesto Orozco López

_____

_____

Doctorand:

Adam Hospital Gasch

_____

I. Acknowledgments ..... *iii*
II. List of Figures ..... *v*
III. Abbreviations ..... *vii*

1. Introduction *1*

1.1. Structural Bioinformatics *1*
1.2. Macromolecular Structures *2*
1.3. Macromolecular Dynamics simulation techniques *3*
1.3.1. Molecular Dynamics *3*
1.3.2. Coarse-Grained Dynamics *4*
1.4. MD & High Performance Computing (HPC) *5*
1.5. Biomolecular Data Storage *6*
1.6. Web-Based Bioinformatics *7*
1.6.1. Web Services *7*
1.6.2. Semantic Web Services *8*
1.6.3. Data Ontologies *9*
1.6.4. Web Services Workflows *10*
1.6.5. BioMOBY *11*
1.7. Biological Databases *11*
1.7.1. Structural Databases *11*
1.7.2. Non-structural Databases *14*
1.8. Section Bibliographic References *18*

2. Objectives *24*

3. Methods *27*

3.1. Macromolecular Dynamics simulation techniques *27*
3.1.1. Molecular Dynamics (MD) *27*
3.1.2. Coarse-Grained Dynamics (CG) *37*
3.2. Trajectory Files *43*
3.2.1. Topology Formats *44*
3.2.2. Trajectory Formats *45*
3.3. Trajectory Analyses *49*
3.3.1. Standard Cartesian Analyses *49*
3.3.2. Solvent Accessible Surface Area (SASA) *52*
3.3.3. Hydrogen Bonds (HB) *52*
3.3.4. Principal Component Analysis *53*
3.3.5. NMR Observables *54*
3.3.6. Protein-specific Analyses *56*
3.3.7. Nucleic Acid-specific Analyses *58*
3.3.8. Solvent-specific Analyses *64*
3.4. Biological Databases *68*
3.5. On-line Tools *68*
3.5.1. PHP *68*
3.5.2. JavaScript *68*
3.5.3. MySQL *69*
3.5.4. Jmol *69*
3.5.5. BioMOBY Web Services *70*
3.6. Section Bibliographic References *71*

4. PhD Advisor Report …… *75*

5. Results …… 77

5.1. Automatic extraction of interesting information from available structural data … *77*
5.1.1. Synopsis …… *77*
5.1.2. **Results:** Structural databases framework ……. *78*

5.2. Generation of a protein dynamics library of MD trajectories …… ***83***
5.2.1. Synopsis …… *83*
5.2.2. **Paper 1:** MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories ……. *84*

5.3. High throughput analysis of hydration solvent from MD simulations …… ***115***
5.3.1. Synopsis …… *115*
5.3.2. **Paper 2**: Water-omics: High throughput analysis of protein-solvent interactions from MD simulations …….. *116*

5.4. Development of data types and informatic tools to port macromolecular dynamics methods to the high-throughput regime …… ***156***
5.4.1. Synopsis …… *156*
5.4.2. **Paper 3:** MDWeb & MDMoby an integrated web-based platform for molecular dynamics simulations ……. *157*
5.4.3. **Paper 4:** FlexServ: An integrated tool for the analysis of protein flexibility …… *171*

5.5. Development of a tool for structure generation, dynamic simulation and trajectory analysis of nucleic acids ……... ***174***
5.5.1. Synopsis …… *174*
5.5.2. **Paper 5:** NAFlex: A web server for the study of nucleic acids flexibility …… *175*

5.6. Section Bibliographic References …… *194*

6. Global Discussion …… *195*

7. Conclusions …… *213*

8. Summary (Spanish) …… *215*

9. List of publications …… *235*

Annex I: High-throughput MD simulations review …. *237*
Annex II: Semantic Web and Web Services Definitions …. *253*

I. Acknowledgments

Em temo que necessitaria tot un capítol d'aquesta tesis per no deixar-me a ningú que, de manera activa o passiva, n'ha format part. Han sigut 10 anys. Molta il·lusió, moltes hores, molt d'aprenentatge, però també molts companys, amics, i experiències que mai oblidaré. Tenia moltes ganes d'escriure aquest apartat, perquè estic convençut que molts anys endavant, quan torni a llegir aquest manuscrit, la bioinformàtica existent ja no tindrà res a veure amb la descrita en els capítols següents, però els sentiments acumulats durant aquests anys es mantindran intactes, tot i que amagats, esperant la lectura d'aquests paràgrafs per aflorar de nou. Aquests agraïments doncs, són tant per tota la gent a qui van dirigits, com per a mi mateix.

Començant pel principi, gràcies a la colla de bioinformàtics, liderats per tot un *gentleman*, Xavier de la Cruz, amb qui he tingut la sort de compartir més que feina, i amb el seu gran deixeble, David Talavera, amb qui, a part de compartir ordinador durant un temps (sí sí, un sol ordinador -*Borzoi*-, amb dos teclats i dos monitors), sempre recordaré les apassionants sobretaules literàries, científiques... i greixoses de Geologia. Gràcies al David Piedra i al Sergi Lois per les innumerables estones de diversió i entreteniment, que van arribar a quotes desorbitants amb l'arribada al grup d'una gallega -la Rebe- que desbordava felicitat per tots costats.

En el mateix període de temps, m'agradaria també agraïr l'acollida al grup dels "*veterans*" que vaig tenir la sort de conèixer abans de que emprenguessin el vol a importants grups d'arreu del món. Blas, Carles, Manu, Alberto, Agnès, Antonio, Ramón, Tim, i segur que em deixo algú. Tots em van ensenyar molt, en un moment en què fins i tot les altíssimes dosis d'il·lusió i entusiasme estaven trontollant per la descomunal quantitat de nova informació a processar.

Una època especialment dolça la vull agrair a tots els companys que em van acompanyar a la sala INB. Aprofito per agrair a l'Institut Nacional de Bioinformàtica, l'organisme que ha fet possible aquesta tesis gràcies a la seva aportació econòmica, i al que desitjo un bon futur, després d'haver sobreviscut a la corda fluixa durant 10 anys ja. En ple auge de l'INB, vam arribar a tenir la nostra pròpia sala de treball, on vaig tenir la sort de conviure amb els físics Oliver (ASP d'Aspirina) i Agustí (vigileu l'esmorzar), i els informàtics ("*los Ivanes*" Sanchez i Párraga) i el Jordi Camps. Tots ells, juntament amb els bioinformàtics, i liderats pel Mr. Párraga, vam fer la pinya més gran i ben avinguda que recordo d'aquesta llarga etapa, que ha portat a amistats que espero no perdre mai.

Encara amb el rerefons INB, he d'agrair el grapat de bons moments passats amb *l'INB Team*, el grup de científics i treballadors de l'INB que han viscut a l'altra banda de la diagonal (BSC). Alexis, Laia, Dmitry, Aida, Carles Pons, i sobretot Romina, la cap del grup, una *currante* incansable, capaç de molt més del que ella mateixa creu. Gràcies per les innumerables reunions INB i jornades de bioinformàtica per tot el país, per acompanyar-me sempre, fins i tot quan al meu grup INB ja només hi quedava jo.

Voldria agrair també la proximitat sense condicions que em van demostrar una parella d'italians que van arribar un dia al grup, i que gairebé sense adonar-nos-en, vam acabar formant part del seu casament i del naixement del Pau (català 100%). Gràcies als dos, però sobretot gràcies al Marco per ensenyar-me un altre tipus de ciència i a confiar cegament en els companys.

Una època que recordo amb alegria i a la vegada amb tristesa és la compartida amb el Juan Fernández-Recio. En aquells moments, un Ramón y Cajal arribat de grups molt importants amb ganes de començar nous projectes, amb una vitalitat que encomanava a tothom. Vaig tenir la gran sort de treballar una temporada amb ell i d'aprendre moltíssim del seu camp, tot i que sempre em quedarà l'espina de no haver pogut portar cap dels projectes a bon port. Gràcies Juan per tot el que vam compartir, no només en el camp científic, sinó també per les xerrades amb el David.

Ja més recentment, he d'agrair els *brainstormings* amb "*los Ramones*", el grup format pel que antigament havia sigut company de grup, el Ramón Goñi. Gràcies pels consells informàtics del Carles Fenollosa, i sobretot gràcies al Pau, per ser tan bon company de feina com bona persona. Estic convençut que us espera un gran futur en el camp més informàtic de la bioinformàtica.

Gràcies també al Guillem, company de pàtria, per ensenyar-me tot el que sé sobre Gromacs.

Quan vaig començar la carrera d'informàtica a Barcelona recordo que la meva mare em va dir: nen, tu ajunta't amb gent intel·ligent i tot serà més fàcil. No se com m'ho vaig fer, però ho vaig aconseguir i sincerament creia que ja no podia superar el límit marcat pels companys de carrera. Però la vida dóna moltes sorpreses, i un dia va arribar al grup un noi anomenat Oscar Flores, que amb només 3 anys va ser capaç d'engegar tot sol una branca nova d'investigació dins el grup, i que no content amb això va compaginar-ho amb un eMBA durant el seu últim any de tesis. Acabo aquest treball just quan està a punt d'engegar la seva primera empresa. El temps dirà. Jo no el perdria de vista. Sigui com sigui, gràcies per les teves sorprenents reflexions que no ens deixaven mai indiferents, i per les nostres classes d'*spinning*.

El món de la bioinformàtica requereix grans recursos computacionals, i la tendència continua essent exponencial. Vaig començar la tesis, com ja he esmentat, treballant amb un ordinador compartit, i amb un *cluster* de càlcul de 8 màquines. Avui treballem conjuntament amb el BSC, amb màquines que en aquell moment no podíem ni somiar. És difícil veure la gran importància dels administradors de sistemes dins d'aquest engranatge, però us asseguro que és vital. La feina feta primer per l'Ivan Sanchez i després pel Jose, totalment en solitari, és espectacular. El número d'usuaris del grup, capacitat de càlcul i d'emmagatzematge segueix pujant, i tot està, en la mesura del possible, controlat. Felicitats als dos per la feina feta. I per acollir-me a mi com a persona, fora de la feina, i obrir-me la vostra vida.

Més sorpreses de la vida, un dia qualsevol d'estiu, quan la gent ja estava més pendent de les vacances que de la feina, va arribar un noi de Saragossa, per passar l'estiu al parc. El Nacho va entrar sense saber res de programació, i ara és tot un doctor, que fa just dos dies m'explicava noves tècniques de dinàmica molecular. I d'aquí un mes se'ns casa amb l'Eva, una científica belga feta a mida per ell. Allà estaré, i ja aprofito per desitjar-los una llarga i feliç vida en la ciència i en el matrimoni. Mai li podré agrair prou les xerrades, sortides, dinars... i el més important, que m'ensenyés, d'una vegada per totes, a parlar el castellà!

Durant tots aquests anys, tal i com dicta el món de la ciència, la rotació de companys ha sigut constant. Però he tingut la gran sort de tenir un punt d'ancoratge que sempre ha estat allà. Els continus dinars setmanals amb les mútues explicacions de la vida, tant personal com professional, m'han ajudat a entendre els assoliments i les equivocacions, i tinc la sensació de que, poc a poc, hem anat creixent junts. Gràcies Sergi, espero que passi el que passi en un futur, trobem la manera de continuar aquestes trobades.

Agrair a un director de tesis acostuma a ser una obligació de protocol. Però deixeu-me dir que aquest no és el cas. Vull donar les gràcies, i de tot cor, a la persona que m'ha ensenyat gran part del que se sobre la bioinformàtica, que m'ha fet costat en tots els projectes, que m'ha acompanyat a totes les xerrades, i que finalment ha perdut moltes estones de la seva vida llegint i rectificant documents, articles i finalment aquesta tesis. I molt especialment vull agrair la seva ajuda durant aquest últim any, un any on malauradament ha passat per unes situacions personals extremes, i que malgrat això, sempre ha aconseguit trobar uns moments per mi i la meva feina. Gràcies Josep Lluís per aquests 10 anys al meu costat. Et puc assegurar que aquest sentiment d'agraïment és compartit amb tots els teus "*nens*" de l'INB.

Gràcies també al Modesto, per creure en mi durant tots aquests anys i permetre'm allargar la meva aventura fins avui.

I finalment, necessito donar les gràcies als que han permès que arribés fins aquí, educant-me, ensenyant-me, fent-me costat, donant tot el que tenien i més perquè pogués estudiar a Barcelona, recolzant-me amb totes les meves decisions. Gràcies a tota la meva família, perquè diuen que la família no es tria, però jo he tingut molta sort.

I per descomptat, gràcies a qui porta mitja vida aguantant-me, ensenyant-me tot el que jo mai hagués arribat a fer sol, demostrant-me que la vida s'ha de viure i que les sorpreses no deixaran d'aparèixer, sempre que la tingui al meu costat. Gràcies per compartir la meva vida, en els bons i els mals moments, sempre. Per tu, Judith.

II. List of figures

Figure 1.1.- *Protein Data Bank yearly growth* *2*

Figure 1.2.- *Macromolecular dynamics time scales* *3*

Figure 1.3.- *Example of a Coarse-Grained DNA representation* *5*

Figure 1.4.- *Biological process Gene Ontology* *9*

Figure 1.5.- *Microarray Taverna workflow* *10*

Figure 1.6.- *Crystal asymmetric unit* *13*

Figure 3.1.- *Histidine ionization states* *34*

Figure 3.2.- *Weighted RMSd* *50*

Figure 3.3.- *Solvent Accessible Surface Area* *52*

Figure 3.4.- *Hydrogen Bond interactions* *53*

Figure 3.5.- *DNA sugar H1'-2H2' J-Coupling* *55*

Figure 3.6- *Ribose ring torsion angles* *58*

Figure 3.7.- *Sugar puckering pseudo-rotational circle* *59*

Figure 3.8.- *Nucleic acid backbone torsion angles* *59*

Figure 3.9.- *BI-BII conformations* *60*

Figure 3.10.- *Axis base pairs* *60*

Figure 3.11.- *Intra-base pairs* *61*

Figure 3.12.- *Inter-base pairs* *61*

Figure 3.13.- *Nucleic acid major and minor grooves* *62*

Figure 3.14.- *Hydrogen bonding in Watson-Crick base pairing* *62*

Figure 3.15.- *HB/Stacking interactions* *63*

Figure 3.16.- *Protein-Water HB dynamics graph* *67*

Figure 3.17.- *Jmol applet molecule viewer* *69*

Figure 5.1.- *PDB derived metadata database tables ……. 81*

Figure 5.2.- *Structure important parts schema and database tables …….. 82*

Figure 5.3.- *Protein active sites database tables ……. 83*

Figure 5.4.- *MMB PDB Mirror …… 84*

Figure 6.1.- *MoDEL Coverage …… 199*

Figure 6.2.- *MoDEL Web Server …… 201*

Figure 6.3.- *MDWeb & FlexServ web servers …… 202*

Figure 6.4.- *NAFlex Web Server …… 203*

Figure 6.5.- *MDWeb & NAFlex structure checking ……. 204*

Figure 6.6.- *MDWeb & NAFlex internal configuration ……. 205*

Figure 6.7.- *MDWeb usage statistics …… 206*

Figure 6.8.- *FlexPortal: INB Integrated platform for macromolecular flexibility … 207*

III. Abbreviations

3D: Three-dimensional

ASCII: American Standard Code for Information Interchange

BD: Brownian Dynamics

BioUnit: Biological Unit

BMRB: Biological Magnetic Resonance Bank

BSC: Barcelona Supercomputing Center

CASP: Critical Assessment of Protein Structure Prediction

CG: Coarse-Grained

CGI: Common Gateway Interface

CPU: Central Processing Unit

CSS: Cascade Style Sheet

CSV: Comma Separated Value

DMD: Discrete Molecular Dynamics

EBI: European Bioinformatics Institute

EC: Enzyme Commission (Number)

ED: Essential Dynamics

EMBL: European Molecular Biology Laboratory

FPGA: Field-Programmable Grid Array

GO: Gene Ontology

GPU: Graphics Processing Unit

GUI: Graphical User Interface

HB: Hydrogen Bond

HDF: Hierarchical Data Format

HMM: Hidden Markov Model

HPC: High Performance Computing

HT: High Throughput

HTML: HyperText Markup Language

INB: Instituto Nacional de Bioinformática

JPEG: Joint Photographic Experts Group

JSON: JavaScript Object Notation

LRT: Linear Response Theory

MD: Molecular Dynamics

mmCIF: Macromolecular Crystallographic Information File

MPI: Message Passing Interface

MRT: Mean Residence Time

MST: Mean Square Displacement

NA: Nucleic Acid

NCBI: National Center for Biotechnology Information

NetCDF: Network Common Data Form

NFS: Network File System

NMA: Normal Mode Analysis

NIH: National Institute of Health (US)

NMR: Nuclear Magnetic Resonance

NOE: Nuclear Overhäuser Effect

OWL: Web Ontology Language

PCA: Principal Component Analysis

PC: Personal Computer

PDB: Protein Data Bank

PDBML: Protein Data Bank Markup Language

Pfam: Protein Family Database

PHP: Hypertext PreProcessor (Recursive Acronym)

PIR: Protein Information Resource

PNG: Portable Network Graphics

PSF: Protein Structure File

PSI: Protein Structure Initiative

RCSB: Research Collaboratory for Structural Bioinformatics

RDF: Radial Distribution Function

REST: Representational State Transfer

RMSd: Root Mean Square deviation

SADI: Semantic Automation, Discovery and Integration

SASA: Solvent Accessible Surface Area

SAWSDL: Semantic Annotations for WSDL

SCOP: Structural Classification of Proteins

SG: Structural Genomics

SIB: Swiss Institute of Bioinformatics

SOAP: Simple Object Access Protocol

TNG: Trajectory Next Generation

UniMES: UniProt Metagenomic and Environmental Sequences

UniProt: Universal Protein Resource

UniProtKB: UniProt KnowledgeBase

UniRef: UniProt Reference Clusters

URI: Uniform Resource Identifier

XML: Extensible Markup Language

WS: Web Services

WSDL: Web Services Description Language

wwPDB: Worldwide Protein Data Bank

WWW: World Wide Web

W3C: World Wide Web Consortium

1. Introduction

The discovery of the DNA helix structure in 1953 by James Watson and Francis Crick [1] and the resolution of the 3-D structure of the first protein by Max Perutz and John Kendrew in 1958 [2] filled the gap between chemistry and biology and opened the door to the understanding of life from basic physical principles. More recently, the sequencing of the human genome [3, 4] provided with the basic information on the cell proteins, and raised several structural genomic projects that aimed to obtain the 3-D structure of the entire proteome [5, 6], resulting in a spectacular growth in the number of protein structures solved. However, as macromolecular structure field was evolving, new clues and discoveries pointed out that structure is not enough to understand protein function, but information about the dynamics of the macromolecules should be considered. Biological macromolecules are large and flexible entities that perform their function through the recognition of other molecular entities. Coupled to the recognition process, mutual conformational adjustments in the interaction partners occur. Proteins have an intrinsic ability to undergo functionally relevant conformational changes under native state conditions. Well known examples are enzymes, mobility proteins, receptors, transporters or protein channels [7].

Flexibility is not only a unique property of proteins. Nucleic Acid (NA) flexibility is known to be important since the determination of the first double-stranded DNA crystal structure by Drew & Dickerson in 1981 [8]. The crystallized structure came with a number of irregularities when compared with the canonical double helix, with bends, non-planar base pairs, variations in groove widths, etc., being the first experimental evidence to reflect the amount of flexibility that DNA could show [9]. Current studies about DNA flexibility covers a wide range of fields, from high resolution studies such as DNA wrapping around nucleosomes [10] and protein induced nucleic acid flexibility [11] to low resolution mesoscopic DNA dynamics [12].

Intense efforts are being made to obtain experimental information about protein and nucleic acids flexibility. However, despite encouraging advances particularly in NMR spectroscopy, we are far from achieving a complete description of the flexibility of a molecular system by experimental methods, and theoretical approaches become typically the only possible alternative. One of the most used theoretical techniques to account for dynamic information of structures is Molecular Dynamics (MD) [13, 14]. Unfortunately, the practical use of MD has been severely limited by its computational cost and by the problems found in the setup and analysis of simulations. As computer power increases, problems related to the setup and analysis of trajectories is becoming more prevalent.

1.1. Structural Bioinformatics

With millions of proteins now sequenced, the next obvious step is trying to understand the function of each of those gene products. The number of solved protein structures is growing exponentially, and new computational methods are being able to predict the structure of a protein molecule in those cases where it is not known experimentally. For these reasons, a new discipline called *Structural Bioinformatics* has appeared, trying to rationalize and classify the information contained in the three dimensional structures of molecules, and to derive functional information from such structural data [15].

1.2. Macromolecular Structures

The main interest of this thesis is focused on the study of macromolecular dynamics. For this reason, the starting point of the studies is typically the molecular three-dimensional (3D) structure derived by modelling, or by experimental sources: *X-Ray Crystallography* and *Nuclear Magnetic Resonance* (NMR).

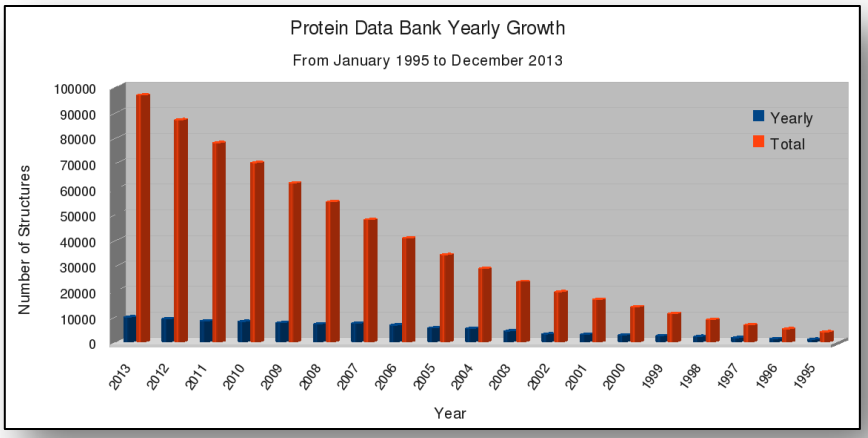


Fig 1.1.- Protein Data Bank Yearly Growth. *Blue cylinders represent the number of structures deposited per year, whereas red cylinders represent the total number of structures stored in the PDB.*

Since the early 1970s, solved structures are stored in a worldwide repository called *Protein Data Bank* (PDB) [16]. The PDB was started in 1971 at *Brookhaven National Laboratory* (New York) and originally contained 7 structures. Nowadays, with the explosion of *High Throughput* (HT) crystallographic initiatives such as the *Structural Genomics* (SG), the growth rate is about to reach a number of 10,000 new structures per year (Fig. 1.1). *Structural Genomics (SG)* was an initiative started in the late 1990s willing to create a foundation for a systematic worldwide effort to determine the structures of proteins [5]. This enterprise ended up with the establishment of major efforts in the United States with the *Protein Structure Initiative* (PSI), in Europe with the *Protein Structure Factory* and *SPINE* projects and in Japan with the *National Project on Protein Structural and Functional Analyses* and the *RIKEN Structural Genomics/Proteomics* initiative. During the past decade, many additional groups have joined the SG effort. From all these initiatives, the most successful one has been PSI which has provided thousands of protein structures, many of them of biomedical impact [6].

In spite of the impressive growth in the number of structures deposited in the PDB, the existent gap between known sequences and structures is still discouraging. Large-scale sequencing projects have increased the available protein sequence information, far exceeding the ability to characterize 3D protein structures. During the recent years, several computational approaches have been developed trying to predict three-dimensional structures in the HT-regime. The prediction of the structure of a protein from its amino acid sequence information is not a straight-forward process, and we are still far from having methods that could generate accurate high resolution structures. However the field is very active, and a biannual competition on *Critical Assessment of Protein Structure Prediction* (*CASP*) is being performed since 1994,

joining the best scientists working on this problem, and contributing to sustained improvement in the methods [17].

1.3. Macromolecular Dynamics Simulation Techniques

During the last decade, the most used and well-known theoretical technique for the prediction of macromolecular flexibility has been Molecular Dynamics (MD). Present MD covers a wide range of time scales, although at the expense of still requiring different levels of representations, from the atomistic one to the, so called, Coarse-Grained, popular in the first days of the study of protein folding, and recovered now (Fig. 1.2).

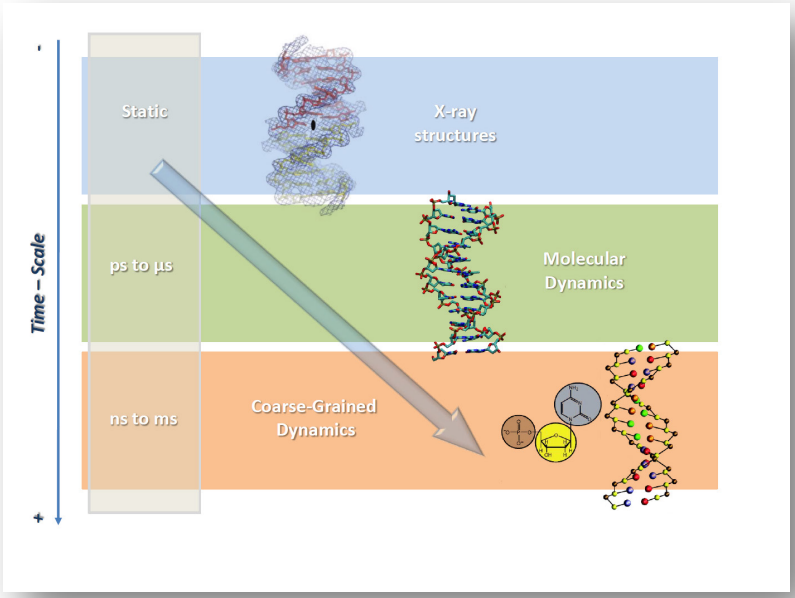


Fig 1.2.- Macromolecular Dynamics Time Scales. *Temporal coverage of theoretical techniques for going from a static structure (e.g. X-ray) to a dynamic representation of macromolecular structures.*

1.3.1. Molecular Dynamics

Molecular Dynamics is nowadays a mature technique that is being used for more than 30 years. The first study on bio-macromolecules developed with the help of MD came to light in 1977, by the group of Martin Karplus [13, 18], who studied the dynamics of the bovine pancreatic trypsin inhibitor (BPTI), a small globular protein. Although it was done in vacuum (without any kind of solvent surrounding the protein structure) and lasted only 9.2 ps, it was enough to change the early view of proteins as static structures to a new one where proteins had to be seen as dynamical systems, with internal motions playing a functional role [19]. Much has been done since then in the field of molecular dynamics, and the number of studies published using MD methods is still increasing today.

Nowadays, MD simulations can work with systems having biological relevance, as entire proteins in solution (with explicit solvent representation), membrane-embedded proteins, or large macromolecular complexes like nucleosomes [20] or ribosomes [21, 22]. However, in spite of all its success, MD technique has been conceptually limited, apart from the problems associated with the setup and analysis of

trajectories previously introduced, by a couple of algorithm-related factors: it is very computational demanding, and the physical models used (force-fields, see Methods section) are still not accurate enough. Algorithm improvements in MD codes during the past 5 years have led to an increase in performance by over three orders of magnitude. The fact that in the recent years a higher transistor density has translated to more processor cores per chip, instead of to faster individual processor cores has pushed MD field to work hardly on parallelism, using many processors for a single simulation. Improvements in parallel scalability and efficiency have been possible thanks to a number of algorithmic innovations, particularly in methods for reducing the communication requirements between processors. Coupling of these new methods with the now available *High Performance Computing* (HPC, see section 1.4) has turned out into increasing simulation capabilities. Impressive advances are also emerging from the use of MD-specific chips, from the modification of MD codes to take advantage of Graphical Processing Units (GPUs), and even from the possibility to perform ensemble-MD simulations in GRID platforms (section 1.4).

On the other hand, commonly used biomolecular force-fields have been improved, in part thanks to the increase of computational power that has led to longer simulations, exposing hidden force-field caveats. Large efforts have been recently directed towards nucleic acid [23, 24] and protein [25, 26] physical description. We can expect further improvements related to the use of more flexible force-fields going beyond the pair-additive functionals.

1.3.2. Coarse-Grained Dynamics

Most of the relevant macromolecule dynamics and interactions within cells, protein-protein docking, ligand binding, biochemical reactions or even protein folding, occur beyond the microsecond scale, beyond the time scale currently feasible with all-atom MD simulations [27]. This, and the need to study larger macromolecular systems are the main reasons for the development of approximate coarse-grained (CG) models, where, in order to increase computer efficiency, a certain loss of accuracy is accepted with a significant reduction of the complexity of the system, and thus in its structural resolution. CG methods use a simplified representation of the system, usually reducing the number of particles, compressing sets of atoms in pseudo-particles (usually called beads) (Fig. 1.3). The representation of the solvent environment is also generally simplified, either simulated as a continuum medium or represented by particles formed by clusters of solvent molecules.

The most common level of simplification for CG in proteins involves the representation of every residue by a single particle located at the C_α. In general, as the number of beads decreases, the simulation becomes less expensive and larger systems and longer trajectories are possible. However, these methods, as MD, work with parameterized force-fields to describe the physical properties of the system, and they have to be re-parameterized for every CG resolution, which becomes increasingly difficult as the CG resolution decreases, because more specific interactions must effectively be included in fewer parameters and functional forms. For this reason, most currently popular CG models are parameterized based on a single reference configuration and the dynamics they reproduce are strongly biased towards it [28].

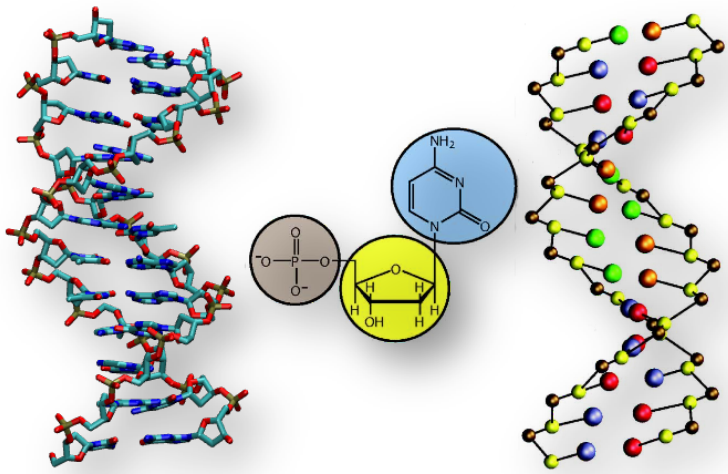


Fig 1.3.- Example of a Coarse-Grained DNA Representation. *Coarse-Grained 3-beads representation of a DNA molecule: every nucleotide is divided in three beads, joining together all atoms belonging to sugar (yellow bead), base (blue bead) and phosphate (grey bead).*

1.4. MD & High Performance Computing (HPC)

The present generation of computers takes benefit of parallelism and accelerators to speed up simulations. The most used MD programs: *AMBER* [29], *CHARMM* [30], *GROMACS* [31] and *NAMD* [32], have been long ago compatible with the *Message Passing Interface* (*MPI*), a protocol for computer-to-computer communication that permits work-sharing between processors. When a large number of computer cores can be used simultaneously, MPI can greatly reduce the computation time. With the current availability of large supercomputers, usually built-up with more than 10,000 processors, MPI has become the most popular technique to run MD simulations.

On the other hand, with just a few years, a new hardware resource coming from the field of computer gaming has risen as the best platform to perform MD simulations: *Graphical Processing Units* (GPUs). These processors, specifically designed to accelerate the generation of frames per second in 3D-games, have found to be extremely efficient for running MD algorithms. They can deliver over an order of magnitude more floating-point operations per second than classical central processing units (CPUs). Although a major redesign of the MD code is necessary to take full advantage of GPU power, the majority of the most popular MD programs offer today a version specifically designed to work with this kind of hardware. New MD programs have also appeared written from scratch to take a maximum profit from the GPU power (e.g. *ACEMD*) [33]. Unfortunately, parallelization of processes across GPUs is difficult, because communication between GPUs remains slower than communication between classical processors, although different groups, particularly the AMBER one, are making advances toward direct GPU-GPU communication, which would certainly improve scalability. HPC strategies to work with MD are thus going to a combination of MPI and GPU processes.

D. E. Shaw Research in New York chose a new way to deal with MD drawbacks. They built a completely new, special-purpose parallel supercomputer for molecular dynamics: Anton. Each Anton chip, containing application-specific integrated circuits (ASICs) specifically designed to perform MD calculations, is able to compute MD particle interactions hundreds of times faster than a common CPU processor. The result is a machine that performs all-atom protein simulations over one hundred times longer than any one published before [34]. Thanks to this machine, they were able to publish the first 1ms-length MD simulation of BPTI protein [35]. Unfortunately, Anton is a private resource, and only one of such machines is available for public (US) research at the *National Resource for Biomedical Supercomputing* at the *Pittsburgh Supercomputing Center*.

In a completely opposed way to the supercomputing approach, large initiatives have raised using Distributed Computing (DC) or Grid Computing (GC). These approaches use a collection of computer resources from multiple locations to perform ensemble simulation, assuming that the system moves in the Ergodic regime, and ensemble and time average are equivalent. The idea is surprisingly easy: when the PC was not executing anything, a screensaver was started, triggering the execution of the analysis. This way, dividing a huge amount of data into little independent pieces and sending them to volunteers, they were able to overtake the performance of a supercomputer. This divide-and-conquer approach has tackled one of the main challenges in molecular science: protein folding (http://folding.stanford.edu). Similar initiatives have been recently emerged, with spectacular success, using also GPUs (*GPUGRID*) [36] or even Sony Playstation3 Cell processors (*PS3GRID*) [37].

1.5. Biomolecular Data Storage

Storage and access to the data is a major issue in current MD projects. As computer power increases, larger simulations are run, and thus larger trajectory files are produced. Several file formats exist for storing coordinates information, the majority of them optimally binary-encoded. However, even with the best data encoding, a raw 10 ns length MD trajectory, with explicit water solvent and an average protein size, still needs more than 5Gb of disk space. Expanded to the current time scales and HPC methodologies, MD data storage and retrieval is clearly an obstacle to confront, which has originated the development of several compressing strategies (see Methods section). In addition to its size, efficient data storage and mining requires for data standards. The *Scalalife* project (http://www.scalalife.eu) wished to tackle these problems, trying to develop: an efficient data storage system based on new cloud databases paradigm; a standard file format for MD trajectory data; a standard XML-like format to store metadata related to the MD simulations (time, temperature, ensemble, etc.).

A last data-related problem arises from the need to organize the huge amount of trajectories in databases to allow an efficient re-use of the collected information. One of the first attempts back on 2006 was the project *BioSimGrid*, which published a grid-enabled biomolecular simulation data storage and analysis system [38]. The platform worked using a dual approach: metadata was stored in relational databases, whereas raw trajectory data was stored in a disk-based structure. For the flat files, a distributed middleware called *Storage Resource Broker* (*SRB*) was used, providing efficient distributed data transmission in addition to security issues as redundancy and replication. A relatively similar storage approach has been used in *MoDEL*, presented and discussed as part of this thesis (Section 5.2). A slightly different approximation of

scientific data repository for big amounts of data was presented in the *Dynameomics* project [39], with trajectory coordinates loaded into a set of transactional (not relational) databases. More recently, *iBIOMES* project redefined the dual approach used by *BioSimGrid*, implementing a distributed solution to data storage and sharing across research laboratories [40].

1.6. Web-Based Bioinformatics

The number of on-line bioinformatics tools in the field of life sciences is growing at an incredible speed. As for 2014, *Bioinformatics Links Directory* of *Nucleic Acid Research* (http://bioinformatics.ca/links_directory) contains an impressive number of 623 databases and 1,472 web server tools registered [41]. The same *NAR* journal publishes every year a couple of specific issues dedicated to Web Servers and Databases. As for 2014, 13 different web server issues have been published, presenting an average of 100 on-line tools each. The corresponding database part is even more well-established, having published last year (2013) the 20th annual database issue. Only in this issue, 176 bioinformatics databases have been published, half of which describing new online molecular biology databases and the other half providing updates on previously featured databases [42]. And that is just a small number of the whole web-based projects available, most of them advertised only via publications, laboratory web pages or even existing in relative obscurity.

Although these interactive resources have been of enormous benefit to the scientific community over the years, there is still a growing demand for programmatic interfaces allowing the linkage of databases and on-line tools in automated analysis pipelines. A technology that allows this linkage by definition is becoming increasingly popular in life sciences: *Web Services* (*WS*). WS can be easily accessed from most programming languages, and joined together to build complex workflows. A compilation of technical definitions for *Semantic Web* and *Web Services* technologies, used in the next sections, is available in Annex II.

1.6.1 Web Services

Web Services model is a framework for communication between computer applications over the *World Wide Web* (*WWW*). WS usually have an interface describing input, output, function and location, traditionally written in *Web Services Description Language* (*WSDL*) format. Due to their network-based definition, WS are commonly short pieces of software, doing very specific work, such as information retrieval or fast analysis. Web services are nowadays divided in two big worlds: SOAP and REST:

- *SOAP (Simple Object Access Protocol)* has defined the WS standard specifications since 2007 (Version 1.2; http://www.w3.org/TR/soap/). SOAP is a protocol specification for exchanging structured information data in *Web Services* communications. The message format is codified in the web standard XML language, and is wrapped in an envelope together with error and status information, application-related information, and a variety of details providing standards for security and reliability.

- *REST (Representation State Transfer)* is a relatively recent software architecture (although it was firstly introduced in the year 2000) designed to be used in the development of web services. During the last years, many computer scientists and companies have taken this approach as an alternative to SOAP, to the point of being named the "second generation WS". REST services rely on resources, identified by URIs (*Uniform Resource Identifiers*); the transfer of information is done through representation of these resources, which can be codified in many different formats (XML, JSON, CSV, PNG, etc.), using HTTP protocol. As REST web services do not require the amount of information included in SOAP XML envelope, and also because of the ability to transfer binary data, they can be considerably more lightweight. For that same reason, REST WS use much less bandwidth and are usually faster than SOAP ones.

The quantity and heterogeneity of data in the life sciences field has given rise to thousands of WS that provide methods for its analysis, retrieval and integration [43]. Large bioinformatics services providers, such as the *European Bioinformatics Institute* (*EBI*) and the *National Center for Biotechnology Information* (*NCBI*) offer WS access to their resources. These services can then be combined into analysis pipelines or workflows.

Despite their power, WS still have one important limitation: while the WSDL is machine-readable, the meaning of input and output of the operations (the semantics of the service) are hidden to the final user, usually being encoded into an *Extensible Markup Language* (*XML*) file. That makes the building of workflows joining different WS more difficult, as the user has to map the output of one service into the input of the next, losing then all the automation possibilities. This problem was tackled by designing the so-called *Semantic WS*.

1.6.2 Semantic Web Services

Semantic WS can be defined as a network-based infrastructure that comprises not only machine understandable content, but also a set of semantically linked data, therefore easy to be interrogated and retrieve information from [44]. They are built around universal standards for the interchange of semantic data, which makes it easy for programmers to combine data from different sources and services without losing meaning. Semantic descriptions of the methods implemented in WS allow a fully automatic service discovery, whereas input/output semantic data, usually represented as ontologies, allow the automatic building of pipelines and workflows directly chaining sets of WS.

The world of semantic WS is currently being introduced in the whole life sciences field, from medical communities [45] to chemical informatics [46]. Distribution, management, access and sharing of biological data are crucial aspects that semantic web can afford by integrating heterogeneous sources of chemical, biochemical, biological and medical information. However, standardization of semantic web services is still an unresolved issue. While semantic metadata for SOAP WS can be codified through *Semantic Annotations for Web Services Description Language* (*SAWSDL*) files, REST WS still do not have an associated semantic description file with standard specifications. A new technical document for an updated version of WSDL (WSDL Version 2.0) has been submitted to the *World Wide Web Consortium* (W3C),

adding important capabilities associated to semantic web, such as the ability to describe RESTful WS and ontologies (http://www.w3.org/TR/wsdl20/).

This possibility to choose between different WS technical specifications and formats make even harder the task to collect and use together the huge offer in biological WS. Initiatives such as *BioCatalogue* [43] try to gather a big amount of life sciences WS in a central repository, offering services discovery through rich semantic annotations. Another approximation, developed at *Barcelona Supercomputing Center* (*BSC*), achieved the difficult goal of centralize, in a single registry, WS developed using different protocols, with different semantic annotation technologies, and different life science ontologies: *BioSWR* [47].

1.6.3 Data Ontologies

A data ontology is a formal system for representing knowledge [48]. Ontologies help managing complexity in the information processing, as knowledge is broken down into clear concepts that can be considered independently. Such structured data, once linked, form the basis of the semantic web.

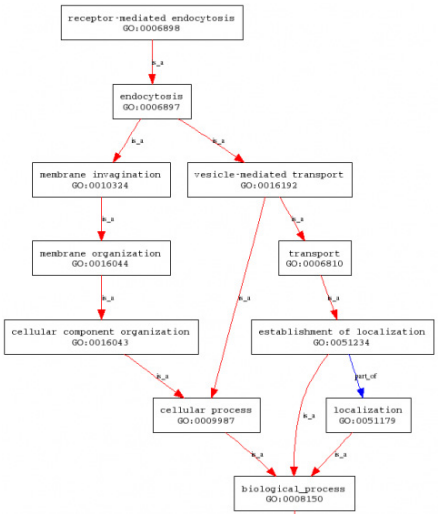


Fig 1.4.- Biological Process Gene Ontology. *Ontology example from the well-known Gene-Ontology database: SORL1 biological process. Red lines indicate that the process higher in the hierarchy is a member of the box directly connected to it and blue lines indicate the process is a part of the larger process it is directly connected to.*

Ontologies are usually composed of classes, properties (relations between different classes) and restrictions. Classes are built in a hierarchical way, where child classes are more specific than, and inherit the properties of, parent classes. Classes can have properties (attributes) which could be related to other classes, and restrictions (constraints) defining valid values (Fig. 1.4).

The big impact of this knowledge management system in life science is demonstrated in the *EBI Ontology Lookup Service* (http://www.ebi.ac.uk/ols), a web portal providing several means to query, browse and navigate biomedical ontologies. The last version (1.21) from May 2014 contained 93 different ontologies [49]. The continuous increase in the number and scope of biological ontologies has triggered several projects trying to join those ones belonging to the same field. *Protein Standards Initiative Molecular Interaction (PSI-MI)* wants to develop a schema to enable the description of interactions between a wider range of molecular types, nucleic acids,

chemical entities and molecular complexes [50]. The project aims to standardize the big number of interactions data sets currently available, represented in many different forms and database schemas. A more ambitious project wants to develop a standard for pathway data sharing: *Biological Pathway Exchange (BioPAX)*. Representing biological pathways at the molecular and cellular level to facilitate the exchange of pathway data implies the combination of already existent different ontologies: metabolic and signaling pathways, gene regulatory networks, molecular interactions (with the above-mentioned PSI-MI) and genetic interactions [48]. This standardization will support easily visualization, analysis and biological discovery of pathway data.

1.6.4 Web Services Workflows

One of the strongest points of WS is their capacity to be chained together to form complex workflows. Thanks to the capacity of the semantic web to automatically discover WS fulfilling desired needs, workflows can be designed in a semi-automatic way. Some *Graphical User Interfaces* (*GUI*) to build workflows from WS have been designed during the past years: *Taverna* [51], *jORCA* [52], *Galaxy* [53] or *Kepler* [54]. Taverna is the most well-known and widely used WS GUI [51]. Taverna workbench allows users to identify and combine services by dragging and dropping them onto a workflow design panel, joining them together connecting suitable outputs and inputs. Taverna workflows can be composed not only by WS, but also from a mixture of WS, local scripts and other service types (like *BioMart* queries). Workflows can then be executed, letting Taverna suite to communicate with all the WS, sending corresponding inputs, executing, and receiving outputs until the workflow is completed. *jORCA* is a similar tool developed by the *Spanish National Institute of Bioinformatics* (INB), and was designed to tackle the main WS and workflow usability problems related with data input types/formatting/connecting.

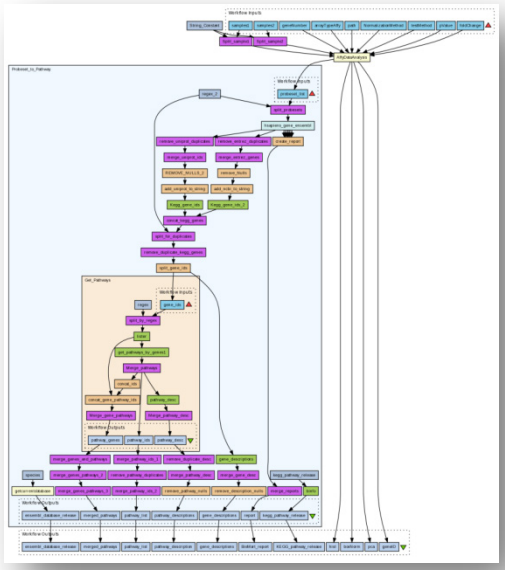


Fig 1.5.- Microarray Taverna Workflow. *Workflow registered in myExperiment entitled Human Microarray CEL file to candidate pathways represented with Taverna suite.*

A repository of public workflows is available at http://www.myexperiment.org. *myExperiment* is an online research environment that supports the social sharing of

bioinformatics workflows [55]. Since its release in 2007, myExperiment currently contains more than 2,500 workflows registered (Fig. 1.5). myExperiment users are developers interested in contributing their workflows into the repository for sharing them with the scientific community and also scientists wishing to discover workflows to be reused in their own research. *myExperiment* currently has over 7,500 members, showing the great interest from the scientific community for bioinformatics workflows.

1.6.5 BioMOBY

BioMOBY is an open source research project from the *Genome Canada Bioinformatics Platform*, and adopted by the *Spanish National Institute of Bioinformatics*, which aims to generate an architecture for the discovery and distribution of biological data through semantic web services, using a biological ontology (www.biomoby.org) [56].

The main goals of BioMOBY projects were:

- Create an ontology of bioinformatics data-types.
- Define a serialization of this ontology (data syntax).
- Create an open API over this ontology.
- Define Web Services inputs and outputs using the ontology.
- Register WS in an ontology-aware registry.

BioMOBY defined an ontology-based messaging standard through which a client is able to automatically discover and interact with task-appropriate biological data and analytical service providers, without requiring manual manipulation of data formats as data flows from one provider to the next. In this way, thanks to the semantic information added by the ontology, complex workflows formed by joining many BioMOBY WS can be almost automatically generated. All the BioMOBY ontology and registered WS can be accessed directly using GUIs like Taverna or jORCA (see previous section). Complex biological workflows can be built combining and connecting available WS with just a few mouse clicks.

Recently, a new project also from the *Genome Canada Bioinformatics Platform* called *Semantic Automation, Discovery and Integration (SADI)* has emerged as the successor of BioMOBY, applying the same semantic and ontology approaches but using new web standard tools and improving the platform thanks to the expertise gained with BioMOBY [57].

1.7 Biological Databases

Biological databases are key pieces in the bioinformatics field. They store and make available crucial information for the scientific community, and, thanks to the WWW, they can be accessed anytime and anywhere.

1.7.1 Structural Databases

1.7.1.1 PDB

The *Protein Data Bank* (*PDB*; http://www.rcsb.org/pdb) is the single worldwide archive of structural data of biological macromolecules [58].

The PDB was started in 1971 at *Brookhaven National Laboratory* (New York) and originally contained 7 structures. Nowadays (May 2014), the number of deposited structures counts 100,147, and the growth rate is about to reach a number of 10,000 new structures per year.

PDB's 3D structures are the basis for the majority of structural bioinformatics studies. They can be downloaded directly from the Protein Data Bank home page or from the different existing mirrors, and with different file formats. The PDB format provides a standard representation for macromolecular structure data mostly derived from X-ray diffraction and NMR studies. It contains the information about atoms, residues and 3D coordinates in plain text, and is the most used and well-known structure format. Other formats are the *macromolecular Crystallographic Information File* (*mmCIF*) and the *Protein Data Bank Markup Language* (*PDBML*) in XML, which add more information contents, at a cost of being less human-readable.

In the year 2003, an international collaboration project to manage the PDB archive was founded: The *Worldwide Protein Data Bank* (*wwPDB*). It consists of organizations that act as deposition, data processing and distribution centers for PDB data. Its members are: *Research Collaboratory for Structural Bioinformatics – PDB* (*RCSB-PDB*) (USA), *PDBe* (PDB Europe), *PDBj* (PDB Japan), and *Biological Magnetic Resonance Bank* (*BMRB*) (USA). The wwPDB's mission is to maintain a single PDB archive of macromolecular structural data freely and publicly available to the global community. One of the initiatives of wwPDB is to create a "remediated" PDB [59], which improve current PDB format in different points, such as the detailed chemical description of ligands and atom nomenclature standardization, among many others. This "remediation" process launched the current PDB format version: 3.30 (http://www.wwpdb.org/documentation/format33/v3.3.html).

The PDB database also contains information about biological assemblies or biological units (*biounits*): macromolecular assemblies that have either been shown to be or are believed to be the functional form of the molecule [60]. The primary coordinate file of a crystal structure typically contains just one crystal asymmetric unit and may or may not be the same as the biological assembly (Fig. 1.6). Depending on the particular crystal structure, symmetry operations consisting of rotations, translations or their combinations may need to be performed in order to obtain the complete biological assembly. Alternately, a subset of the deposited coordinates may need to be selected to represent the biological assembly. Thus, a biounit may be built from:

- One copy of the asymmetric unit.
- Multiple copies of the asymmetric unit.
- A portion of the asymmetric unit.

The biounit part of the PDB builds the corresponding PDB files for each of the biological assemblies, either splitting the original file or applying the necessary symmetry operations.

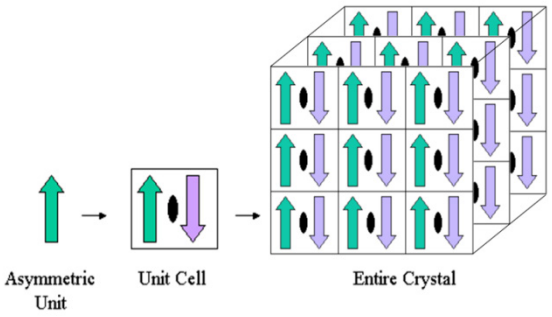


Fig 1.6.- Crystal Asymmetric Unit. *The asymmetric unit (green arrow) is rotated 180 degrees to produce a second copy (purple arrow). Together the two arrows comprise the unit cell. The unit cell is then repeated in three directions to make a 3D crystal.*

1.7.1.2 CATH

Class, Architecture, Topology and Homologous superfamily (CATH; www.cathdb.info*) Protein Structure Classification* database is a hierarchical classification of protein domain structures, using manual curation aided by a variety of classification and prediction algorithms. It was built and published in 1997 by Christine Orengo and Janet Thornton at the *University College, London* [61].

CATH hierarchy is divided in four major levels:

- Class: Structures are classified according to their secondary structure composition (mostly alpha, mostly beta, mixed alpha/beta or few secondary structures).
- Architecture: Structures are classified according to their overall shape as determined by the orientations of the secondary structures in 3D space, ignoring the connectivity between them.
- Topology (fold family): Structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures.
- Homologous superfamily: This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous.

CATH database is in continuous update, and the latest version (v3.5, 2013) contains 173,536 domains, 2,626 homologous superfamilies and 1,313 fold groups. An interesting point is that looking at the Structural Genomics new structures, the number of new folds in the new CATH version is slightly less than for previous releases, suggesting that we have already obtained the majority of exiting folds, or, at least, those that are easily accessible to structure determination [62].

1.7.1.3 SCOP

Structural Classification of Proteins (*SCOP;* http://scop.mrc-lmb.cam.ac.uk/scop) is a database that aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the PDB. SCOP classification of proteins was originally constructed manually by visual inspection and comparison of structures, with

the assistance of tools to make the task manageable. The first version was published in 1994 in the *Laboratory of Molecular Biology* in Cambridge (UK), by Alexey G. Murzin [63]. According to SCOP, proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold:

- Family (Clear evolutionarily relationship): Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater.
- Superfamily (Probable common evolutionary origin): Proteins having low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable.
- Fold (Major structural similarity): Proteins having the same major secondary structures in the same arrangement and with the same topological connections.

During the last years, and mainly due to the growth in available 3D structures, SCOP authors have built automated classification methods that are currently being benchmarked against prior SCOP versions, to check for consistency with previous manual curation. In addition, SCOP has joined ASTRAL database compendium [64] for analyzing protein structures and their sequences, adding an extra value in the information offered. Recently, a prototype of a new version (SCOP 2.0) has been released [65]. It is a major reclassification of the database, with a new complex network of hierarchical and non-hierarchical relationships represented by a directed acyclic graph. This network allows users to represent the structural and most likely evolutionary relationships between proteins in more detail.

Very recently, a new collaborative project called *Genome3D*, integrating UK-based structural resources has been launched to provide a unique perspective on sequence-structure-function relationships, containing the first official mapping between CATH and SCOP databases, thus obtaining useful information from the two resources at various degrees of consensus [66].

1.7.2 Non-structural Databases

1.7.2.1 UniProt

The *Universal Protein Resource* (*UniProt*; www.uniprot.org) is a comprehensive resource for protein sequence and annotation data. It is maintained by the Uniprot consortium since 2002, when the *European Bioinformatics Institute* (*EBI*), the *Swiss Institute of Bioinformatics* (*SIB*) and the *Protein Information Resource* (*PIR*) joined together to share their independent data sets in a single platform [67].

Uniprot resource is built up of four different databases:

- **UniProt Knowledgebase** (UniProtKB): UniProtKB is the core UniProt database, containing a collection of protein sequences derived from the translation of the coding sequences submitted to the public nucleic acid databases. In addition to the amino acid sequence information, UniProtKB stores as much annotation data as possible for each protein entry, from protein name and description to biological ontologies and citation information.

UniProtKB if formed by two different resources:

- ***UniProtKB/Swiss-Prot***: Manually-annotated records with information extracted from literature and curator-evaluated computational analysis (reviewed).
- ***UniProtKB/TrEMBL***: Computationally analyzed records awaiting for a full manual annotation (unreviewed).

UniProt release 2014_05 (May 2014) contains 545,388 entries in manually reviewed UniProtKB/Swiss-Prot, and 56,010,222 entries in the unreviewed UniProtKB/TrEMBL.

- **UniProt Reference Clusters** (UniRef): Provide clustered sets of sequences from the UniProtKB and UniParc databases, removing redundant sequences. UniRef database contains in turn three main clusters:

  - ***UniRef100***: Identical sequences and subfragments with 11 or more residues from any organism are merged into a single UniRef entry, displaying the sequence of a representative protein (usually the longest sequence with biologically relevant information available).
  - ***UniRef90***: Formed by sequences from UniRef100 sharing at least 90% of sequence identity and having an 80% overlap with the longest sequence of the cluster.
  - ***UniRef50***: Formed by sequences from UniRef90 sharing at least 50% of sequence identity and having an 80% overlap with the longest sequence in the cluster.

  UniProt release 2014_05 (May 2014) contains 36,473,742 entries in UniRef100, 21,903,386 entries in UniRef90, and 10,335,613 entries in UniRef50.

- **UniProt Archive** (UniParc): Non-redundant database containing most of the known and publicly available protein sequences. Only aminoacidic sequences are stored, with a given stable and unique identifier, from which all the additional information can be retrieved using cross-references to other existing databases. UniProt release 2014_05 (May 2014) contains 63,875,797 entries in UniParc.

- **UniProt Metagenomic and Environmental Sequences** (UniMES): Repository for metagenomic and environmental data, currently containing only data from the *Global Ocean Sampling Expedition* (*GOS*). Entries stored in UniMES cannot be found in UniProtKB and UniRef databases. UniProt release 2014_05 (May 2014) contains 6,028,191 entries in UniMES (Release 1.0).

1.7.2.2 Protein Family – Pfam

Protein Family (*Pfam*) database (http://pfam.sanger.ac.uk) is a large collection of protein families, defined as sets of regions sharing a significant degree of sequence similarity represented by multiple sequence alignments based on *Hidden Markov Models* (*HMM*) [68]. Pfam is currently produced at the *Wellcome Trust Sanger Institute* and *Howard Hughes Janelia Farm Research Campus* and contains two types of families: high quality, manually curated Pfam-A families and automatically generated Pfam-B families. Pfam-A families are built following a four-step process:

1. Building a high-quality multiple sequence alignment (seed alignment).
2. Constructing a profile Hidden Markov Model (HMM) from the seed alignment.
3. Searching the profile HMM against the UniProtKB sequence database.
4. Choose family-specific sequence and domain gathering thresholds. All sequence regions scoring above the thresholds are included in the full alignment for the family.

Pfam also provides matches for the *National Center for Biotechnology Information* (*NCBI*) non-redundant database, as well as a collection of metagenomic samples. The latest version of Pfam (27.0 as of March 2013) contains 14,831 different families, with a nearly 80% of UniProtKB coverage.

1.7.2.3 Enzyme Database

Since 1955, protein enzymes are identified by an *Enzyme Commission* (*EC*) number code. This number numerically classifies enzymes based on the chemical reactions they catalyse. Each code consists of a set of four numbers separated by periods, defining the classes and sub-classes of enzymes, each number giving a more specific definition than the last. The first number defines the main six enzyme classes: oxidoreductases (EC 1.), transferases (EC 2.), hydrolases (EC 3.), lyases (EC 4.), isomerases (EC 5.) and ligases (EC 6.) [69].

Information relative to the nomenclature of enzymes can be found in the *SwissProt Enzyme database* (http://enzyme.expasy.org) [70]. For each known enzyme, information that can be queried is:

- EC number.
- Recommended name and alternative names (if any).
- Catalytic activity.
- Cofactors.
- Links to Swiss-prot protein sequence and human diseases associated with a deficiency of the enzyme.

1.7.2.4 Gene Ontology – GO

The *Gene Ontology* (*GO*) project (www.geneontology.org) is a collaborative effort to address the need for consistent descriptions of gene products in different databases. Within the GO project, a major bioinformatics initiative was created with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data [71]. The GO

project began as a collaboration between three model organism databases, *FlyBase* (Drosophila), *Saccharomyces Genome Database* (SGD) and *Mouse Genome Database* (MGD), in 1998. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes.

GO project has designed three species-independent ontologies to describe gene products in terms of their:

- Associated biological processes.
- Cellular components.
- Molecular functions.

The use of GO terms by collaborating databases facilitates uniform queries across them. Moreover, ontologies are designed in a way that permit flexible queries (at different levels, joining the different ontologies, etc.) and can be easily extended with new properties.

1.8. Section Bibliographic References

[1] J. D. Watson and F. H. Crick, "The structure of DNA," *Cold Spring Harb Symp Quant Biol,* vol. 18, pp. 123-131, 1953.

[2] J. Kendrew, G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff and D. Phillips, "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis," *Nature,* vol. 181, pp. 662-666, 1958.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. S. Granger and H. O. Smith, "The sequence of the human genome," *Science,* vol. 291, no. 5507, pp. 1304-1351, 2001.

[4] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, 2001.

[5] T. C. Terwilliger, D. Stuart and S. Yokoyama, "Lessons from Structural Genomics," *Annu. Rev. Biophys.,* no. 38, p. 371–383, 2009.

[6] G. T. Montelione, "The Protein Structure Initiative: achievements and visions for the future," *F1000 Biology Reports,* vol. 4, p. 7, 2012.

[7] A. Bidon-Chanal, M. A. Marti, D. A. Estrín and F. J. Luque, "Dynamical regulation of ligand migration by a gate opening molecular switch in truncated hemoglobin-N from Mycobacterium tuberculosis.," *J. Am. Chem. Soc.,* vol. 129, pp. 6782-6788, 2007.

[8] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. E. Dickerson, "Structure of a B-DNA dodecamer-conformation and dynamics.," *Proc. Natl. Acad. Sci. USA,* vol. 78, pp. 2179-2183, 1981.

[9] C. A. Laughton and S. A. Harris, "The atomistic simulation of DNA," *Wires Comput. Mol. Sci.,* no. 1, pp. 590-600, 2011.

[10] F. Battistini, C. A. Hunter, E. J. Gardiner and M. J. Packer, "Structural mechanics of DNA wrapping in the nucleosome," *J. Mol. Biol.,* vol. 396, no. 2, pp. 264-279, 2010.

[11] A. Rubio-Cosials and M. Solà, "U-turn DNA bending by human mitochondrial transcription factor A.," *Curr. Opin. Struct. Biol.,* vol. 23, pp. 116-124, 2013.

[12] R. Collepardo-Guevara and T. Schlick, "Insights into chromatin fibre structure by in vitro and in silico single-molecule stretching experiments," *Biochem. Soc. Trans.,* vol. 41, no. 2, pp. 494-500, 2013.

[13] J. A. McCammon, B. R. Gelin and M. Karplus, "Dynamics of folded proteins," *Nature,* vol. 267, pp. 585-590, 1977.

[14] C. L. Brooks, M. Karplus and B. M. Pettitt, "Proteins: A theoretical perspective of dynamics, structure and thermodynamics.," Cambridge: Cambridge University Press, 1987.

[15] Y. Kalidas and N. Chandra, "Structural Bioinformatics: Transforming protein structures into biological insights," *Journal of the Indian Institute of Science,* vol. 88, no. 2, pp. 107-129, 2008.

[16] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne and H. M. Berman, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, pp. 235-242, 2000.

[17] A. Kryshtafovych, A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede and A. Tramontano, "Assessment of the assessment: Evaluation of the model quality estimates in CASP10," *Proteins,* p. Ahead of Print, 2013.

[18] M. Karplus, M. Levitt and A. Warshel, "www.nobelprize.org," 2013. [Online]. Available: www.nobelprize.org/noble_prizes/chemistry/laureates/2013/.

[19] M. Karplus, "Molecular dynamics of biological macromolecules: a brief history and perspective," *Biopolymers,* vol. 68, pp. 350-358, 2003.

[20] D. Roccatano, A. Barthel and M. Zacharias, "Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation," *Biopolymers,* vol. 85, pp. 407-421, 2007.

[21] I. J. Tinoco and J. D. Wen, "Simulation and analysis of single-ribosome translation," *Phys. Biol.,* vol. 6, p. 025006, 2009.

[22] R. Brandman, Y. Brandman and V. S. Pande, "A-site residues move independently from P-site residues in all-atom molecular dynamics simulations of the 70S bacterial ribosome," *PLoS One,* vol. 7, p. e29377, 2012.

[23] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton and M. Orozco, "Refinement of the AMBER force-field for nucleic acid simulations. Improving the representation of α/β conformations.," *Biophys. J.,* vol. 92, pp. 3817-3829, 2007.

[24] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham and P. J. Jurecka, "Refinement of the Cornell et al. nucleic acid force field based on reference quantum chemical calculations of torsion profiles of the glycosidic torsions.," *J. Chem. Theory Comput.,* vol. 7, p. 2886–2902, 2011.

[25] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins,* vol. 65, pp. 712-725, 2006.

[26] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins,* vol. 78, pp. 1950-1958, 2010.

[27] D. A. Potoyan, A. Savelyev and G. A. Papoian, "Recent successes in coarse-grained modeling of DNA," *WIREs Comput. Mol. Sci.,* vol. 3, pp. 69-83, 2013.

[28] V. Tozzini, "Coarse-grained models for proteins," *Current Opinion in Structural Biology,* vol. 15, pp. 144-150, 2005.

[29] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang and K. M. Merz, "AMBER 12," *San Francisco, California: University of California.,* 2012.

[30] B. R. Brooks, C. L. 3. Brooks, A. D. J. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels and S. Boresch, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.,* vol. 30, pp. 1545-1614, 2009.

[31] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation.," *J. Chem. Theory Comput.,* vol. 4, pp. 435-447, 2008.

[32] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kale, R. D. Skeel and K. Schulten, "NAMD: a parallel, object oriented molecular dynamics program.," *Int. J. Supercomput. Appl. High Perf. Comput.,* vol. 10, pp. 251-268, 1996.

[33] M. Harvey, G. Giupponi and G. De Fabritiis, "ACEMD: accelerated molecular dynamics simulations in the microseconds timescale," *J. Chem. Theory Comput.,* vol. 5, pp. 1632-1639, 2009.

[34] R. O. Dror, R. M. Dirks, J. P. Grossman, X. Huafeng and D. E. Shaw, "Biomolecular simulation: a computational microscope for molecular biology," *Annu. Rev. Biophys.,* vol. 41, pp. 429-452, 2012.

[35] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," *Science,* vol. 330, pp. 341-346, 2010.

[36] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson and G. De Fabritis, "High-throughput All-Atom molecular dynamics simulations using distributed computing," *J. Chem. Inf. Model,* vol. 50, pp. 397-403, 2010.

[37] M. Harvey, G. Giupponi, J. Villà-Freixa and G. De Fabritiis, "PS3GRID.NET: Building a distributed supercomputer using the PlayStation 3.," *Distributed and Grid Computing-Science Made Transparent for Everyone. Principles, Applications and Supporting Communities.,* 2007.

[38] M. Hong Ng, S. Johnston, B. Wu, S. E. Murdock, K. Tai, H. Fangohr, S. J. Cox, J. W. Essex, M. S. P. Sansom and P. Jeffreys, "BioSimGrid: Grid-enabled biomolecular simulation data storage and analysis," *Future Generation Computer Systems,* vol. 22, pp. 657-664, 2006.

[39] A. M. Simms, R. D. Toofanny, C. Kehl, N. C. Benson and V. Daggett, "Dynameomics: design of a computational lab workflow and scientifc data repository for protein simulations," *Protein Eng. Des. Sel.,* vol. 21, no. 6, pp. 369-377, 2008.

[40] J. C. Thibault, J. C. Facelli and T. E. Cheatham III, "iBIOMES: Managing and Sharing Biomolecular Simulation Data in a Distributed Environment," *J. Chem. Inf. Model.,* p. Ahead of print, 2013.

[41] M. D. Brazas, D. Yim, W. Yeung and B. F. Francis Ouellette, "A decade of web server updates at the bioinformatics links directory: 2003-2012," *Nucleic Acids Research,* vol. 40, pp. W3-W12, 2012.

[42] X. M. Fernández-Suárez, D. Rigden and M. Galperin, "The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection," *Nucleic Acids Research,* vol. 42, no. D1, pp. D1-D6, 2014.

[43] J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, J. Orlowski, M. Roos, K. Wolstencroft, S. Aleksejevs, R. Stevens, S. Pettifer, R. Lopez and C. A. Goble, "BioCatalogue: a universal catalogue of web services for the life sciences," *Nucleic Acids Research,* vol. 38, pp. W689-W694, 2010.

[44] E. Antezana, M. Kuiper and V. Mironov, "Biological knowledge management: the emerging role of the Semantic Web technologies," *Briefings in bioinformatics,* vol. 10, no. 4, pp. 392-407, 2009.

[45] G. Falkman, M. Gustafsson, M. Jontell and O. Torgersson, "SOMWeb: A semantic Web-Based system for supporting collaboration of distributed medical communities of practice," *J. Med. Internet Res.,* vol. 10, no. 3, p. e25, 2008.

[46] J. G. Frey and C. L. Bird, "Cheminformatics and the Semantic Web: adding value with linked data and enhanced provenance," *WIREs Comput. Mol. Sci.,* p. Early View, 2013.

[47] D. Repchevsky and J. Gelpí, “BioSWR: Semantic Web Services Registry for Bioinformatics,” *In prep.,* 2014.

[48] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer and I. Martinez-Flores, "The BioPAX community standard for pathway data sharing," *Nature Biotechnology,* vol. 28, no. 12, pp. 935-942, 2010.

[49] R. Côté, F. Reisinger, L. Martens, H. Barsnes, J. A. Vizcaino and H. Hermjakob, "The Ontology Lookup Service: bigger and better," *Nucleic Acids Research,* vol. 38, pp. W155-W160, 2010.

[50] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. J. Salama and S. Moore, "Broadening the horizon -- level 2.5 of the HUPO-PSI format for molecular interactions," *BMC Biology,* vol. 5, no. 44, 2007.

[51] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall and A. Hardisty, "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," *Nucleic Acids Research,* vol. 41, pp. W557-W561, 2013.

[52] V. Martín-Requena, J. Rios, M. García, S. Ramírez and O. Trelles, "jORCA: easily integrating bioinformatics Web Services," *Bioinformatics,* vol. 26, no. 4, pp. 553-559, 2010.

[53] J. Goecks, A. Nekrutenko and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.," *Genome Biol.,* vol. 11, p. R86, 2010.

[54] T. Stropp, T. McPhillips, B. Ludascher and M. Bieda, "Workflows for microarray data processing in the Kepler environment," *BMC Bioinformatics,* vol. 13, p. 102, 2012.

[55] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li and D. De Roure, "myExperiment: a repository and social network for the sharing of bioinformatics workflows," *Nucleic Acids Research,* vol. 38, pp. W677-W682, 2010.

[56] The BioMoby Consortium, “Interoperability with Moby 1.0 - It's better than sharing your toothbrush!,” *Briefings in Bioinformatics,* vol. 9, no. 3, pp. 220-231, 2008.

[57] M. D. Wilkinson, B. Vandervalk and L. McCarthy, "The Semantic Automated Discovery and Integration (SADI) Web Service Design-Pattern, API and Reference Implementation," *Journal of Biomedical Semantics,* vol. 2, p. 8, 2011.

[58] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.,* vol. 28, no. 1, pp. 235-242, 2000.

[59] K. Henrick, Z. Feng, W. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin and H. Berman, "Remediation of the Protein Data Bank Archive," *Nucleic Acids Res.,* vol. 36, pp. D426-D433, 2008.

[60] E. Krissinel and K. Henrick, “Inference of macromolecular assemblies from crystalline state,” *J. Mol. Biol.,* vol. 372, no. 3, pp. 774-797, 2007.

[61] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, "CATH - a hierarchic classification of protein domain structures," *Structure,* vol. 5, no. 8, pp. 1093-1109, 1997.

[62] I. Sillitoe, A. L. Cuff, B. H. Dessailly, N. L. Dawson, N. Furnham, D. Lee, J. G. Lees, T. E.

Lewis, R. A. Studer, R. Rentzsch, C. Yeats, J. M. Thornton and C. A. Orengo, "New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures," *Nucleic Acids Research,* vol. 41, pp. D490-D498, 2013.

[63] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, "SCOP: A structural classification of proteins databaase for the investigation of sequences and structures," *J. Mol. Biol.,* vol. 247, pp. 536-540, 1995.

[64] S. E. Brenner, P. Koehl and M. Levitt, "The ASTRAL compendium for sequence and structure analysis," *Nucleic Acids Research,* vol. 28, pp. 254-256, 2000.

[65] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha and A. Murzin, "SCOP2 prototype: a new approach to protein structure mining," *Nucleic Acids Research,* vol. 42, no. Database Issue, pp. D310-314, 2014.

[66] T. E. Lewis, I. Sillitoe, A. Andreeva, T. L. Blundell, D. W. Buchan, C. Chothia, A. Cuff, J. M. Dana, I. Filippis, J. Gough, S. Hunter, D. T. Jones, L. A. Kelley, G. J. Kleywegt, F. Minneci, A. Mitchell, A. G. Murzin, B. Ochoa-Montaño, O. J. Rackham, J. Smith, M. J. Sternberg, S. Velankar, C. Yeats and C. Orengo, "Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains," *Nucleic Acids Research,* vol. 41, pp. D499-D507, 2013.

[67] The UniProt Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013," *Nucleic Acids Research,* vol. 41, pp. D43-D47, 2013.

[68] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, "The Pfam protein families database," *Nucleic Acids Research,* vol. 40, pp. D290-D301, 2012.

[69] International Union of Biochemistry and Molecular Biology (IUBMB), Enzyme Nomenclature, San Diego, California: Academic Press, 1992.

[70] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research,* vol. 28, pp. 304-305, 2000.

[71] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinsky, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genetics.,* vol. 25, no. 1, pp. 25-29, 2000.

2. Objectives

The main global objective of our work is focused in the comprehension of the influence of the dynamics properties in the structure to function relationships in biological macromolecules. The pretended contribution would be to generate a number of tools to ease the analysis of such properties. Specific objectives for the present work are:

1. To generate tools for the automatic extraction of information from the available structural data, through the design and implementation of relational databases for the storage and efficient retrieval of structural data.

2. Develop a library of protein molecular dynamics simulations. This work will focus on the generation of the appropriate datasets, and the design and implementation of protocols for the preparation and analysis of structures, and for proper data management.

3. Use the created library to explore the dynamics of the proteins and their surrounding solvent molecules in a proteome scale.

4. Getting benefit from the experience in building a large simulation database, the last objective is to develop informatics tools to help porting these simulation methods to the high-throughput regime, on the one side, and to popularize the access to MD for the non-experts, on the other.

3. Methods

In this chapter we outline the main methods used in this thesis. Further details on the particular usage of methods presented here can be found in *Results* chapter.

3.1. Macromolecular Dynamics simulation techniques

Along the different existing techniques to obtain macromolecular dynamic information, the most popular is atomistic *Molecular Dynamics (MD)*. However, due to its high computational cost, *Coarse-Grained dynamics* (CG) algorithms are showing an increasing use.

3.1.1. Molecular Dynamics (MD)

3.1.1.1. Classical Mechanics and Force Fields

Classical mechanics represent atoms as spheres of a given radii, charge and mass. Compared with quantum mechanics, classical mechanics allows us to increase the size of the system that can be computed, as well as the time scale of the trajectory. Unfortunately, the use of classical mechanism implies that some features are lost, and processes such as the bond formation/rupture or the electron transfer cannot be reproduced. *Classical mechanics* use *force-fields* to calculate the potential energy of a system. *Force-fields* are a set of parameters (charges, masses, radii, bond lengths, bond dihedrals, etc.) together with their functional term (energy functional).

Force-fields functional terms are generally built by the sum of two components: bonded terms and non-bonded terms. Bonded terms are those describing bond, angle and dihedral interactions, whereas non-bonded terms describe electrostatics and van der Waals interactions between non-bonded particles. The most popular force-fields for MD simulations (protein or nucleic acids) are *AMBER* [1], *CHARMM* [2], *OPLS* [3] and *GROMOS* [4]. They differ from each other especially in the parameters, but also (slightly) in the functional terms, which, to increment the computational efficiency, are usually modified to fit the calculus to the algorithms used. That precludes a straightforward transfer of force-field parameters. In the next lines, AMBER force-field functional will be briefly described as an example. A number of studies have been published comparing different force-fields [5, 6, 7, 8, 9].

The potential energy can then be written as the sum of bonded and non-bonded terms:

$$E_{pot} = E_{bonded} + E_{non-bonded} \qquad (3.1),$$

where

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \qquad (3.2),$$

and

$$E_{non-bonded} = E_{elec} + E_{vdw} \qquad (3.3),$$

with the different terms shortly described in the following paragraphs:

- Bond term:

  Describe the energy related to the bond length between two atoms. As covalent bonds vibrate around a given bond length, harmonic potentials are used to approximate this behavior:

$$E_{bond} = \sum_{bonds} k_r(r - r_0)^2 \qquad (3.4),$$

  where r is the observed bond length, r_0 is the atom pair bond reference (equilibrium) length and k_r is the bond force constant.

- Bond angle term:

  Describe the angle observed between two adjacent bonds in a molecule. As in the bond length, a harmonic oscillator is used to estimate the energy corresponding to a deviation of the observed bond angle with the equilibrium value:

$$E_{angle} = \sum_{angle} k_\theta(\theta - \theta_0)^2 \qquad (3.5),$$

  where θ is the observed angle, θ_0 is the reference (equilibrium) bond angle and k_θ is the angular force constant.

- Torsions:

  Describe rotation energetic barriers of an atom pair bond. They are defined using the pair of atoms involved in the bond, and their adjacent atoms (four atoms in total). As the previous terms, dihedral angles cannot be described by one simple harmonic due to their periodicity; a truncated cosine Fourier expansion is used instead.

$$E_{dihedral} = \sum_{dihedral} \sum_{n} \frac{V_n}{2}(1 + cos(n\omega - \gamma)) \qquad (3.6),$$

  where V_n is the height of the barrier (for every term), n the periodicity (usually truncated to 3), ω is the observed dihedral angle, and γ is the phase angle.

  In some cases, an additional torsion term is required to reproduce out of plane bending frequencies, i.e. the capacity of an atom to be out of the plane formed by the other three atoms involved in the dihedral. These cases are called improper dihedral angles, whereas the previous ones are called proper dihedral angles, for differentiation. AMBER force-field has just one term for

both dihedral types in the functional, but GROMOS and CHARMM add a new term specifically related to improper dihedral angles.

- Electrostatic energy:

Contrary to quantum mechanics which explicitly accounts for electron distribution and nuclear charges, classical mechanics approximate the electrostatic effect by means of partial point charges assigned to each atom center. Coulombs law is used to estimate the electrostatic energy of two interacting particles:

$$E_{elec} = \sum_{i=1} \sum_{j=1} \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \qquad (3.7),$$

where q_i and q_j are the atomic point charges of atoms i and j, ε is the medium dielectric constant and r_{ij} is the Euclidean distance between atoms i and j.

The dielectric constant ε is equal to the vacuum dielectric constant (ε_0) if explicit solvent is used. Explicit solvent calculations consider each solvent molecule explicitly, and thus it is the most exact representation of the medium, but unfortunately it is also very computational demanding, as the system number of particles increases significantly. Implicit solvent tries to approximate the medium via the dielectric constant ($\varepsilon \neq \varepsilon_0$). Different implicit representations of solvent in classical mechanics simulations exist. Numerical approaches to the *Poisson-Boltzmann* equation are widely used, as they have been proven to provide similar results than explicit solvent simulations and experiments. However, due to its high computational cost, various simpler approximations have emerged, the most popular ones the *Generalized Born* models [10].

- Van der Waals energy:

Van der Waals term describes the behavior of two atoms *i)* when approaching each other without forming a covalent bond (repulsion), and *ii)* when they are near to an internuclear optimum distance (attraction). The final van der Waals term is then a sum of the attractive and repulsive forces between the pair of atoms, usually described through a Lennard-Jones potential:

$$E_{vdw} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\varepsilon_{ij} \left[\left(\frac{{R^0}_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{{R^0}_{ij}}{r_{ij}} \right)^{6} \right] \right) \qquad (3.8),$$

where ε_{ij} is the depth of the potential well, ${R^0}_{ij}$ the distance at which the potential is zero and r_{ij} the distance between atoms i and j. The factor of 2 ensures that the equilibrium distance is ${R^0}_{ij}$.

3.1.1.2. Molecular Mechanics and Dynamics

Based on classical mechanics, MD algorithms use Newton's second law of motion to generate the trajectories along time for a set of particles:

$$a_i = \frac{d^2 x_i}{d_t^{\,2}} = \frac{F_i}{m_i} \qquad (3.9),$$

where a_i, x_i, F_i and m_i are the acceleration, position, force and mass of a particular particle *i* respectively, and *t* is the time. A force *F* acting on a particle *i* results in its acceleration a_i, which modifies the instant velocity v_i and position r_i within a time step Δt. In principle, system particle coordinates during time (trajectory) could be obtained analytically solving equation 3.9. However, due to particle (atom) couplings when dealing with macromolecules, an analytical solution is not possible. Instead, an iterative, step by step numerical integration is used to obtain an approximate solution.

- Finite difference integration approaches.

Finite difference techniques are used to generate molecular dynamics trajectories with continuous potential models, breaking down integration into many small stages, each separated in time by a fixed time Δt (integration step). The total force on each particle in the configuration (e.g. molecule atom) at a time *t* is calculated as a vector sum of its interactions with other particles, which are then combined with the positions and velocities at a time *t* to calculate the positions and velocities at a time $t + \Delta t$ [11]. For this, the force is assumed to be constant during the time step, making the election of Δt crucial. Different algorithms for integrating the equations of motion exist, based on Taylor series expansions. The *verlet* algorithm and its variants *leap-frog* and *velocity verlet* are the most used in MD simulation programs. *Verlet* algorithm [12] calculates the atomic positions *r* at time $t + \Delta t$ from the actual positions *r(t)*, the last steps positions $r(t - \Delta t)$ and the accelerations *a(t)* calculated from the force-field:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \Delta t^2 a(t) \qquad (3.10).$$

Velocities are not explicitly calculated with *verlet* algorithm, but can be extracted using different simple approaches.

A variation of the *verlet* algorithm, named *leap-frog*, implicitly calculates the new positions together with the new velocities [13]:

$$r(t + \Delta t) = r(t) + \Delta t v\left(t + \frac{1}{2}\Delta t\right) \qquad (3.11),$$

$$v\left(t + \frac{1}{2}\Delta t\right) = v\left(t - \frac{1}{2}\Delta t\right) + \Delta t a(t) \qquad (3.12).$$

As can be seen in equations 3.11 and 3.12, positions and velocities are not synchronized; positions are computed at time $t + \Delta t$ whereas velocities are computed at time $t + \frac{1}{2}\Delta t$.

The *velocity verlet* method [14] allows calculating positions, velocities and accelerations at the same time:

$$r(t\ +\Delta t) = r(t) + \Delta t v(t) + \frac{1}{2}\Delta t^2 a(t) \qquad (3.13),$$

$$v(t\ +\Delta t) = v(t) + \frac{1}{2}\Delta t[a(t) + a(t+\Delta t)] \qquad (3.14).$$

- Integration Step

As forces are not recomputed within the integration step, the election of Δt is crucial. During this integration time, the system keeps following the actual movement, with risk of instabilities if working with too high integration steps. On the other hand, working with too small integration steps avoids instabilities, but increases the cost of sampling conformational space.

In the context of MD simulations, the integration step has to be smaller than the fastest motion of the system. Highest frequencies are known to occur in the hydrogen-involving bonds, requesting an integration step below 1fs. As the effect of bond vibrations in macromolecular systems are not extremely relevant for the final result, some constraints can be inserted in determined atomic bonds, maintaining the atomic distances within a certain range. The most used algorithms for this kind of constraints are *Shake* [15] (used by AMBER package), *Rattle* [16] (used by NAMD package, and including velocity constraints) and *LINCS* [17] (used by GROMACS package). The use of these algorithms allows longer integration steps, and thus an important gain in efficiency.

- Simulation Conditions

Current MD packages can deal with different simulation conditions:

  - N: Number of particles
  - V: Volume of the system
  - T: Temperature of the system
  - P: Pressure of the system
  - E: Energy of the system

Prior to running a MD simulation, an ensemble of conditions must be chosen. Common ensembles are the NVE (microcanonical ensemble), NPT (isothermic-isobaric ensemble) and NVT (canonical ensemble). NPT is currently the most widely used as it can be directly compared with experimental data.

3.1.1.3. MD System Setup

A necessary previous step before running a MD simulation is the setup of the structure. This is especially relevant in proteins, because of their molecular complexity. Ideally, a series of steps have to be performed to go from a 3D structure to a system completely prepared to be simulated. The most basic steps are:

- Structure curation

The first thing to do is decide whether to keep structure information such as crystallographic water molecules or hydrogen atoms, or let the MD programs add them *a posteriori*. A very important decision is to keep or remove ligands, as it could

drastically change the dynamic obtained. Normally if parameterized, keeping the ligand is a good option.

- Missing atoms / hydrogen addition

Adding missing hydrogen atoms or even missing heavy atoms is a mandatory process in the preparation of the system. Most experimental structures lack hydrogen atoms, as they cannot be obtained from X-ray crystallography. Moreover, and especially in proteins, there are structures with missing side chain atoms, or even missing parts of the backbone. For the first case, programs exist capable of adding the needed atoms to complete the residues, but unfortunately, filling missing parts of a protein in not so easy. An option is to use comparative modeling programs to first obtain a complete molecule. However, even though these kinds of programs are quickly improving, there is still not possible to exactly determine the original protein configuration.

Another important issue to deal with in this step is the correct ionization of titratable residues. The typical case is the histidine residue (His). Due to its pKa value near the physiological pH (7.0), His can appear in both the acid and base forms, and, in addition, can exist as two different tautomers in the base form, with a proton in the *Nε* or with a proton in the *Nδ* (Fig. 3.1).

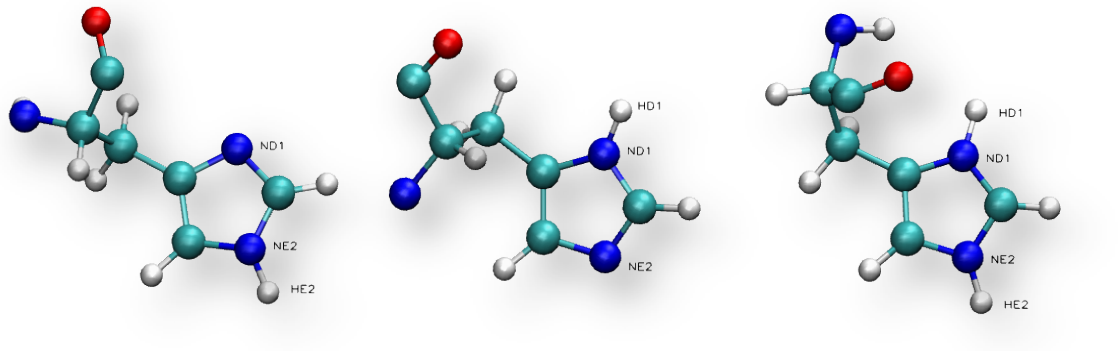


Fig 3.1.- Histidine ionization states. *Histidine residue can be found in three different ionization states at physiological pH (from left to right):Histidine Epsilon (Nε protonated), Histidine Delta (Nδ protonated) and Protonated Histidine (Both nitrogen atoms protonated).*

- Structure minimization

The addition of missing atoms is done taking into account the position of the surrounding atoms, but the resulting structure is usually stressed by unfavourable contacts or distortions in the covalent structure. To obtain an energetically better conformation, it is recommended to run two energetic minimizations on the structure. The first one usually involves only the hydrogen atoms, keeping the rest of the molecule atoms fixed, whereas the second one involves the heavy atoms, restraining the movement to just allow the system to relax. These minimizations are done in vacuum. Different methods exist and the most popular algorithms (*Conjugate gradient* and *Steepest descent*) are implemented in the majority of MD packages.

- Solvent & counterions addition

The next step in the MD structure setup is the addition of the solvent and counterions. Solvent molecules are usually integrated on the structure surface in two steps:

  - Structural waters: A first shell of water molecules is commonly added in the energetically most favorable positions on the surface of the structure. It is a computationally expensive process and is usually reduced to just tens of water molecules (depending on the structure size).

  - Solvent box: A box of solvent molecules is created surrounding the original structure. Different types of boxes have been used in the MD algorithms (cubic, octahedron, dodecahedron, etc.), being the truncated octahedron the most used nowadays.

Counterions are also usually added to the system in two steps:

  - Neutralization ions: A set of counterions are added until reaching system electroneutrality. This is especially important when working with nucleic acids, due to their high negative charge.

  - Ionic concentration: After neutralizing the system, a certain ionic concentration is added, commonly using chloride (Cl^-) and sodium (Na^+) ions, to mimic the physiological ionic strength.

- System equilibration

The last step in the MD system setup is crucial for obtaining a correct simulation. An incorrect equilibration of the whole system (structure, ions and solvent) can led to an unstable simulation and eventually to a failed process. A proper equilibration results in a relaxed system, ready to be used as an input to a MD algorithm.

A theoretically correct equilibration process involves at least 5 steps:

  - Heating system to a desired temperature (usually from 0 to 300K), with the atoms of the solute restrained.
  - Once reached the desired temperature, relax the system for a certain time, reducing the force constant of the solute restrains.
  - Limit restraints to just backbone atoms, reducing a little bit more the force constant.
  - Reduce restraints to a minimum force constant.
  - Completely remove restraints and let the system move freely.

The time length for each of these steps varies, depending on the system size, but they are commonly in the order of hundreds of picoseconds, eventually reaching the nanosecond. Typically an extended nanosecond-scale post-equilibration is done before production runs.

These MD setup steps will be extended in the corresponding thesis section, when explaining the setup automatization process implemented in MDWeb (section 5.4).

3.1.1.4. MD Packages

The number of available MD packages is currently increasing mainly due to the hardware specialization (GPUs and parallel/grid supercomputing). From all the different possibilities, just the ones used during the development of this thesis, and actually the most popular and widely-used, will be enumerated and briefly introduced in the following lines: *AMBER*, *NAMD* and *GROMACS*.

3.1.1.4.1. AMBER

AMBER (Assisted Model Building with Energy Refinement) is a MD program originally developed by Peter Kollman's group at the *University of California*, San Francisco, and is maintained by an active collaboration between David Case at *Rutgers University*, Tom Cheatham at the *University of Utah*, Tom Darden at *NIEHS*, Ken Merz at *Florida*, Carlos Simmerling at *Stony Brook University*, Ray Luo at *UC Irvine*, and Junmei Wang at *Encysive Pharmaceuticals*. AMBER MD package is divided in two main parts: Ambertools and AMBER. Ambertools is a completely free package of programs for preparing molecules to run MD simulations and for the subsequent trajectory analysis. AMBER, on the other hand, is the package of programs to run MD simulations. The first part is freely distributed, whereas the last part is not free, and is distributed under a site-license agreement. Very recently the MD engine of AMBER has been completely re-written to take advantage of the new GPU architectures, showing extremely large computer efficiency.

http://ambermd.org [18]

3.1.1.4.2. NAMD

NAMD (Not [just] Another Molecular Dynamics program) is a MD simulation program based on Charmm++ parallel objects, developed at the *Theoretical and Computational Biophysics Group* in the *University of Illinois at Urbana-Champaign*, group directed by Prof. Klaus Schulten. NAMD is distributed free of charge (non-exclusive, non-commercial license) with source code. It uses the popular molecular graphics program *VMD (Visual Molecular Dynamics)* for simulation setup and trajectory analysis, but is also file-compatible with AMBER, CHARMM and X-PLOR. NAMD greatest advantage is its great parallelism, superior to the remaining codes.

http://www.ks.uiuc.edu/Research/namd/ [19]

3.1.1.4.3. GROMACS

GROMACS (GROningen MAchine for Chemical Simulations) is a MD software package originated in the group of H. Berendsen at the *Biophysical Chemistry department* of *University of Groningen, The Netherlands*, including an impressive set of small programs to prepare and run MD simulations and analyze the resulting trajectories. Each of these programs contain a brief help about what are they doing, how to run them, and information about parameters, inputs and outputs. The entire GROMACS package is free software, licensed under the *GNU Lesser General Public License*. GROMACS is now mostly supported by E. Lindahl and B. Hess in Sweden and is probably the fastest CPU implementation of a MD algorithm.

http://www.gromacs.org [20]

3.1.2. Coarse-Grained Dynamics (CG)

Many of the relevant dynamics and interactions within cells occur on the timescale of milliseconds to seconds, and involve large macromolecular aggregates. In this scenario, MD technique, with its computational cost, is clearly not the most appropriate tool. That motivated the development of approximate coarse-grained models, where, in order to increase computer efficiency, a certain loss of accuracy is accepted, with a significant reduction in structural resolution. Then, using CG algorithms, larger macromolecules and larger timescales can be simulated, reaching the mesoscopic scale.

CG algorithms consist of a sum of *i)* coarse-grained potentials, defining the resolution and functionals to compute the potential energy and *ii)* a dynamic sampling technique, defining the method used to obtain dynamic information. Many different combinations of both parts can be used, and new methods are published regularly. Here we describe only the algorithms (i.e. the engines used to sample the system) used in this thesis. For additional information, please refer to [21, 22, 23].

3.1.2.1. Protein CG Algorithms

3.1.2.1.1. Brownian Dynamics (BD)

Brownian Dynamics (BD) is a CG algorithm based on *Langevin Dynamics* sampling technique [24]. It considers the protein molecule to be immersed in a stochastic bath that keeps the temperature constant. The Brownian motion of a particle (of mass *m*) in this bath is due to the molecule-thermal agitation of the surrounding solvent (which lead to random collisions on the particle, $\vec{\xi}$) and a dispersive force accounting for the viscous resistance the particle feels on going through the fluid ($-\gamma\vec{v}$) at velocity $\vec{v}$:

$$m\vec{a}_i = \vec{F}_i - \gamma\vec{v}_i + \vec{\xi}_i(t) \quad (3.15),$$

where *m* stands for the effective mass of the particle (usually Cα), *v* is the velocity, *a* is the acceleration, and *F* represents the force. γ parameter is the inverse of a characteristic time at which the particle loses its energy in a given solvent. The stochastic (random term) $\vec{\xi}(t)$ is considered a white noise (with Gaussian distribution with zero mean), with autocorrelation function:

$$\langle\vec{\xi}_i(t)\vec{\xi}_i(t')\rangle = \sigma^2\delta_{ij}\delta(t-t') \quad (3.16),$$

where δ_{ij} is the *Kronecker's delta* (1 when i = j, 0 otherwise) and *δ(t - t')* is the *Dirac's delta* (∞ when t - t' = 0, 0 otherwise), and σ^2 is the standard deviation (also named noise intensity) associated with the Gaussian process $\vec{\xi}(t)$:

$$\sigma^2 = 2m_i k_B T_\gamma \quad (3.17),$$

which means that the standard deviation of noise can be expressed as a function of the mass of the particle, the temperature of the thermal bath, and the factor of the dissipation force.

The equation of motion (eq. 3.15) is integrated using *Verlet*'s algorithm (section 3.1.1.2), with resulting velocities after time Δt:

$$\overrightarrow{v_i} = e^{-\frac{\Delta t}{\tau}}\overrightarrow{v_i^0} + \frac{1}{\gamma}\left(1 - e^{-\frac{\Delta t}{\tau}}\right)\overrightarrow{F_i^0} + \Delta\overrightarrow{v_i^G} \qquad (3.18),$$

and positions:

$$\overrightarrow{r_i} = \overrightarrow{r_i^0} + \tau\left(1 - e^{-\frac{\Delta t}{\tau}}\right)\overrightarrow{v_i^0} + \frac{\Delta t}{\gamma}\left(1 - \frac{\tau}{\Delta t}\left(1 - e^{-\frac{\Delta t}{\tau}}\right)\right)\overrightarrow{F_i} + \Delta\overrightarrow{r_i^G} \qquad (3.19),$$

where $\tau = m\,\gamma^{-1}$ is the characteristic time, and $\Delta\overrightarrow{r_i^G}$, $\Delta\overrightarrow{v_i^G}$ are the changes in position and velocity induced by the stochastic term.

The potential energy used to compute forces assumes a Cα-only CG model of the protein and a quasi-harmonic representation of the interactions:

$$U_{ij} = \frac{1}{2}C\left(\frac{r^*}{\left|\overrightarrow{r_{ij}^0}\right|}\right)^6\left(\overrightarrow{r_{ij}} - \overrightarrow{r_{ij}^0}\right)^2 \qquad (3.20),$$

where $\overrightarrow{r_{ij}} = \vec{r_i} - \vec{r_j}$ stands for the vector connecting Cα atoms *i* and *j*, and $\overrightarrow{r_{ij}^0}$ are the relative vectors taken from the initial (native) structure and r* is the mean distance between two consecutive alpha carbons, set to 3.8 Å.

3.1.2.1.2. Discrete Molecular Dynamics (DMD)

Discrete Molecular Dynamics (DMD) considers the molecule as a system of beads interacting through a discontinuous potential (flat wells). Outside the discontinuities, potentials are considered constant, thereby implying a ballistic regime for the particles (constant potential, constant velocity) in all conditions, except at such time as when the particles reach a potential discontinuity (a "collision"). At this time, velocities of the colliding particles are modified by imposing conservation of the linear momentum, angular momentum and total energy [24]. Since the particles were constrained to move within a configurational space where the potential energy is constant, the kinetic energy remains unchanged and therefore all collisions are assumed to be elastic.

In contrast to MD technique, DMD does not require the integration of the equations of motion at fixed time steps. Instead, the calculation progresses from event to event, being the time between events (depending on the number of particles), the rate-limiting step of the method. The equations of motion, corresponding to constant velocity, are solved analytically:

$$\vec{r_i}(t + t_c) = \vec{r_i}(t) + \vec{v_i}(t)t_c \qquad (3.21),$$

where t_c is the minimum among the collision times t_{ij} between each pair of particles *i* and *j*, given by:

$$\left|\vec{r_i}(t) - \vec{r_j}(t)\right| = d \qquad (3.22),$$

from which, solving the modulus and converting the result to a quadratic equation we can obtain:

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{b_{ij}^2 - v_{ij}^2\left(r_{ij}^2 - d^2\right)}}{v_{ij}^2} \qquad (3.23),$$

where r_{ij} is the square modulus of $\overrightarrow{r_{ij}} = \vec{r_j} - \vec{r_i}$, v_{ij} is the square modulus of $\overrightarrow{v_{ij}} = \vec{v_j} - \vec{v_i}$, $b_{ij} = \overrightarrow{r_{ij}} \cdot \overrightarrow{v_{ij}}$, and d is the distance corresponding to a discontinuity in the potential (signs + and – before the radical are used for particles approaching one another and moving apart, respectively).

The collision between particles i and j is associated with a transfer of linear momentum in the direction of the vector $\overrightarrow{r_{ij}}$. The conservation equations are:

$$m_i\vec{v_i} = m_i\overrightarrow{v_i'} + \Delta\vec{p} \qquad (3.24),$$

$$m_j\vec{v_j} + \Delta\vec{p} = m_j\overrightarrow{v_j'} \qquad (3.25),$$

where variables after the event are identified by a prime index.

To calculate the change of velocities occurring upon collision, the velocity of each particle is projected in the direction of the vector $\overrightarrow{r_{ij}}$ so that the conservation equations become one-dimensional along the interactomic coordinate. Thus, conservation of linear momentum is defined by:

$$m_iu_i + m_ju_j = m_iu_i' + m_ju_j' \qquad (3.26),$$

and the conservation of energy, from the previous equality, and $E_c = \frac{1}{2}mv^2$:

$$\frac{1}{2}m_i{u_i}^2 + \frac{1}{2}m_j{u_j}^2 = \frac{1}{2}m_iu_i'^2 + \frac{1}{2}m_ju_j'^2 + \Delta V \qquad (3.27),$$

where u_i, u_j are the projections of the velocities v_i, v_j along the direction $\overrightarrow{r_{ij}}$ and ΔV stands for the height of the step in the interatomic potential.

Again, working with the conservation equations 3.24 to 3.27 and converting the result to a quadratic equation we can obtain:

$$\Delta p = \frac{m_im_j}{m_i + m_j}\left\{\sqrt{\left(u_j - u_i\right)^2 - 2\left(\frac{m_i + m_j}{m_im_j}\right)\Delta V} - \left(u_j - u_i\right)\right\} \qquad (3.28).$$

In the case of infinite walls (approach used in the calculations outlined in this thesis), the term affected by the interatomic potential can be removed, and taking the negative solution of the root, the transferred linear momentum is determined as:

$$\Delta p = \frac{2m_im_j}{m_i + m_j}\left(u_i - u_j\right) \qquad (3.29).$$

In the implemented approach (*MDWeb* and *FlexServ* web servers, section 5.4), interaction potentials are defined as infinite square wells, such that the particle-particle distances vary between $d_{min} = (1 - \sigma)\, r_{ij}^0$ and $d_{max} = (1 + \sigma)\, r_{ij}^0$, being r_{ij}^0 the distance in the native conformation and 2σ the width of the square well. Interaction potentials are

defined only for the particles at a distance smaller than a cut-off radius r_c in the native conformation, where $r_c = 8$ *Å* and $\sigma = 0.1$ for non-sequential contacts and $\sigma = 0.05$ for the sequential ones. The rest of the contacts are represented by simple hardcore potentials.

3.1.2.1.3. Elastic Network Model - Normal Mode Analysis (ENM-NMA)

Elastic Network Model coupled to *Normal Mode Analysis (ENM-NMA)* is a very popular method for the study of the "near-equilibrium" dynamics of proteins [25]. The basic assumption of the model is that the protein response to changes in equilibrium geometry is harmonic, and deformation modes can then be derived by diagonalization of the Hessian matrix. ENM-models are typically used in conjunction with Cα representations of the protein. Such Cα atoms act as network nodes which are connected by harmonic springs. The number of springs is directly connected to the number of residues in the protein, and therefore, direct application of ENM will result in an artifactual over-rigidification of the protein as the protein size is increased. This problem is corrected using distance-dependent cutoffs that annihilate the interactions between remote residues, leading to an energy functional as:

$$E_{ij} = \sum_{i \neq j} K_{ij}\left(r_{ij} - r_{ij}^0\right)^2 \qquad (3.30),$$

where r_{ij} stands for the distance between residues i and j, r_{ij}^0 is the equilibrium conformation, and K_{ij} stands for the spring constant. The force of the spring restricting the motion of the ij residue pair is computed as:

$$K_{ij} = \frac{1}{2}\kappa\Gamma_{ij} \qquad (3.31),$$

where κ is a phenomenological constant (in energy/distance^2 units) and Γ is Kirchhoff topology matrix of inter-residue contacts, where ijth element for $i \neq j$ is equal to -1 if residues i and j are within the cutoff distance r_c or 0 otherwise, and the diagonal elements *(iith)* are equal to the coordination number or residue connectivity:

$$\Gamma_{ii} = -\sum_{k|k \neq i}^{N} \Gamma_{ik} \qquad (3.32).$$

The potential function in eq. 3.30 is used to build a *Hessian* matrix *(H)*, a *3N x 3N* matrix with *N* being the number of particles, defined as *N x N* submatrices H_{ij} containing the second derivatives of the energy with respect to the coordinates of each particle. Diagonalization of the *Hessian* yields the eigenvectors (the essential deformation modes) and the associated eigenvalues (stiffness constants).

Despite their simplicity, these functionals are able to provide quite accurate representation of the near-equilibrium dynamics of many proteins but are extremely dependent on the selected cutoff for remote interactions, which can have different optimal values for each protein. This led to the derivation of new methods where the discrete Hamiltonian outlined above is replaced by continuous functions, typically dependent on the inverse exponential of the inter-residue distance.

The method used in this thesis (implemented in *MDWeb* and *FlexServ* web servers, section 5.4) is a hybrid approach developed in the group, that treats differentially the sequential and non-sequential "Cartesian" contacts. For the first M sequential contacts, a fully connected matrix is used, while Cartesian contacts are treated using a continuum, distance-dependent function with a calibrated size-dependent cutoff, which helps to remove artifactual long-range interactions. Therefore, the elements of the topology matrix are defined as:

$$\Gamma_{ij}\begin{cases} if\ S_{ij} \le M, = -1 \\ otherwise \begin{cases} = -1\ if\ r_{ij} \le r_c \\ = 0\ otherwise \end{cases} \end{cases} \quad (3.33),$$

and the matrix Γ has always 2M+1 non-zero diagonal entries defining neighbor chained contacts. Accordingly, the force constants K_{ij} are dependent, not only on the Cartesian but also on the sequential distance:

$$K_{ij}\begin{cases} = C^{seq}/S_{ij}^{ns}, if\ S_{ij} \le M \\ otherwise \begin{cases} = \left(C^{cart}/r_{ij}\right)^{nc}, if\ r_{ij} \le r_c \\ = 0\ otherwise \end{cases} \end{cases} \quad (3.34),$$

where values for all terms are: $ns = 2$, $C^{seq} = 60$ Kcal/(mol Å^2), $nc = 6$ and $C^{cart} = 6$ Kcal/(mol Å^2), and were obtained by fitting to apparent force constants and structural variance profiles taken from a large number of atomistic MD simulations [25]. A value of M = 3 is used for sequential interactions based on MD simulations, which were also instrumental to define the cutoff radii r_c, that is computed using an empirical logarithmic relationship with the size of the protein. This formalism guarantees sequential contacts which decay quickly with the number of connecting bonds, and a continuum decay of the strength of Cartesian contacts up to a cutoff.

3.1.2.2. Nucleic Acid CG Models

3.1.2.2.1. Elastic Mesoscopic Model

The elastic mesoscopic model provides an intermediate level of resolution (base pairs) and potential energy complexity. It assumes that DNA deformations can be approximated as the addition of harmonic distortions of equilibrium base-pair step geometries. Three rotational (*twist, roll and tilt*) and three translational (*slide, shift and rise*) degrees of freedom (see figure 3.12) are considered, thereby allowing us to define the Hamiltonian as:

$$E = \Xi(\Delta X)^2 \ \ with\ \ \Xi = k_B T C^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{rw} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{tw} & k_{tr} & k_t & k_{ts} & k_{tl} & k_{tf} \\ k_{sw} & k_{sr} & k_{st} & k_s & k_{sl} & k_{sf} \\ k_{lw} & k_{lr} & k_{lt} & k_{ls} & k_l & k_{lf} \\ k_{fw} & k_{fr} & k_{ft} & k_{fs} & k_{fl} & k_f \end{pmatrix} \quad (3.35),$$

where k_B is the Boltzman constant, T is the absolute temperature, E is the energy associated with the deformation ΔX, and k_{XY} stands for the different stiffness constants defined by the 36 elements of the stiffness matrix (Ξ) (*twist (w), roll (r), tilt (t), rise (s), slide (l) and shift (f)*). The Ξ can be calculated (see eq. 3.35) by inversion of the covariance matrix obtained from either analysis of MD trajectories (at dinucleotide or tetranucleotide level [26, 27, 28]) or from the analysis of dinucleotide step variability in the crystal structures of DNAs and DNA-protein complexes [29, 30]. In all the cases, sampling is obtained using Monte Carlo simulations in the helical coordinate space.

3.1.2.2.2. Worm-like Chain Model (WLC)

In Worm-like chain model, the DNA is interpreted at the lowest level of resolution [31]. The combined *Worm-Like Chain* (*WLC*) features with those of a bead model (approximation used in this thesis, see section 5.5) include a mechanical and an electrostatic potential [32, 33]. In this treatment, a DNA segment is represented as an elastic chain of N beads, each bead comprising M base pairs (one bead x M base pair steps), connected by N-1 inter-bead segments of average length l_0=1.36nm. To account for the torsional deformation, a local coordinate system is fixed to each DNA bead, and a set of Euler angles (α, β, γ) that transform from the coordinate frame of bead i to that of bead $i+1$ are defined.

The total potential energy is defined by a simple Hamiltonian consistent of a sum of stretching, bending, torsion and electrostatic contributions:

$$E = E_S + E_B + E_T + E_{ele} \qquad (3.36),$$

where the stretching energy (E_S) is computed as:

$$E_S = 0.5K_S \sum_{i=1}^{N-1} (l_i - l_0)^2 \qquad (3.37),$$

where K_s is the stretching constant, l_i is the actual distance between beads, and l_0 is the optimum bead-bead distance. Based on [33, 32], the value of K_s was set at $100k_BT/l_0^2$ (as it reproduces the correct DNA bond variance), with k_B the Boltzmann constant and T the temperature.

The bending energy is determined as:

$$E_B = 0.5K_B \sum_{i=1}^{N-2} {\beta_i}^2 \qquad (3.38),$$

where K_B is the bending constant and β_i is the Euler bending angle between the local coordinate systems of two consecutive beads; and the torsion potential is defined by:

$$E_T = 0.5K_T \sum_{i=1}^{N-1} (\propto_i + \gamma_i - \phi_0)^2 \qquad (3.39),$$

where K_T stands for the torsional rigidity, and the sum of α and γ Euler angles defines the torsion between the local coordinate systems of two consecutive beads. ϕ is a parameter that gives the mean DNA twist and it depends on the bead-bead equilibrium distance and the helical repeat. The values of the bending and torsional rigidity constants are related to the DNA bending (P) and twist (C) persistence lengths, respectively, and the bead-bead equilibrium distance:

$$K_B = \frac{Pk_BT}{l_0} \qquad (3.40),$$

and

$$K_T = \frac{Ck_BT}{l_0} \qquad (3.41).$$

The electrostatic repulsion energy (E_{ele}), the only non-local term in the Hamiltonian, is determined using Debye-Hückel potential:

$$E_{ele} = \frac{\nu^2 l_0^2}{D} \sum_{i,j} \frac{e^{-\kappa r_{ij}}}{r_{ij}} \qquad (3.42),$$

where ν is the salt-dependent Stigter's effective DNA linear charge density [34], D is the dielectric constant of water, κ is the inverse Debye length, and r_{ij} is the distance between two beads. Alternatively, one can define a value for the effective DNA-bead charge, q_{DNA}, and account for the electrostatic potential as follows:

$$E_{DH} = \frac{{q_{DNA}}^2}{4\pi\epsilon\epsilon_0} \sum_i \sum_{j \neq i} \frac{e^{-\kappa r_{ij}}}{r_{ij}} \qquad (3.43),$$

where ε_0 is the electric permittivity of vacuum, and ε is the dielectric constant (set to 80).

3.2. Trajectory Files

Outputs from MD (or any sampling simulation technique) are usually represented using a couple of data files: topology and trajectory.

A typical MD-topology file contains the static information about a molecular system that is needed for a molecular simulation. Items in this file include, among many others, a list of atoms, their non-bonded parameters for van der Waals and electrostatic interactions, and the complete connectivity in terms of bonds, angles and dihedrals. Commonly, solvent molecules are defined only once in the topology file, even though many solvent molecules are usually included in the actual molecular system.

Trajectory files contain the dynamic information about a molecular system; typically the Cartesian 3D coordinates of all the atoms throughout the simulation. Depending on the format used, more information about the number of atoms, snapshots, system size, etc. can be also found.

Topology and trajectory files both are needed for visualization and analysis, as they complement each other. Every MD package has historically worked with a particular topology and coordinate file format. The most well-known formats are revised in the next sections. Please note that it is not pretended to be an exhaustive list, only the ones used during the development of this thesis are going to be pointed out.

3.2.1. Topology Formats

Each MD simulation program uses a particular topology and trajectory file format. The majority of trajectory file formats are easily interconvertible, but topology files cannot be directly translated to another format because of the differences in the functional forms of the force-field potentials. For example, if a given force-field has an additional term in the potential energy function, it has to be added to the topology file, and that can only be made if the topology format is designed to accept this new term. The same thing happens with the usage of different scaling factors or parameter units.

The most used topology file formats are *AMBER Prmtop*, *X-Plor PSF* and *GROMACS topology top* files.

3.2.1.1. AMBER Prmtop

The *Parameter-Topology* file format used extensively by the AMBER software suite is referred to as the *Prmtop* file for short. The format specification of the AMBER *Prmtop* file was written a decade ago, and a recent drastic update took place with the release of version 7 of AMBER package in 2004 (known as *Prmtop7*). *Prmtop* file is a text file divided in several sections designed to be parsed easily using simple Fortran code. Each section contains particular topology information, such as atom name, charge, mass, angles, dihedrals, etc. It can be modified manually, but as the size of the system increases, the hand-editing becomes increasingly complex.

3.2.1.2. Protein Structure File (PSF)

X-Plor Protein Structure Files (*PSF*) are used by NAMD and CHARMM molecular simulations programs. Trajectory analysis program ptraj (included in Ambertools package) is also able to read *PSF* file topologies. The high similarity in the functional form of the two potential energy functions used by AMBER and CHARMM force-fields gives rise to the possible use of one force-field within the other MD engine. Therefore, the conversion of *PSF* files to AMBER Prmtop format is possible with the use of AMBER *chamber* (CHARMM ↔ AMBER) program. The *PSF* contains six main sections of interest: atoms, bonds, angles, dihedrals, improper dihedrals (force terms used to maintain planarity) and cross-terms.

3.2.1.3. GROMACS topology

A *top* file defines the entire system topology, either directly, or by including *itp* files in GROMACS. *Itp* files are used to define individual (or multiple) components of a topology as a separate file. This is particularly useful if there is a molecule that is used frequently, and also reduces the size of the system topology file, splitting it in different parts. There is currently no tool available for conversion between GROMACS topology format and other formats, due to the internal differences in both approaches. There is, however, a method to convert small molecules parameterized with AMBER force-field

into GROMACS format, allowing simulations of these systems with GROMACS MD package.

3.2.2. Trajectory Formats

American Standard Code for Information Interchange (*ASCII*) files are text files that can be human-readable. ASCII files basically contain letters, numbers, carriage returns and punctuation marks, so they can be visualized with any common text editor. At the contrary, binary files contain information encoded in binary form, aimed to the efficient storage and communication between computer machines, and thus, not human-readable.

3.2.2.1. ASCII (text) formats

3.2.2.1.1. PDB Models

The *Protein Data Bank* file format (PDB) can be used to save trajectory information using the *MODEL* tag, originally added to store in a single PDB file the ensemble of structures extracted from an NMR study. The tag *MODEL* is accompanied by a number, thus snapshot order can be maintained.

It is a user-friendly format in the sense that it is written in the standard PDB format that can be directly read by any scientist familiar with 3D structures, and visualized with any molecular graphics computer application. Drawbacks about this model are the limit in the number of frames, residues and atoms associated with the PDB format (9,999 models/residues, 99,999 atoms), the limit also in the information that can be stored in a PDB file, restricted by the file format specification, and the huge size of the trajectory produced, as information regarding to atoms and residues are replicated for every snapshot.

http://www.wwpdb.org/documentation/format33/v3.3.html

3.2.2.1.2. XYZ format

The *XYZ* chemical file format is widely supported by many programs, although no formal specification has been published. Consequently, many slightly different *XYZ* file formats coexist (*Tinker XYZ*, *UniChem XYZ*, etc.).

XYZ files are structured in this way:

- First line contains the number of atoms in the file.
- Second line contains a title, comment, or filename.
- Remaining lines contain atom information. Each line starts with the element symbol, followed by x, y and z coordinates in angstroms separated by whitespace.

Multiple molecules or frames can be contained within one file, so it supports trajectory storage. *XYZ* files can be directly represented by a molecular viewer, as they contain all the basic information needed to build the 3D model.

3.2.2.1.3. Atom Coordinates format

ASCII files containing only Cartesian atom coordinates were originally used by AMBER package [18] to save trajectory as well as restart atom positions (or velocities) in 3D space. They consist on a line of chained 3D coordinates (x,y,z) for all atoms of the structure and for every simulation snapshot.

The main difference between atom coordinates formats is the number of coordinates stored per line and their decimal format. The most common trajectory files are the ones used by AMBER package: *AMBER coordinate/restart* file with 6 coordinates per line and decimal format F12.7 (fixed point notation with field width 12 and 7 decimal places), and the *AMBER trajectory* (also called *mdcrd*), with 10 coordinates per line and format F8.3 (fixed point notation with field width 8 and 3 decimal places).

These formats are very efficient in comparison to the PDB files, because there is no replication of data, but they need to be accompanied by a topology file with the description of atom, residues and connections between them to be properly visualized with a computer graphic program (usually a simple PDB file can serve as a topology). This type of trajectory files are commonly used by Coarse-Grained algorithms, as they are very easy to write, are human-readable and can be opened with the majority of trajectory visualizers and analysis programs.

3.2.2.2. Binary formats

3.2.2.2.1. BinPos format

Scripps Research Institute BinPos format is a binary formatted file to store atom coordinates. It is basically a translation of the ASCII atom coordinate format to binary code. The only additional information stored is a magic number that identifies the *BinPos* format and the number of atoms per snapshot. The remainder is the chain of coordinates binary encoded.

The major drawback of this format is its architecture dependency. Integers and floats codification depends on the architecture, thus it needs to be converted if working in different platforms (little endian, big endian).

3.2.2.2.2. DCD format

X-Plor DCD format, and its variant CHARMM DCD are binary trajectory file formats produced by CHARMM and NAMD MD packages. The file specification is quite simple, having just a header consisting of three metadata lines with information about the trajectory (title, number of atoms, timestep, etc.). The only difference between CHARMM and NAMD DCD formats is the presence of the number of coordinate sets in the metadata section in the case of CHARMM DCD. The remaining part of the file contains just the set of coordinates. The binary content is also architecture specific, so it must be converted when transferring between different platforms; the developers provide a tool to solve this problem: *FlipDCD*.

3.2.2.2.3. XTC format

The *XTC* [35] format is a portable binary format for trajectories produced by GROMACS package. It uses the *External Data Representation* (*xdr*) routines for writing and reading data which were created for the *Unix Network File System* (*NFS*).

Trajectories are written in *XTC* files using a reduced precision (lossy) algorithm which works multiplying the coordinates by a scaling factor (typically 1000), so converting them to pm (GROMACS standard distance unit is nm). This allows an integer rounding of the values. Several other tricks are performed, such as making use of atom proximity information: atoms close in sequence are usually close in space (e.g. water molecules). That makes *XTC* format the most efficient in terms of disk usage, in most cases reducing by a factor of 2 the size of any other binary trajectory format, whereas it is still overcame in I/O velocity by the *netCDF* format (see below).

3.2.2.2.4. NetCDF format

A binary file format for trajectory data based on *Network Common Data Form* (*NetCDF*) library is supported by AMBER package from version 9. This binary format is characterized by:

- Efficient input and output.
- Compact, high-precision representation of data.
- Portability of data files across different machine architectures.
- Extensibility of the format (ability to add additional data without rewriting parsers).
- Compatibility with existing tools and formats.

The file format is based on the NetCDF set of software libraries developed by Unidata (http://www.unidata.ucar.edu/software/netcdf/), specifically designed for representation of array-oriented scientific data [36]. NetCDF uses also *xdr* routines (see *XTC* format). The format is self-describing, there is a header which describes the layout of the rest of the file, in particular the data arrays, as well as arbitrary file metadata in the form of name/value attributes. The format is also platform independent. NetCDF is currently considered the most efficient trajectory file format to deal with MD data.

3.2.2.2.5. HDF5 format

HDF5 [37] is a data model, library, and file format for storing and managing data, based on *Hierarchical Data Format* (*HDF*). It is used as a molecular dynamics trajectory file, and is currently supported by the Python *MDTraj* package. *HDF* is the name of a set of file formats and libraries designed to store and organize large amounts of numerical data, originally developed at the *National Center for Supercomputing Applications* at the *University of Illinois*. *HDF* is currently supported by many commercial and non-commercial software platforms such as *Java, MATLAB/Scilab, Octave, Python* and *R*.

The *HDF5* design goals include:

- Trajectories should be small and space efficient on disk.
- Trajectories should be fast to write and read.

- Data format should support flexible read options: random access to different frames, reading trajectory dimensions without loading the file into memory, load directly a subset of the atoms or snapshots, etc.
- Trajectory format should be easily extensible in a backward compatible manner.

HDF5 is able to join the best parts of the already existent trajectory file formats *netCDF* and *XTC*: the efficiency in I/O of the former as well as the compressed size of the latter. Moreover, *HDF5* is designed to contain not only the trajectory data but also the topology information, thus allowing direct analysis and visualization. The last version of *netCDF* libraries (version 4) actually uses *HDF5* libraries to take profit of the advantages offered by this numerical data package.

3.2.2.2.6. TNG format

Trajectory Next Generation (TNG) [38, 39] is a recently developed format for storage of molecular simulation data. It is designed and implemented by the GROMACS development group, and it is called to be the substitute of the aforementioned XTC format (and according to the developers, a future standard file format).

Among all the requirements taken into account, the most relevant are:

- Fully architecture-independent format, regarding both endianness and the ability to mix single/double precision trajectories and I/O libraries.
- Self-sufficient, it should not require any other files for reading, and all the data should be contained in a single file for easy transport.
- Temporal compression of data, improving the compression rate of the previous *XTC* format.
- Possibility to store meta-data with information about the simulation.
- Direct access to a particular frame.
- Efficient parallel I/O.

Compression algorithms used [39] give impressive results, but remove the "generality" of the format, as they can only be used in MD-specific numerical data, in contrast to *netCDF* or *HDF5* which are general data formats.

The source code is designed to be very small (just 0.2MB) and easy to support and extend, in order to fulfill the future GROMACS requirements to run on nonstandard platforms. *TNG* is announced to be the default trajectory output format in the new GROMACS releases (from v. 5.0).

3.2.2.2.7. PCAZip format

PCAZip format is a binary compressed file to store atom coordinates based on *Essential Dynamics* (*ED*) and *Principal Component Analysis* (*PCA*) [40]. The compression is made projecting the Cartesian snapshots collected along the trajectory into an orthogonal space defined by the most relevant eigenvectors obtained by diagonalization of the covariance matrix (PCA).

In the compression/decompression process, part of the original information is lost, depending on the final number of eigenvectors chosen. However, with a reasonable

choice of the set of eigenvectors the compression typically reduces the trajectory file to less than one tenth of their original size with very acceptable loss of information. Compression with *PCAZip* can only be applied to unsolvated structures.

3.3. Trajectory Analyses

3.3.1. Standard Cartesian Analyses

3.3.1.1. Root Mean Square deviation (RMSd)

Root Mean Square deviation (RMSd) is a standard magnitude to calibrate the deviation of a structure with respect to a reference conformation. It is computed as:

$$RMSd = \sqrt{\frac{\sum_i^n d_i^2}{n}} \qquad (3.44),$$

where *n* is the total number of atoms/residues considered in the calculation and d_i :

$$d_i = (x_i - x_i^0)^2 + (y_i - y_i^0)^2 + (z_i - z_i^0)^2 \qquad (3.45),$$

where the two structures compared have been previously aligned to rigid body motions

RMSd is the most common method used to analyze conformational deviations in MD simulations from an experimental structure. It is also commonly used to compute fluctuations in single molecules and groups of atoms or residues, the reference structure being then the average one.

3.3.1.2. Weighted RMSd (wRMSd)

The standard alignment process in RMSd is optimal in the sense that it minimizes the variance between ensembles of structures, but presents some limitations when applied to anisomorphic and flexible molecules, where we can observe flexible regions like loops or hinges which generate large and localized movements in specific parts of the molecule. The standard alignment process would distribute the variance in all residues, providing a non intuitive flexibility picture. For example, in situations like the one shown in figure 3.2, a human expert will not perform a global alignment (as RMSd does), but a local alignment considering only the large domain, focusing all the structural variability into the small one. Approaches like *Gaussian RMSd (gRMSd)* and *Template Modeling Score (TM-Score)* are going to this direction.

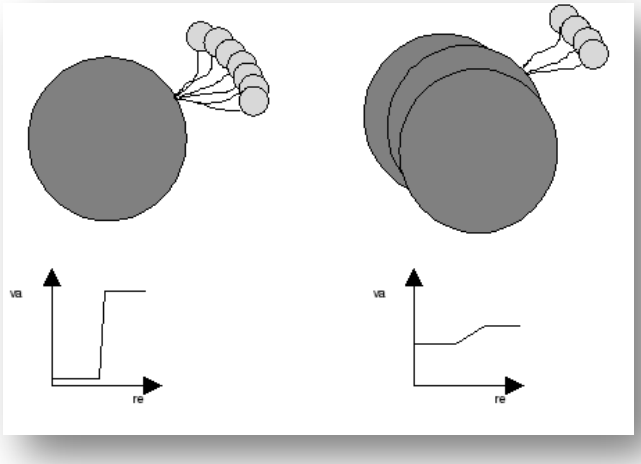


Fig 3.2.- Weighted RMSd. *Movement of a small domain with respect to a large one with the associated variance plot. Left picture represents a Weighted RMSd alignment where the big domain is fully aligned in each structure of the ensemble, whereas right picture represents a normal RMSd, where big domain superposition is penalized by the small domain flexibility.*

Weighted RMSd can be defined as:

$$wRMSd = \sqrt{\frac{\sum_i^n w_i d_i^2}{W}} \qquad (3.46),$$

where w_i is the weight of coordinate i and W the averaged sum of all weights.

In the *Gaussian RMSd* approximation, an iterative algorithm based on the procedure used in the robust linear regression is applied [41]. The weight factor w_i is determined with a gaussian term based on the distance between each residue/atom pair, assigning high weightings ($w_i \sim 1$) to static domains and lower weightings ($w_i \sim 0$) to more flexible regions like loops or hinges:

$$w_i = e^{-{d_i}^2/c} \qquad (3.47),$$

with c being an empirical scaling factor (2-5 Å^2).

Note that equations 3.46 and 3.47 are interdependent via the alignment and accordingly need to be solved iteratively until a convergence criterion is met (ΔwRMSd $< 10^{-6}$ Å). A residue in a flexible loop will have originally a weight of one, leading to an illogical alignment, whereas with this approximation, eq. 3.47 will detect that the residue is quite mobile and will reduce its weight in the gRMSd function, and in parallel its weight into the new alignment. Subsequent cycles will maximize this effect, leading to an overall alignment constructed considering the rigid blocks of the protein as the ones with more relevance.

TM-Score is an algorithm designed to obtain good protein structure templates in the process of molecular modelling or threading [42]. It also computes structural similarity weighting the close atom pairs stronger than the distant matches. The main

difference is that TM-Score results are normalized by a protein size dependent scale so that the average score of random protein pairs has no bias to the target protein's length.

TM-Score scoring function is determined as:

$$TM - Score = Max\left[\frac{1}{L_N}\sum_{1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \quad (3.48),$$

where L_N is the length of the particular structure, L_T is the length of the aligned residues to the reference structure, d_i is the distance between the i^{th} pair of aligned residues and d_0 is a scale to normalize the match difference. The *Max* operation denotes the maximum value after optimal spatial superposition. Results lay between 0 and 1, being 1 a perfect superposition. A value greater than 0.4 is assumed to be significant.

The d_0 scale is defined as:

$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8 \quad (3.49),$$

which drops, for example, from 6.4 to 2.3Å when L_N changes from 300 to 50 residues, thus making the scoring function independent of protein size for the random structure pairs.

3.3.1.3. Radius of Gyration (Rgyr)

Radius of gyration is a measure used in molecular simulations to study the compactness of a structure. Rgyr is defined as the mass weighted distance of each atom from its center-of-mass (COM):

$$Rgyr = \sqrt{\frac{\sum_i^n d_i^2 m_i}{\sum_i^n m_i}} \quad (3.50),$$

where m_i is the mass of the particle i, and d_i is as defined in eq. 3.45.

Rgyr is an interesting shape indication, because it can be determined experimentally by techniques such as *Small Angle Neutron/X-ray Scattering*, thus allowing direct comparison with theoretical methods.

3.3.1.4. B-factor fluctuations

B-factor fluctuation term (also called Debye-Waller factor), although originally used in X-ray crystallography to describe the attenuation of x-ray scattering or coherent neutron scattering caused by thermal motion, is nowadays extensively used as a standard measure of local residue/atom flexibility. It is determined from the oscillations of a residue with respect to its equilibrium position:

$$B - factor = \frac{8}{3}\pi^2\langle\Delta r^2\rangle \quad (3.51),$$

where $\langle \Delta r^2 \rangle$ stands for the oscillations of atoms/residues around equilibrium positions. The more flexible an atom is, the larger the displacement from the mean position will be.

B-factor is usually found in PDB files, and molecules in 3D are typically represented in graphics programs colored according to these values, with high B-factors in red color (hot) and low B-factors in blue color (cold). An inspection of a PDB structure with such coloring scheme will immediately reveal regions with high flexibility. Values of the B-factors are normally ranging between 15 to 30Å^2, but more flexible regions can show higher values. Theoretically computed B-factor profiles can be, in principle, directly compared with experimental X-ray data, but caution is needed in comparison, as low values in PDB B-factors can come from crystal lattice effects that rigidify exposed protein residues, and high values can also indicate errors in the model building, or simply regions were no clear density signal is found.

3.3.2. Solvent Accessible Surface Area (SASA)

Solvent accessible surface area is the region of the protein surface exposed enough to be able to interact with solvent molecules. It is usually defined as a surface built by the delineation drawn by the center of a sphere (rough representation of a solvent molecule, usually of 1.4 Å of radii; i.e. a water) rolling over the molecular surface (Fig. 3.3).

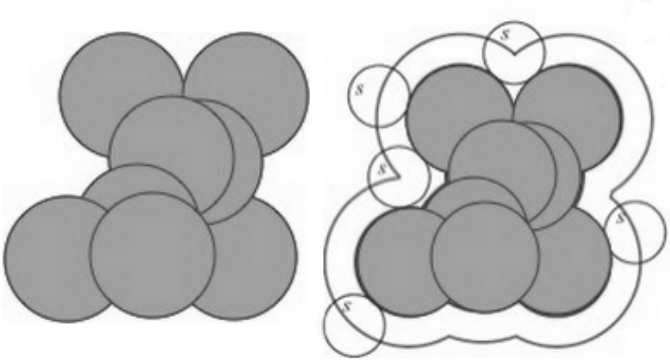


Fig 3.3.- Solvent Accessible Surface Area. *Atoms of the protein are represented as van der Waals spheres, and the solvent accessible surface area is defined by the center of a rolling ball (S) representing the solvent molecule while going over the protein surface.*

Variations in SASA values throughout the whole trajectory raise important information about conformational changes in the protein surface, with the resulting differences in solvent interactions. The total SASA is sometimes divided into that due to apolar and polar residues, which allows the identification of protein hot-spots, surface regions of the protein with higher probability to be found in protein-protein or protein-ligand interactions. SASA values are computed in this thesis using the well-known software called *NACCESS* [43], developed by Simon Hubbard in Manchester.

3.3.3. Hydrogen Bonds (HB)

Hydrogen bonds (HBs) are electrostatic attractive interactions between a proton in one molecule and an electronegative atom in the other. The resulting interaction is

weaker than a covalent bond, but stronger than a van der Waals interaction. HBs are very important for the final 3D structure adopted by a macromolecule, as it is usually maintained by intramolecular HBs (Fig. 3.4). These bonds are crucial to understand protein and water structure.

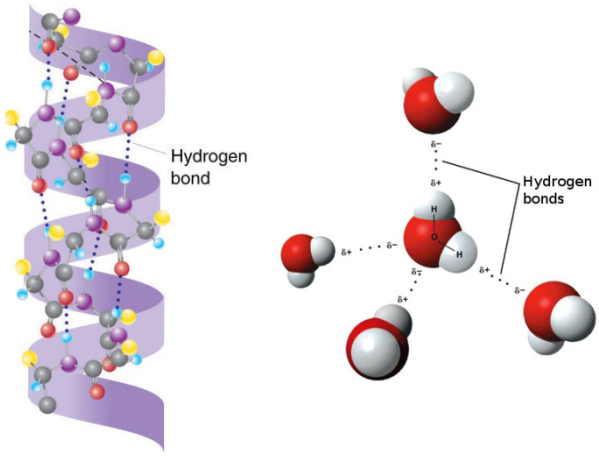


Fig 3.4.- Hydrogen Bond interactions. *Classical examples of Hydrogen Bond interactions: intramolecular HBs helping in the formation of an alpha-helix secondary structure (left part) and intermolecular HBs forming the tetrahedral arrangement between water molecules (right part).*

HBs are usually defined with the pair "*polar acceptor atom – hydrogen donor atom*". The typical parameters used to define a HB are the following: distance between the heavy atoms below a certain cutoff (usually 3.5Å) and angle between the acceptor hydrogen and the donor atom below a certain cutoff (usually 120º). Studies of HB dynamics (breaking/formation) during a simulation give quantitative information about conformation reshapes, from changes in secondary structures or percentage of native HBs lost.

3.3.4. Principal Component Analysis

The extraction of relevant motions from MD simulation data is not straightforward as the majority of the dynamic information when analysing a trajectory comes from irrelevant local fluctuations. The use of *Principal Component Analysis* (PCA), a multivariate statistical analysis to obtain collective variables on the atomic positional fluctuations, helps to separate the configurational space in two subspaces: an "essential" subspace containing relevant motions, and another one containing irrelevant local fluctuations. The analysis of molecule dynamics using PCA applied to the atomic positional fluctuations is known as *Essential Dynamics* (ED) [44, 45].

PCA is able to reduce the complexity (dimensionality) of the system obtaining a complete and orthogonal basis set, by means of diagonalizing the Cartesian covariance matrix C containing the atomic positional fluctuations in all 3 coordinate axes:

$$C = cov(X) = \langle \Delta X \Delta X^T \rangle \qquad (3.52),$$

with

$$\Delta X = X - x_{ref} \quad (3.53),$$

where x_{ref} is a reference value (usually the average structure), ΔX^T is the transpose of ΔX and the angle brackets represent averaging over the distribution.

Therefore, the transformation matrix T for the diagonalization of the correlation (covariance) matrix, provides:

$$\Lambda = T^T C T \quad (3.54),$$

where Λ is the diagonalized correlation matrix with eigenvalues λ, and the i^{th} column of the transformation T corresponds to the eigenvector belonging to λ_i.

Prior to the PCA analysis, a least-square fit (either RMSd or gRMSd based) removing the overall translational and rotational effects should be applied. Then, when a sufficient number of independent configurations typically 100-1000 (the number of degrees of freedom), C is diagonalized to obtain 3N eigenvalues, with the ones (at least 6) representing overall translational and rotation being nearly zero. The set of eigenvectors define a new complete basis set, with each eigenvalue indicating the variance explained by each essential movement of the molecule (each eigenvector).

For proteins, typically, more than 90% of the total atomic fluctuation is described by ~20% of the principal components, involving the majority of the atoms, thus defining an essential subspace containing the main conformational flexibility information.

While high frequency modes are local and show small amplitude, the lower frequency modes (those explaining larger variance percentages) exhibit larger amplitudes and often a collective character, i.e, they involve correlated motions of a large number of atoms [46]. The degree of collectivity of a mode reflects the number of atoms which are significantly affected by that mode:

$$\kappa_i = \frac{1}{N} exp\left\{-\sum_{n=1}^{N} u_{i,n}^2 \log u_{i,n}^2\right\} \quad (3.55),$$

with

$$u_{i,n}^2 = \alpha\left(\frac{Q_{i,3n-2}^2 + Q_{i,3n-1}^2 + Q_{i,3n}^2}{m_n}\right) \quad (3.56),$$

where N is the number of atoms, Q is the corresponding eigenvector, α is a term used to normalize the value ($\sum_{n=1}^{N} u_{i,n}^2 = 1$) and m_n is the atom mass.

3.3.5. NMR Observables

Nuclear Magnetic Resonance (NMR) spectroscopy is, together with X-ray spectroscopy, the most important source of structural information on bio-macromolecules. NMR studies the magnetic features of atomic nuclei possessing a particular property called nuclear spin, obtaining structural and dynamic information from such features. Thus, studies on the chemical environment of a nucleus by NMR give rise to a set of observables such as the *Nuclear Overhäuser Effect* (NOE), and the *Chemical Shift* and its derivative *Spin-Spin Coupling* or (*J-Coupling*). Both observables

have been proved to be highly useful for characterizing and refining organic chemical structures. NOE is characterized by being a "space observable", because the atoms that are in close proximity to each other are the ones that give a signal, whereas J-Coupling is observed only when the atoms are connected by 2 or 3 chemical bonds.

NMR observables can be coupled to MD simulations to derive experimental models of structures. Alternatively, NMR observables can be used for validation of MD trajectories. Particularly, in this thesis we have implemented ways to determine J-Coupling and Nuclear Overhäuser Effect (NOE) values from MD ensembles, to be used in the validation of Nucleic Acid trajectories (Chapter 5.5).

3.3.5.1. J-Coupling

J-Coupling or *Scalar Coupling* is a magnetic interaction (coupling) between nuclei with non-zero spin, connected through chemical bonds. It is extensively used to obtain information about bond distance and angles, and about the connectivity of molecules: vicinal ^{1}H-^{1}H Coupling constants (^{3}J-Couplings) are spin-spin couplings between protons located three bonds away (H-C-C-H); vicinal J-Couplings are empirically correlated with the dihedral angle (H-C-C-H) through the so-called Karplus equation:

$$J(\emptyset) = A\cos\emptyset + B\cos\emptyset + C \qquad (3.57).$$

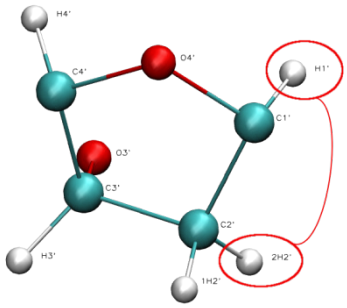


Fig 3.5.- DNA sugar H1'- 2H2' J-Coupling . *Scalar coupling between two protons located three bonds away in a sugar base can be computed using the dihedral angle H-C-C-H and the Karplus equation.*

3.3.5.2. Nuclear Overhäuser Effect (NOE)

Nuclear Overhäuser Effect (NOE) is the transfer of magnetization from one nuclear spin to another via cross-relaxation. The intensity of NOE cross-peaks between two particular protons depends on their relative distance (NOE intensity = 1 / d^{6}). Proton-proton distances derived from NOEs are the most useful NMR parameters for structure elucidation, since they provide information on structural proximity of sequence-distant residues.

NOEs intensities can be represented by a matrix A, the elements a_{ij} of which contain the NOE intensity between the protons i and j. These intensities can be related to the molecular geometry using the following equation:

$$A = exp[-R\tau_m] \qquad (3.58),$$

where τ_m is the experiment mixing time and R represents the relaxation matrix, with elements ρ_{ij} related to the molecule geometry by:

$$\rho_i = k\sum_{\substack{j=1\\ j\neq i}}^{N}\left(\frac{1}{r_{ij}^6}\right)[6\mathrm{J}_2(\omega) + 3\mathrm{J}_1(\omega)+\mathrm{J}_0(\omega)] \qquad (3.59),$$

with

$$\sigma_{ij} = k\left(\frac{1}{r_{ij}^6}\right)[6\mathrm{J}_2(\omega) - \mathrm{J}_0(\omega)] \qquad (3.60),$$

where r_{ij} is the interprotonic distance and $J(\omega)$ are the so-called spectral densities, dependent of the molecule dynamics:

$$J_n(\omega) = \frac{\tau_c}{1 + n^2\omega^2\tau_c^2} \qquad (3.61),$$

where ω is the spectrometer frequency and τ_c is the rotational correlation time.

3.3.6. Protein-specific Analyses

3.3.6.1. Lindemann Coefficient

The Lindemann Coefficient is an estimate of the solid-liquid behavior of proteins based on the root-mean-squared atomic fluctuations. Values below 0.15 are considered to be solid, while values over 0.15 can be considered liquids [47]. The index is computed as:

$$\Delta_L = \frac{\sqrt{\sum_i \langle {\Delta r_i^2}/{N}\rangle}}{a'} \qquad (3.62),$$

where N is the number of atoms, a' is an empirical constant (the most probable non-bonded near-neighbor distance) and Δr_i^2:

$$\Delta r_i^2 = (r_i - \langle r_i\rangle)^2 \qquad (3.63),$$

where r_i is the position of atom i.

3.3.6.2. Hinge Prediction

Protein hinges are localized regions where large changes in main-chain torsional angles occur, producing protein domain movements similar to rotations around an articulated joint. Hinge regions are mechanistically informative areas of the structure of great importance in mediating cooperative motions with functional relevancy. They can be predicted using dynamic information determining residue patches around which large protein movements are organized. Some of the existing methods are:

- Bfactor slope change: using a wRMSd algorithm (see section 3.3.1.2) to superpose the snapshots of a trajectory, hinge regions will display low Bfactors in fixed domains and large Bfactors in floppy domains. Hinge regions can be located examining slope changes in Bfactor landscape.
- Force constant method: computing the force constant for each residue, dependent upon the distances between the residues along the trajectory [48]:

$$k_i = \frac{1}{\langle \left(d_i - \langle d_{ij} \rangle\right)^2 \rangle} \qquad (3.64),$$

where

$$d_i = \langle d_{ij} \rangle_{j*} \qquad (3.65),$$

and d_{ij} is the distance between residue i and residue j, and $j*$ refers to all the residues except $j-1$, j and $j+1$. Hinge regions can be predicted identifying peak values in the force constants landscape.

- Dynamic domain detection: hinge regions can also be identified clustering residues according to its correlation [49]. Regions between clusters can be classified as belonging to hinge regions.

3.3.6.3. Residue Correlation

Residue correlation analysis helps to determine connections between the movements of residues. Strong correlations are usually found between residues sharing trivial relationships such as sequence proximity, but in other cases, a significant correlation can signal important connections.

The residue correlation matrix C contains in every position (C_{ij}):

$$C_{ij} = \frac{\sum_{k=1}^{N} (x_{ik} - \mu_i)(x_{jk} - \mu_j)}{(N-1)\,\sigma_i \sigma_j} \qquad (3.66),$$

where i and j are two protein residues, x_{ik} and x_{jk} are parameter values (e.g. distance), μ_i and μ_j are the corresponding parameter expected (mean) value, σ_i and σ_j are the corresponding parameter standard deviation and N is the number of observations (e.g. MD frames).

A common post-processing task involves the exclusion of sequence and structural neighbors, applying straightforward filters. Another interesting post-processing analysis studies the correlation transfer from one residue to the rest of the protein, by iteratively following the different correlations with structural neighbors (chained correlations). This approach allows finding distant correlations that are not easily perceived just by visual inspection.

3.3.6.4. Apparent Stiffness

The force constant acting between two residues in the case of completely disconnected oscillators can be defined in the harmonic limit from the variance in the inter-residue distance:

$$\kappa_{ij} = \frac{k_B T}{\langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle} \qquad (3.67),$$

where κ_{ij} is the apparent stiffness between residues i and j, k_B is the Boltzmann constant and T is the temperature.

Apparent stiffness is useful in detecting stronger than expected interactions between residues, which might indicate physically intense direct contacts or strong chain-related interactions.

3.3.7. Nucleic Acid-specific Analyses

Nucleic acids (NA) can be studied using Cartesian analysis as described above, but it is often useful to use helical parameters. In this thesis we have followed standard EMBO definitions [50] as implemented in the Curves+ program [51].

3.3.7.1. Helical Parameters

A list of NA helical parameters as defined in *Curves+* algorithm [51] are presented in this section.

3.3.7.1.1. Backbone Torsions

The three major elements of flexibility in the backbone are:

- **Sugar Puckering**

Ribose ring conformation can be described by sugar pucker phase angle (P) and sugar pucker amplitude (τ_m), obtained from the highly correlated torsion angles of the five bonds in the ribose ring (Fig. 3.6, eq. 3.68, 3.69).

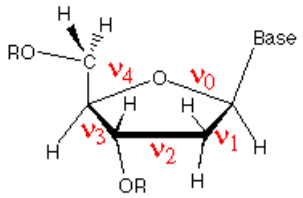


Fig 3.6.- Ribose Ring Torsion Angles. *Sugar puckering is calculated from the five torsion angles (ν_0.. ν_4) formed by the bonds in the ribose ring.*

$$\tau_m = \sqrt{(a^2 + b^2)} \qquad (3.68),$$

and

$$P = \cos^{-1}\left(a/\tau_m\right) \qquad (3.69),$$

where:

$$a = 0.4 \sum_{i=0}^{4} v_i \cos[0.8\,\pi\, i] \qquad (3.70),$$

and

$$b = -0.4 \sum_{i=0}^{4} v_i \sin[0.8\,\pi\, i] \qquad (3.71).$$

Sugar Puckering annotation is done by dividing the pseudo-rotational circle in four equivalent sections (Fig. 3.7), according to the sugar pucker angle *P*:

- North: 315:45º
- East: 45:135º
- South: 135:225º
- West: 225:315º

These four conformations are those dominating sugar conformational space, in agreement with all available experimental data. The most found conformations are the *C2'-endo* (144º-180º) of the B-helices and the *C3'-endo* (0º-36º) of the A-helices.

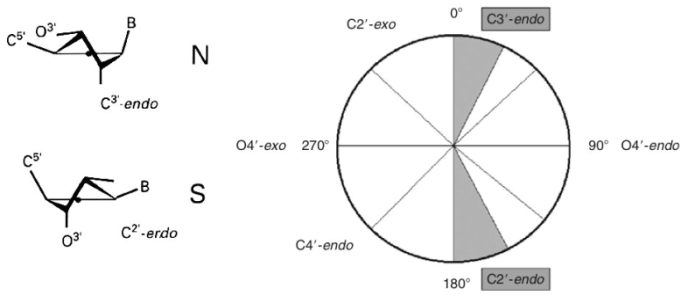


Fig 3.7.- Sugar Puckering Pseudo-rotational circle. *The Pseudo-rotational circle is divided in four main conformations (north, south, east and west). Common B-DNA explored region (C2'-endo, S-SE) and A-DNA/RNA explored region (C3'-endo, N-NE) are highlited.*

- **Alpha-Gamma torsions**

The torsion angles in the backbone of nucleic acid chains are designated with Greek letters (Fig. 3.8). Rotations around *α/γ* torsions out of the usual values generate non-canonical local conformations leading to a reduced twist and they have been reported as being important in the formation of several protein-DNA complexes.

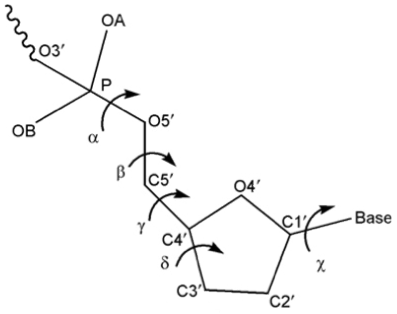


Fig 3.8.- Nucleic Acid Backbone Torsion Angles. *The set of NA backbone torsion angles are identified by greek letters. Rotations around α/γ torsions generate non-canonical local conformations.*

- **ε and ζ torsions**

ε and ζ torsion angles determine the position of the phosphate with respect to the sugar of the preceding nucleotide (Fig. 3.9). The concerted rotation around *ζ/ε* torsions generates two major conformers: *BI and BII*, which are experimentally known to co-exist in a ratio around 80%:20% (BI:BII) in B-DNA. The difference between ε and ζ angles is commonly used to characterize the BI/BII conformational state.

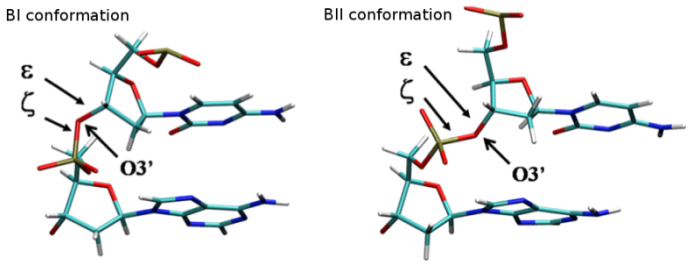
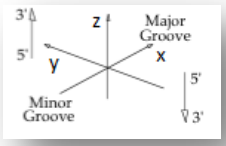


Fig 3.9.- BI-BII Conformations. *ε* and ζ torsion angles from the bonds between O3'atom and its neighbors define the phosphate group conformation: BI or BII.

3.3.7.1.2. Base Pair-Axis Parameters

Translational (*x/y-displacement*) and rotational (*inclination, tip*) parameters related to a dinucleotide Base Pair with respect to a reference helical axis (Fig. 3.10).

- **X-displacement:** Translation around the X-axis.
- **Y-displacement:** Translation around the Y-axis.
- **Inclination:** Rotation around the X-axis.
- **Tip:** Rotation around the Y-axis.



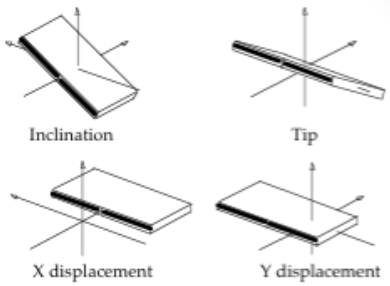
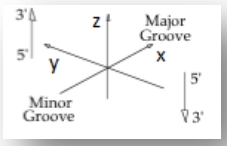


Fig 3.10.- Axis Base Pairs. *Dinucleotide Base Pair translational and rotational parameters with respect to a reference helical axis.*

3.3.7.1.3. Intra-Base Pair Parameters

Translational (*Shear, Stretch, Stagger*) and rotational (*Buckle, Propeller, Opening*) parameters related to a dinucleotide Intra-Base Pair (conformation of one base against its pair complementary) (Fig. 3.11). Zero values of these parameters describe canonical Watson-Crick base pairs in B-DNA and non-zero values describe alternative conformations.

- **Shear:** Translation around the X-axis.
- **Stretch:** Translation around the Y-axis.
- **Stagger:** Translation around the Z-axis.
- **Buckle:** Rotation around the X-axis.
- **Propeller:** Rotation around the Y-axis.
- **Opening:** Rotation around the Z-axis.



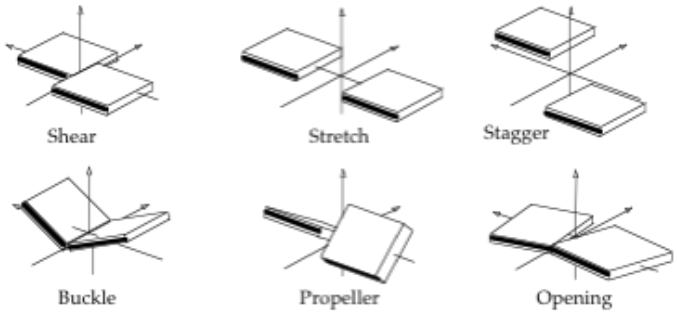
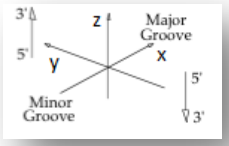


Fig 3.11.- Intra-Base Pairs. *Translational and rotational parameters of one base against its pair complementary in a nucleic acid base pair.*

3.3.7.1.4. Inter-Base Pair Parameters

Translational (*Shift, Slide, Rise*) and rotational (*Tilt, Roll, Twist*) parameters related to a dinucleotide Inter-Base Pair (conformation of one base pair against the successive base pair, also called *base pair step*) (Fig. 3.12).

- **Shift:** Translation around the X-axis.
- **Slide:** Translation around the Y-axis.
- **Rise:** Translation around the Z-axis.
- **Tilt:** Rotation around the X-axis.
- **Roll:** Rotation around the Y-axis.
- **Twist:** Rotation around the Z-axis.



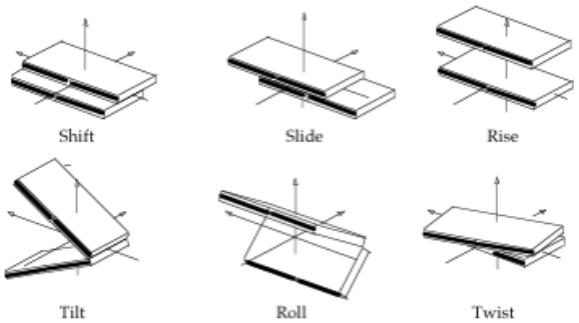


Fig 3.12.- Inter-Base Pairs. *Translational and rotational parameters of one base pair against the successive base pair in a nucleic acid base pair step.*

3.3.7.1.5. Grooves

Nucleic Acid structure's strand backbones appear closer together on one side of the helix than on the other. This creates a *major groove* (where backbones are far apart) and a *minor groove* (where backbones are close together) (Fig. 3.13). *Depth* and *width* of these grooves can be measured giving information about the different conformations that the nucleic acid structure can achieve.

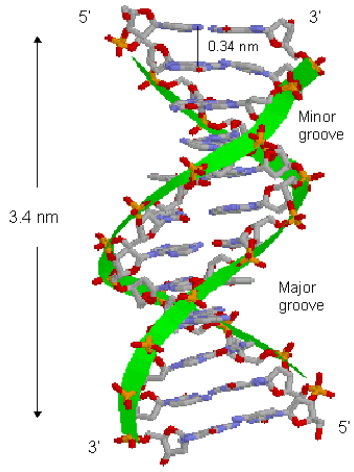


Fig 3.13.- Nucleic Acid Major and Minor Grooves. *Representation of major and minor grooves in a DNA double helix.*

3.3.7.2. HB/Stacking Interactions

Nucleic acid bases interact with each other through a number of different types of interactions. The most common way of base-base recognition is the one formed by hydrogen bond interactions proposed by Watson & Crick (WC), where three HB are formed between *Cytosine* and *Guanine* (C·G) whereas only two HB are formed between *Thymine* (*Uracil*) and *Adenine* (T(U)·A) (Fig 3.14).

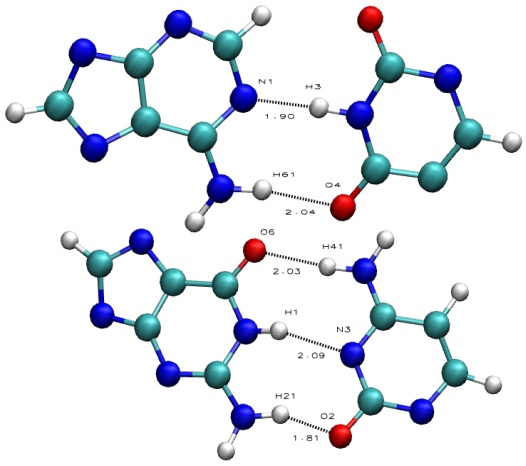


Fig 3.14.- Hydrogen Bonding in Watson-Crick Base Pairing. *Adenine and Thymine form two HB whereas Cytosine and Guanine form three HB, buiding a more stable interaction.*

Other base-base recognition modes different from the WC exist, involving up to two HB interactions, although they are less frequent. The more important ones are the *reverse Watson-Crick, the normal and reversed Hoogsteen* and the *Wobble* pairs.

Apart from HB interactions, energetically favourable stacking interactions between nucleic acid bases in water are believed to play an important role in determining and stabilizing the secondary and tertiary structures of DNA and RNA [52]. In these kinds of interactions, two bases placed one above the other interact noncovalently through the aromatic rings (*π-stacking*; Fig. 3.15).

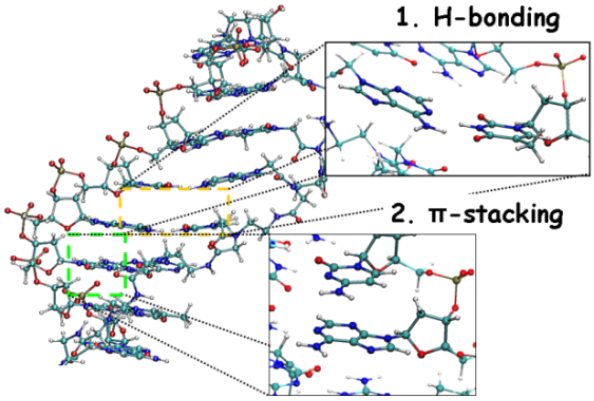


Fig 3.15.- HB/Stacking Interactions. *The most common forces stabilizing the NA helices: HB interactions (1) and Stacking interactions (2).*

3.3.7.3. Inter-Base Pair Stiffness Constants

Essential dynamics provides important information on the deformability of molecules (e.g. DNA) along their low-frequency modes. Unfortunately, this information is often difficult to manipulate, since essential movements are a complex combination of Cartesian movements. A different approach for nucleic acid molecules uses the helical inter-base pair parameters (*twist (w), roll (r), tilt (t), rise (s), slide (l) and shift (f)*) to compute a covariance matrix in helical space, obtaining then the force constants representing the energetic response of the molecule to deformation along helical coordinates.

The stiffness matrix can be determined by inversion of the covariance matrix C computed in the helical space, containing in every position (C_{ij}):

$$C_{ij} = \frac{\sum_{k=1}^{N}(x_{ik} - \mu_i)(x_{jk} - \mu_j)}{N - 1} \qquad (3.72),$$

where i and j are two of the k variables (i.e. the six helical parameters), x_{ik} and x_{jk} are inter-base pair parameter values, μ_i and μ_j are the corresponding helical parameter mean values and N is the number of observations (e.g. MD frames).

The stiffness matrix Ξ is then defined as:

$$\Xi = k_B T C^{-1} = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{rw} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{tw} & k_{tr} & k_t & k_{ts} & k_{tl} & k_{tf} \\ k_{sw} & k_{sr} & k_{st} & k_s & k_{sl} & k_{sf} \\ k_{lw} & k_{lr} & k_{lt} & k_{ls} & k_l & k_{lf} \\ k_{fw} & k_{fr} & k_{ft} & k_{fs} & k_{fl} & k_f \end{pmatrix} \quad (3.73),$$

where *k* stands for the different stiffness constants defining the 36 elements of the stiffness matrix. Diagonal elements (k_{ij}) correspond to pure helical stiffness constants, whereas non-diagonal ones (k_i) correspond to coupling terms.

3.3.8. Solvent-specific Analyses

3.3.8.1. Atom/Residue Type Water Preference

Solvent analysis is very useful to understand protein properties, but also complex due to the large number of molecules and their high mobility. In order to study particular preferences for water molecules to certain atom or residue types, each particular water molecule should be followed, getting results from a water molecule point of view. Water molecules tracking is represented by this formula:

$$Pref_{resType} = \frac{\sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} numSnapshotsBound_j}{SASA_i} \right)}{N * n}, \quad (3.74)$$

where i=*1..n* are all the residues of type *resType* (e.g. Arg) in the trajectory (or dataset); *j=1..m* are all the atoms of a given residue; $SASA_i$ is the *Solvent Accessible Surface Area* (absolute value computed with NACCESS program [43]) of the corresponding residue side chain or backbone (usually considering only exposed residues with *SASA* > 10Å^2). $numSnapshotsBound_j$ is the number of snapshots where protein atom *j* is in contact with a particular water molecule (being *j* the nearest protein atom for this specific water molecule). Finally, *N* is the total number of snapshots. Final units are *contacts/ps*Å*^{2}.

Similarly, atom type preferences from a water molecule point of view are calculated using:

$$Pref_{atType} = \frac{\sum_{i=1}^{n} \left(\frac{numSnapshotsBound_i}{SASA_i} \right)}{N * n}, \quad (3.75)$$

where *i=1..n* are all the atoms of type atType (e.g. alpha carbon) in the trajectory; $SASA_i$ is the *Solvent Accessible Surface Area* (absolute value) of the corresponding atom (usually considering only exposed atoms with $SASA > 10Å^2$); $numSnapshotsBound_i$ is the number of snapshots where the corresponding protein atom *i* is in contact with a particular water molecule (being *i* the nearest protein atom for this specific water molecule). Finally, *N* is the total number of snapshots. Units are $contacts/ps*Å^2$.

3.3.8.2. Residue Type Hydration

Residue type hydration values are computed by accounting the number of contacts with water molecules normalized by the SASA:

$$Hydration_{resType} = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m} numContactWats_{ij}}{SASA_{ij} * f}\right)}{n * m}, \qquad (3.76)$$

where *i=1..n* are all the snapshots (representative set) from the trajectory; *j=1..m* are all the residues of type *resType* (e.g. Arg) in the protein (or dataset); $SASA_{ij}$ is the *Solvent Accessible Surface Area* (absolute value) for the residue *j* in the snapshot *i* (usually considering only exposed atoms with $SASA > 10$ Å^2); $numContactWats_{ij}$ is the number of water molecules attached to the corresponding residue *j* in the snapshot *i* (considering attached when the distance between water molecule and any protein atom belonging to the residue *j* is smaller than 3.5Å); *f* is the hydration factor, the maximum number of water molecules that can be fitted in one Å^2 ($f = 0.127$ $wats/Å^2$).

3.3.8.3. Mean Residence Time

A wide range of different definitions for mean residence times exist in the literature, giving rise to different interpretations of the results and generating some discrepancies [53, 54]. The approach used in this thesis follows the definition postulated by Impey et al. [55], although some variants have been introduced enriching the resulting analysis. The approach consists of a calculation of the relaxation of water molecules in the hydration layers around a protein atom by means of a survival probability function $P_{i,j}(t,t+t';t^*)$, where the function adopts a value of 1 if the water molecule labelled *j* has been attached (using a certain cutoff distance) to the particular protein atom *i*, from time *t* to time *t+t'*, without escaping by period longer than *t** in this time interval, and 0 otherwise:

$$P_{i,j}(t) = \sum_{t'} P_{i,j}(t', t'+t; t^*), \qquad (3.77)$$

Taking this formula as a base, we can obtain mean residence times following water molecules during a MD trajectory time window. This way, a range of MRT variants can be computed:

- Number of consecutive snapshots where a particular water molecule is attached to the same residue (making parameter *i* be a residue instead of a single atom), allowing the water molecule to change contacts between different residue atoms.
- Total number of snapshots where a particular water molecule is attached to the same protein atom, allowing the water molecule to escape for a while (*t**) and return to the original contact.
- Total number of snapshots where a particular water molecule is attached to the same protein residue (making parameter *i* be a residue instead of a single atom), allowing the water molecule to escape for a while (*t**) and return to the original contact.

3.3.8.4. Radial Distribution Function (RDF)

Radial distribution functions represent relative water density as a function of distance from the protein:

$$g_{PS}(r) = \frac{1}{\rho\ 4\pi r^2 dr} \sum_S \langle \delta(r - |r_S - r_P|) \rangle \ , \quad (3.78)$$

being ρ the average water density. The summatory corresponds to the number of solvent molecules within a spherical shell of radii between *r* and *r* + *dr*, measured from the reference site *P* in the protein. Index *S* denotes either the water hydrogen or the water oxygen atoms [56].

RDFs for globular proteins usually have a classic shape with a great peak placed at 2.8Å away from the protein, corresponding to the first hydration layer, and another smaller peak around 5Å, corresponding to the second hydration layer.

3.3.8.5. 3-Dimensional Water Density

3-D water density is typically computed using grid approximations. The MD system (protein + solvent) is placed in a grid with specified resolution, representing density points as a measure of the number of water molecules inside a grid cell. Average density values from a MD trajectory can be computed using our *CMIP* program [57]. Water density values got from MD trajectories must be taken with caution, as fluctuations in the exposed protein residues result in grid cells near the protein surface with low water density values, contrary to the expected output. They are however very useful to investigate hydration properties in specific regions, as for example protein cavities.

3.3.8.6. Water Diffusion

Water diffusion coefficient can be obtained from time-dependent mean square displacements (MSD) of the solvent. The translational self-diffusion coefficient *D* can be calculated from:

$$6Dt = \langle|\vec{r}(t) - \vec{r}(0)|^2\rangle \quad t \to \infty, \qquad (3.79)$$

where $\vec{r}(t)$ is the position vector of the solvent molecule at time t, and the brackets indicate that the average is taken over both the time origins and solvent molecules.

3.3.8.7. Protein-Water HB dynamics

In order to follow the dynamics of creation/destruction of intra-protein HB and the role of hydration waters in it, we developed a simple method based on graph theory. Graphs are built for a particular protein HB, following the dynamics throughout the whole MD trajectory, with a 1 ps-resolution. The different possible states (graph nodes) are characterized by the presence or absence of protein HBs, and presence or absence of intervening water molecules. Depending on the resolution we want to obtain, different state fingerprints can be built. The most simple description correspond to just 2-bits fingerprint (presence of intra-protein HB and presence of a water bridge), but more informative ones can be defined, like for example a 4-bits fingerprints (adding to the previous one the presence of solvating water molecules in both protein atoms). State transitions during the trajectory are depicted as the graph edges. A 2-bits fingerprint HB graph is shown in figure 3.16, where, as an example, *0,1* state would mean protein hydrogen bond broken and presence of a water bridge molecule. In this example, the probability of losing the HB without water interference is really low (from *1,0 to 0,0* – 0.074), whereas a previous water bridge formation has a probability of 0.216 (from *1,0 to 1,1*).

The population distribution and frequency of transition along the MD trajectory can be used as the source of thermodynamics and kinetics information. In this work, to evaluate HB significance, *ΔG* of HB stability has been derived from population differences taking as reference states that corresponding to the *00NN* fingerprint (i.e. no protein HB or water bridges, and full solvation of the intervening atoms) (section 5.3).

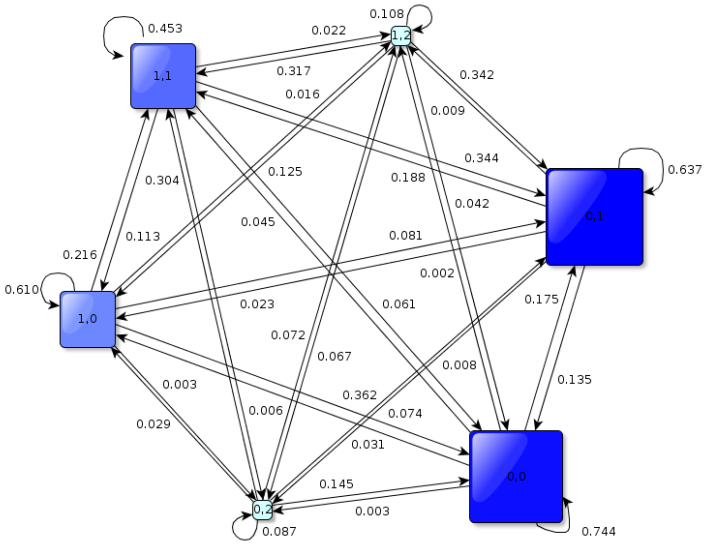


Fig 3.16.- Protein-Water HB dynamics graph. *This particular example shows a 2-bit fingerprint graph, where the first bit represents presence or absence of protein hydrogen bond and the second bit represents presence or absence of a water bridge. Nodes represent states (with size of nodes representing the number of occurrences) and edges represent transition between states with their associated probability.*

3.4. Biological Databases

Biological databases containing information on macromolecular sequences and structures are essential for this research, being the Protein Data Bank (PDB) the most important one. PDB structures are the central point of structural bioinformatics, and the starting point for all of our projects. Additional databases mined in this thesis include well-known structural (CATH and SCOP) and sequential (Pfam) domain databases, a database focused specifically on enzyme data (EC), and a database linking protein descriptions with function, biological processes and cell location (GO). Uniprot database, together with its reference clusters based on sequence identity (UniRef) has been our principal sequence database.

3.5. On-line Tools

The explosion of the *World Wide Web* has provided a fantastic platform for the scientific community to share data information and make useful tools publicly available. The main objectives of this thesis are focused on offering macromolecular flexibility data and analysis tools publicly and freely available through the internet. For this reason, a set of on-line programming tools have been used. They are listed below.

3.5.1. PHP

PHP (*Hypertext Preprocessor*) is a programming language specifically designed for web development (www.php.net). It is extensively used to add interactive and dynamic content in web pages. PHP is a server-side language, the code is executed on the server, and not on the final user browser, dynamically generating and returning information in plain *HyperText Markup Language* (HTML) compatible with any internet browser. It is an interpreted scripting programming language, it doesn't need to be compiled because it is interpreted at runtime by the computer processor. The last PHP version is v5.5.13 (May 2014) is very rich in useful libraries, that allow the user direct connectivity with databases and batch queuing systems. Nowadays is typically used in combination with another programming web tools as *JavaScript* and *Cascade Style Sheets* (*CSS*).

3.5.2. JavaScript

JavaScript, like PHP, is an interpreted programming language specifically designed for web development. The main difference with PHP is that JavaScript is a client-side language, the code is executed on the client browser, rather than in the server machine. Thanks to that, JavaScript is able to interact with the user, communicate with the server and alter the document content. All modern browsers support JavaScript standard (*ECMAScript*), and the majority of current web pages are using JavaScript to add functionality, validate inputs, communicate with the server, etc., thus transforming a static web page to a totally dynamic one. As with PHP, JavaScript is always used in combination with other programming languages, mostly HTML and CSS.

Recently (2005), a new technology named *Ajax* (*Asynchronous JavaScript and XML*) appeared, used to describe the technology behind emerging services like Google Maps: the capacity to reload part of the web page instead of the whole one. Actually, it is not a new technology but a group of interrelated web development techniques joined together using JavaScript. With Ajax, web applications can send and retrieve data asynchronously (in the background) without interfering with the display and behavior of

the existing page. The impact of Ajax in the Web 2.0 time has given even more popularity to the JavaScript language, to a point of being stated as "*the scripting language of the Web*".

3.5.3. MySQL

MySQL (www.mysql.com) is one of the most used relational database management system in the world. It is a popular choice of database for use in web applications, as free software and open-source projects requiring a full-featured database management system often use MySQL, a system now owned by *Oracle Corporation*. The freely downloadable version (*Community Edition*) offers a complete package with different storage engines and integrity checks, whereas the commercial editions (*Standard, Enterprise and Cluster Carrier Grade Edition*) offer additional security features and full 24x7 support.

The popularity of MySQL triggered the appearance of a set programming language libraries for direct connection to its relational databases. PHP, Perl, Java and many other programming languages already offer the possibility to directly connect and query a MySQL database, converting the platform in a perfect candidate to use in bioinformatics tools requiring the power of a relational database.

3.5.4. Jmol

Jmol is a free, open source molecule viewer (*Jmol: an open-source Java viewer for chemical structures in 3D*; http://www.jmol.org/). It is platform-independent and can be executed in different operative systems like Windows, Linux/Unix and Mac OS. It has become popular in the last years thanks to its java applet that can be easily integrated in web pages.Jmol is able to read a complete set of different 3D atomic coordinates file formats, including the standard PDB format. It is able to read and identify multiple models in a PDB file, building and animating trajectories. AMBER topology files and mdcrd trajectory files can be also directly loaded into Jmol.

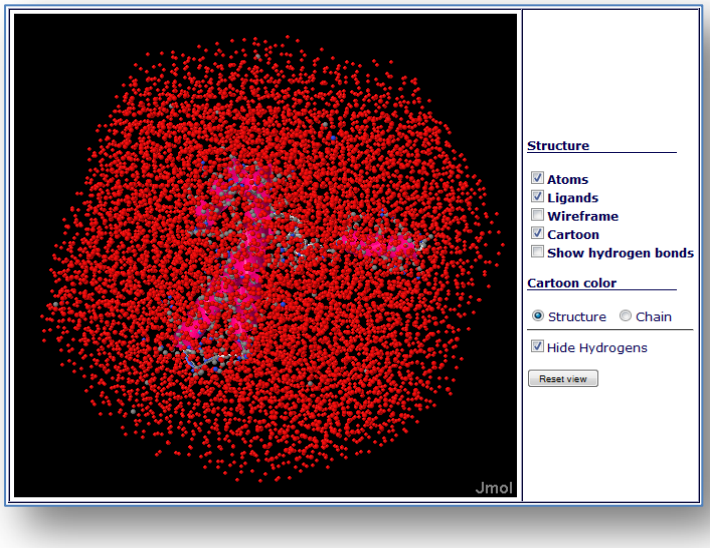


Fig 3.17.- Jmol applet molecule viewer. *Representation of a MD system formed by a protein surrounded by water molecules in a truncated octahedric box with Jmol applet.*

The evidence of the impact of Jmol in the structural bioinformatics field is given by the fact that every important web server containing structural information (PDB, PDBsum, CATH, etc.) offer the possibility to view the molecule in 3D through a Jmol applet viewer. For the times coming, the Jmol team have already implemented a new version of Jmol written in what is believed to be the new standard in web encoding: HTML5 (*JSmol: an open-source HTML5 viewer for chemical structures in 3D*; http://wiki.jmol.org/index.php/JSmol#JSmol). That new version will open up the use of Jmol in any kind of systems, including tablets and smartphones.

3.5.5. BioMOBY Web Services

BioMOBY WS semantic technology (introduced in section 1.6.5) was chosen by the Spanish *Instituto Nacional de Bioinformática* (*INB*; http://www.inab.org/) to implement a portfolio of more than 300 interoperable web services related to life sciences: sequence analyses, biochemistry, structural studies, text mining, etc. The complete set of available WS can be effectively integrated thanks to the BioMOBY technology, allowing the building of complex analysis workflows. INB BioMOBY central catalog stores information about all the registered INB WS, both syntactic and semantic: their inputs, outputs, service descriptions and identifiers and metadata information. The set of INB BioMOBY web services was included in the *EMBRACE* web services collection [58].

A few of the web services implemented by the INB institute are used in a couple of projects of this thesis. Moreover, a set of new web services related to Molecular Dynamics and Coarse-Grained simulations, as well as protein flexibility analyses were designed and implemented, and were finally added to the INB BioMOBY web service collection (see chapter 5.4).

3.6. Section Bibliographic References

[1] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell and P. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.," *J. Am. Chem. Soc. ,* vol. 117, no. 19, pp. 5179-5197, 1995.

[2] A. MacKerell Jr., B. Brooks, C. Brooks III, L. Nilsson, B. Roux, Y. Won and M. Karplus, "CHARMM: The energy function and its parameterization," in *The encyclopedia of computational chemistry*, P. Schleyer, P. Schreiner, N. Allinger, T. Clark, J. Gasteiger, P. Kollman and H. Schaefer III, Eds., Chichester, England, John Wiley & Sons, 1998, pp. 271-277.

[3] W. Jorgensen and J. Tirado-Rives, "The OPLS (optimized potentials for liquid simulations) potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *J. Am. Chem. Soc. ,* vol. 110, no. 6, pp. 1657-1666, 1988.

[4] W. van Gunsteren and H. Berendsen, "GROMOS (GROningen MOlecular Simulation package)," *BIOMOS B.V. Nijenborgh 4, 9747 AG Groningen, The Netherlands.*

[5] E. A. Cino, C. Wing-Yiu and M. Kaarttunen, "Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations," *Journal of Chemical Theory and Computation,* vol. 8, pp. 2725-2740, 2012.

[6] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, A. Hospital, J. Gelpí and M. Orozco, "A consensus view of protein dynamics," *Proceedings of the National Academy of the United States of America,* vol. 104, no. 3, pp. 796-801, 2007.

[7] J. Huang and A. J. Mackerell, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data," *J. Comput. Chem.,* vol. 34, no. 25, pp. 2135-2145, 2013.

[8] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins,* vol. 78, pp. 1950-1958, 2010.

[9] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham and P. J. Jurecka, "Refinement of the Cornell et al. nucleic acid force field based on reference quantum chemical calculations of torsion profiles of the glycosidic torsions.," *J. Chem. Theory Comput.,* vol. 7, p. 2886–2902, 2011.

[10] M. Orozco and F. Luque, "Theoretical methods for the description of the solvent effect in biomolecular systems," *Chem. Rev. ,* vol. 100, pp. 4187-4225, 2000.

[11] A. R. Leach, Molecular Modelling, principles and applications, Pearson Education Limited, 1996.

[12] L. Verlet, "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules," *Phys. Rev.,* vol. 159, no. 1, pp. 98-103, 1967.

[13] R. W. Hockney, The potential calculation and some applications. Volume 9., New York / London: Academic Press, 1970.

[14] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *Journal of Chemical Physics,* vol. 76, pp. 637-649, 1982.

[15] J. P. Ryckaert, G. Ciccotti and H. J. Berendsen, "Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.," *J. Comput. Phys.,* vol. 23, no. 3, pp. 327-341, 1977.

[16] H. C. Andersen, "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations.," *J. Comput. Phys.,* vol. 52, no. 1, pp. 24-34, 1983.

[17] B. Hess, H. Bekker, H. J. Berendsen and J. E. Fraaije, "LINCS: A Linear Constraint Solver for Molecular Simulations," *Journal of Computational Chemistry,* vol. 18, no. 12, pp. 1463-1472, 1997.

[18] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods, "The Amber biomolecular simulation programs," *Computat. Chem. ,* pp. 1668-1688, 2005.

[19] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD," *Journal of Comput. Chem.,* vol. 26, pp. 1781-1802, 2005.

[20] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.,* vol. 4, pp. 435-447, 2008.

[21] V. Tozzini, "Coarse-grained models for proteins," *Current Opinion in Structural Biology,* vol. 15, pp. 144-150, 2005.

[22] M. G. Saunders and G. A. Voth, "Coarse-Graining Methods for Computational Biology," *Annu. Rev. Biophys.,* vol. 42, pp. 73-93, 2013.

[23] D. A. Potoyan, A. Savelyev and G. A. Papoian, "Recent successes in coarse-grained modeling of DNA," *WIREs Comput. Mol. Sci.,* vol. 3, pp. 69-83, 2013.

[24] A. Emperador, O. Carrillo, M. Rueda and M. Orozco, "Exploring the suitability of Coarse-Grained techniques for the representation of protein dynamics," *Biophysical Journal,* vol. 95, pp. 2127-2138, 2008.

[25] L. Orellana, M. Rueda, C. Ferrer-Costa, J. R. Lopez-Blanco, P. Chacón and M. Orozco, "Approaching Elastic Network Models to Molecular Dynamics Flexibility," *J. Chem.*

Theory Comput., vol. 6, pp. 2910-2923, 2010.

[26] A. Pérez, F. Lankas, F. J. Luque and M. Orozco, "Towards a consensus view of B-DNA flexibility," *Nucleic Acids Res.,* vol. 36, pp. 2379-2394, 2008.

[27] I. Faustino, A. Pérez and M. Orozco, "Towards a consensus view of duplex RNA flexibility," *Biophysical Journal,* vol. 99, pp. 1876-1885, 2010.

[28] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. A. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Pérez, N. Spackova and J. Sponer, "A systematic molecular dynamics study of the nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA," *Nucleic Acid Research,* vol. 38, pp. 299-313, 2010.

[29] P. Dans, A. Pérez, I. Faustino, R. Lavery and M. Orozco, "Exploring polymorphisms in B-DNA helical conformations," *Nucleic Acids Res.,* vol. 40, pp. 10668-10678, 2012.

[30] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," *Proc. Natl. Acad. Sci. USA,* vol. 95, no. 19, pp. 11163-11168, 1998.

[31] S. A. Allison, "Brownian dynamics simulation of wormlike chains. Fluorescence depolarization and depolarized light scattering.," *Macromolecules,* vol. 19, p. 118, 1986.

[32] H. Jian, A. V. Vologodskii and T. Schlick, "A combined wormlike-chain and bead model for dynamic simulations of long linear DNA," *J. Comp. Physics,* vol. 136, pp. 168-179, 1997.

[33] H. Jian, T. Schlick and A. V. Vologodskii, "Internal motions of supercoiled DNA: brownian dynamics simulation of site juxtaposition," *J. Mol. Biol.,* vol. 284, no. 2, pp. 287-296, 1998.

[34] D. Stigter, "Interactions of highly charged colloidal cylinders with applications to double-stranded DNA," *Biopolymers,* vol. 16, no. 7, pp. 1435-1448, 1977.

[35] D. Green, K. Meacham, M. Surridge, F. van Hoesel and J. Berendsen, in *Methods and Techniques in Computational Chemistry: METECC-95*, E. Clementi and G. Corongiu, Eds., Cagliari, STEF, 1995, pp. 435-463.

[36] R. Rew and G. Davis, "NetCDF: An interface for scientific data access," *IEEE Comput. Graph. Appl.,* vol. 10, pp. 76-82, 1990.

[37] The HDF Group, "Hierarchical data format, version 5," 2000-2010. [Online]. Available: http://www.hdfgroup.org/HDF5.

[38] M. Lundborg, R. Apostolov, D. Spångberg, A. Gärdenäs, D. van der Spoel and E. Lindahl, "An efficient and extensible format, library, and API for binary trajectory data from molecular simulations," *J. Comput. Chem. ,* vol. 35, pp. 260-269, 2014.

[39] D. Spångberg, D. Larsson and D. van der Spoel, "Trajectory Next Generation," *J. Mol. Model.,* vol. 17, pp. 2669-2685, 2011.

[40] T. Meyer, C. Ferrer-Costa, A. Pérez, M. Rueda, A. Bidon-Chanal, F. J. Luque, C. A. Laughton and M. Orozco, "Essential Dynamics: A tool for efficient trajectory compression and management," *J. Chem. Theory Comp.,* vol. 2, pp. 251-258, 2006.

[41] K. L. Damm and H. A. Carlson, "Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures," *Biophys. J.,* vol. 90, no. 12, pp. 4558-4573, 2006.

[42] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins,* vol. 57, no. 4, pp. 702-710, 2004.

[43] S. Hubbard and J. Thornton, "'NACCESS' computer program," Department of Biochemistry and Molecular Biology, University College London, 1993.

[44] I. Daidone and A. Amadei, "Essential dynamics: foundation and applications," *WIREs Comput. Mol. Sci.,* vol. 2, pp. 762-770, 2012.

[45] A. Amadei, B. M. Linssen and H. J. Berendsen, "Essential Dynamics of Proteins," *Proteins: Structure, Function and Genetics,* vol. 17, pp. 412-425, 1993.

[46] R. Brüschweiler, "Collective protein dynamics and nuclear spin relaxation," *J.Chem.Phys.,* vol. 102, no. 8, pp. 3396-3403, 1995.

[47] Y. Zhou, D. Vitkup and M. Karplus, "Native proteins are surface-molten solids: application of the lindemann criterion for the solid versus liquid state," *J. Mol. Biology,* vol. 285, no. 4, pp. 1371-1375, 1999.

[48] S. Sacquin-Mora and R. Lavery, "Investigating the local flexibility of functional residues in hemoproteins," *Biophysiscal,* vol. 90, no. 8, pp. 2706-2717, 2006.

[49] I. Navizet, F. Cailliez and R. Lavery, "Probing protein mechanics: residue-level properties and their use in defining domains," *Biophysical,* vol. 87, no. 3, pp. 1426-1435, 2004.

[50] R. E. Dickerson, "Definitions and nomenclature of nucleic acid structure components," *Nucleic Acid Research,* vol. 17, pp. 1797-1803, 1989.

[51] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+," *Nucleic Acids Research,* vol. 37, no. 17, pp. 5917-5929, 2009.

[52] R. Luo, H. S. Gilson, J. M. Potter and M. K. Gilson, "The physical basis of nucleic acid base stacking in water," *Biophysical Journal,* vol. 80, pp. 140-148, 2001.

[53] B. Schoenborn, A. Garcia and R. Knott, "Hydration in protein crystallography," *Prog. Biophys. molec. Biol.,* vol. 64, no. 2/3, pp. 105-119, 1995.

[54] V. Makarov, B. Andrews, P. Smith and B. Montgomery Pettitt, "Residence times of water molecules in the hydration sites of myoglobin," *Biophysical Journal,* vol. 79, pp. 2966-2974, 2000.

[55] V. Lounnas and B. Pettitt, "A Connected-cluster of hydration around myoglobin: Correlation between molecular dynamics simulations and experiment," *Proteins: Structure, Function and Bioinformatics,* vol. 18, no. 2, pp. 133-147, 1994.

[56] C. Schröder, T. Rudas, S. Boresch and O. Steinhauser, "Simulation studies of the protein-water interface. I. Properties at the molecular resolution," *The Journal of Chemical Physics,* vol. 124, p. 234907, 2006.

[57] J. Gelpí, S. Kalko, X. Barril, J. Cirera, X. de la Cruz, F. Luque and M. Orozco, "Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins.," *Proteins,* vol. 45, no. 4, pp. 428-437, 2001.

[58] S. Pettifer, J. Ison, M. Kalas, D. Thorne, P. McDermott, I. Jonassen, A. Liaquat, J. M. Fernández, J. M. Rodriguez, INB-Partners, D. G. Pisano, C. Blanchet, M. Uludag, P. Rice, E. Bartaseviciute, K. Rapacki, M. Hekkelman, O. Sand, H. Stockinger, A. Clegg, E. Bongcam-Rudloff, J. Salzemann, V. Breton, T. Attwood, G. Cameron and G. Vriend, "The EMBRACE web service collection," *Nucleic Acids Research,* vol. 38, no. Web Server Issue, pp. W683-W688, 2010.

4. PhD Advisor Report

Paper 1:

MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories.

Tim Meyer*, Marco D'Abramo*, **Adam Hospital***, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluís Gelpí, Modesto Orozco. (* *These authors contributed equally to this work)*

Structure. (2010) 18(11), 1399-1409.

This paper reports the first release of MoDEL, the largest European database containing molecular dynamics trajectories. The foundations of the project, and results concerning the reference dataset were published in an earlier work in Rueda et, al.: *A consensus view of protein dynamics; Proc. Natl. Acad. Sci. (2007) 104, 796-801*, with the participation of Adam Hospital. The first initial release presented in 2010 in this paper was published in Structure, journal with an Impact factor of 5.994 (Q1). Adam Hospital is one of principal authors of the work, being in charge of the preparation of the datasets, the design and implementation of the data management structure and some of the analysis performed.

Paper 2:

Water-omics: High throughput analysis of protein-solvent interactions from MD simulations.

Adam Hospital, Modesto Orozco, Josep Lluís Gelpí.

In preparation (2014)

The paper describes a high-throughput analysis of the solvent-protein interaction out of the MoDEL database. The availability of such amount of information allows obtaining global conclusions although it has a significant difficulty. Adam Hospital has been the principal author of the work, including the design and implementation of the necessary tools. The paper is in its final stage of preparation and will be submitted by the time of thesis defense.

Paper 3:

MDWeb & MDMoby an integrated web-based platform for molecular dynamics simulations.
Adam Hospital, Pau Andrio, Carles Fenollosa, Damjan Cicin-Sain, Modesto Orozco and Josep Lluís Gelpí.

Bioinformatics (2012), 28(9), 1278-1279.

The paper describes a platform and a series of tools designed to perform molecular dynamics simulations, starting from plain structures and performing a full setup, preparing the necessary environment for simulation and providing a basic analysis interface. The paper has been published in Bioinformatics, with an Impact Factor of 5.32 (Q1). Bioinformatics was chosen not only due to its impact in the community but also as having a broader reader's community than more specialized journals. Adam Hospital is the main author of the work. He was the responsible of the design of the datatypes ontology, the implementation of the internal engine, and the web-services interface MDMoby, and most of the web interface. Adam also is the responsible of users support and prepared the help and tutorials section.

Paper 4:

FlexServ: An integrated tool for the analysis of protein flexibility

Jordi Camps, Oliver Carrillo, Agustí Emperador, Laura Orellana, **Adam Hospital**, Manuel Rueda, Damjan Cicin Sain, Marco D'Abramo, Josep Lluís Gelpí, Modesto Orozco.

Bioinformatics (2009), 25, 1709-1710.

FlexServ was the first integrated tool released in our group, seeking to combine in a single server a set of tools providing information about protein flexibility. The work was published in Bioinformatics, with an impact factor of 5.32 (Q1), being the most popular journal in the field. Adam Hospital contributed with the implementation of different analysis tools and part of the Web server regarding to data management.

Paper 5:

NAFlex: A web server for the study of nucleic acids flexibility.

Adam Hospital, Ignacio Faustino, Rosana Collepardo-Guevara, Carlos González, Josep Lluís Gelpí, Modesto Orozco.

Nucleic Acids Research (2013), 41(W1), W47-W55,
NAR Featured Article June 2013.

With the experience of MDWeb, and based in its internal engine, NAFlex is a specific server to study nucleic acid flexibility. The paper was published in the Web Servers special issue of Nucleic Acid Research, with an impact factor of 8.278 (Q1), being selected as Featured Article by the Editorial Board. Adam Hospital was the main author of the work, being responsible of the adaptation of the MDWeb internal engine, the Web Server and the implementation of a series of NA specific analysis software.

5. Results

5.1 Automatic extraction of information from available structural data.

5.1.1 Synopsis

To be able to perform high throughput analysis of macromolecular structures, we designed and developed a bioinformatics infrastructure consisting on hardware (web server machines, storage systems) and software (mirrors, databases, web portal) to automatically save and extract information from PDB and other structural databases. Stored data contains structural information as obtained from the PDB (PDB code, resolution, source, etc.), but also specifically computed information such as protein active site regions. External biological databases were also incorporated to increase the power of the whole framework. The final infrastructure allows us to make complex queries retrieving information in an efficient way, and has served as a starting point for many HT studies.

Although the project doesn't have an associated publication, it constitutes the necessary starting point of this thesis, required in most of the HT studies performed in the group, and in particular in all the projects presented in the following sections.

5.1.2 Structural databases framework.

Web servers were considered the best approach to facilitate structure metadata, i.e., any additional information other than atomic coordinates. Many great servers offering structural information with graphical support exist, among them *RCSB PDB* www.rcsb.org and the *European Bioinformatics Institute PDBSum* www.ebi.ac.uk/pdbsum portals [1]. These servers offer general information for every structure (authors, resolution, chains, etc.), useful links to other databases (Uniprot, SCOP, CATH, PFam, GO, etc.), and graphical support, usually in the form of Jmol interactive viewer or secondary structure representation. Moreover, they allow the user to search PDB structures from a large number of parameters: name of the molecule, experimental method used to obtain the structure, resolution, etc.

Another family of web servers' offers additional structural information that is not directly contained in the PDB files: protein-protein [2] and protein-ligand [3] interactions information, structural and domain cluster families [4, 5], crystallographic ligand-bound waters [6], etc. These servers make publicly available structural related information obtained with research studies on their own. The majority of these tools are designed to give a graphical view of structural information through a user-friendly web page. They provide very useful data, but they are usually not designed for the high-throughput (HT) information extraction. HT studies requires fast access to stored information data instead of graphical representations, and needs for the possibility to exploit different information that is split through different web servers or databases, making the access to this data not efficient.

The goal of the project was, then, to design and implement a set of in-house relational databases storing structural data. The data stored can be divided in three main parts:

- **PDB derived metadata:** PDB code, resolution, residue/nucleotide sequence, experimental type, and all the information that can be directly extracted from the PDB files.
- **Structure important parts:** Information about *parts* of a protein/nucleic acid that could be of interest for a posterior analysis. Structure *parts* are defined as proteic/nucleic, non-proteic (ligands) and water molecules. Size, composition and atomic interaction between them are examples of information stored.
- **Protein active sites:** Groups of residues possibly involved in ligand recognition for every protein in the PDB. Clustering of these active centers has been also computed, to ease the extraction of sets of proteins with similar ligand recognition sites.

The first part of the project involved the development of a PDB mirror site, to have a local copy of the complete PDB database. That allowed us to efficiently work (open, read, search) with PDB files, as they will be in our local storage disks. It is a

crucial point if working with HT studies, as the file accession time is usually a bottleneck. To do that, we reproduced the scripts available at the original RCSB PDB web portal to obtain a local copy of the whole database. This set of scripts is executed automatically by the system every week, updating the flat files.

To have an efficient and direct access to PDB metadata, we designed a relational database for storage of (in our point of view) the most important information related to the PDB structure files. This relational database is also automatically updated each time the PDB mirror scripts are executed.

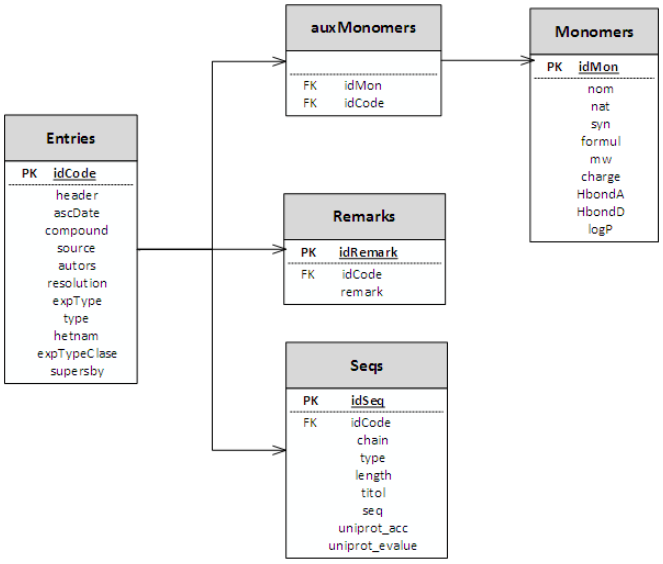


Fig.5.1.- PDB derived metadata database tables.

PDB derived metadata database contains basically 5 interconnected tables (Fig. 5.1). *Entries* table stores information related to PDB structure such as PDB code, file header, date of deposition, source organism, resolution, etc.; *Monomers* table contains information about ligands found in the structures; *auxMonomers* links tables entries and monomers, that is, contains information about which ligand/s has/have been solved together with a given PDB structure; table *Remarks* contain all the REMARK tags information included in the PDB files, such as missing residues, references, data collection details, etc.; and finally table *Seqs* contains information about the aminoacidic/nucleotidic sequences associated to the PDB chains.

Similarly, information regarding non-redundant clusters of PDB structures is stored in a table named *ClusterRef*. These clusters are built using programs *blastclust* [7] and *cd-hit* [8], joining together in a single cluster family protein chains sharing a given sequence similarity, choosing one of them as a representative. Program and sequence similarity thresholds identify a specific cluster (as defined in the RCSB PDB):

- **bc-30 / bc-40 / bc-50 / bc-70 / bc-90 / bc-95 / bc-100**: computed with *blastclust*, sharing 30, 40, 50, 70, 90, 95 or 100 sequence identity respectively.

- **clusters50 / clusters70 / clusters90 / clusters95**: computed with *cd-hit*, sharing 50, 70, 90 and 95 sequence identity respectively.

This is useful information, as the PDB database is known to be biased towards pharmaceutical interesting proteins, i.e. the same structure can be found with many different point mutations having different PDB entry codes. *ClusterRef* table joins together all of these structures assigning a representative one, thus removing redundancy.

Protein and nucleic acids structures were split in pieces according to an internal definition: proteic (or nucleic), non-proteic (ligands) and water molecules. Information extracted from this division was stored in three new tables: *parts*, *atoms* and *contacts* (Fig. 5.2).

Each one of the PDB files was analyzed, extracting information for each of the parts: type (*A* for proteic/nucleic, *H* for non-proteic and *W* for water molecules), chain id, initial atom, final atom and molecular weight. While looking for the different parts forming the structure, contacts between them (atoms belonging to different parts being at a maximum distance of 5Å) were also computed and stored in the database.

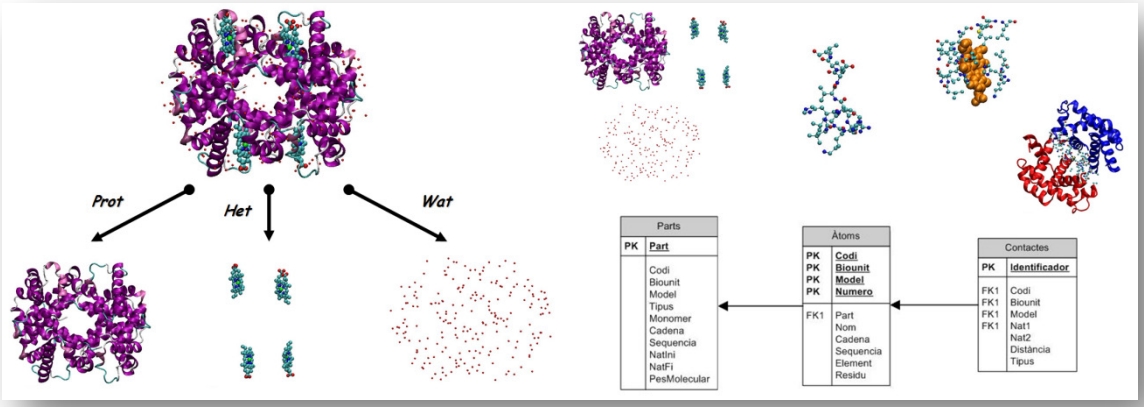


Fig.5.2.- Structure important parts schema and database tables. *Molecule structures are split into different parts forming them. Then, tables Parts, Contacts and Atoms store information about important structure regions and interactions between them.*

During the design of the database we decided to work with biological units (*biounits*) instead of with direct PDBs. These biological unit files are available at the RCSB PDB portal, and are included in the previously mentioned mirror. When working with biounits, we can have for each PDB:

- One copy of the PDB coordinates: PDB structure coincide with the biological assembly.

- Multiple copies of the PDB coordinates: PDB structure is just a portion of the biological assembly. Biounit file contains multiple copies of the PDB coordinates, with symmetry operations applied. These copies are stored as different models within the biounit file, so they must be distinguished from the NMR generated models.

- A portion of the PDB coordinates: PDB structure has multiple copies of the biological assembly, usually coming from crystallization conditions. Biounit file contains just a portion of the PDB coordinates, known or believed to be the functional form of the molecule. In this case, more than one biounit file is generated from a single PDB, identified by a sequential number (biounit 1, 2, 3, etc.).

Protein active sites, groups of residues possibly involved in ligand recognition for every protein in the PDB were also computed and stored in the database. Active centers were also clustered, to easily extract sets of proteins with similar ligand recognition sites.

Binding sites tables are built upon a set of drug-like ligands, chosen by human inspection, following the well-known *Lipinski's rule of 5* [9]. Reference active center is defined as a group of protein residues being 5Å away from a certain ligand. From a given reference active center, the cluster of similar ligand recognition sites is computed taking all the protein belonging to the correspondent protein cluster family (sharing at least 90% of sequence identity), and mapping the group of residues forming the reference active center. Thus, an active site cluster can be formed by different proteins sharing a similar ligand recognition site, and different ligands that can be identified by this site. Tables *ActiveCenterCluster, ActiveCenter, Ligand* and *Contact* store this information (Fig. 5.3).

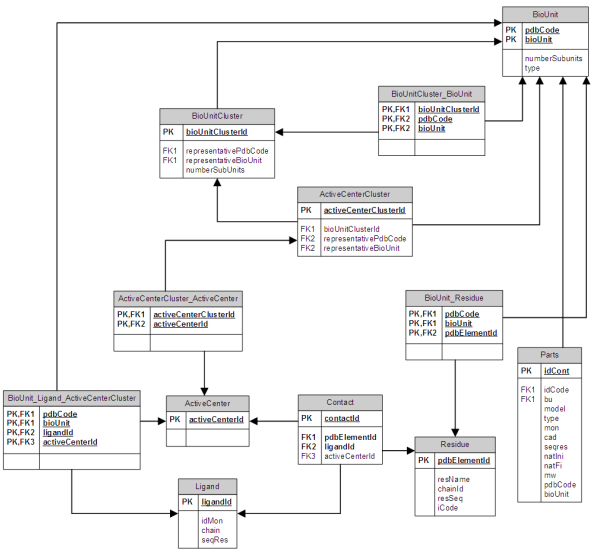


Fig.5.3.- Protein active sites database tables. *Tables ActiveCenterCluster, ActiveCenter, Ligand and Contact store information about similar ligand recognition sites.*

All of the different tables introduced in the previous sections were interconnected by relational primary and foreign keys, allowing complex queries involving all the information stored in the database. As a complement to this database infrastructure, we created a web server interface to easily obtain information and PDB datasets on-line, without having to deal with database commands (Fig. 5.4).

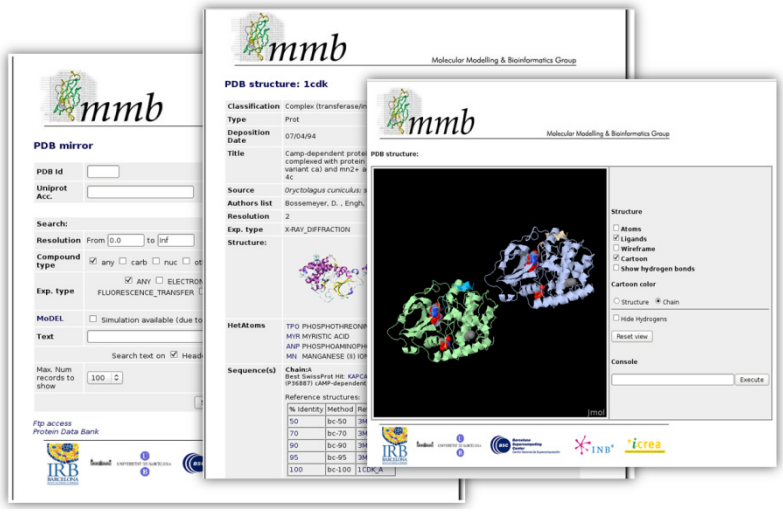


Fig 5.4.- MMB PDB mirror. *Web server for local PDB structure information retrieval and visualization.*

The MMB PDB mirror server offers:

- Multiple search possibilities: by PDB code, resolution, compound type, experimental type and keyword.
- Information shown: classification, type, deposition date, title, source, authors, resolution, experimental type, ligands (if any), and sequences by chain, with links to its clusters families (clusters 50, 70, 90, 95 and 100).
- Visualization: *Chime* and *Jmol* applet interactive visualization of the asymmetric unit and the biological functional unit.
- Downloading: files offered to download are the original PDB code and the biological unit (biounit).
- Links to other databases: links to external databases (RCSB and PDBSum) and to other internal databases (clusters, active sites, ligand information).

5.2 Generation of a protein dynamics library of MD trajectories.

5.2.1 Synopsis

Taking advantage of the HT framework presented in the past section, a large library of protein MD trajectories (MoDEL – Molecular Dynamics Extended Library) was generated. Within the MoDEL project 1,595 different protein structures were simulated. The set of proteins was selected to maximize the PDB coverage, using cluster90 to remove redundancy.

The result of a massive HT effort is a large set of flexibility and MD trajectories of proteins, stored in a complex warehouse specially designed for the project. The warehouse consists in standard disk files maintained in its original formats, as well as a relational database to efficiently store and retrieve simulations metadata and flexibility information. The complete package is publicly accessible through a web server (http://mmb.irbbarcelona.org/MoDEL), where analyses are graphically presented with 2D and 3D plots, simulations can be visualized as videos or through the interactive Jmol application and trajectories can be downloaded in a compressed format.

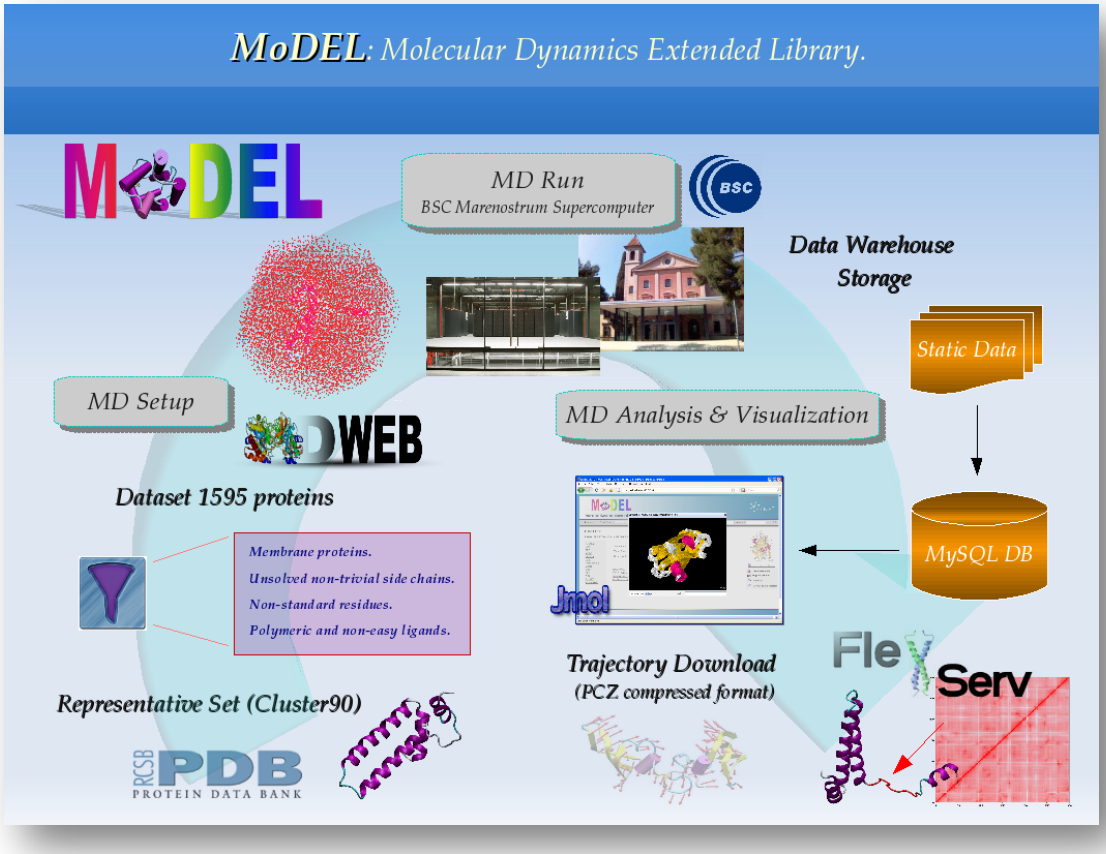
As one of the first authors of the paper reproduced in the next section, I was in charge of the dataset selection, database design and implementation, and I took part in the final analyses.

5.2.2 Paper 1

MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories.

Tim Meyer*, Marco D'Abramo*, **Adam Hospital***, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluís Gelpí, Modesto Orozco.
(* *These authors contributed equally to this work*)

Structure. (2010) 18(11), 1399-1409.



Structure
Ways & Means

MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories

Tim Meyer,[1,2,5] Marco D'Abramo,[1,5] Adam Hospital,[1,3,5] Manuel Rueda,[1] Carles Ferrer-Costa,[1] Alberto Pérez,[1,2] Oliver Carrillo,[1] Jordi Camps,[1,2,3] Carles Fenollosa,[1,3] Dmitry Repchevsky,[1,2,3] Josep Lluis Gelpí,[1,2,3,4] and Modesto Orozco[1,2,3,4,*]

[1]Joint IRB-BSC Computational Biology Programme, Institute of Research in Biomedicine, Parc Científic de Barcelona, Baldiri Reixac 10, Barcelona 08028, Spain
[2]Barcelona Supercomputing Center, Jordi Girona 31, Edifici Torre Girona. Barcelona 08034, Spain
[3]National Institute of Bioinformatics, Parc Científic de Barcelona, Baldiri Reixac 10, Barcelona 08028, Spain
[4]Departament de Bioquímica i Biología Molecular, Facultat de Biología, Avgda Diagonal 645, Barcelona 08028, Spain
[5]These authors contributed equally to this work
*Correspondence: modesto@mmb.pcb.ub.es
DOI 10.1016/j.str.2010.07.013

SUMMARY

More than 1700 trajectories of proteins representative of monomeric soluble structures in the protein data bank (PDB) have been obtained by means of state-of-the-art atomistic molecular dynamics simulations in near-physiological conditions. The trajectories and analyses are stored in a large data warehouse, which can be queried for dynamic information on proteins, including interactions. Here, we describe the project and the structure and contents of our database, and provide examples of how it can be used to describe the global flexibility properties of proteins. Basic analyses and trajectories stripped of solvent molecules at a reduced resolution level are available from our web server.

INTRODUCTION

Proteins are large and flexible molecules. Under physiological conditions, they adopt an ensemble of conformations. Flexibility patterns of proteins have been carefully refined by evolution to optimize functionality (Ma and Karplus, 1998; Kuhlman and Baker, 2000; Daniel et al., 2003; Qian et al., 2004; Leo-Macias et al., 2005; Karplus and Kuriyan, 2005; Henzler-Wildman et al., 2007; Goldstein, 2008; Yang et al., 2009). The similarity of the structural variation found in protein families with that spontaneously sampled during molecular dynamics simulations strongly suggests that protein evolution has used the intrinsic pattern of physical flexibility of proteins when designing new proteins (Leo-Macias et al., 2005; Velazquez-Muriel et al., 2009). In summary, protein evolution and function is difficult to understand if flexibility is ignored. This explains the intense efforts currently being made to obtain experimental descriptions of protein flexibility. However, despite encouraging advances (Lindorff-Larsen et al., 2005), we are far from achieving a full experimental analysis of proteome flexibility, and therefore theoretical approaches are necessary. In this respect, coarse-grained (CG) models coupled to ultrasimplified (pseudo) harmonic potentials have been widely used to obtain rough descriptions of the deformability of proteins (Tirion, 1996; Tozzini, 2005; Bahar and Rader, 2005; Yang et al., 2009; Rueda et al., 2007a; Emperador et al., 2008a); however, in general, the information derived is of low resolution and tends to overestimate the harmonic nature of equilibrium fluctuations. In principle, more accurate descriptions can be obtained from the use of atomistic molecular dynamics (MD), where atomic-resolution trajectories of proteins are derived from the application of Newton's equations of motion and physical potential energy functions (McCammon et al., 1977; Brooks et al., 1987). Unfortunately, the practical use of MD has been severely limited by its computational cost and by the problems encountered in the automatic setup of simulations. These limitations would explain why MD is traditionally used to study individual proteins.

During the last half of this decade, The development of new and more efficient simulation engines and the availability of state-of-the-art supercomputer (or GRID) platforms has led several laboratories to add a fourth dimension (time) to structural databases by running atomistic MD simulations on the deposited proteins (or at least in a selected set of highly representative structures). Of the many initiatives started, two have crystallized in extended databases: one in the US: Dynameomics (Beck et al., 2008; Simms et al., 2008; Kehl et al., 2008; Day et al., 2003) developed by Daggett's group, and another in Europe: MoDEL (Molecular Dynamics Extended Library), which we present here. These large platforms now offer structural biologists a unique tool to analyze the dynamics of proteins.

OVERVIEW OF THE MODEL PROJECT

The main objective of MoDEL is to provide information on the multinanosecond scale dynamics of proteins in near-physiological conditions. This information can then be used for many purposes, ranging from evolutionary studies to biophysical analysis and drug-design processes. In addition, MoDEL is an excellent reference set for calibration, refinement, and validation

Structure
MoDEL: Molecular Dynamics Extended Library

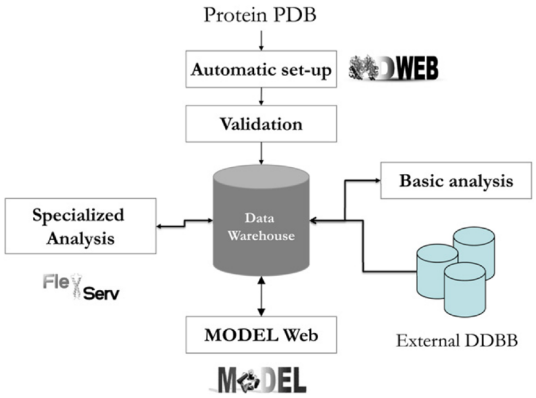


Figure 1. General Flowchart of the MoDEL Platform
The automatic setup tools prepare and run a trajectory from the structure in PDB format. Before storing the results, the trajectory is validated and later analyzed with our analysis tools. MODEL data are available through our public MODEL web server at http://mmb.pcb.ub.es/MoDEL.

of coarse-grained methods of flexibility (Rueda et al., 2007a; Emperador et al., 2008a) and for the benchmarking of force fields, computer programs, and simulation procedures (Rueda et al., 2007a). MoDEL is an ongoing project whose maintenance and extension is one of the main commitments of our group.

MoDEL (Molecular Dynamics Extended Library) is an acronym that defines a complex infrastructure of software and databases that we have developed over several years (Figure 1). It is divided into the following five main blocks: (1) tools for the automatic setup of MD simulations; (2) tools for validation of trajectories and error detection; (3) data warehouse, comprising a relational database and the underlying trajectories database; (4) tools for basic and advanced analysis; and (5) web server and related web applications. All tools have been built using in-house software combined with external software modules (see Table S1 available online) organized and integrated through a software platform. System preparation, simulation, and analysis modules are also available as web services following the framework of the Spanish National Institute of Bioinformatics (Biomoby, BioMoby Consortium, 2008 [www.inab.org]). The modular nature of the software allows combining all operations in fully automated and highly configurable workflows, thereby minimizing human intervention and facilitating maintenance and update. Also, the web services platform allows the integration with the wide offer of bioinformatics services in the community. Raw data are maintained in their original format in order to maximize compatibility with the software designed by third parties. The MoDEL platform is linked directly to a battery of tools for "in-depth" analysis of trajectories and to our FlexServ platform, (http://mmb.pcb.ub.es/FlexServ) (Camps et al., 2009), which includes a variety of flexibility analyses from MD ensembles as well as from a variety of CG representations using either normal modes, Brownian Go-like dynamics or Discrete Molecular Dynamics (dMD) (Rueda et al., 2007a; Emperador et al., 2008a).

Simulations in MoDEL are labeled internally following four criteria: (1) simulated structure; (2) length of the trajectory; (3) force field; and (4) solvent environment. Only cytoplasmatic monomeric proteins selected by diversity criteria (see below) are currently available in the database, but extensions of the database to membrane proteins and specific protein families are now under way. At the time of writing this report, the MoDEL data warehouse contained more than 1700 protein trajectories, ranging from 10 ns (the shortest) to 1 μs (the longest). The raw trajectories collected represent nearly 18 Tb of data corresponding to around 250,000 residues, 4.5 million protein atoms, and around 19 million water molecules. The computational effort required for the derivation of MoDEL required massive use of the *MareNostrum* supercomputer at the Barcelona Supercomputing Center (www.bsc.es) and local platforms in our group, and took more than 4 years to reach its current completion state.

TARGET SELECTION

A number of reasonable protocols for the selection of target proteins have been proposed (Day et al., 2003, Ng et al., 2006). Here, we adopted a very simple diversity approach intended to select nonhomologous proteins covering the largest possible portion of the PDB. The starting point was the release of the PDB in October 2005 (Berman et al., 2000), from which we selected Cluster-90 proteins (i.e., we considered in the following only those proteins with less than 90% sequence identity with other proteins selected for simulation). From this reduced list we then removed the following: (1) all membrane proteins; (2) proteins with gaps in the structure; (3) nonmonomeric proteins (on the basis of biological assembly definitions found in PDB, Krissinel and Henrick, 2007); (4) proteins with nonstandard residues (except Se-Met); and (5) proteins containing polymeric or nonconstitutive ligands difficult to parameterize by automatic procedures (see below). This screening produced a final list of 1595 proteins, which then entered the simulation workflow (see Figure 1). Trajectories that failed standard quality checks (see below) were manually analyzed for potential errors in setup and then either repeated or, if no technical errors were found, labeled as potentially artifactual, on the basis of either local or global criteria. A number of replicates for several proteins (typically corresponding to different simulation times or force fields; see below) were obtained, thus yielding a total of 1875 trajectories, which were then submitted to the analysis workflows and stored in the MoDEL data warehouse. The proteins selected contained from one to four domains and ranged in size from 19 to 994 residues (a distribution plot of protein sizes is shown as Figure S1). A small subset of MoDEL with 30 representative proteins (Day et al., 2003) was created for benchmarking and exploratory studies (this subset is referred to as μMoDEL in the rest of the paper). Additional benchmark and validation was done considering five selected proteins: 1cqy, 1kte, and 1opc as representatives of the three CATH major classes, and two proteins for which very large amount of experimental information on flexibility is available: 1ubq and 2gb1; this ultrasmall set is named nMoDEL in the rest of the paper and was again used for validation purposes. A complete list of proteins (and PDB codes) in the μMODEL and nMODEL sets is shown in Table S2.

Structure
MoDEL: Molecular Dynamics Extended Library

FORCE-FIELD SELECTION

The selection of the force field is a crucial issue in any MD project and there is no clear indication as to which of the many available force fields is the best for protein analysis. Polarizable force fields are promising tools for a careful description of interactions in the future, but they have not been extensively tested to date and they slow down simulations quite significantly. Thus, researchers use standard nonpolarizable force fields. Force fields are in continuous evolution; however, at the time the project was started the following four force fields were the most popular: OPLS-AA (Jorgensen et al., 1996), GROMOS-96 (Hermans et al., 1984; Ott and Meyer, 1996) CHARMM-98 (MacKerell et al., 1995, 1998) and AMBER parm99 (Cornell et al., 1995). Before launching all MoDEL simulations, we evaluated the performance of these four force fields in the μMODEL subset (Rueda et al., 2007b). The data collected demonstrate that these force fields yield similar trajectories, which provide a good reproduction of the structural and dynamical data experimentally available at that time, including residual dipolar coupling (RDC) and order parameter (S^2) measures for selected proteins (Rueda et al., 2007b). Additional calculations on the μMODEL set performed with more recent force fields (parm2003 and parm99sb) confirmed that there is a reasonable consensus between force fields for trajectories started from native structures. This observation suggests that for the time length considered in our project, the considered force fields should provide similar results. Calculations on the entire MoDEL set were then performed using the complementary AMBER parm99 and GAFF force fields, for ease of ligand parameterization. For coherence with parm99 the popular TIP3P model (Jorgensen et al., 1983) was used to represent water molecules. Future revisions of MoDEL will incorporate results obtained with newly developed force fields and local refinements of existing ones. The reader is referred to Rueda et al. (2007b) for detailed discussion on the performance of MD simulations with different force fields.

SIMULATION SETUP AND TRAJECTORY PRODUCTION

One of the biggest challenges in the project was to define robust, flexible, and automatic procedures for the high-throughput setup of MD simulations. The process should be fast and flexible, mimicking the human-based process of preparing and launching a simulation. The refined setup process is detailed in the Supplemental Experimental Procedures section. It was based on a modular and highly flexible workflow structure that could be easily adapted to user requirements. The pipeline allows the user to launch the simulation at the end of the process, by distinct MD codes (at present time: AMBER [Case et al., 2004], NAMD [Phillips et al., 2005], and GROMACS [Hess et al., 2008]). In addition, an independent web application (MDWeb; A.H., M.O., J.L.G., unpublished data) that includes all functionalities has been developed as a side product of the MoDEL project to help in the automatic (but flexible) setup of MD simulations for nonexpert users.

MD simulations were produced in the isothermal-isobaric ensemble (T = 300K, p = 1 atm). Trajectories for the entire MoDEL solution data set were extended for 10 ns (after equilibration). The 30 protein μMoDEL data set was extended to 0.1 μs and up to 1 μs for the nMoDEL subset. These long simulations were used for benchmarking purposes and to check the validity of the 10 ns trajectories to represent the local dynamics of proteins around native structures (see below). Additionally, gas phase simulations in the isothermal ensemble (T = 300 K) were performed (0.1 μs long for the μMoDEL subset; and 1 μs long for the nMoDEL subset). Detailed simulation settings are included in the Supplemental Experimental Procedures section.

TRAJECTORY CONTROL

MD simulations are numerical simulations based on a large series of simplifications that can generate nonnegligible uncertainties in the results. Errors are expected to increase as a result of the automatic setup procedure required in high-throughput (HT) production, which implies that careful and critical checking of trajectories is needed. In our experience, the main sources of errors in simulations are related to the following: (1) incorrect decisions during the setup, particularly wrong ionic states, poorly placed solvent, or wrong description of the ligand; (2) errors in the equilibration and heating procedure; (3) technical problems along equilibrated trajectory (problems with SHAKE, extreme velocities, thermal coupling, etc.); and (4) force-field problems. Deviations of trajectories from experimental models might also arise for other reasons, such as local uncertainties in the experimental models, and varying environmental conditions in the simulation and in the experiment (for example: different pH, different ionic strength or protein concentration). Inspection of trajectories allows us to recognize errors derived from technical factors (setup/equilibration/heating/integration/coupling). However, it is not so easy to determine between deviation caused by force-field problems and that caused by other factors (experimental uncertainties, discrepancies between simulated and experimental conditions, etc.). Thus, our strategy was to scan trajectories for anomalous behavior using simple metrics (see Table S3). This was achieved by inspection of trajectories to identify anomalies caused by technical issues (that can typically be corrected) and those that may arise because of nontechnical reasons. In the first case (35 trajectories in total), simulations were repeated and when the anomalous behavior persisted they were removed from the database, while in the second approach, simulations were labeled as "anomalous" but were maintained in the database since these trajectories can be of interest to some users, and are relevant, for example, in force-field validation and in the discussion of potential local uncertainties in experimental structural models.

Thus, all trajectories were analyzed for global descriptors (see Supplemental Experimental Procedures and Table S3), such as the absolute and relative rmsd, the TM-score$_{rmsd}$ (Zhang and Skolnick, 2004) the radii of gyration and solvent accessible surface (SAS). They were also analyzed for local descriptors, the number of native contacts, and the secondary structure (see Table S3). Trajectories were analyzed after the first nanosecond to check for technical problems in the setup (these usually lead to anomalous diffusion or velocities in protein, ligand, or solvent), which were rare and were easy to correct in most cases. At the end of the simulation, quality analysis was

Cell PRESS

Structure

MoDEL: Molecular Dynamics Extended Library

repeated and a trajectory was labeled "suspicious" in one of three categories on the basis of the checklist and thresholds shown in Table S3: (1) potential errors in local structure; (2) potential errors in global structure; and (3) potential errors in both local and global structure. Less than 3% of trajectories in MoDEL display one or several warnings, which the user should not ignore.

ANALYSIS WORKFLOW

The mining of 18 Tb of raw data is complex and requires automation of analytical tools and further incorporation of results in a relational database (see below). Two types of calculations can be done on raw trajectories: (1) general/basic analysis, which can be performed without previous knowledge of user requirements; and (2) specialized analysis, which requires user specifications and often the development of specific software. The modular nature of the analysis workflow allows the integration of any kind of analysis (for an explanation of commonly used descriptors, see Supplemental Experimental Procedures). Basic analysis includes information on global and local structure, such as rmsd, TM-score$_{rmsd}$ (Zhang and Skolnick, 2004), radius of gyration, total and partial SASAs, collision cross sections, native contacts, secondary structure, and hydrogen-bond pattern. Dynamic descriptors determined by default include fluctuations in all structural values, B factors, Lindemann's indexes (Zhou et al., 1999), frequencies (derived from diagonalization of the mass-weighted covariance matrix), entropies (Schlitter, 1993; Andricioaei and Karplus, 2001; Harris et al., 2001) and all the information derived from principal component analysis (PCA) as described in essential dynamics framework (ED; Amadei et al., 1993; Orozco et al., 2003, Noy et al., 2006) (for detailed information, see Supplemental Experimental Procedures). All analyses were done with a battery of in-house codes and external analytical tools (see Table S1), which were organized in modular workflows, thereby allowing the incorporation of additional analytical tools to the pipeline.

Specialized modules for the data mining of trajectories are in constant evolution in the group and currently include routines for the analysis of the following: solvent environment (structure and dynamics of water shells); fitting of MD simulations to mesoscopic models of motion, determining hinge points and correlated motions (Camps et al., 2009); finding cavities and escape channels in protein ensembles based on ensemble Brownian dynamics (Carrillo and Orozco, 2008); ensemble docking tools (Gelpí et al., 2001); methods for the prediction of potential protein-protein interaction sites (Fernández-Recio et al., 2005); and many others.

STRUCTURE OF THE MODEL DATA WAREHOUSE AND MANAGEMENT SOFTWARE

The data management of MoDEL involves the handling of a large number of structures, linkage to publicly available databases, accessing a wide repertoire of analyses for each simulation, and storage of the trajectories in a way that facilitates efficient analysis. Although valid attempts to fully integrate this complex set of data have been reported (Berrar et al., 2005; Simms et al., 2008), the MoDEL data warehouse (see Figure 2A) has been designed using a conservative approach in order to be fully compatible with available software. MoDEL combines the following two approaches: (1) a central relational database and (2) a disk-based raw data repository. The former stores structures, simulation details, analytical results, and references to bioinformatics databases, while the latter stores the trajectories in both AMBER (native trajectory formats for other programs are also supported) and compressed PCZ formats, as well as advanced analytical data. The relational database is designed not only to show the data available but to query for additional analysis or simulations. The relational database powers the MoDEL web server, which acts as an interface for access to the analyses. The file system layout of the repository is designed to maximize the efficiency of data retrieval, exploiting hardware parallelism on access to data when possible.

The relational database comprises four main sections (Figure 2A): structure selection, simulation, fragment selection, and analysis. Structure selection includes data for the simulated systems linked to the necessary sections of the PDB (Berman

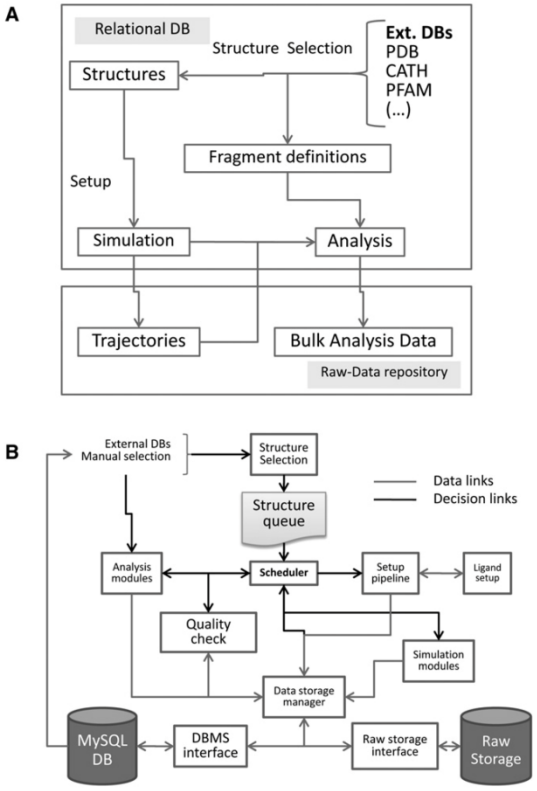


Figure 2. General Structure of the MoDEL Data Warehouse and Management Software

(A) General scheme of MoDEL data warehouse.

(B) Diagram of MoDEL management software.

See also Figures S2–S4, and Table S1.

Structure
MoDEL: Molecular Dynamics Extended Library

et al., 2000), CATH (Pearl et al., 2005), UniProtKb (The UniProt Consortium, 2010), and through the latter to other available databases (Table S1). Simulation details are stored in the Simulation section, which includes references to the software used, force fields and solvent, trajectory parameters, and quality-control data.

Trajectory analyses can be performed with a wide set of criteria, not necessarily known at the time of the design of the database, and storing them efficiently is not trivial. Analysis data are centered in the two last sections: fragment selection and analysis block. The central object for analysis storage (analysisSet) (see Figure S2) is the combination of simulation, the structure fragment analyzed, and the portion of the trajectory to be analyzed. This scheme allows us to store a wide variety of results from a simple collection of trajectory snapshots to a specific combination of analyses done over several parts of the trajectory or restricted to a specific domain. Again, structure fragments can be defined using a series of database data, like our in-house active sites database (A.H., M.O., J.L.G., unpublished data), domain (PFAM; Finn et al., 2008) or fold (CATH) (Pearl et al., 2005) (SCOP) (Murzin et al., 1995) databases, and also functional (Gene Ontology) (The Gene Ontology Consortium, 2000) data (Table S1). Setup and analysis software is adapted to extract that information from the database and perform new simulations and analyses on the basis of the desired criteria (see below). The MoDEL relational database is powered by MySQL 5.1 database manager. A complete Entity relationship schema of the database can be found in Figure S2.

The management software is a fully integrated platform (Figure 2B) with a highly modular core mostly written in PERL, combined with preexisting and third-party software (Table S1). To preserve compatibility with third-party software and eventually to allow the inclusion of new software packages, data are handled in well-known MD formats (amber native, and NetCDF, http://www.unidata.ucar.edu/software/netcdf/). Modules from the platform have been also wrapped to conform to the BioMoby web services framework (MDMoby, A.H., M.O., J.L.G., unpublished data). The central component of the MoDEL management software is the scheduler (Figure 2B). The scheduler module is fed by a queue of structures selected on the basis of a variety of criteria. It selects the operation to be performed, calling, in turn, structure setup, simulation, quality control, and analysis modules. The scheduler also takes care of checking the data warehouse to detect unfinished or faulty simulations or analyses and resuming the appropriate operations accordingly. Data from the different modules are handled by a common data manager module. The software platform is modular and multiarchitectural to take advantage of the computational infrastructure available (see Figure S3 for a description of the flow of data and the computer architectures involved). Data among the different hardware platforms are synchronized at the storage level and system calls are done through standard RPC technologies.

WEB-SERVER STRUCTURE

The MoDEL web server (http://mmb.pcb.ub.es/MoDEL) (see also Figure S4 for screenshots) is designed to allow access to the MoDEL project from several levels: to raw trajectory data for further in-house analysis, to simulation details, and to previously performed analyses. The server is organized into three sections. The first acts as an entry level and is intended for structure selection. The user can either browse the entire set or search for a specific structure. In addition, the database can be browsed following the CATH fold classification. The search criteria implemented include PDB and UniProt Ids, and keyword searches. It is also possible to search from nonstructural descriptors using a sequence comparison module, based on standard BLAST (Altschul et al., 1990) with settings selected to assure that only highly homologous structures are obtained. Using Blast-based sequence comparison with a limit E-value of 10^{-5}, our website currently provides access to simulations covering around 40% of PDB structures, 8% of UniProtKB sequences, 29% of Human UniProtKB sequences and 33% of DrugBank (Wishart et al., 2006) targets.

Once a structure is selected, the system offers a list of available simulations. Simulations can be downloaded, sent to additional tools either open like FlexServ (Camps et al., 2009), or restricted like MDWeb (Hospital et al., to be published), MDGRID (Carrillo and Orozco, 2008), CMIP (Gelpí et al., 2001), to other programs for further analysis, or instead, data previously analyzed can be retrieved. The web also provides videos and 3D animations of the trajectories for visual analysis and projections on the first five principal components to check the nature of the major deformation movements. All the analysis data (see above) are presented as table values, 1D and 2D plots and 3D data using a Jmol applet (http://www.jmol.org). The MoDEL web server is powered by a Jboss application server and is linked to an appropriate database manager and software (see above).

COMPRESSION AND TRANSFER OF DATA

The management and transfer of data included in the relational database do not need specific software infrastructure, while the access, storage, management and transfer of raw trajectories are (due the amount of the data) complex problems. The original trajectories with all solvent molecules and atomistic details require storage, but most analyses are done by taking intermediate files created by removing solvent molecules. Dry trajectories are compressed to obtain smaller files that can be transferred with high efficiency through the internet. The compression is done using our PCAzip technology (Meyer et al., 2006), which is based on three main steps: (1) principal component analysis of the original trajectory; (2) determination of the reduced set of eigenvectors explaining a given variance threshold (90% by default in MoDEL); and (3) projection of the original Cartesian coordinates into the essential eigenvector space. PCAZip splits the original trajectory into two components: the essential eigenvectors and their projections onto the trajectory. This results in a 5- to 10-fold compression of the Cartesian data since a reduced number of eigenvectors is enough to represent a large percentage of variance (Meyer et al., 2006). Note that the compression procedure does not require the assumption of harmonicity in the trajectory and that the original data can be recovered (with the desired accuracy) by simple back-projection to the Cartesian space (Meyer et al., 2006). MoDEL offers (through its webpage, see above) the possibility to download compressed files (90% variance accuracy for heavy atoms). As described elsewhere (Meyer et al., 2006),

Cell PRESS

Structure

MoDEL: Molecular Dynamics Extended Library

compressed files at 90% accuracy provide results that are, for many purposes, indistinguishable from original trajectories (few tenths of Å in most cases from real structures). The largest deviations appear for proteins displaying conformational changes along the trajectory, where a large percentage of variance is then explained by a single mode. The PCAZip program required for compression/decompression can be downloaded from our website http://mmb.pcb.ub.es/software/pcasuite, both as source code or precompiled executables.

RELIABILITY OF MD SIMULATIONS

A first point of concern in our project was the validation of the MD trajectories deposited in our database. This was done in three stages: (1) convergence in force fields; (2) convergence in simulation time; and (3) similarity between MD results and those derived from the experimental structural model. The first point has been checked in a previous paper (Rueda et al., 2007b), which found that the AMBER-parm99 force field appears to show sufficient reliability for the time window considered in MoDEL (see discussion above). Concerns on the time convergence of trajectories were addressed by comparing simulations on 10, 100, and 500 ns trajectories for a reduced number of highly representative proteins (see above). The results summarized in Figure 3A demonstrate the good agreement between the structures sampled during 10 and 100 ns trajectories for the μMoDEL subset both in local and global terms (the same is found for 500 ns trajectories in nMoDEL). Interestingly, not only structural descriptors but also parameters informative on protein flexibility (such as intramolecular entropy) are very similar in short and long trajectories (Figure 3A). This observation confirms that although 10 ns is too short for full protein relaxation, it is long enough to obtain a reasonable representation of the dynamics of proteins around their equilibrium conformation, even in cases of relatively large proteins (see data for GTPase activation protein [1gnd; a protein with 447 residues], in Figure S5 and also in Figure 3A). Finally, given that the typical relaxation times of waters are in the picosecond range (the slowest interchanging waters found have residence times <5 ns), MoDEL simulations should provide a complete sampling of the equilibrium solvent atmosphere around proteins.

Our final concern before accepting the utility of MD simulations was the capacity of trajectories in MoDEL to reproduce the known experimental behavior of proteins. Analysis on a reduced set of proteins (Rueda et al., 2007b) suggested that parm99 simulations provide reasonable approaches to structural models derived from NMR and X-ray data, to B factor profiles, and, when available, to direct NMR dynamic data (see above). The results in Figure 3B, obtained from a large set of proteins, confirm our previous claims and demonstrate that MD simulations accurately reproduce global structural descriptors of proteins, such as the solvent accessible surface area or the radii of gyration. Rmsd between simulated and experimental models are in 80% of cases below <3 Å, which is not far from the range of uncertainty expected from the normal structural variation found for proteins in water at room temperature. Furthermore, most deviations between MD ensembles and data obtained from experimental models are located in loops (where greater flexibility and larger uncertainties caused by lattice effects are expected in the experimental models), as noted in the low values of TM scores (100% simulations show TM scores <3 Å; see Figure 3B). Very encouraging, not only is global structure well preserved but local geometry is also maintained, as noted for example in conservation above 90% in the native contacts for around three-quarters of the database and the small losses of secondary structure (for additional discussion on the quality of MD simulations, see Rueda et al., 2007b).

In summary, although caution is always necessary when analyzing MD results, we are quite confident that the MD trajectories stored in the MoDEL database provide a reasonable approximation of the equilibrium conformational ensemble of proteins.

EXAMPLES OF MODEL DATA MINING FACILITIES

The MoDEL database allows a powerful analysis of average and time-dependent (in the multinanosecond scale) properties of proteins and their solvent environment at various levels of resolution (trace, backbone, heavy atoms, and all atoms) and considering the entire system or parts of it. All the analyses can be crossed with internal data in MoDEL or information in other databases that are linked to it. These features thus allow us, for example, to perform a given analysis restricted to a family in CATH or SCOP, to a given domain in PFAM, to structures with some functional annotation in Swissprot or TrEMBL (http://www.uniprot.org), or to protein families with a specific annotation or specific characteristics in the PDB. As noted above, the MoDEL web server gives access to some general analyses, but the MoDEL data warehouse is accessible for many additional ones, which might require specific input from the user. It is not our purpose here to describe the full proteome dynamics; however, below we give a few examples to illustrate the type of information that can be retrieved from our database. A detailed analysis of dynamic information on proteins that can be extracted from MoDEL will be described elsewhere.

Family-Specific Analysis of Protein Dynamics

The MoDEL relational database allows us to analyze family-dependent structural and flexibility properties, using a wide and flexible definition of the concept "family." This is efficiently done by querying the database against an internal or external descriptor. For example, the data in Figure 4A show how MoDEL provides information on the relative flexibility (as measured by Lindemann's index) of equivalent thermophylic and mesophylic proteins. Global analysis reveals that thermophylic proteins display 90% of the global flexibility of mesophylic protein but that this global change in flexibility is not equally distributed throughout all the regions of the protein. Thus, the largest rigidification in thermophylic compared with mesophylic proteins is located in the backbone (especially in β sheets), while the flexibility of side chains (especially in α helices) is not reduced in the former compared with the latter. Another example of MoDEL data mining is shown in Figure S6, which demonstrate that (1) 40%–90% of the variance in this particular set of proteins can be explained by only five essential deformation movements; (2) no major differences are found in the complexity of the flexibility space when considering distinct CATH families; and (3) large proteins do not necessarily have a more complex flexibility

Structure
MoDEL: Molecular Dynamics Extended Library

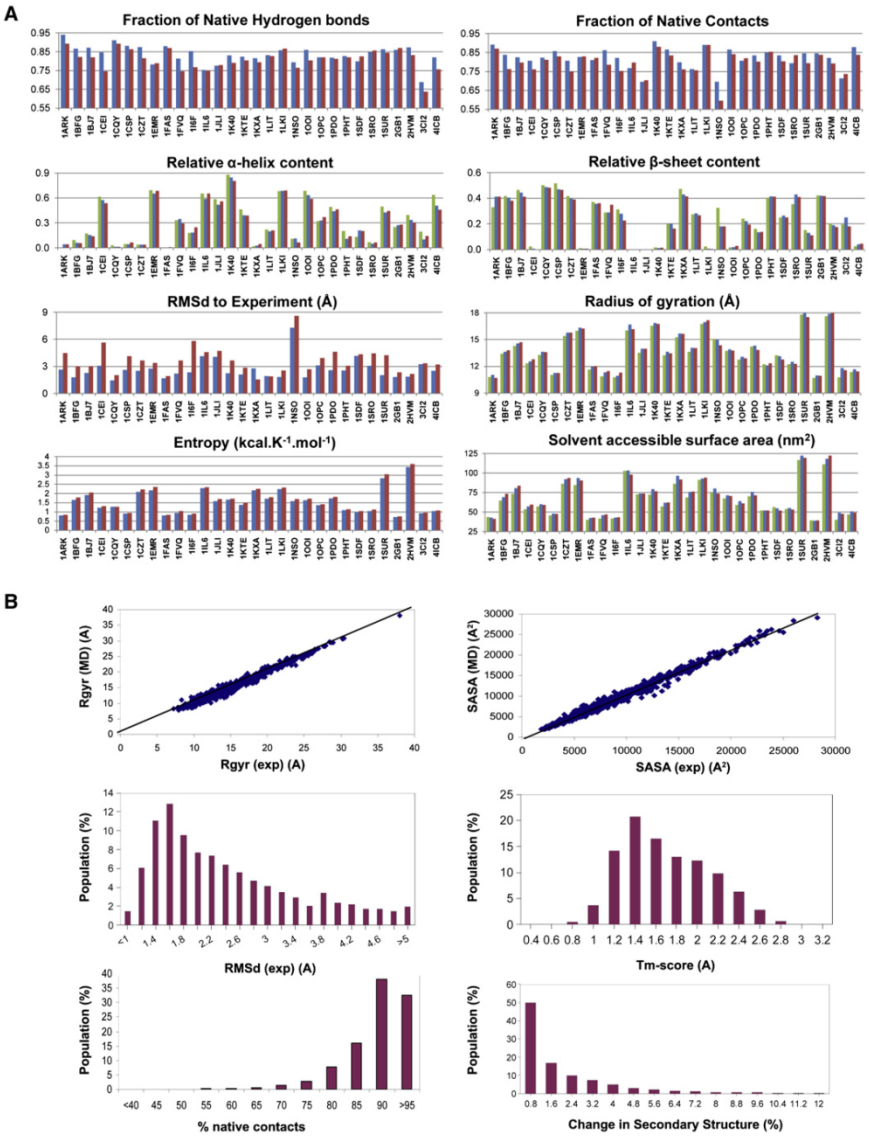


Figure 3. Quality of Simulations in MoDEL

(A) Different average descriptors for MD simulations in the μMoDEL subset. Blue: 10 ns trajectories, red: 100 ns trajectories, green: experimental data. Content in secondary structure is referred to unity.

(B) Comparison of structural parameters obtained from MD simulations and from experimental models (see text for details). We consider no change in the secondary structure when the secondary structure element of the starting structure is still very represented (at least for 0.8 ns) in the last nanosecond of simulation. The Rgyr(exp) and SAS(exp) are calculated using the experimental coordinates as found in the PDB.

See also Figure S5, and Table S4.

space that small ones, thereby indicating that variance in large proteins is often organized around a limited number of well-defined massive deformations (for example, large loop oscillations or rotations around hinge points).

Analysis of the Essential Deformability of Proteins

The MoDEL database has precomputed the essential dynamics (ED) of proteins, which facilitates the study of protein flexibility by reducing the complexity of the deformability space (Amadei

Structure
MoDEL: Molecular Dynamics Extended Library

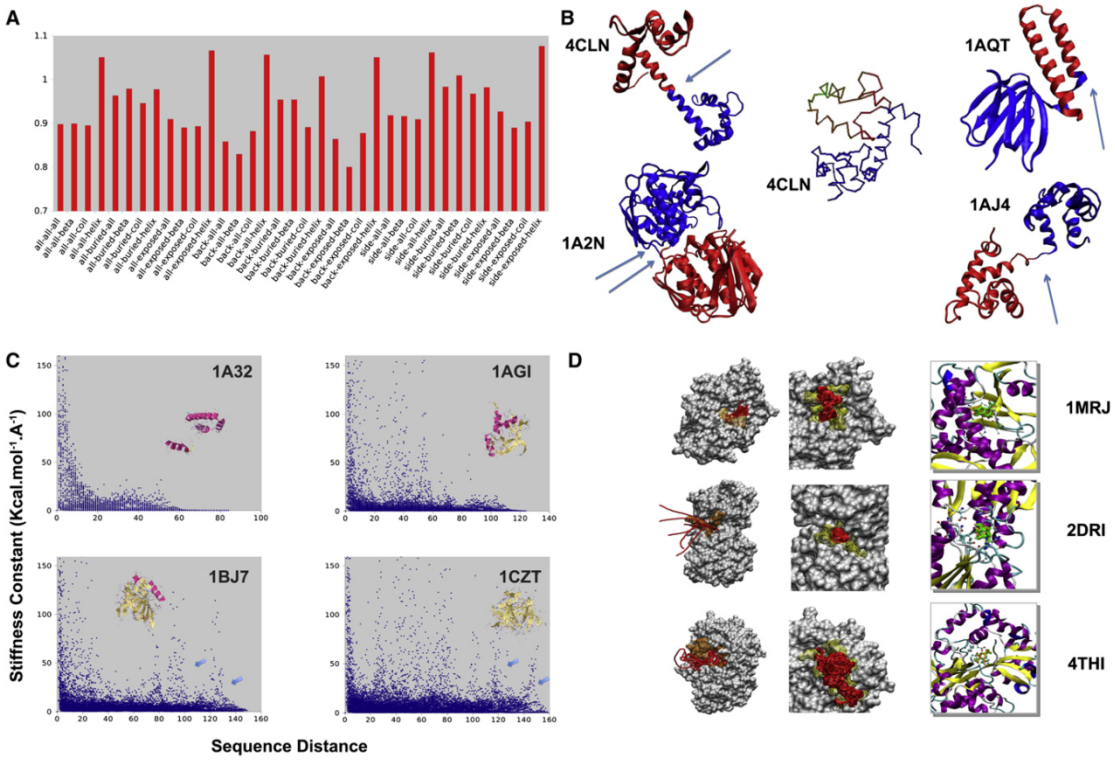


Figure 4. Examples of Data Mining in MoDEL

(A) Relative Lindemann's indexes between protein heavy atoms in thermophylic and mesophylic proteins (see Supplemental Experimental Procedures). To gain extra information, the index is computed for different groups of atoms. The nomenclature XYZ in x axis refers to X: side chains/backbones/all, Y: exposed/buried/all and Z: α helix/β sheet/coil/all. The number of thermophylic proteins is 30; the remaining proteins present in MoDEL are mesophylic.

(B) Examples of dynamics domain definition and hinge-point location, using Lavery's dynamic method, see http://mmb.pcb.ub.es/FlexServ, for four proteins (each dynamic domain is colored differently). Central plot corresponds to the pathway of correlated movements in a protein perturbed at one random residue (color code ranges from green r = 1 to red r = 0.5; blue means no correlation). The search for correlated motions was done with a width of three residues and a depth of four iterations (see Supplemental Experimental Procedures and the FlexServ help (http://mmb.pcb.ub.es/FlexServ) for additional details).

(C) Apparent C_α-C_α stiffness constants for four proteins with increasing percentage of β sheet (from left-top to right-bottom). The significant decay of stiffness constants with increased sequence distance is clear, indicating the local (in sequence) nature of interresidue contacts. However, the presence of long-range effects that lead to important contacts between distant (in sequence) residues is clear. The magnitude or remote interresidue contacts become especially clear in β sheet proteins, where the secondary structure forces H-bond-mediated contacts between distant residues. Some of these remote contacts are marked with arrows in the figure.

(D) Results of using MDGrid and CMIP docking on MoDEL ensembles for three randomly selected diverse proteins: (1MRJ) Ribosome-inactivating protein in complex with Adenosine (ADN); (2DRI) Sugar transport protein in complex with Ribose (RIB), and (4THI) Transferase, Thiaminase I in complex with 2,5-dimethyl-pyrimidin-4-ylamine (PYD). Plots in the first column show channel as red tubes, with the corresponding cavity in orange (only 1 of every 10 routes computed are displayed for clarity). Second column shows drugability measures performed considering true ligands as probes, "drug cavities" are shown in yellow and "hot spots" (regions accumulating 90% of the population of the drug center of mass) are shown in red. The third column shows CMIP best-scored docking poses (green ligands) with a reference to the known crystal structure (orange ligand), where relevant residues at the binding site are displayed with CPK representation.

See also Figures S6 and S7 and Table S5 for additional examples.

et al., 1993; Orozco et al., 2003, Meyer et al., 2006; Noy et al., 2006). Following the ED formalism, after diagonalization of the MD covariance matrix, a set of eigenvectors and another of eigenvalues are obtained, the first gives information on the nature of essential deformation movements, while the second informs on the variance associated with each of these movements. The eigenvectors/eigenvalues can be manipulated in many ways, from simple visualization to complex comparison metrics. Access to external analysis tools, such as PCAzip (http://mmb.pcb.ub.es/software/pcasuite) or FlexServ (http://mmb.pcb.ub.es/FlexServ) (Camps et al., 2009), allows interesting additional analysis, such as the determination of the degree of anharmonicity in the MD simulation, (determined by comparison of ED eigenvectors and those derived from

Structure
MoDEL: Molecular Dynamics Extended Library

diagonalization of a Hessian matrix defined by a simple residue-residue harmonic potential (elastic network model description)). It is also possible, for example, to compare the similarity between the deformability pattern of a set of related proteins, or to analyze the similarity between physical deformability (as defined by the MD-derived eigenvectors) and the evolutionary deformability derived from the analysis of the structural changes in protein families (see Velazquez-Muriel et al., 2009 for discussion). An example of the type of information derived from mining MoDEL with these tools is displayed in Table S5.

Advanced Analysis of Protein Flexibility

The MoDEL database is linked with advanced analysis tools implemented in FlexServ (http://mmb.pcb.ub.es/FlexServ) which allows a complete analysis of protein flexibility. Graphical examples in Figure 4B illustrate how trajectories in MoDEL allow the determination of hinge points, dynamics partition of domains and pathways of concerted motions (see Camps et al., 2009 for details). Several mesoscopic descriptors of protein deformability can be derived from these analyses, such as the apparent harmonic force-constants acting on the C_α of proteins with different relative content of α helix and β sheet (see Figure 4C). This type of information can be efficiently used to derive more realistic CG models of protein flexibility, of general or family-specific use (Emperador et al., 2008a; Rueda et al., 2007a; Camps et al., 2009; Emperador et al., 2008b). Many more analyses, like those described here, are possible through an intuitive interface, which provides the user with an accurate definition of the desired type of query or analysis.

Solvent Analysis

The MoDEL data warehouse contains structural and dynamic information on the solvent atmosphere around protein, which can also be subject to advanced analysis. For example, we can query our database to determine the number of water molecules in close contact with protein residues, to determine water residence times, diffusion properties, preferred solvation sites, and much more information that can also be determined for any given protein family or group of residues. As an example, Figure S7 summarizes some results obtained from the analysis of the first solvation shell around (sixty) representative proteins of CATH families 1 (α-) and 2 (β-). It was found that all the proteins considered here were well solvated with a typical water density around 0.07 to 0.08 waters/Å^2 (in SASA), which compares with a maximum theoretical density (around 0.1 water/Å^2 for ideally packed waters). Interestingly, our data show that β-proteins have more water molecules in their vicinity than α-proteins, even when the water population is corrected by the solvent accessible surface of the proteins (see Figure S7). This observation demonstrates that there is a quite sizeable amount of water around secondary β sheets, even they are traditionally considered hydrophobic structures. Note that analysis similar to that outlined here can be done considering not the entire bulk of solvent but only distinguished water molecules, for example, those placed in crystal positions or cavities, or those with very slow or fast interchange between first and second solvation shells. In other words, MoDEL allows a complete characterization of the solvent atmosphere around proteins.

Channel and Cavity Detection

Advanced analysis tools coupled to MoDEL allow the determination of channels and cavities taking the dynamics of the protein into account. It is therefore possible to detect channels or transient cavities, which are present only on small fractions of the trajectory and, accordingly, might not be detectable in the X-ray structure. The procedure is based on our MDGRID algorithm (Carrillo and Orozco, 2008), combined with the use of classical probe particles, which can be as generic as a "soft sphere" or as specific as a full drug. As explained in detail elsewhere (Carrillo and Orozco, 2008), MDGRID takes the snapshots collected along the trajectory, projects them in a common rectangular grid and precomputes the forces that the protein atoms will exert on basic particles (positive charge, negative charge, different van der Waals atoms, etc.) placed at the grid points. These forces are then Boltzmann-averaged and used to determine precomputed accelerations within a Brownian dynamics algorithm. Graphical examples of the type of information derived for a few proteins are provided in Figure 4D (first column). These examples clearly illustrate the power of the technique to trace not only the boundaries of the binding site but also the pathways for interchange of ligand with the environment. Note that since forces are precomputed MDGRID calculations are extremely fast (multimicrosecond long exploration of channels and cavities in a few minutes in a small desktop personal computer).

Drugability and Ligand Docking

The MDGRID protocol outlined above can be used with small changes to determine the "drugability" of a protein (i.e., the capacity of a protein to bind small molecules with drug-like properties). This type of calculation can be done by taking small drug-like molecules from our local molecular database, or alternatively by using known drugs for the targeted proteins. In the first case, the study provides a direct measure of protein drugability, while in the second case information is obtained on the ability of a protein to interact with a family of drug-like compounds. In both cases a secondary product is the definition of major binding sites in target proteins. Information is retrieved considering not static pictures of proteins but dynamic ensembles, which might make accessible cavities which are not visible in a single X-ray structure. Figure 4D (second column) contains a few examples of drugability plots for three randomly selected proteins known to bind small drug-like ligands, and illustrates how the method detects that both will bind ligands and locate the primary binding cavity.

For binding sites of known pharmacological targets the use of docking programs such as CMIP (Gelpí et al., 2001), can yield potential structures of drug-protein complexes (see some examples in Figure 4D, last column). These are obtained explicitly using the flexibility information on the protein contained in the original MD simulation.

FINAL REMARKS

Initiatives such as Dynameomics and MoDEL provide access to molecular dynamics data at the proteome level. Expert and nonexpert users can access trajectories and a variety of analyses that may be difficult to reach by other means, thus saving

Structure
MoDEL: Molecular Dynamics Extended Library

them months of work and computer time. Large MD databases provide a proteome-level view to the molecular physics of proteins, something that is impossible to achieve by other means. Furthermore, the databases and integrated analysis tools can be useful for both the benchmarking of force fields and the development of new CG methods. Last, but not least, the research effort devoted to performing and analyzing MD trajectories in the high-throughput regimen has generated an extended software platform that allows straightforward, automatic, and robust access to the technique, and to a variety of analysis tools. Initiatives like that presented here are a step forward in the popularization and rationalization of MD simulations, bringing the technique closer to meeting the new needs of the postgenomic era.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Experimental Procedures, eight figures, and four tables and can be found online at 10.1016/j.str.2010.07.013.

ACKNOWLEDGMENTS

MoDEL is a massive effort involving, directly or indirectly, a large part of the Molecular Modeling and Bioinformatics group at IRB Barcelona and the BSC. We are also indebted to Dr. Sergi Girona and the MareNostrum support team for making this project possible. Helpful comments from Prof. F. J. Luque and many colleagues at IRB Barcelona and the BSC are gratefully acknowledged. This work was supported by the Spanish Ministry of Science (CTQ2005-09365-C02-02, BIO2009-10964), INB-Genoma España, the Consolider E-science project, EU-ScalaLife project), the COMBIOMED RETICS project and the *Fundación Marcelino Botin*.

Received: January 23, 2010
Revised: July 19, 2010
Accepted: July 27, 2010
Published: November 9, 2010

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Amadei, A., Linssen, A.B., and Berendsen, H.J. (1993). Essential dynamics of proteins. Proteins *17*, 412–425.

Andricioaei, I., and Karplus, M. (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. J. Chem. Phys. *115*, 6289–6292.

Bahar, I., and Rader, A.J. (2005). Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol. *15*, 586–592.

Beck, D.A., Jonsson, A.L., Schaefer, R.D., Scott, K.A., Day, R., Toofanny, R.D., Alonso, D.O.V., and Daggett, V. (2008). Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. Protein Eng. Des. Sel. *21*, 2038–2050.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Berrar, D., Stahl, F., Silva, C., Rodrigues, J.R., Brito, R.M., and Dubitzky, W. (2005). Towards data warehousing and mining of protein unfolding simulation data. J. Clin. Monit. Comput. *19*, 307–317.

BioMoby Consortium. (2008). Interoperability with Moby 1.0—it's better than sharing your toothbrush! Brief. Bioinform. *9*, 220–231.

Brooks, C.L., III, Karplus, M., and Pettitt, B.M. (1987). Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics (Cambridge: Cambridge University Press).

Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Gelpí, J.L., and Orozco, M. (2009). FlexServ: an integrated tool for the analysis of protein flexibility. Bioinformatics *25*, 1709–1710.

Carrillo, O., and Orozco, M. (2008). GRID-MD—a tool for massive simulation of protein channels. Proteins *70*, 892–899.

Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., III, Ross, W.S., Simmerling, C.L., Darden, T.L., Marz, K.M., Stanton, R.V., Cheng, A.L., et al. (2004). AMBER 8 Computer Program (San Francisco: University of California).

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. *117*, 5179–5197.

Daniel, R.M., Dumm, R.V., Finney, J.L., and Smith, J.C. (2003). The role of dynamics in enzyme activity. Annu. Rev. Biophys. Biomol. Struct. *32*, 69–92.

Day, R., Beck, D.A.C., Armen, R.S., and Dagget, V. (2003). A consensus view of fold space: Combining SCOP, CATH and the Dali Domain Dictionary. Protein Sci. *12*, 2150–2160.

Emperador, A., Carrillo, O., Rueda, M., and Orozco, M. (2008a). Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. Biophys. J. *95*, 2127–2138.

Emperador, A., Meyer, T., and Orozco, M. (2008b). United-atom discrete molecular dynamics of proteins using physics-based potentials. J. Chem. Theory Comput. *4*, 2001–2010.

Fernández-Recio, J., Totrov, M., Skorodumov, C., and Abagyan, R. (2005). Optimal Docking Area: a new method for predicting protein-protein interaction sites. Proteins *58*, 134–143.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, J.S., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. (2008). The PFAM protein families databases. Nucleic Acids Research *36*, D281–D288.

Gelpí, J.L., Kalko, S.G., Barril, X., Cirera, J., de La Cruz, X., Luque, F.J., and Orozco, M. (2001). Classical molecular interaction potentials: improved setup procedure molecular dynamics simulations of proteins. Proteins *45*, 428–437.

Goldstein, R.A. (2008). The structure of protein evolution and the evolution of protein structure. Curr. Opin. Struct. Biol. *18*, 170–177.

Harris, S.A., Gavathiotis, E., Searle, M.S., Orozco, M., and Laughton, C.A. (2001). Cooperativity in drug-DNA recognition: a molecular dynamics study. J. Am. Chem. Soc. *123*, 12658–12663.

Henzler-Wildman, K.A., Lei, M., Thai, V., Kerns, S.J., Karplus, M., and Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature *450*, 913–916.

Hermans, J., Berendsen, H.J.C., Van Gunsteren, W.F., and Postma, J.P.M. (1984). A consistent empirical potential for water-protein interactions. Biopolymers *23*, 1513–1518.

Hess, B., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput. *4*, 435–447.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. *118*, 11225–11236.

Karplus, M., and Kuriyan, J. (2005). Molecular dynamics and protein function. Proc. Natl. Acad. Sci. USA *102*, 6679–6685.

Kehl, C., Simms, A.M., Toofanny, R.D., and Daggett, V. (2008). Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data. Protein Eng. Des. Sel. *21*, 379–386.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774–797.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA *97*, 10383–10388.

Structure
MoDEL: Molecular Dynamics Extended Library

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A.R. (2005). An analysis of core deformations in protein superfamilies. Biophys. J. *88*, 1291–1299.

Lindorff-Larsen, K., Best, R.B., Depristo, M.A., Dobson, C.M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. Nature *433*, 128–132.

Ma, J., and Karplus, M. (1998). The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. Proc. Natl. Acad. Sci. USA *95*, 8502–8507.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature *267*, 585–590.

MacKerell, A., Jr., Wiorkiewicz-Kuczera, J., and Karplus, M. (1995). An all-atom empirical energy function for the simulation of nucleic acids. J. Am. Chem. Soc. *117*, 11946–11975.

MacKerell, A.D., Jr., Bashford, D., Bellott, M., Dunbrack, R.L., Jr., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Meyer, T., Ferrer-Costa, C., Pérez, A., Rueda, A., Bidon-Chanal, A., Luque, F.J., Laughton, C.A., and Orozco, M. (2006). Essential dynamics: a tool for efficient trajectory compression and management. J. Chem. Theory Comput. *2*, 251–258.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Ng, M.H., Johnston, S., Wu, B., Murdock, S.E., Tai, K.H., Fangohr, H., Cox, S.J., Essex, J.W., Sansom, M.S.P., and Jeffreys, P. (2006). BioSimGrid: Grid-enabled biomolecular simulation data storage and analysis. Future Gener. Comput. Syst. *22*, 657–664.

Noy, A., Meyer, T., Rueda, M., Ferrer, C., Valencia, A., Perez, A., de la Cruz, X., Lopez-Bes, J.M., Pouplana, R., Fernández-Recio, J., et al. (2006). Data mining of molecular dynamics trajectories of nucleic acids. J. Biomol. Struct. Dyn. *23*, 447–456.

Orozco, M., Pérez, A., Noy, A., and Luque, F.J. (2003). Theoretical methods for the simulation of nucleic acids. Chem. Soc. Rev. *32*, 350–364.

Ott, K.H., and Meyer, B. (1996). Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. J. Comput. Chem. *17*, 1068–1084.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., et al. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. *33*, D247–D251.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Qian, B., Ortiz, A.R., and Baker, D. (2004). Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc. Natl. Acad. Sci. USA *101*, 15346–15351.

Rueda, M., Chacón, P., and Orozco, M. (2007a). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. Structure *15*, 565–575.

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., Gelpí, J.L., and Orozco, M. (2007b). A consensus view of protein dynamics. Proc. Natl. Acad. Sci. USA *104*, 796–801.

Schlitter, J. (1993). Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem. Phys. Lett. *215*, 617–621.

Simms, A.M., Toofanny, R.D., Kehl, C., Benson, N.C., and Daggett, V. (2008). Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Protein Eng. Des. Sel. *21*, 369–377.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

The UniProt Consortium. (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. *40*, D142–D148.

Tirion, M.M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. *77*, 1905–1908.

Tozzini, V. (2005). Coarse-grained models for proteins. Curr. Opin. Struct. Biol. *15*, 144–150.

Velazquez-Muriel, J.A., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., and Carazo, J.M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. BMC Struct. Biol. *17*, 6.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668–D672.

Yang, L.,, Song, G., and Jernigan, R.L. (2009). Protein elastic network models and the ranges of cooperativity. Proc. Natl. Acad. Sci. USA *106*, 12347–12352.

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.

Zhou, Y., Vitkup, D., and Karplus, M. (1999). Native proteins are surface-molten solids: application of the lindemann criterion for the solid *versus* liquid state. J. Mol. Biol. *285*, 1371–1375.

Structure *18*

Supplemental Information

MoDEL (Molecular Dynamics Extended Library):

A Database of Atomistic Molecular Dynamics

Trajectories

Tim Meyer, Marco D'Abramo, Adam Hospital, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluis Gelpí, and Modesto Orozco

Suppl. Figure S1

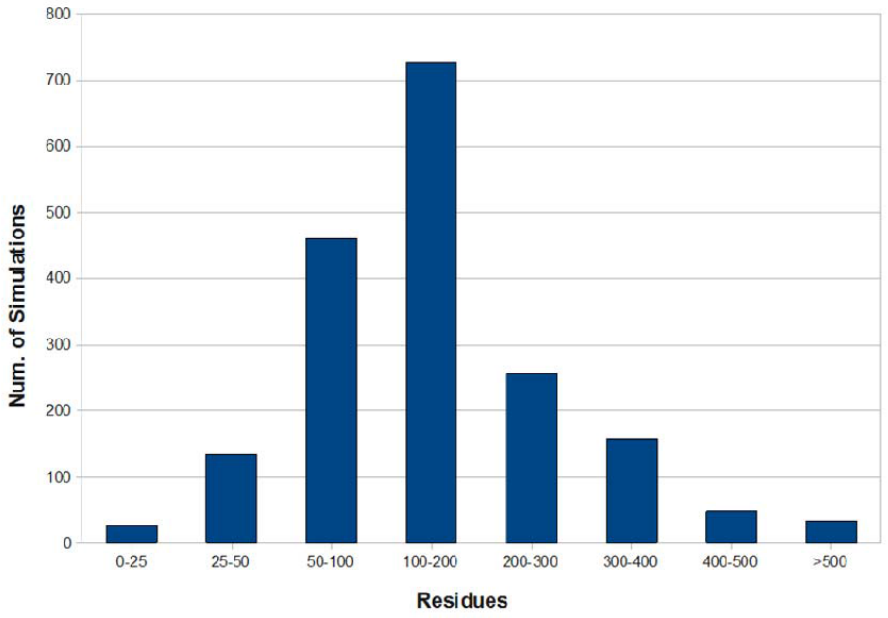


Figure S1. Size distribution of all proteins in the MoDEL database

Suppl. Figure S2

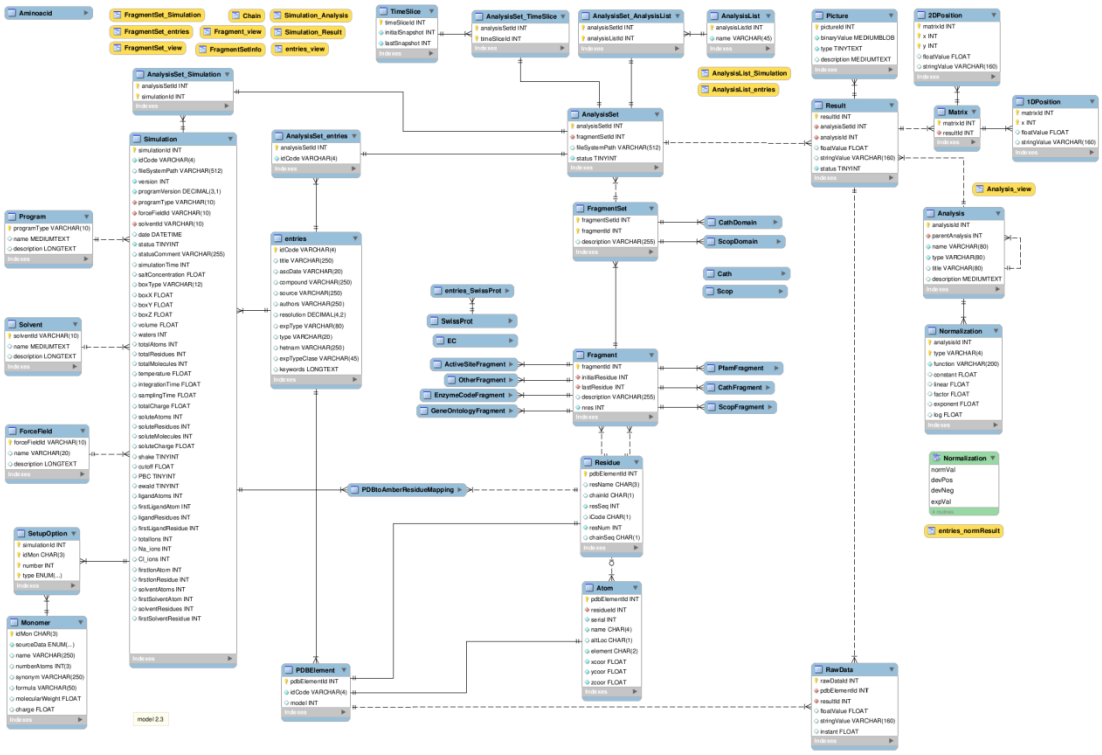


Figure S2. Schematic representation of the MoDEL database schema. External databases are underlaid green, experimental structures retrieved from the PDB are stored in tables underlaid red. Normalization parameter for quality control and to identify dynamic outlier are kept in the tables underlaid yellow.

Suppl. Figure S3

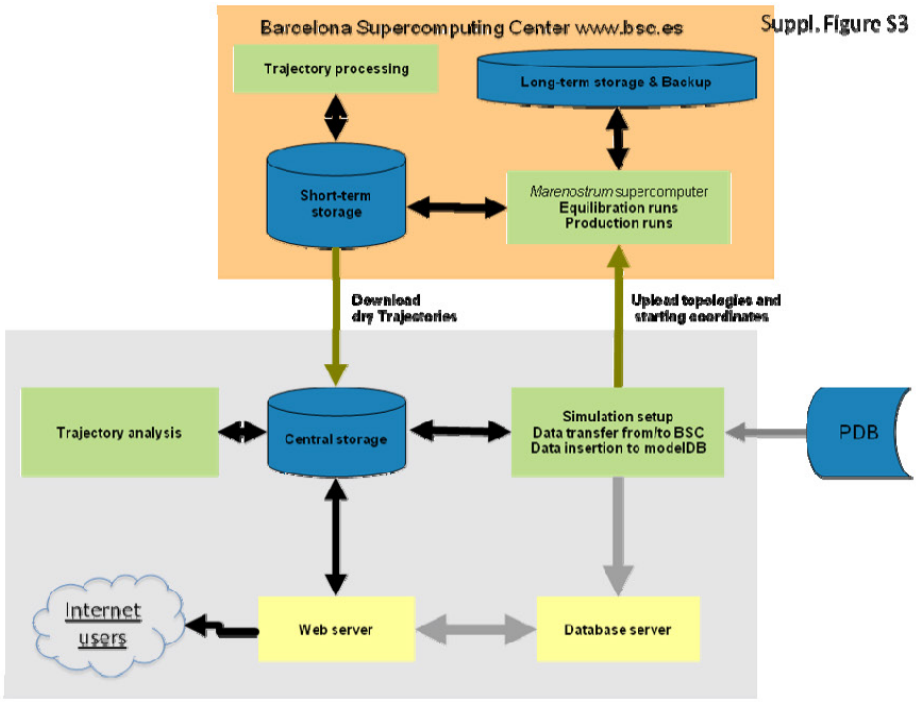


Figure S3. Schema of flow of data in MoDEL

Supp . Figure S4

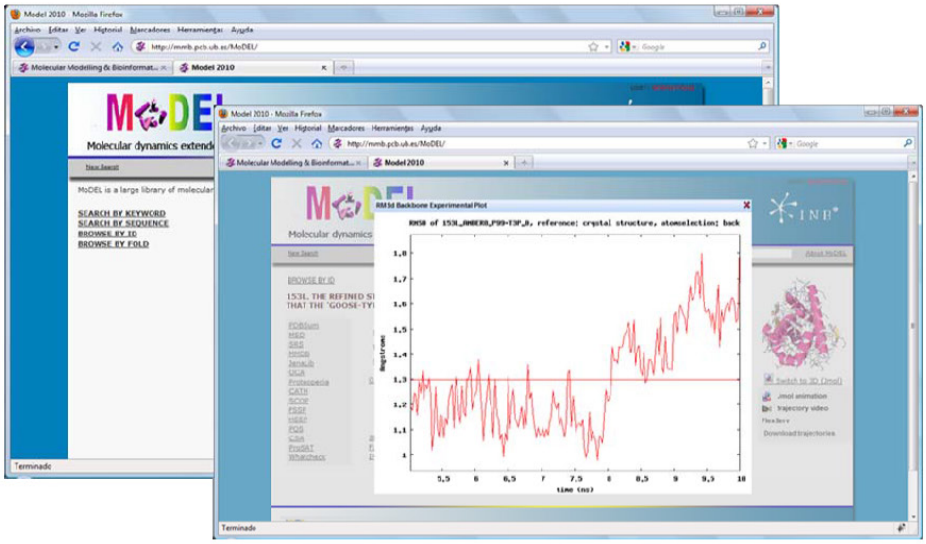


Figure S4. Examples of screenshots of the MoDEL web server

Suppl. Figure S5

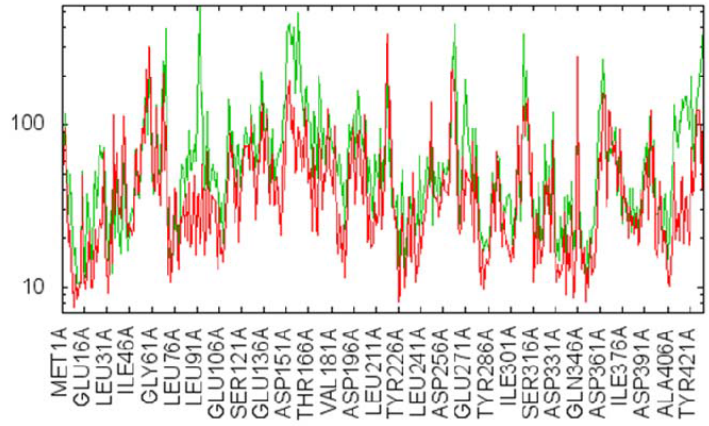


Figure S5. B-factor comparison for GTPase activation protein (1gnd), the largest protein in the µMoDEL subset. The 100ns simulation (green) shows especially in loop regions increased flexibility as compared to the 10ns (red) simulation but the overall pattern is maintained.

Figure S6

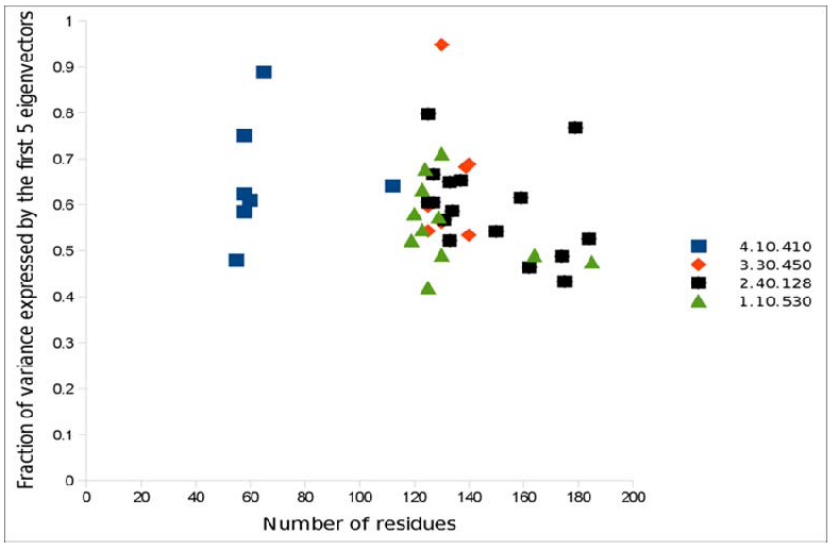


Figure S6. Measure of the complexity of flexibility space determined as the amount of total variance explained by a given number of essential deformation movements (in this case 5). The Higher the variance explained, the simpler the flexibility space. Values are displayed for all the CATH class levels (in CATH classification the first number indicates the class: 1 refers to mainly alpha protein, 2 to mainly beta, 3 to alpha-beta and 4 to protein domains which have low secondary structure content. See text and Suppl. Methods for additional details.

Figure S7

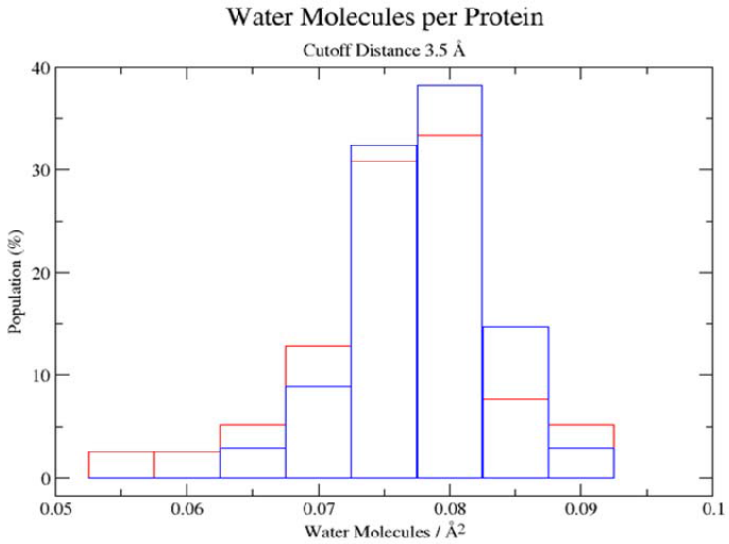


Figure S7. Water atmosphere around proteins from two CATH classes (red: mainly alpha, and blue: mainly beta). The number of waters molecules in close contact to protein residues (distance cutoff 3.5 Å) was normalized by the absolute solvent accessible surface (SASA) of the protein computed with NACCESS program. Analysis was done with 30 randomly selected proteins for both CATH 1 and CATH 2 families.

Figure S8

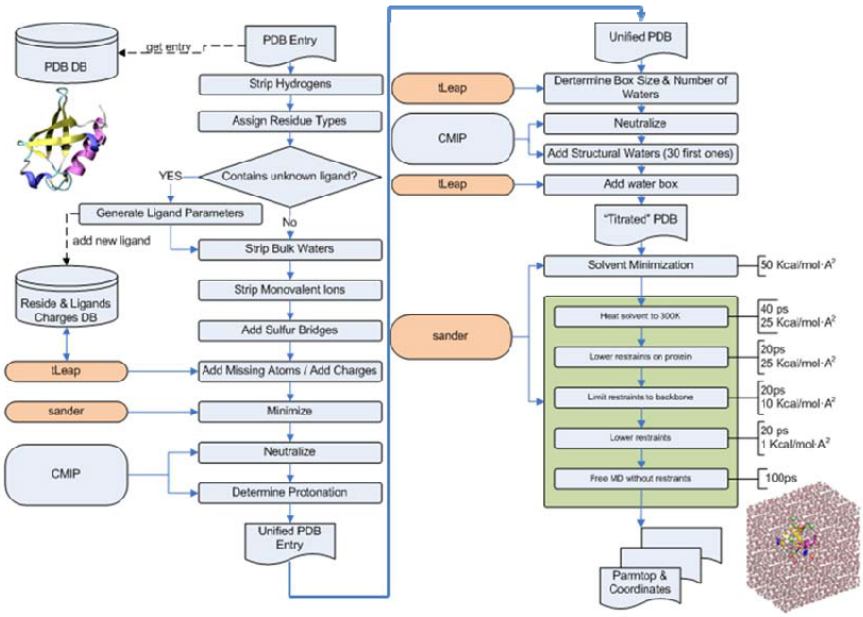


Figure S8. General structure of the automatic set-up procedure implemented for MoDEL. This procedure leads to a topology file (in the case shown here in Amber format) and an equilibrated file that can be launched to different MD codes.

Table S1. Major external tools used in the analysis of MoDEL and major external databases which are linked to MoDEL analysis workflows

External analysis tools	Major external linked databases
Ptraj (*http://ambermd.org*) PCAsuite (*http://mmb.pcb.ub.es/software/pcasuite*) Gnuplot (*http://www.gnuplot.info/*) Naccess (*http://www.bioinf.manchester.ac.uk/naccess/*) Stride (*http://webclu.bio.wzw.tum.de/stride/*) Procheck (*http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/*) FlexServ (http://*mmb.pcb.ub.es/FlexServ*) PDL (*http://pdl.perl.org/*)	Protein Data Bank (*http://www.pdb.org*) UniprotKb (*http://www.uniprot.org*) CATH (*http://www.cathdb.info/*) Enzyme (*http://expasy.org/enzyme/*) DSSP (*http://swift.cmbi.ru.nl/gv/dssp/*) Gene Ontology (*http://www.geneontology.org/*) PFAM (*http://pfam.sanger.ac.uk/*) SCOP (*http://scop.mrc-lmb.cam.ac.uk/scop/*)

Table S2. MicroModel and NanoModel Datasets.

1agi	Bovine Angiogenin
1bfg	Basic Fibroblast Growth Factor
1bj7	Bovine Lipocalin Allergen BOS D 2
1bsn	F1-ATP synthase (epsilon subunit) from *E. coli*
1chn	Bacterial Chemotaxis Protein Chey
1cqy *	Starch Binding Domain of *Bacillus cereus* Beta-Amylase
1csp	*Bacillus subtilis* Major Cold Shock Protein, CSPB
1czt	C2 Domain of Human Coagulation Factor V
1emr	Human Leukemia Inhibitory Factor (LIF)
1fas	Fasciculin 1 From Green Mamba Snake Venom
1fvq	Yeast Copper Transporter Domain CCC2A
1gnd	Guanine Nucleotide Dissociation Inhibitor, Alpha-Isoform
1i6f	Insect-Specific Neurotoxin Variant 5 (Cse-V5) From the Scorpion
1il6	*Centruroides sculpturatus*
1j5d	Human Interleukin-6
1jli	Oxidized Paramagnetic Cu(Ii) Plastocyanin from *Synechocystis* Pcc6803
1k40	Human Interleukin 3 (Il-3) Mutant
1kte *	Fat Domain of Focal Adhesion Kinase
1kxa	Thioltransferase
1lit	Sindbis Virus Capsid, (Wild-Type) Residues 106-264
1lki	Human Lithostathine
1nso	Leukemia Inhibitory Factor
1ooi	Folded Monomer of Protease From Mason-Pfizer Monkey Virus
1opc *	LUSH from *Drosophila melanogaster*
1pdo	OMPR DNA-Binding Domain, *Escherichia coli*
1pht	Phosphoenolpyruvate-Dependent Phosphotransferase
1sdf	Phosphatidylinositol 3-Kinase P85-Alpha Subunit Sh3 Domain
1sp2	Stromal Cell-Derived Factor-1 (Sdf-1)
1sur	Zinc Finger Domain From Transcription Factor Sp1f2
2hvm	Phospho-Adenylyl-Sulfate Reductase
1ubq **	Hevamine
2gbl **	Ubiquitin
	Immoglobulin binding domain of streptococcal protein G

* Proteins included in both the Micro and NanoModel datasets.
** Proteins included in only the NanoModel Dataset.

Table S3. Descriptors used to check for potential problems in MD simulations that are not clearly caused by technical problems. Trajectories failing to fulfil two (or more) global and/or one (or more) local criteria are labelled with warnings in the database and although they are accessible, interpretation should be made with care.

Global descriptor	Threshold	Local descriptor	Threshold
RMSd	< 8 Å	lost native contacts[%]	< 30%
relative RMSd[$]	< 0.2 Å/res	sec struct. changes[&]	< 10%
Tmscore	< 2.5 Å		
relative Tmscore[$]	< 0.06 Å/res		
relative Rgyr[$]	< 0.4 Å/res		
relative SAS[$]	< 100 Å^2/res		

[$] Values referred to the total number of residues in the protein.

[%] Compared with the experimental model

[&] To reduce thermal noise, a comparison is made here between the first and last ns of trajectory.

Suppl. Material. Meyer et al., 29/10/2010

Table S4. Detailed comparison of 10ns versus 100ns trajectories of 430 residue GTPase activation protein (1gnd), the larges protein in the μMoDEL subset.

Descriptor	**10 ns**	**100 ns**
RMSd to crystal (Å)	2.96±0.20	3.01±0.13
RMSd to average (Å)	1.56±0.18	1.84±0.20
Solvent accessible SA ($Å^2$)	20'418±395	20'517±416
Radius of Gyration (Å)	22.81±0.12	22.98±0.20
Schlitter Entropy ($kcal*mol^{-1}*K^{-1}$)	5.762	6.027
Bfactor vs Experiment (Spearman)	0.75	0.68
$PCA^{C\alpha}$ Total Variance ($Å^2$)	497.54	801.99
$PCA^{C\alpha}$ Complexity*	91	71
Native contacts	1'568 (85%)	1'533 (82%)
Native H-bonds	272 (98%)	256 (93%)

*number of vectors that sum 90% of variance

Suppl. Material. Meyer et al., 29/10/2010

Table S5. Hess's similarity index (1 identical trajectory, 0 completely dissimilar) between NMA and essential dynamics eigenvectors and anharmonicity index (in percentage) considering a space of 10 eigenvectors for a reduced set of representative proteins (2 of each class: α, β, α+β and few secondary structures). These measures indicate the similarity between MD and a trajectory generated using a purely harmonic elastic network model. See Supplementary Methods for detailed discussion on the different metrics.

Protein		CATH class	Similarity index	Anharmonicity
1A68	Potassium channel	α+β	0.28	72%
1ABA	Electron transport	α+β	039	61%
1AE2	DNA-binding protein	β	0.28	72%
1BPI	Proteinase inhibitor (trypsin)	Few	0.43	57%
1E9M	Iron-sulfur protein	α+β	0.32	68%
1ERG	Complement factor	β	0.46	54%
1HDP	DNA-binding protein	α	0.42	58%
1HNR	DNA-binding protein	Few	0.37	63%
1K5K	Transcription	Few	0.19	81%
1LAC	Transferase (acyltransferase)	β	0.46	54%
1P1A	DNA binding protein	α+β	0.29	71%
1TAC	Transcription regulation	Few	0.39	61%

Suppl. Material. Meyer et al., 29/10/2010

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

SIMULATION SET-UP

The final automated procedure for setting up simulations in water comprises a number of stages (see Figure S8). First, original structures retrieved from the PDB were stripped of hydrogen atoms, monovalent ions, and all water molecules that did not belong to the first solvation shell of a divalent ion ($d(O\text{-}M^{2+}) < 6$ Å). Non-covalent ligands were kept and parameterized in the GAFF force-field using standard procedures (Wang et al., 2006). Finally, missing side-chain atoms and hydrogen atoms were added from Amber residue libraries, and disulfide bonds were placed between cysteine residues where $d(S\gamma\text{-}S\gamma) < 2.6$ Å. All added atoms were relaxed by a short restrained energy minimization (20 steps steepest descent, 80 steps conjugate gradient, heavy atoms with a 20 Kcal/mol $Å^{-2}$ restrain to the initial structure).

The protonation state of titrable residues was adjusted to the situation at neutral pH using Classical Molecular Interaction Potential (CMIP; Gelpí et al., 2001) calculations with Poisson-Boltzmann potentials (see Orozco & Luque, 2000 for a review). For the exposed residues, protonation energies obtained from experimental pK_as and solvation free energies obtained from MD simulations of the free side chains were used as references.

The systems were neutralized by addition of Na^+ and Cl^- ions placed at the most favored positions using an iterative Poisson-Boltzmann procedure, as implemented in CMIP (Gelpí et al., 2001), until reaching a final 50 mM extra salt concentration. Structural water molecules (i.e. the 2% of water displaying the strongest contacts with the protein) were added by iterative CMIP calculations (Gelpí et al., 2001), while bulk water was added from pre-equilibrated TIP3P water boxes, filling truncated octahedral simulation cells with a minimum solute-box distance of 12Å. The all-atoms simulated systems ranged from 8946 (1rpb) to 269575 (1t5s) particles. All systems were then minimized, heated, and equilibrated in a multi-stage process (see Figure S8), followed by 0.4-ns post-equilibration by unrestrained MD under production conditions (see below).

The MoDEL database also contains information for trajectories collected in the gas phase for the μMoDEL subset (Meyer et al. 2009). These simulations are useful to study the behavior of proteins when subjected to mass spectrometry or X-free electron laser (XFEL) experiments (Neutze 2004). Starting conformations for gas-phase simulations were taken from the 10^{th} ns of the simulation in water and were then manipulated to generate two gas-phase conditions without altering geometry: i) ideally mild vaporization conditions, and ii) vaporization conditions in a typical electrospray experiment (Aebersold & Goodlett 2001; Benesch & Robinson 2006; Meyer et al. 2009). Optimal protonation sites were then determined from Poisson-Boltzmann calculations and experimental proton affinities, as described in Meyer et al. (2009). The resulting gas-phase systems were then subjected to a multi-stage equilibration and heating (T=300 K) process similar to that used for solution simulations (see above).

Suppl. Material. Meyer et al., 29/10/2010

TRAJECTORY PRODUCTION

MD simulations in water were performed in the isothermal (T=298K)-isobaric (P=1 atm) ensemble, using SHAKE (Ryckaert et al., 1977) to eliminate the vibration of chemical bonds involving hydrogen atoms, thereby allowing the use of a 2-fs integration step. Periodic boundary conditions and Particle Mesh Ewald (Darden et al., 1993) were used with standard PMEMD defaults (8 Å cutoff for direct/reciprocal space calculation, an order 4 spline interpolation (spacing 1 Å), distant van der Waals terms approximated using continuum correction) to account for distant interactions. In order to prevent potential periodicity artifacts (one of the possible problems of PME), unit boxes were created to allow at least 12 Å of water from the protein to the sides of the box. By default, MD runs were carried out using the parallel PMEMD module of AMBER8, employing 8 to 32 processor cores, depending on system size. Trajectories were generated on the *MareNostrum* supercomputer at the Barcelona Supercomputing Center (http://www.bsc.es) and local computers in our laboratory and were stored with 1-ps spacing.

METRICS AND ANALYSYS METHODS

MoDEL contains a variety of metrics of the characterization of the structure and dynamics of proteins. The most important ones are summarized below:

Root mean square deviation (RMSd). It is the standard magnitude to calibrate the deviation of a structure with respect to a reference conformation. It is computed as:

$$RMSd = \left(\frac{\sum_{N} \left(R_i - R_i^0 \right)^2}{N} \right)^{1/2} \qquad (1)$$

where N is the total number of atoms/residues considered in the calculation, R_i stands for the vector position of particle i in the snapshot or in the reference conformation (with the 0 superscript), computed after alignment of the structure to the reference conformation to maximize the overlap. The RMSd can be computed using restricted sets of atoms, or mass- weighting the different particles

Gaussian RMSd. It is interesting alternative for very flexible proteins since here the overlap is performed in an iterative way as to maximize the correspondence of rigid parts of the protein (1), focussing then all the movement into the flexible region:

$$gRMSd = \left(\frac{\sum_{i=1}^{N} \varsigma_i d_i}{\sum_{i=1}^{N} \varsigma_i} \right)^{1/2} \qquad (2)$$

where d_i is the distance between the position of a particle in the snapshot and in the reference structure and ζ_i is a weight factor ranging from 0 to 1 which modulates the impact in the gRMSd of the i-particle and is computed as:

$$\zeta_i = \exp(-d_i / c) \qquad (3)$$

where considering the similarity between the compared structures the scaling constant c is taken as 2 Å (Damm and Carlson; 2006). Note that equations 2 and 3 are interdependent via the alignment and accordingly need to be solved iteratively. During this process the weight of flexible residues in the alignment is reduced

Tm-score is an additional metrics for measuring similarity, which was developed to compare very different structures in a robust manner. In its general formalism (Zhang and Skolnick; 2004) the index ranges from 1 to 0 and is defined as:

$$Tm-score = \max\left[\frac{1}{L_N}\sum_{i=1}^{Lt}\frac{1}{1+\left(\frac{d_i}{d_o}\right)^2}\right] \qquad (4)$$

where the max function refers to the situation found after optimum superposition, L_N is the length of the protein, L_t is the length of the aligned segment and d_0 is a normalization factor depending on the size of the protein: $d_0 = 1.24(L_N - 15)^{1/3} - 1.8$. Note that the RMSd associated to the Tm-score can be computed from:

$$Tmscore_{rmsd} = \left(\frac{1}{L_t}\sum_{i=1}^{Lt} d_i^2\right)^{1/2} \qquad (5)$$

Hydrogen bonds. We determined that two residues are hydrogen bonded when donor and acceptor are at less than 3.5 Å and acceptor-H-donor angle is less than 120 degrees. We consider that a hydrogen bonds is maintained in a portion of the trajectory when it is detected in more than 50% of the collected snapshots

Secondary structure assignment was done using Kabsch and Sander algorithm (Kabsch and Sander; 1983). We consider that a secondary structure element was maintained at a given position of the protein if it was the predominant one in the collected snapshots.

Native Contacts. A contact in a snapshot of the simulation is recorded when the C_α of two residues are at less than 7Å. When the same contact was present in the experimental structure, such contact is defined as "native". We consider that a native contact is lost when it is not preserved in more than half of the collected snapshots.

Solvent accessible surface (SAS) was calculated by means of the NAccess software package (Hubbard and Thornton; 1993), using as rolling probe radius the same radius of a water molecule, i.e. 1.4 Å.

B-factors. They are the standard measure of residue/atom harmonic flexibility. They are computed from the oscillations of a residue/atom with respect to its equilibrium position, assuming that these oscillations are isotropic.

$$B_{factor} = \frac{8}{3}\pi^2 \left\langle \Delta r^2 \right\rangle \qquad (6)$$

where $\left\langle \Delta r^2 \right\rangle$ stands for the oscillations of residues around equilibrium positions.

B-factor profiles represent the distribution of residue harmonic oscillations. They can be compared with X-ray data, but caution is needed, since crystal lattice effects tend to rigidify exposed protein residues. Very large-B-factors should be taken with caution since indicate very flexible residues that might display conformational changes along the trajectory, which is difficult to follow within the harmonic approximation implicit to B-factor analysis.

Apparent inter-residue stiffness. It is defined (Camps et. al. 2009) as the force-constant acting between two (i and j) residues in the case of completely disconnected harmonic oscillators.

$$K_{ij}^{app} = \frac{k_B T}{\left\langle \left(R_{ij} - \left\langle R_{ij} \right\rangle \right)^2 \right\rangle} \qquad (7)$$

where k_B is the Boltzman's constant and T is the temperature. The averages are computed using the MD ensemble.

The index helps to detect strong interactions between residues, which might indicate physically-intense direct contacts or strong chain-related interactions.

Chained correlations. This analysis is performed to detect correlations between the movement of distant residues. It is done by a post-processing of the residue-residue correlation matrix, where irrelevant (ex. i→i+1, or non significant correlations were removed. Then a root residue must be selected and residues with a large correlation with it are recorded. These first-generation correlated residues are then used for a new iteration. The user selects the width (the number of correlated residues to be considered) and the depth (the number of iterations) of the search (Camps et al. 2009).

Hinge point predictions. The hinge points are residues or protein segments around which large movements are detected along the trajectory. Analysis of hinge points are performed by three different methodologies, each one with its own implementation using trajectories re-centred using a Gaussian RMSd fitting protocol (Damm and Carlson;

2006, Camps et al. 2009): a) looking at changes in the B-factor slope, b) looking at residues with peak inter-residue force constants (Sacquin-Mora and Lavery; 2006) and c) by clustering residues based on their correlations (Navizet et al. 2004).

Lindemann indexes were introduced for determining whether an infinite system is solid-like or liquid-like (Zhou et al; 1999), were calculated via:

$$\Delta_L = \frac{\sqrt{\sum_i \left\langle \Delta r_i^2 \right\rangle / N}}{a'} \qquad (8)$$

where N is the number of atoms and a' is the most-probable non-bonded near-neighbor distance, r_i is the position of atom i, $\Delta r_i^2 = \Delta r_i^2 = (r_i - \langle r_i \rangle)^2$, and <> denotes configurational averages.

Variance-related metrics. The total (and relative to the number of residues) variance is computed to take a rough estimate of the degree of movement in the protein along the dynamics. The number of essential movements needed to explain a given variance threshold is indicated to define the complexity of the simulation space. The amount of variance explained by 5 essential movements is used to evaluate the quality of the essential deformation space. Finally, the dimensionality of the deformation space is defined by the rank order of the first essential movement with (at room temperature) associated variance lower than 1 $Å^2$ (see below).

Strand entropies. They are computed using a pseudoharmonic approach following either Schlitter (1993; see eq. 8) or Andricioaei-Karplus (2001, see eq. 10) methods and the mass-weighted covariance matrix derived from the MD-ensemble. Values are extrapolated to infinite simulation time using optimized exponential relationships (Harris et al. 2001, see eq. 11). To gain additional insights partial estimates are obtained, by considering, for example only the trace of the protein.

$$S \approx 0.5k \sum_i \ln\left(1 + \frac{e^2}{\alpha^2}\right) \qquad (9)$$

$$S = k \sum_i \frac{\alpha_i}{e^{\alpha_i} - 1} - \ln(1 - e^{-\alpha_i}) \qquad (10)$$

where $\alpha_i = \hbar\omega_i / kT$, ω being the eigenvalues obtained by diagonalization of the mass-weigthed covariance matrix, and the sum extends to all the non-trivial vibrations.

$$S(t) = S_\infty - \frac{a}{t^\beta} \qquad (11)$$

where t is simulation time α and β are fitted parameters.

Suppl. Material. Meyer et al., 29/10/2010

Essential dynamics. Following essential dynamics algorithm (ED, Amadei et al; 1993), the orthogonal movements describing the variance of a system are obtained by diagonalization of the covariance matrix derived from the MD simulation. The result of the analysis is the generation of a set of eigenvectors (the modes or the principal components), which describe the nature of the deformation movements of the protein and a set of eigenvalues, which indicate the stiffness associated to every mode:

$$K_\lambda = \frac{k_B T}{\lambda} \qquad (12)$$

where λ stands for an eigenvalue (in distance^2 units)

Sorting the eigenvectors by their associated eigenvalues, the largest part of the variance will be explained by the first few eigenvectors. Since the eigenvectors represent a full-basis set, the original Cartesian, trajectory can be always projected into the eigenvectors space, without lost of information. Furthermore, if a restricted set of eigenvectors is used information is lost, but the level of error introduced in the simplification is always on user-control by considering the annihilated variance (the residual value between the variance explained by the set of the eigenvectors considered and the total variance). Inspection of the atomic components of the most important eigenvalues helps to determine the contribution of different residues to the key essential deformations of the protein. Visualization of the ED modes allowed us to gain insight into the nature of the most prevalent protein movements. Analysis of the variance contribution of the different modes helped to quantify the size and complexity of the deformation space.

Similarity indexes helped to determine the similarity between the intrinsic deformation patterns sampled in trajectories of the same (A=B) or different proteins (A≠B). Different metrics based on the dot product can be envisaged to determine this similarity such as Hess's index (Orozco et al. 2003; Eq. 13) or the RMSIP (Amadei et al. 1999; Eq. 14).

$$\gamma_{AB} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}(\nu_i^A \bullet \nu_j^B)^2 \qquad (13)$$

where n stands for the minimum number of eigenvectors which explains more than a given threshold of the variance of the trajectories (A and B) and ν is an eigenvector.

$$RMSIP = \left(\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\nu_i^A \bullet \nu_j^B\right)^2\right)^{1/2} \qquad (14)$$

Note that it is often useful to obtain relative, rather than absolute similarity indexes, just by referring the similarity index to the expected self-similarity:

$$\Gamma ij = 2\frac{\gamma_{ij}}{(\gamma_{ii} + \gamma_{jj})} \qquad (15)$$

where i and j stand here for different trajectories, and the self-similarity indexes are obtained by comparing two sub-trajectories of the same length of the same molecule.

Using dot-product metrics the anharmonicity of a trajectory can be determined by looking at the overlap between the normal modes determined using a fully harmonic potential and those obtained from diagonalization of the MD covariance matrix:

$$\Theta_a = 1 - \gamma_{MD-NMA} \qquad (15)$$

SUPPLEMENTAL REFERENCES

Aebersold, R., and Goodlett, D.R. (2001) Mass spectrometry in proteomics. Chem. Rev. *101*, 269-295.

Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. Essential dynamics of proteins. Proteins: Struct. Funct. Gen. 17:412–425, 1993

Amadei, A., de groot, B. L., Ceruso, M. A., Paci, M., Di Nola, A. & Berendsen, H. J. Proteins. 1999. 35, 283-292

Andricioaei, I., and Karplus, M. (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. J. Chem. Phys. *115*, 6289-6292.

Benesch, J.L., and Robinson, C.V. (2006) Mass spectrometry of macromolecular assemblies: preservation and dissociation. Curr. Opin. Struct. Biol. 16, 245-251.

Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Gelpí, J.L., and Orozco, M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. Bioinformatics. *25(13)*, 1709-1710

Darden, T., York, D., and Pedersen, L. (1993) Particle Mesh Ewald-an N.Log(N) method for Ewald sums in large systems. J Chem Phys *98*, 10089-10092

Damm K.L. and Carlson H.A.. "Robust-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures". Biophysical Journal. (2006), 90, 4558-4573.

Harris, S.A., Gavathiotis, E., Searle, M.S., Orozco, M., and Laughton, C.A. (2001). Cooperativity in drug-DNA recognition: a molecular dynamics study. J. Am. Chem. Soc. *123*, 12658-12663

Hubbard,S.J.and Thornton, J.M. (1993), 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.".
Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637.

Meyer, T., de la Cruz, X., and Orozco, M. (2009) An atomistic view to the gas phase proteome. Structure. *17(1)*, 88-95

Navizet, I.; Cailliez, F. and Lavery,R. Probing Protein Mechanics: Residue-Level Properties and Their Use in Defining Domains (2004). Biophysical Journal 87:1426-1435.

Neutze, R., Huldt, G., Hajdu, J., and van der Spoel, D. (2004) Potential impact of an X-ray free electron laser on structural biology. Radiat. Phys. Chem. *71*, 905-916.

Orozco, M., and Luque, F.J. (2000) Theoretical methods for the description of the solvent effect in biomolecular systems. Chem. Rev. *100*, 4187-4226

Orozco, M., Pérez, A., Noy, A., and Luque, F.J. (2003) Theoretical methods for the simulation of nucleic acids". Chem.Soc.Rev. *32*, 350-364

Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys *23*, 327-341

Sacquin-Mora,S., and Lavery,R. Investigating the Local Flexibility of Functional Residues in Hemoproteins. (2006) Biophysical Journal 90:2706-2717.

Schlitter, J. (1993). Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem. Phys. Lett. *215*, 617-621.
Wang, J., Wang, W., Kollman P. A.; and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. Journal of Molecular Graphics and Modelling, *25*, 247-260.

Zhang,Y, and Skolnick,J. "Scoring Function for automaied assessment of protein structure template quality". Proteins (2004), 57, 702-710.

Zhou, W., Vitkup, D., and Karplus, M. (1999) Native proteins are surface-molten solids: application of the lindemann criterion for the solid *versus* liquid state. J. Mol. Biol., *285*, 1371-1375

5.3 High Throughput computational study of hydration molecules.

5.3.1 Synopsis

In this project, a HT study involving more than 16 million of water molecules was performed using solvent dynamic information from our library of protein molecular dynamics simulations MoDEL. Presence of distinguishable water shells, mean residence times, residue type preferences or hydrogen bond formation/destruction cycle was studied. Results obtained suggest a highly dynamic behavior of hydration water molecules, with very low residence times and few pure structural waters. The direct participation of water molecules in the kinetics of formation and destruction of hydrogen bonds on the protein surface is also demonstrated, obtaining thermodynamic information suggesting that about 60% of these hydrogen bonds are not more favorable that the corresponding interactions to water.

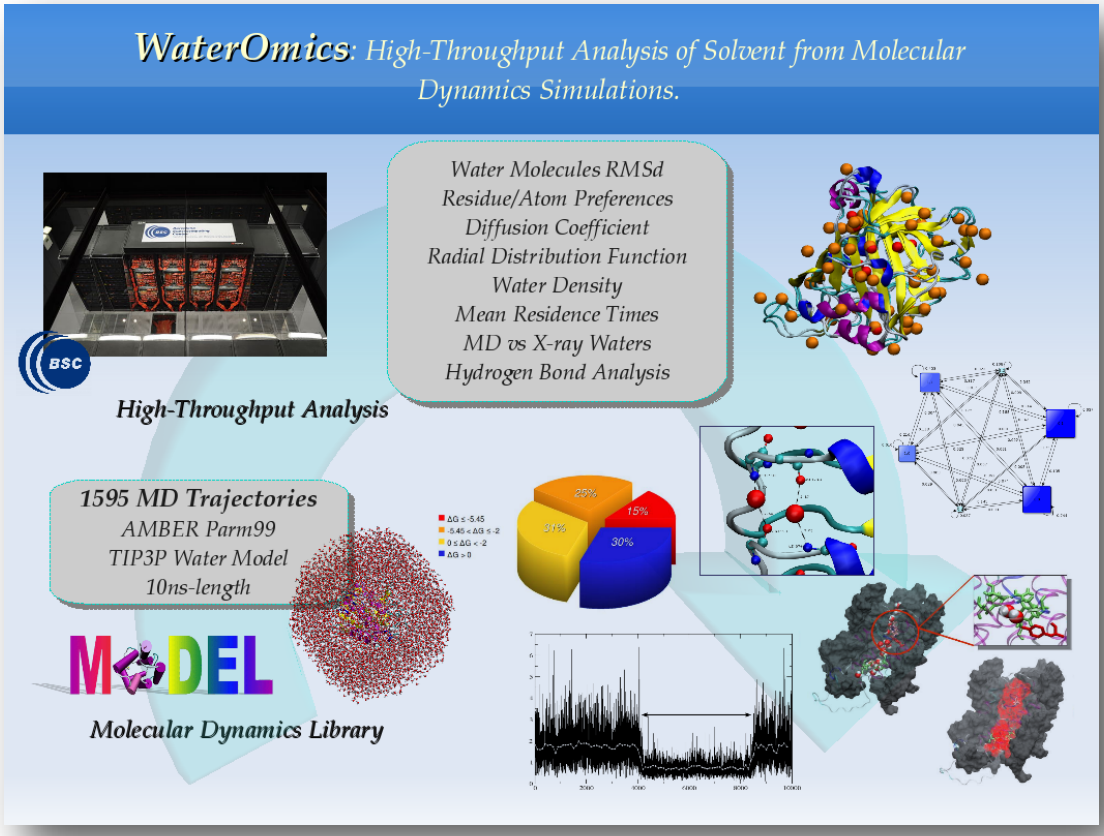
This study, which took several years of work, is to our knowledge, the first high throughput analysis of protein-solvent interactions using molecular dynamics simulations, and has served as a proof of concept to our MD library.

5.3.2 Paper 2:

Water-omics: High throughput analysis of protein-solvent interactions from MD simulations.

Adam Hospital, Modesto Orozco, Josep Lluís Gelpí.

In preparation.



Water-omics: high throughput analysis of protein-solvent interaction from molecular dynamics simulations

Adam Hospital[1,2,3,4], Modesto Orozco[1,2,3,4,6*] and Josep Lluís Gelpí[1,2,3,5,6*]

Hydration around protein is described from the mining our library of around 1,800 molecular dynamics trajectories of proteins in explicit solvent (MoDEL). Analysis of the trajectories of more than 16 million of water molecules on the most representative protein folds provide a picture of unprecedented quality of the solvent environment around folded proteins. Results suggest a much more dynamic behavior of the solvent-protein interactions than expected. For instance, there is a very limited correlation between low mobility water molecules in the simulation, and crystallographic water molecules, usually taken as a reference for structural water. Analysis of mean residence times agrees with that conclusion. MRT generally lay in the range of picoseconds, only reaching the nanosecond in those water molecules placed within protein pockets. Water molecules do interact closely with protein atoms but they only remain in the first solvent layer for a limited amount of time. On the other hand, solvent also plays a significant role in the stability and dynamics of hydrogen bonds. The general conclusion of the HB stability study reveals that only a 40% of them can be considered as stable (formation ΔG better than -2 kcal/mol). Water acts, also, actively in the breaking of protein HBs, mainly by solvating the individual atoms, and in a much lesser extend by forming water bridges. Overall, this study provides a systematic methodology to undergo the analysis of solvent-protein interactions that can be generally applied. Two examples of proteins where water plays a significant and well known functional role are presented to illustrated the use of such methodology.

[1] Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain
[2] Joint BSC-CRG-IRB Program on Computational Biology, Barcelona, Spain
[3] Barcelona Supercomputing Center, Barcelona, Spain
[4] Structural Bioinformatics Node, National Institute of Bioinformatics, Barcelona, Spain
[5] Computational Bioinformatics Node, National Institute of Bioinformatics, Barcelona, Spain
[6] Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain

* Correspondence to Modesto Orozco modesto.orozco@irbbarcelona.org or Josep Lluis Gelpí gelpi@ub.edu

1.-Introduction

Water is a main molecule defining life and is then involved in key biomolecular processes such as protein folding, ligand recognition, enzymatic activity and is crucial to define the structure and stability of proteins and nucleic acids. However, after several decades of work the questions formulated by Levitt and Park twenty years ago regarding waters interacting with proteins (1) remain unanswered: Where are they? How long do they stay there? How strongly do they interact with the protein? How do they affect protein structure and stability?

More than 40 years (2) of intense experimental research have outlined key characteristics of strongly bound water molecules, but the description hydration atmosphere generated by more mobile and fluctuating water molecules remained more challenging for experimental techniques (2-6). Dielectric Relaxation (DR) was one of the first methods used to study the dynamics of protein solutions (6,7). Neutron Scattering technique revealed the mean square displacement of water protons on the picosecond time scale (8). Later on, Magnetic Relaxation Dispersion (MRD) provided data such as number, residence times and mean rotational correlation times of protein-bound water molecules (9,10). More recently, new methods like Time-Dependent Stokes' Shift (TDSS) (also called Solvent Dynamics or Time-Resolved Fluorescence Spectroscopy) were able to give information on full time-resolved solvent relaxation (11,12). Finally, in the last years Optical Kerr-effect (13) and 2D-IR Spectroscopies (14) have provided complementary information on the hydration dynamics. However, the use of these experimental techniques has not provided yet a complete proteome-scale representation of water dynamics, justifying the use of simulation techniques, particularly molecular dynamics (15,16), which has been used by different groups to study hydration of individuals or a limited set of proteins (17-19). In this study, we present the first High-Throughput analysis of hydration water molecules using almost 1,800 MD simulations from our MoDEL database (21). We report a systematic analysis of protein-solvent interaction properties with special interest in their dynamics. The methodology used here has been tested and can be easily applied to other systems and longer simulations.

2.-Methods

Trajectories for 1,595 proteins were taken from our Molecular Dynamics Extended Library (MoDEL) database (21-23). Simulation boxes range from 102,000 $Å^3$ (pdb 1sop) to 1,890,000 $Å^3$ (pdb 1i49) with an average volume around 102,000 $Å^3$ (average box radii around 38 Å (see Suppl. Fig.1 for a summary of size distributions). Results presented here correspond to the analysis of 16,232,408 individual water molecules (unless otherwise noted, at least 10^3 snapshots for each water molecule corresponding to the last ns of MoDEL simulations), and imply mining of more than 20 Tb of data.

Water molecules dynamics around the simulation box were studied by means of their root mean square deviation (RMSd) in controlled time windows, which provides information on water-protein interactions, and the distribution and dynamic behavior of water molecules near the protein surface. In the study, trajectory snapshots were superimposed to remove translation and rotation of the protein molecule. In such conditions, solvent movements relative to a static protein are measured. As extreme cases, water molecules permanently attached to protein atoms are expected to have a near to 0 RMSd, whereas the bulk water molecules are expected to have a high, uniform, RMSd value, as they are moving throughout the solvent simulation.

A large variety of analysis was performed for the entire protein and selected portions (residues, atoms, functional groups). Structural and dynamic analyses were performed using web-tools in MoDEL and FlexServ webservers (21,24), standard analysis packages in AMBER (25) and GROMACS (26), MD-linear response theory (MD-LRT) programs (27), and other external and internal analysis codes (see Suppl. Material for a complete description of the analysis done).

3.-Results & Discussion

3.1.- Water Molecules RMSd

We first explored the distribution of RMSd for water molecules in finite time windows (Fig.1 and Suppl. Fig. 2). Taking a window of 200ps, only a few water molecules have had time to travel far away from their initial position, which explains the large peak at RMSd around 8Å, but for time windows greater than 0.5 ns, the first peak disappears and the distribution becomes centered around 30 Å, which implies mostly free water molecules. Interestingly, a small peak (Fig.1 and Suppl. Fig. 2) does not disappear even for 5 ns window. The peak can be assigned to a small population of fixed water molecules, which are either very strongly attached to the protein or trapped in an internal cavity.

Water molecules RMSd distribution in 500 ps time window is particularly interesting, since three different water shells (with Gaussian-like distributions) can be observed: one small centered near 0.5-1 Å, corresponding to tightly bound waters (see discussion above), a second one centered around 12 Å, corresponding to water molecules that are not strongly bound to the protein, but that display a mobility smaller than that expected for bulk water; and finally the major peak (30 Å) corresponding to bulk water (Fig. 2a). Populations corresponding to HB donor (d_{pw} ~ 2.5 Å) and HB acceptor (d_{pw} ~ 2.0 Å) water molecules can be clearly distinguished in the first water shell (i.e. tightly bound waters; Fig. 2b). The second shell (i.e. moderately slow water molecules) is more complex as noted in the three-modal distribution, with two peaks corresponding to waters in direct contact with the protein and a large one centered around 3 Å. This can be interpreted as this shell is composed by water molecules that remain very closely packed to the protein for a limited time, whereas at some point diffuse away from the protein (up to 3Å), but without reaching the bulk water region.

The largest peak in the RMSd distribution corresponds to bulk water: water molecules that are freely traveling around the solvent box with little interaction with the protein if any. There is a clear correlation between the mean RMSd of bulk water molecules and the average radii of the MoDEL simulation boxes, showing their mobility limited by the size of the box (data not shown). Indeed, the analysis of the time needed for water molecules to explore MD solvent box (Suppl. Fig 3a and below) confirms that 10 ns-length simulations are long enough for bulk water to be equilibrated.

Fitting the RMSd distributions to Gaussian curves, the complete set of MoDEL water molecules can be systematically classified in three sets: i) the 1st shell defined by those waters with RMSd lower than 1.2 Å (0.1% of total MoDEL waters), which correlate with regions of extreme water density (around 15g/cm^3; see Suppl. Fig 3b), ii) the 2nd shell contains water molecules with an RMSd between 9.7 Å and 16.3 Å (~10% of total MoDEL water molecules), and iii) bulk water (~90% of total MoDEL water molecules). The following analyses were performed in the first two sets: the 1st and 2nd shells of hydration.

3.2.- Atom/Residue Type Water Preference

Water molecules in the 1st shell prefer backbone atoms rather than residue side chains (Figure 3a, black columns). In contrast to what is usually suggested (2,17; see discussion below), water molecules in the 1st sell prefer to act as HB acceptors, as can be deduced from the high frequency of residues with hydroxyl groups (Thr, Tyr and Ser) and amine groups (Arg, Trp, Asn, His and Gln). Lys residue has a lower preference than expected, probably penalized against Arg due to its higher mobility. HB acceptor residues have surprisingly low contacts/ps*Å^2, especially Asp and Glu. Finally, with a few exceptions, such as Trp, hydrophobic residues are, as expected, the least preferred for hydration water molecules.

Similar picture is obtained from the analysis of atom type preferences (Fig. 3b, black columns), with Arg nitrogen atom in the first position, followed by the uncharged backbone nitrogen atoms and hydroxyl groups. As with the residue type study, Lys nitrogen and charged oxygen atoms of Glu and Asp show less contacts, indicating that they rarely coordinate long-living water molecules. Again, as expected, Sulfur group and aromatic and aliphatic carbons are those with the smallest number of contacts with hydration waters.

In contrast to the 1st water-hydration shell, water molecules belonging to the 2nd water shell prefer to act as a HB donor (Fig. 3a, white columns), which agrees with the more accepted picture of protein hydration (2,17). Protein backbone is still one of the regions preferred for water molecules, but now Arg and Lys residues are the ones with highest number of contacts/ps*Å^2. Asp and Glu have increased the preference in comparison with the first water shell, and HB donor residues have in general decreased the number of contacts. Similar picture can be obtained looking at the atom type

preferences, where oxygens of Glu and Asp side chains are clearly the preferred ones by 2nd shell water molecules, followed by the terminal nitrogen atom of Lys and the uncharged nitrogen atoms of the protein backbone (Fig. 3b, white columns). The general conclusion of both analyses is that water acting as HB acceptor is more suitable to form stable (low RMSd) interactions, especially when rigid protein residues are involved, whereas that more mobile residues and interactions where water acts as HB donors appear in the second RMSd shell.

3.3.- Structural Solvent.

The concept of structural or resident water molecules is typically used in X-ray crystallography to label those waters showing strong electron densities in diffraction maps, and which are believed to play an important structural role. From the point of view of simulations the analysis of structural water behavior and even its existence requires the analysis of their dynamic properties (mean residence time (MRT) seems a particularly suitable metrics of such a dynamics; see Methods).

Based on the analysis of MRT (Fig. 4 and Suppl. Fig. 4) we can conclude that the majority of water molecules don't stay attached to the same protein atom more than 200 ps (Suppl. Fig. 4a). However, when looking to the residue MRT, results suggest that 1st shell water molecules remained bound longer periods to the same residue (Fig. 4a and 4b), although rarely more than 3 ns. Only a few pure structural waters with a MRT greater than 4ns (Fig. 4b, Suppl. Fig. 4b) were found. These water molecules are generally placed in buried protein cavities. To gain a more precise picture of the solvation of internal cavities, we analyzed all internal cavities with an average volume greater than 100 $Å^3$ during the dynamics (from the available dataset, 114 proteins shown these internal cavities; see methods). Even for those internal cavities water appears very dynamic, and just in a few cases trapped waters were detected (Suppl. Table 1). Results also show that even the most static water molecules placed in deep internal cavities are rarely attached to a single atom; instead, they prefer to move from one atom to another throughout the cavity (Fig. 4c, Suppl. Fig 4c), which probably helps to reduce the entropy cost of cavity trapping.

Waters classified as being in the 2nd water shell stay bound to the same protein atom only for a very short period of time, usually less than 50 ps, and residue MRTs never overpass the 200 ps (Fig. 4d, Suppl. Fig. 4d). These results are consistent with the hypothesis of 2nd RMSd shell water molecules to form a surrounding shell near the 1st shell, entering it and staying attached to a protein atom for a while, and then moving to the vicinities, where they are more mobile.

3.4.- Crystallographic waters

There is a controversy in the literature between the exact nature of crystal waters, and several results indicate a poor overlap between crystallographic waters and places for especially favorable water-protein interactions found using new developed techniques (28).

To analyze whether or not crystal water molecules correspond to the ones detected as low-mobility in MD simulations we defined a restricted set of 567 proteins (X-ray solved proteins having crystallographic water molecules and resolution better than 2 Å; see Suppl. Methods) and compute accuracy values against RMSd-defined water shells, and against the totality of crystal waters. Figure 5a show the ability of low mobility water molecules to match crystallographic water molecules. Almost half of the first RMSd shell was found to correspond to crystallographic water molecules. The accuracy goes down as the water with higher mobility is considered, but at least 25% of water molecules in regions above the first shell still match the X-ray ones. However, the picture is different when looking to the accuracy against the total number of recovered crystallographic water molecules (Figure 5b). In this case, only a 3% of water molecules reported in PDB structures can be recovered in the first mobility shell, and this value decreases below 1% with more mobile molecules. However, we must say that the set of crystal waters matched in this way changes when analyzing different time-windows along the MD simulation, (see InternalinB case below as an example). The combination of both results reflects that the majority of the water molecules as obtained from X-Ray crystallography correspond to transient water-protein interactions, i.e. crystal waters might be pointing to positions where there is a favorable water-protein interaction, but not, in general, to long-living water molecules. Specific crystal contacts or changes in the dynamics of the surrounding residues upon crystallization may contribute to the selection of some of those water molecules to be trapped in the crystal. Therefore, their role as integral components of the protein structure is not fully justified. Indeed, by no means crystallographic waters should be taken as the only representation of water molecules having structural importance in a protein structure.

3.5.- RDF: Radial Distribution Function

The shape of the radial distribution function obtained for all the proteins in the study is almost identical (Suppl. Fig. 5, and Suppl. Methods). A high rdf peak is found at 2.75Å, and a broader but shorter second peak is observed roughly from 4 to 6Å, corresponding to first and second hydration shells. When rdf are calculated taking as reference different types of atoms (Fig. 6a), we can easily detect the different nature of hydration environment around them. If we group the profiles by similarity we find four clusters the first corresponding to backbone nitrogen and charged groups of Lys and acids (cluster 1, Fig. 6b), backbone carbon and oxygen atoms, hydroxyl groups, carbonyl, amine and guanidine groups (cluster 2, Fig. 6c) and finally backbone alpha carbons, hydrophobic atoms, thiols and rings (cluster 3, Fig. 6d). The main difference between clusters can be observed in the first RDF peak. Hydrophilic clusters 1 and 2 have a pronounced peak near 2.8Å (much higher in cluster 1), indicating strong hydrogen bond interactions, whereas hydrophobic cluster 3 has the first peak near 3.6Å, close to the Van der Waals contact distance between water oxygen and protein heavy atoms. Interestingly first peaks in clusters 1 and 2 are very sharp, which illustrate the directionality and stiffness of hydrogen-bond interactions, in contrast to the less orientated van der Waals that dominate interaction in the first rdf peak in cluster 3.

Second rdf hydration peak appears only in cluster 1 and guanidine and carbonyl groups of cluster 2, such peaks indicate that a second hydration layer can be expected for these highly polar atoms. It is worth to note here the excellent agreement between MD estimates and the conclusions derived experimentally from the analysis of a reduced set of proteins (29,30), which give confidence into MD-simulation results.

3.6.- Water Diffusion

Mean square displacements (MSD) and average velocities have been computed for all waters in 1st, 2nd shell and bulk water (see Suppl. Methods). As a reference, values for a simulation of pure water have been computed and represented in the same plots (Fig. 7). First and second shell water molecules show extremely low diffusion coefficient values (1.1 x 10^{-5} cm^2/s, and 1.7 x 10^{-5} cm^2/s, respectively), which are from 1/4 to 1/5 of those of the bulk water (5.2 x 10^{-5} cm^2/s) in the same set of simulations. It is worth noting that the effect of the protein on the diffusion of bulk water is rather small (diffusion coefficient for water obtained in a simulation box containing only waters was 5.7 x 10^{-5} cm^2/s, in perfect agreement with TIP3P accepted value (31)), suggesting that the effect of protein after the second hydration layer is modest (18,32). MD-derived diffusion values can be translated to experimental scale by scaling then down considering the well know overestimation of diffusion in TIP3P water model (2.56 from references 31, 33, 34). From this we can conclude that inner solvent show diffusion constant around 4 x 10^{-6} cm^2/s and those in the 2nd shell around 7 x 10^{-6} cm^2/s, values that roughly corresponds to supercooled water at 262.3 K and 250.8 K respectively (35).

The analysis of MSD for specific water molecules allows us to reveal the highly dynamics nature of the water-protein interactions. The classification of water molecules as 1st or 2nd RMSd shells allows identifying the specific time windows when water molecules approach the protein surface and change their dynamic behavior. Suppl. Fig 8 shows the behavior of some representative water molecules. Some water molecules remain as 1st shell during the complete simulation (Suppl. Fig. 8a), others behave as bulk water and get trapped near the protein surface for a given period (Suppl. Fig. 8b-d), that changed from one water molecule to another. Typically, for 1st shell water molecules, approaching the time window where interaction with the protein is produced, velocities decrease abruptly (100-300 ps, Suppl. Fig 8b-c), reaching values of about 0.8 Å/s, almost 1 Å/s less than the average velocity of the MoDEL water average (1.8 Å/s). 2nd shell water molecules show also a velocity slow down, but in this case the decrease is not so pronounced, around 0.1 Å/s.

3.7.- Influence of solvent in Protein Hydrogen Bond dynamics

Hydrogen bonds, including salt bridges, are key interactions in the interpretation of molecular recognition. Most analyses of the behavior of macromolecular structures are based in the presence or absence of hydrogen bonds between key residues and their interaction partners. However, hydrogen bond is a kind of interaction that is difficult to

characterize, a rigorous determination requiring quantum mechanics calculations. For this reason routine determination of hydrogen bonds in protein structure analysis is usually restricted to geometric analysis, being all interactions within a valid distance qualified as hydrogen bonds. More restrictive analysis may include angular measurements but this is not the rule. Besides, structures, even those coming from MD simulations, are usually minimized previous to the analysis. This scenario leads to a general overestimation of the number and significance of hydrogen bonds, and in any case to the loss of the energetic information related to them. Hydrogen bonds in proteins, as any other structural characteristic are dynamic, and their evolution should be considered in the understanding of their significance. In addition, solvent, particularly water, interferes in the formation and stability of intra-protein hydrogen bonds.

To analyze further this issue we have studied in detail 819,636 intra-protein hydrogen bonds extracted from a subset of 1,392 simulations. Dynamics for these HBs dataset were captured following the 10 ns simulations every 1 ps snapshot. For the sake of simplicity hydrogen bond states have been classified according to the presence of a given HB and the participation of water molecules, either as water bridge or solvating the intervening atoms (see the Suppl. methods section for a detailed explanation). For each state, population and transitions to other states were measured, and the information gathered analyzed to establish stability and dynamics of every protein HB. ΔG energies associated with the formation of HB were computed taking as the common reference state "00NN" (no hydrogen bond, no water bridge, water molecules solvating both atoms). These energies provide a real estimation of the significance of a given HB in the energetic of the protein structure, taking into account its dynamics, and the competition of water molecules. The length of the simulation sets a limit to the energies that can be measured using this procedure. In the present case, the "fixed" HB class (HB maintained during the complete simulation) corresponds to formation ΔG better than –5.45 Kcal/mol. Statistics of such energetic analysis is shown in figure 7 for global values and table 1 split according to the involved chemical groups. Strikingly, global results show that 30% of the hydrogen bonds defined from the usual geometry analysis show unfavorable ΔG formation energies, 15% of the favorable ones correspond to fixed HBs, and 25% show a reasonable stability (ΔG < -2kcal/mol), whereas the most populated class correspond to slightly favorable HBs where the stability of the HB is only marginally better that the interaction with solvent. Considering the dynamics of the structures, this is not unexpected, however when analyzing static structures both crystallization conditions for the experimental ones or energy minimization for theoretical ones tend to reinforce HBs including the less favorable ones, hence leading to artifactual results. Although no clear trend can be identified for favorable HB properties, there is a slight decrease in solvent accessible surface for those favorable HB formed in residues side chains (Table 1). For the backbone, fixed HBs are mainly placed in beta sheet regions, whereas favorable HBs are mainly placed in helix secondary structures. In general, there are a higher percentage of non-fixed bonds found in regions without defined secondary structures.

The dynamic analysis of HBs can be exploited also to gather kinetic information. Table 2 summarizes the effect of water molecules in MoDEL protein hydrogen bonds. In general, the results show that water molecules play an important role in the breaking of HB. The majority of atoms originally involved in protein HB interacts with water molecules after the bond is broken. In all of HB types, from 20 to 50% of the cases, both atoms are interacting with water molecules, and for the remaining, at least one of the atoms does. However, the tendency to interact with the same water molecule, thus becoming a water bridge, is lower, reaching only 10% in the best case (COO-). The probability of having the HB broken without any water molecule interfering is really low, with an average around 5%.

When looking to the mechanics of the loss of HB, the shortest path with higher probability has mainly one single step, involving water molecules either in one or in both of the atoms forming the bond. In the majority of the cases, there is no water bridge formed though. However, considering not just the highest probable shortest path, but a number of more probable paths, the presence of water bridges increase (Table 3). Remarkably, HB formed between COO- and NH3 groups show the higher percentage of water bridges. As a general trend, the participation of water molecules as water bridges is more significant in less favorable HBs.

3.8.- Case Examples

The systematic methodology used in this study has been applied to a couple of examples, with the purpose of both proving the generality of the pipeline and its capacity to describe important features where water molecules are involved. Proteins used in these examples were chosen as representatives of two big and relevant families: proteins with leucine-rich repeated motifs (LRR) (Internalin B), and proteins with membrane channel functions (Aquaporin-0). En both cases water besides of its effect as solvent, play a key and well known role in the protein functionality.

3.8.1.- Internalin B

The major function of LRR fragments is known to be the facilitation of protein-protein recognition processes (36), and proteins containing this motif are widely studied nowadays due to their role in diseases like Parkinsonism (37). Water molecules forming bridging hydrogen bonds between adjacent repeats in LRR proteins are organized in distinct spines along the fragment, that serve as important structural elements. Internalin B (InlB) is a bacterial surface protein found in *Listeria monocytogenes* that helps to activate the bacterium's phagocytotic defense against the mammalian immune system. The leucine-rich repeat motif of InlB, which is common to all proteins of the internalin family, contains a series of stacked loops that are held together by water molecules bridging the peptide chains. These waters are organized into three distinct "spines" through the stack (Fig. 8, with permissions of ref. 38) and are an integral part of the secondary structure (38).

Mean residence times of 1st shell water molecules (according to their RMSd computed in the last ns of the simulations) are only able to recover half of the water molecules forming the different spines. The reason behind that is again the high mobility of hydration water. Even in this extreme case where this set of waters is maintaining the structure of the protein, they continue interchanging with waters belonging to the 2nd shell. This can be demonstrated by reducing the solvent RMSd time window, that allow recovering the remaining water molecules, and thus completing the missing gaps. In figure 8b, the highest MRT found for each water molecule acting as a water bridge in InlB structural positions is indicated (see also table 4).

The high occupancy of the hydration sites is confirmed also by water density analysis. Water density analysis for the complete 10 ns MD simulation reflects perfectly the water molecule spines placed between peptide chains, bridging them together (Fig. 8c). Spine 1 is almost completely reproduced through water density patches. Spines 15 and 13 are also well represented, in spite of being placed in a more mobile region of the protein. A few water density clouds are found having no match with the original x-ray crystallographic waters

Hydrogen bond analysis was computed for all the key residues participating in the secondary structure spines. Results obtained showed clearly a high percentage of water bridges formed between the analyzed residues (Table 4). In these residues where the water bridge percentage is lower, a water molecule is always present in almost one of the two residues involved in the bond. The highly dynamic behavior of hydration water molecules is also present in this study: the number of different water molecules acting as a water bridge is unexpectedly uplifted, suggesting a great interchange of water molecules, in agreement with our previous MRT analysis. For the 1,083 water molecules found acting as a water bridge between protein atoms, the vast majority of them stay less than 100 ps in this position (mean = 104.29, stdev = 304.72), although some of them are clearly out of the mean (Table 5).

3.8.2.- Aquaporin-0 membrane water channels are extremely important for many related diseases (39). From all the protein families acting as membrane channels, Aquaporin family is being recently described as a potential therapeutic target for edema cancer, obesity, brain injury and glaucoma (40,41). The lens-specific water pore aquaporin-0 (AQP0) is a member of the aquaporin family, a family known to form pores that are highly selective for water (42). AQP0 proteins, as all members of the AQP family, are present in the membrane as tetramers. But, contrarily to the ion channels, AQP channel for water permeability does not reside in the center of the tetramer, but in the center of each of the AQP0 monomers forming the tetramer (43). AQP0 water permeability at neutral pH is approximately 40 times lower than that of its closely related family member AQP1 (42). AQP0 pore has seen to be much more constricted than that in AQP1 with many residues substituted with larger and more hydrophobic ones. A couple of constriction sites (CSI and CSII) at the entrance and exit regions of the AQP0 pore are shown to adopt conformations that make the pore too

narrow for water permeation. In addition to these constriction sites, a phenolic barrier is also formed by a Tyr residue not seen in the other known aquaporin structures.

Residues lining the pore in AQP0 are larger and more hydrophobic than the ones forming the pore in the rest of aquaporin family. That makes this AQP0 particular pore narrower, hindering the water molecules permeability. Applying our pipeline to this system, we could clearly find the water channel profile (Fig. 9a, with permissions of ref. 42). Although no structural fixed water molecules could be identified inside the pore, the majority of the 2nd shell water molecules identified (those with lower RMSd) are placed into the water channel. Mean residence times for these 2nd shell molecules show up a couple of interesting stacked water molecules (water cluster 1, see table 6 and inline image of fig. 9b). Both of them are hydrogen bonded at different time windows to the so-called 'phenolic barrier' formed by tyrosine residue 24, thus explaining the reduction in water permeability rate. Two other water molecules have residence times greater than 4ns (water cluster 2, see table 6 and inline image of fig. 9b). They are bonded to residues belonging to the constriction site I (CSI) region (Arg 187 and Asn 184). Another two water molecules have residence times greater than 1 ns (water cluster 3, see table 6 and inline image of fig. 9b), and they are placed just below the CSI. All these results are consistent with the studies claiming that the narrower pore of AQP0, and more specifically the region were the CSI is placed acts as a blockage for the water molecules channel. Our LRT-based algorithm allow us to obtain also the residues (and atoms) involved in the protein-water bonds formed inside the channel, defining its shape perfectly (Fig. 9b). Water density studies show that although we could just identify a small number of water molecules inside the channel with mean residence times up to 1 ns, the positions of the molecules inside the channel are well-defined, again another example of the fast interchanging rate of water molecules.

4.-Conclusions

Water is a key factor in the understanding of protein structure and dynamics, but its study is difficult due to the fast time scale of water rearrangements. Most solvent analysis published to date provided information on average properties of water around proteins (17-20), escaping from the individual analysis of solvent waters around selected proteins. We present here a massive study where more than 16 million of water molecules have been followed for around 1,800 proteins representative of the entire Protein Data Bank. Our study provided new details of the dynamics of protein-water interaction. "Static interactions" are negligible, and even for the most stable water-interaction sites there is a quite fast interchange. Not even water molecules trapped in cavities remain attached to a specific atom for long periods of time. Water molecules retain a high degree of dynamics even in the periods that are attached to the protein surface, as depicted by the fact that the HB acceptor role where most internal mobility of water is maintained is preferred. Only 8% of structural waters, as derived from X-Ray crystallography match low mobility water molecules. Our study also shows water as playing an active role in the modulation of protein interactions. Particularly, hydrogen bonds largely depend on the presence of water, participating directly in the

kinetics of their formation or destruction. Considering the effect of solvent, the real strength of protein hydrogen bonds should be also reconsidered, as about 40% of hydrogen bonds that can be obtained from the usual trajectory analysis procedures are not more favorable that the corresponding interactions to water. Overall, solvent-protein interaction, as obtained from molecular dynamics simulations, show a very dynamic picture where a 1st shell of water molecules do interact closely with protein residues, but at a high exchange rate.

5.-References

(1) Levvit, M., and Park, B.H. (1993) Water: now you see it, now you don't. *Structure* 15, (1) 223-226.

(2) Kuntz, I., D., Kauzmann, W. (1974) Hydration of proteins and polypeptides. *Adv. Protein Chem.,* 28, 239-345.

(3) Fogarty, A.,C., Duboué-Dijon, E., Sterpone, F., Hynesac J. T., Laage, D. (2013) Biomolecular Hydration Dynamics: a jump model perspective. *Chem. Soc. Rev., 42, 5672- 5683*.

(4) Ball, P. (2008) Water as an active constituent in cell biology. *Chem. Rev.* 108, 74-108.

(5) Halle, B. (2004) Protein Hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc. London B,* 359, 1207-1224.

(6) Bagchi, B. (2005) Water dynamics in the hydration layer around proteins and micelles. *Chem. Rev.*, 105, 3197-3219.

(7) Grant, E. H. (1965) The structure of water neighboring proteins, peptides and amino acids as deduced from dielectric measurements. *Ann. N. Y. Acad. Sci.* 125, 418-427.

(8) Settles, M., Doster, W. (1996) Anomalous diffusion of adsorbed water: a neutron scattering study of hydrated myoglobin. *Faraday Discuss.* 103, 269-279.

(9) Mattea, C., Qvist, J., Halle, B. (2008) Dynamics at the protein-water interface from 17O spin relaxation in deeply supercooled solutions. *Biophys. J.* 95 (6), 2951-2963.

(10) Denisov, V., Halle, B. (1996) Protein hydration dynamics in aqueous solution. *Faraday Discuss.* 103, 227-244.

(11) Zhong, D., Kumar Pal, S., Zewail, A.H. (2011) Biological water: a critique. *Chemical Physics Letters* 503, 1-11.

(12) Zhang, L., Wang, L., Kao, Y., Qiu, Y.Y., Okobiah, O., Zhong, D. (2007) Mapping hydration dynamics around a protein surface. *PNAS* 104 (47), 18461-18466.

(13) Mazur, K., Heisler, I.A., Meech, S.R. (2012) Water dynamics at protein interfaces: ultrafast optical Kerr effect study. *J. Phys. Chem. A.,* 116 (11), 2678-2685.

(14) King, J.T., Kubarych, K.J. (2012) Site-specific coupling of hydration water and protein flexibility studied in solution with ultrafast 2D-IR Spectroscopy. *JACS* 134 (45), 18705-18712.

(15) McCammon, J.A., Gelin, B.R., Karplus, M. (1977) Dynamics of folded proteins. *Nature* 267, 585-590.

(16) Brooks, C.L., III, Karplus, M. Pettitt, B. M. (1987) Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics. *Cambridge: Cambridge University Press.*

(17) Pal, S., Bandyopadhyay, S. (2013) Importance of protein conformational motions and electrostatic anchoring sites on the dynamics and hydrogen bond properties of hydration water. *Langmuir* 29, 1162-1173.

(18) Pizzitutti, F., Marchi, M., Sterpone, F., Rossky, P. J. (2007) How protein surfaces induce anomalous dynamics of hydration water. *J. Phys. Chem. B* 111, 7584-7590.

(19) Bizzarri, A.R., Cannistraro, S. (2002) Molecular Dynamics of water at the protein-solvent interface. *J. Phys. Chem. B* 106 (26), 6617-6633.

(20) Schröder, C., Rudas, T., Boresch, S., Steinhauser, O. (2006) Simulation studies of the protein-water interface I. Properties at the molecular resolution. *J. Chem. Phys.* 124, 234907, 1-18.

(21) Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., Repchevsky, D., Gelpí, J.L., Orozco, M. (2010) MoDEL (Molecular Dynamics Extended Library): A database of atomistic molecular dynamics trajectories. *Structure*, 18, 1399-1409.

(22) Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., Gelpí, J.L., Orozco, M. (2007) A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA* 104, 796-801.

(23) Candotti, M., Pérez, A., Ferrer-Costa, C., Rueda, M., Meyer, T., Gelpí, J.L, Orozco, M. (2013) Exploring early stages of the chemical unfolding of proteins at the proteome scale. *PLOS Comput. Biol.* 9 (12) E1003393.

(24) Camps J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Gelpí, J.L., Orozco, M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* 25(13) 1709-1710.

(25) Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J. (2005) The Amber biomolecular simulation programs. *J.Comput.Chem.* 26 (16), 1668-1688.

(26) Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput., 4, 435-447.*

(27) Morreale, A., de la Cruz, X., Meyer, T., Gelpí, J.L., Luque, F.J., Orozco, M. (2005) Partition of protein solvation into group contributions from molecular dynamics simulations. *Proteins*, 58(1), 101-109.
(28) Nucci, N. V., Pometum, M. S., Wand, A.J. (2011) Site-resolved measurement of water-protein interactions by solution NMR. *Nature structural & Molecular Biology,* 18, 245-249.
(29) Schoenborn, B.P., García, A.E., Knott, R. (1995) Hydration in protein crystallography. *Prog. Biophys. Mol. Biol.* 64, 105.
(30) Matsuoka, D., Nakasako, M. (2009) Probability distributions of hydration water molecules around polar protein atoms obtained by a database analysis. *J. Phys. Chem. B.* 113(32), 11274-11292.
(31) Pekka, M., Nilsson, L. (2001) Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A.* 105, 9954-9960.
(32) Ebbinghaus, S., Joong Kim, S., Heyden, M., Yu, X., Heugen, U., Gruebele, M., Leitner, D.M., Havenith, M. (2007) An extended dynamical hydration shell around proteins. *P.N.A.S.* 20749-20752.
(33) Hertz, H.G. (1973) Nuclear magnetic relaxation spectroscopy. *Water: A comprehensive treatise.* Vol 3, F. Franks editor. Plenum Press, New York, 301-395.
(34) Guillot, B. (2002) A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids.* 101/1-3, 219-260.
(35) Price, W.S., Ide, H., Arata, Y. (1999) Self-Diffusion of Supercooled Water to 238 K using PGSE NMR Diffusion Measurements. *J. Phys. Chem. A* 103, 448-450.
(36) Kobe, B., Kajava, A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology,* 11(6), 725-732.
(37) Mills, R.D., Mulhern, T.D., Liu, F., Culvenor, J.G., Cheng, H-C. Prediction of the Repeat Domain Structures and impact of Parkinsonism-associated variations on structure and function of all functional domains of Leucine-rich Repeat Kinase 2 (LRRK2). (2014) *Human Mutation*, 35(4), 395-412.
(38) Marino, M., Braun, L., Cossart, P., Ghosh, P. (1999) Structure of the InlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes. *Mol. Cell* 4, 1063.
(39) Celesia, G.G. Disorders of membrane channels or channelopathies. (2001) *Clinical Neurophysiology,* 112(1), 2-18.
(40) Verkman, A.S., Anderson, M.O., Papadopoulos, M.C. (2014) Aquaporins: important but elusive drug targets. *Nature Reviews Drug Discovery,* 13, 259-277.
(41) Ribatti, D., Ranieri, G., Annese, T., Nico, B. (2014) Aquaporins in cancer. *Biochimica et Biophysica Acta (BBA)- General Subjects,* 1840(5), 1550-1553.
(42) Gonen, T., Sliz, P., Kistler, J., Cheng, Y., Walz, T. (2004) Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature* 429, 193-197.
(43) Ozu, M., Alvarez, H.A., McCarthy, A.N., Grigera, J.R., Chara, O. (2013) Molecular dynamics of water in the neighborhood of aquaporins. *Eur. Biophys. J.* 42, 223-239.

Table 1. Hydrogen-Bond energy classification: Three different categories are defined according to the bond energy: *Fixed HB* ($\Delta G \leq -5.45$ Kcal/mol), *Favorable* HB ($-5.45 < \Delta G \leq -2$ Kcal/mol), *Neutral* HB ($-2 \leq \Delta G < 0$ Kcal/mol), and *Unfavorable* HB ($\Delta G \geq 0$ Kcal/mol). ASA stands for Accessible Surface Area, with units in $Å^2$. SS stands for Secondary Structure, 1 is the first atom of the HB pair, whereas 2 is the second. The secondary structure elements are identified by H: Helix, B: Beta sheet and C: Coil. Only the most representative HB atom pairs are shown in the table.
(*)Values presented as % of states from a total of 10000 snapshots taken from 10ns-length simulations (1ps resolution).

Pair	Category	NumHB	%HB	ASA1	ASA2	SS1-H*	SS1-B*	SS1-C*	SS2-H*	SS2-B*	SS2-C*
BCK-O BCK-N	Fixed	3950	22.32	14.25	14.19	29.11	56.33	14.56	33.27	52.35	14.38
	Fav.	4718	26.66	31.69	36.26	54.81	10.58	34.61	46.74	8.99	44.28
	Neutral	4469	25.25	38.59	40.71	38.38	10.38	51.24	33.07	7.72	59.21
	Unfav.	4560	25.77	45.57	43.95	21.32	10.79	67.89	14.89	11.49	73.62
BCK-O OH	Fixed	48	0.94	12.28	4.59	45.83	8.33	45.83	50.00	20.83	29.17
	Fav.	252	4.95	25.13	20.86	44.84	17.46	37.70	44.44	23.41	32.14
	Neutral	794	15.59	30.61	27.97	38.79	20.28	40.93	39.17	25.06	35.77
	Unfav.	3998	78.52	40.84	38.23	20.36	21.16	58.48	18.83	25.41	55.75
COO- OH	Fixed	48	3.74	9.18	6.88	37.50	16.67	45.83	39.58	12.50	47.92
	Fav.	97	7.55	24.53	21.00	23.71	20.62	55.67	26.80	16.49	56.70
	Neutral	237	18.44	33.25	32.80	33.76	15.61	50.63	25.32	18.14	56.54
	Unfav.	903	70.27	45.50	39.12	30.23	16.94	52.82	23.15	22.81	54.04
CO- BCK-N	Fixed	51	3.77	12.04	19.16	19.61	15.69	64.71	11.76	15.69	72.55
	Fav.	186	13.74	33.06	40.92	15.05	24.73	60.22	23.12	12.37	64.52
	Neutral	375	27.70	41.87	42.95	9.07	21.07	69.87	18.13	15.20	66.67
	Unfav.	742	54.80	48.59	45.88	17.39	15.50	67.12	14.56	15.09	70.35
COO- NH3	Fixed	12	0.80	7.91	10.47	58.33	16.67	25.00	41.67	25.00	33.33
	Fav.	58	3.86	27.15	29.91	53.45	12.07	34.48	55.17	10.34	34.48
	Neutral	348	23.14	41.93	44.65	40.80	20.69	38.51	37.36	26.15	36.49
	Unfav.	1086	72.21	50.63	51.83	36.28	14.36	49.36	40.52	19.80	39.69
NH2 OH	Fixed	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Fav.	9	1.18	9.63	8.58	33.33	0.00	66.67	33.33	0.00	22.22
	Neutral	60	7.86	19.77	19.55	38.33	23.33	38.33	31.67	33.33	35.00
	Unfav.	694	90.96	42.37	38.12	29.11	19.88	51.01	22.48	22.05	55.48

Table 2. MoDEL Protein HB Water interferences. Statistics of water molecules interfering in HB protein atom pairs computed in 669 simulations. Only HB having a broken period of time of at least 1ns (from 10ns simulations) are used in the study. Hydrogen Bonds have been divided in atom pair types. *Number of HBs* corresponds to the number of HB of a particular type found in the complete set of simulations. *Hydrogen Bond* corresponds to the percentage of simulation time where the HB is formed; When the HB is broken, five different situations can be found: *Water bridge* is defined when a particular water molecule is interacting with both protein atoms at the same time; *Interfering water for atom 1/2* corresponds to HB protein atoms interacting with at least one water molecule; and *Interfering water for both atoms* is defined when both HB atoms are interacting with different water molecules. *Without water* corresponds to states where the HB is broken but no water molecules are interfering with protein atoms.
(*) Values presented as % of states from a total of 10000 snapshots taken from 10ns-length simulations (1ps resolution).

HB types		Number of HBs	Hydrogen Bond*	Water Bridge*	Interfering Water Atom1*	Interfering Water Atom2*	Interfering Water Both Atoms*	Without Water*
Atom 1	**Atom 2**							
Bck-O	Bck-N	7836	43.26	4.12	21.71	2.75	22.37	5.77
	Hydroxyl	2193	36.81	5.25	7.90	14.27	29.22	6.53
	COO-	195	41.15	**11.08**	6.18	3.87	37.18	0.50
	CO-	86	34.86	5.24	9.05	6.31	38.34	6.19
	NH3	1078	33.18	6.95	3.50	10.82	44.49	1.05
	NH2	1927	36.83	4.37	5.44	12.87	37.07	3.41
	Guanidine	1531	37.47	8.08	1.20	10.74	41.96	0.54
	Rings	441	37.12	5.95	7.02	10.58	34.82	4.48
Bck-N	Hydroxyl	2724	40.58	1.36	2.47	30.62	18.54	6.42
	COO-	1441	41.27	**6.97**	1.37	24.12	25.38	0.88
	CO-	1228	41.03	2.93	3.52	23.85	23.90	4.77
	NH2	511	32.96	1.85	4.17	31.45	20.93	8.62
	Rings	277	42.50	2.47	2.38	26.88	22.72	3.04
Hydroxyl	COO-	775	39.84	**11.69**	6.42	2.91	38.27	0.86
	CO-	259	36.34	6.87	6.92	7.94	37.11	4.79
	NH3	254	32.75	5.55	3.82	6.97	50.15	0.74
	NH2	716	34.88	4.20	8.76	8.27	39.41	4.47
	Rings	187	38.01	5.60	8.14	7.56	34.85	5.83
	Guanidine	368	35.20	6.71	1.19	9.27	46.29	1.33
COO-	NH3	1508	37.32	**10.36**	3.07	4.04	44.77	0.43
	NH2	601	37.01	**10.05**	5.20	3.90	43.19	0.63
	Rings	195	41.15	**11.09**	6.18	3.87	37.19	0.50
	Guanidine	1152	45.09	**10.21**	0.67	4.82	38.84	0.35
CO-	NH3	258	32.77	7.94	3.79	4.93	50.06	0.49
	NH2	264	34.15	4.71	6.63	5.42	46.15	2.92
	Rings	86	34.86	5.24	9.05	6.31	38.33	6.19
	Guanidine	302	35.29	8.71	1.28	6.40	47.75	0.55
NH3	NH2	95	25.89	4.24	11.22	3.32	54.25	1.07
	Rings	31	33.30	7.47	2.26	2.35	54.17	0.44
NH2	Rings	97	33.73	5.71	10.43	4.55	42.29	3.28
	Guanidine	229	33.72	3.10	1.36	8.27	52.69	0.83
Rings	Guanidine	48	34.08	10.71	2.86	1.70	50.57	0.06

Table 3. Percentage of **water bridges** in the process of breaking a HB.

HB type	Favorable HB WB (%)	Unfavorable HB WB (%)
BCK_O – BCK_N	12.71	21.06
BCK_O – OH	18.21	49.70
BCK_O – NH2	10.97	47.86
OH – BCK_N	11.33	19.58
COO- – BCK_N	**34.29**	**44.44**
CO- – BCK_N	7.36	34.75
NH2 – BCK_N	2.81	24.44

Table 4. Internalin B Water Bridges. Statistics of water molecules interfering in Internalin B hydrogen bonded protein atom pairs. Hydrogen Bond atom pairs shown together with water bridges and water percentage. *Water bridge* is defined when a particular water molecule is interacting with both protein atoms at the same time. *Water* is defined when at least one of the protein atoms involved in the HB is interacting with a water molecule, breaking the protein HB. *# Bridge Waters* is the number of different water molecules found acting as a water bridge for a particular protein HB.
(*)Values presented as % of states from a total of 10000 snapshots taken from 10ns-length simulations (1ps resolution).

	HB residues		Water Bridge*	Waters*	# Bridge Waters
Spine 1	LEU 75_O	ASN 99_N	85.10	14.63	2
	ILE 78_O	ASN 99_N	81.07	18.91	1
	PRO 98_O	ASN 121_N	58.56	41.26	35
	VAL 100_O	ASN 121_N	29.91	70.06	4
	LYS 120_O	LYS 143_N	55.34	43.21	48
	LEU 122_O	LYS 143_N	64.54	35.46	2
	LEU 166_O	LYS 187_N	59.03	26.41	5
	PRO 164_O	LYS 187_N	62.82	34.44	62
	LEU 188_O	LYS 209_N	79.85	19.74	4
	THR 186_O	LYS 209_N	85.89	13.93	66
	LEU 210_O	ASN 231_N	26.30	64.90	5
	THR 208_O	ASN 231_N	15.81	83.69	62
Spines 15/13	VAL 91_N	ASP 112_O	86.16	13.62	20
	LYS 89_O	THR 111_N	85.97	13.99	108
	LYS 89_O	ASP 112_N	78.33	21.64	12
	ILE 113_N	ASP 134_O	25.58	73.93	39
	THR 111_O	LYS 133_N	87.16	10.34	12
	THR 111_O	ASP 134_N	68.70	29.30	11
	LEU 135_N	ASP 156_O	53.32	46.17	61
	LYS 133_O	SER 155_N	74.80	25.13	174
	LYS 133_O	ASP 156_N	43.55	56.32	141
	ILE 157_N	ASP 178_O	78.64	21.24	16
	SER 155_O	THR 177_N	87.43	12.27	96
	SER 155_O	ASP 178_N	75.42	24.27	27
	ILE 179_N	ASP 200_O	48.99	50.79	156
	THR 177_O	SER 199_N	88.34	11.62	135
	THR 177_O	ASP 200_N	71.49	28.45	53
	ILE 201_N	ASP 222_O	29.32	70.00	60
	SER 199_O	SER 221_N	79.62	20.07	155
	SER 199_O	ASP 222_N	69.72	30.03	82

Table 5. Internalin B Water Bridges MRTs. Highest Mean Residence Time found for water molecules placed in the Internalin B secondary structure spines. Residence time shown for MD water molecules interacting with *backbone N-H atom* from residue *Residue Id.* MRT values shown in ps.

	Residue Id	MRT
Spine 1	ASN 99	4839
	ASN 121	599
	LYS 143	4381
	LYS 187	1927
	LYS 209	3788
	ASN 231	1225
Spine 15/13	VAL 91	1281
	THR 111	837
	ILE 113	542
	LYS 133	1910
	LEU 135	367
	SER 155	413
	ILE 157	3900
	THR 177	1001
	ILE 179	510
	SER 199	581
	ILE 201	906
	ASP 222	241
	ALA 225	722

Table 6. Aquaporin-0 channel water molecules MRTs. Highest Mean Residence Time found for water molecules placed inside the AQP0 channel. Residence time shown for MD water molecule residue number *MD Water* interacting with atom *Atom Name* from residue *Residue Id.* MRT values shown in ps.

Residue Id	Atom Name	MD Water	MRT
TYR 24	OH	278 (Cluster 1)	1402
TYR 24	OH	12303 (Cluster 1)	3455
ASN 184	1HD2	286 (Cluster 2)	4909
ARG 187	1HH1	11525 (Cluster 2)	4418
ASN 68	2HD2	284 (Cluster 3)	2463
PHE 141	HZ	22226 (Cluster 3)	1704
ALA-65	C-O	18164 (Cluster 3)	272
ARG 156	2HH1	9903 (Cluster 3)	234

Legends to Figures

Fig1. Water molecules RMSd distributions in different time windows. From top-left to bottom-right: 200ps, 500ps, 1ns and 2ns, respectively. Gradual disappearing of the peak centered at 7Å can be observed, while the first peak corresponding to fixed water molecules (inset plots) is still visible even in 2ns time-window.

Fig2. Water molecules RMSd distribution in 500ps time-window. a) Approximation of RMSd distribution to a sum of Gaussian-like functions corresponding to 1st shell according to water molecules RMSd (*), 2nd shell (**) and bulk water (***). b) Detailed view of 1st shell, representing distribution of distances protein-water, taking into account only the contact distance between a water molecule and its most prevalent contact ("preferred" protein atom) during the simulation. Correlation between these distances and water molecules RMSd is also shown, where both populations corresponding to those water molecules acting as donors and those acting as acceptors can be clearly distinguished. c) Detailed view of 2nd RMSd shell, representing distribution of protein-water distances, taking into account all the contact distances between water molecules and their nearest protein atoms during MD simulation (water promiscuity). 2nd shell can be viewed as a sum of water molecules attached to the protein and a cloud of water molecules surrounding those ones.

Fig3. Atom/Residue type water preferences. a) 1st (black columns) and 2nd (white columns) shell water residue preferences. Base lines representing the average of all residue types preferences are plotted as references. Units are in contacs/ps*Å^2. b) 1st (black columns) and 2nd (white columns) shell water atom preferences. Base lines representing the average of all atom types preferences are plotted as references. Units are in contacs/ps*Å^2.

Fig4. Mean residence times (MRT) of water molecules attached to a particular protein residue. a) MRT for 1st shell water molecules in a 1ns time-window. b) MRT for 1st shell water molecules in a 5ns time-window. c) MRT for water molecules trapped in internal cavities. d) MRT for 2nd shell water molecules in a 1ns time-window.

Fig5. Correlation between X-ray determined water molecules and the ones found to be fixed using our RMSd approach. Matching was defined by a distance criterion (cutoff depending on the particular RMSd shell, from 1.2 Å to 6 Å). Plots represent the percentage of water molecules belonging to our RMSd-defined shells having a corresponding crystallographic water molecule against a) total number of water molecules of a particular RMSd shell and b) total number of crystallographic water molecules.

Fig6. Radial Distribution Function (RDF) for different protein atom groups. a) Representation of RDF plots for all the studied protein groups (see Materials & Methods). Three different clusters can be extracted from this general plot. b) Representation of RDF for cluster 1, formed by Backbone-N, COO- and NH3 groups. c) Representation of RDF for cluster 2, formed by Backbone-C/O, Hydroxyl, CO-, NH2 and Guanidine groups. d) Representation of RDF for cluster 3, formed by Backbone-CA, Hydrophobic, Thiol and Rings groups.

Fig7. Mean Square Displacements (MSD) computed in last nanosecond of MoDEL 10ns-length simulations. Different water subsets are plotted: 1st shell, 2nd shell, and bulk water, with free water MD as a reference. Diffusion coefficients were extracted from the slope values of each MSD line.

Fig8. HB type classification according to our thermodynamic analysis. Only 40% of the hydrogen bonds studied are found to be energetically favorable ($\Delta G < -2$ Kcal/mol).

Fig9. Internalin B (InlB) as a case study. a) Representation of water spines (red balls) as an integral part of the protein secondary structure (image courtesy of ref. 29). b) MRT found for water molecules acting as peptide chain bridges. c) Representation of water density clouds found with the MD analysis, together with the water spines as a reference.

Fig10. Aquaporin 0 (AQP0) as a case study. a) Representation of the internal channel of AQP0 protein, with the narrow constriction site I (CSI) represented in red (image courtesy of ref. 30). b) Water channel profile in AQP0 found using our approach, with detailed screenshots of the so-called 'phenolic barrier' and CSI.

Figure 1

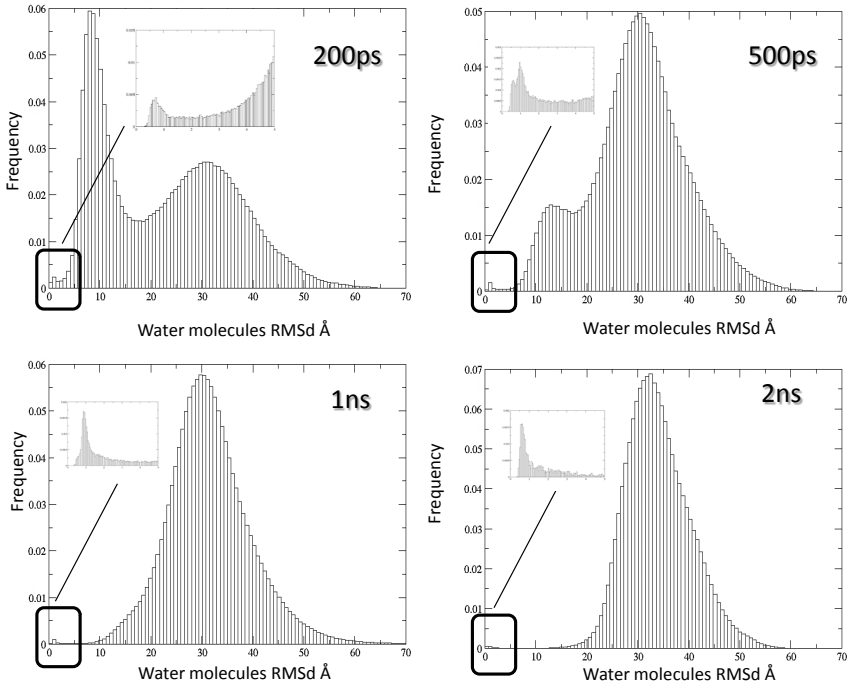


Figure 2

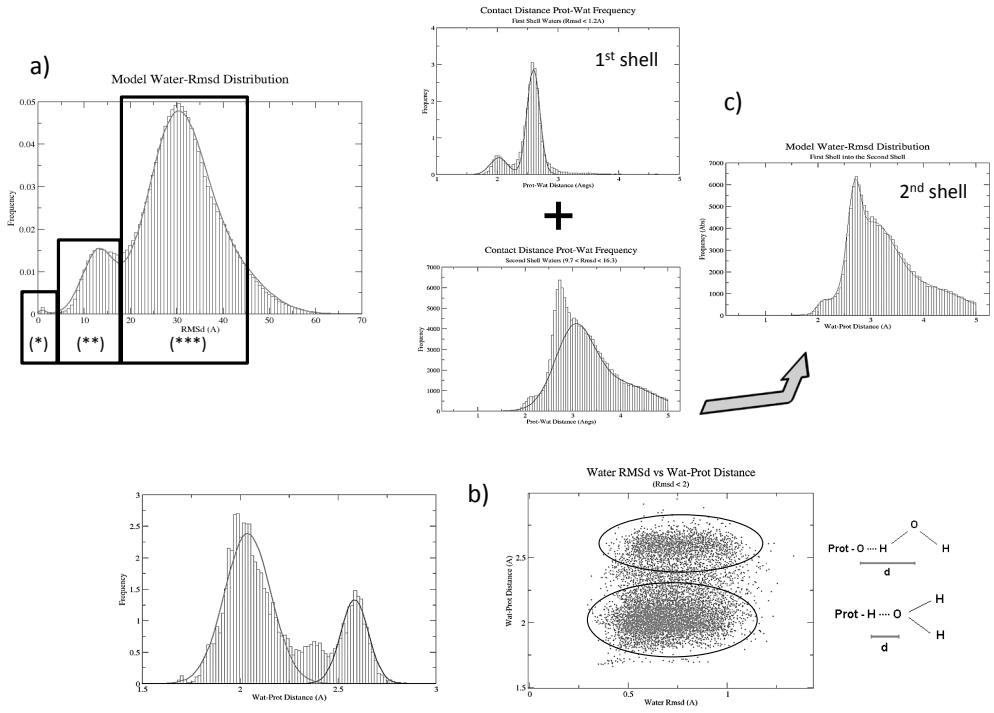


Figure 3

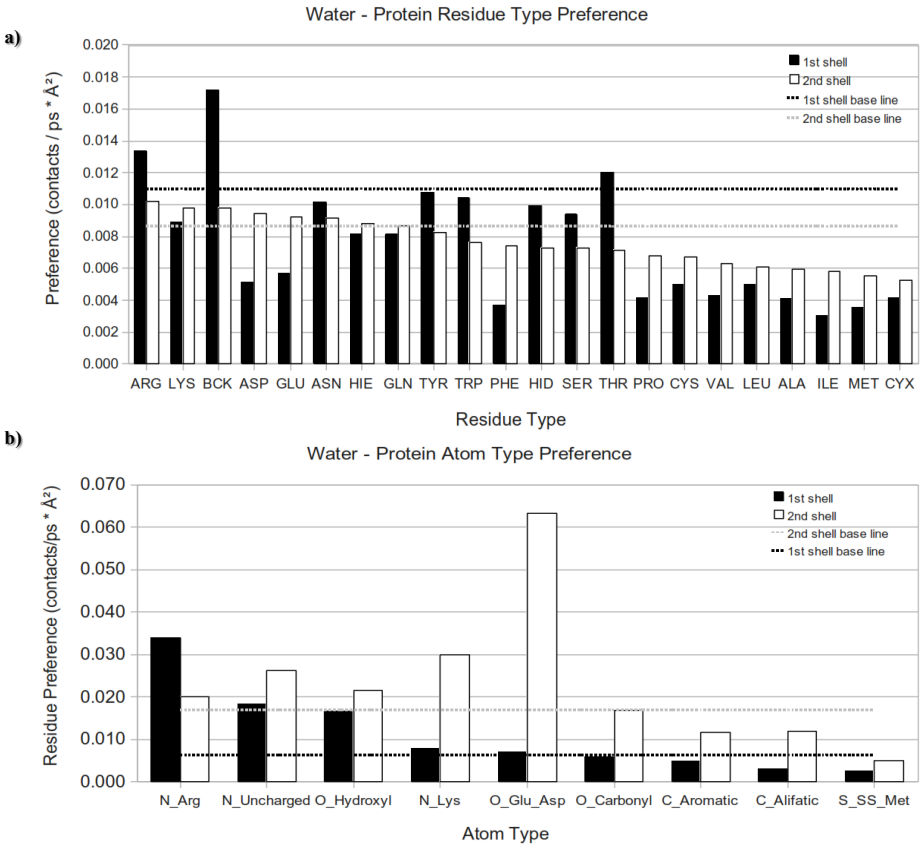


Figure 4

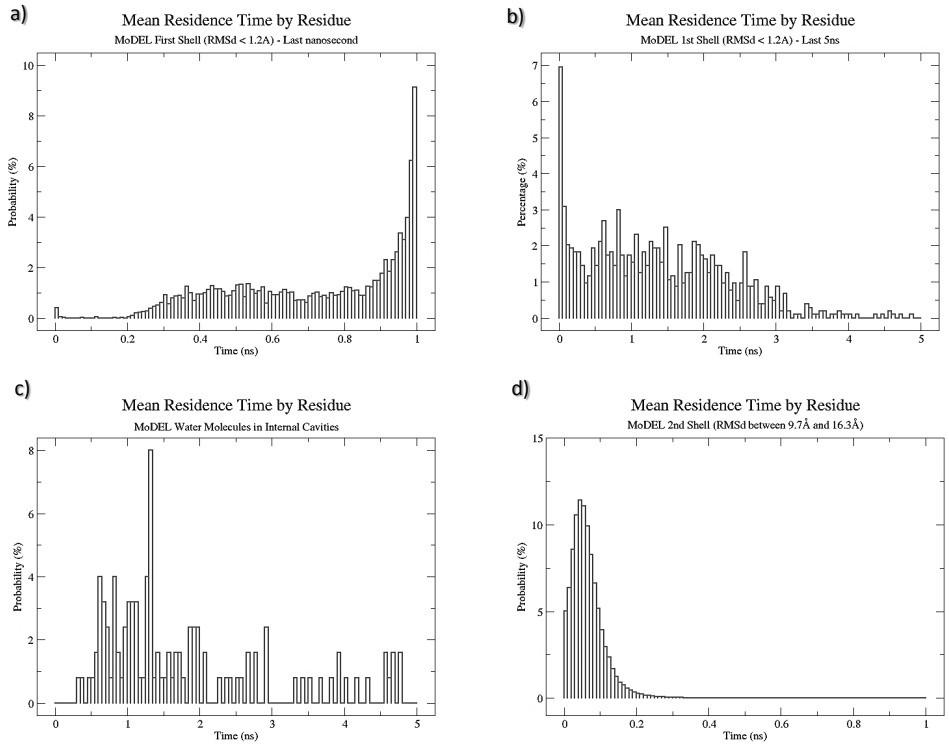


Figure 5

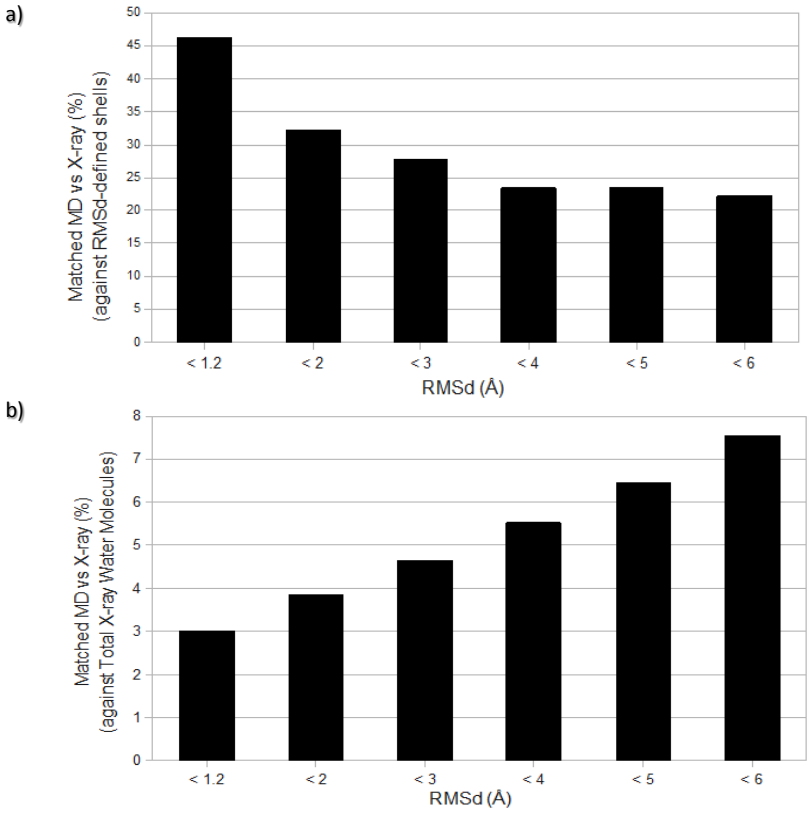


Figure 6

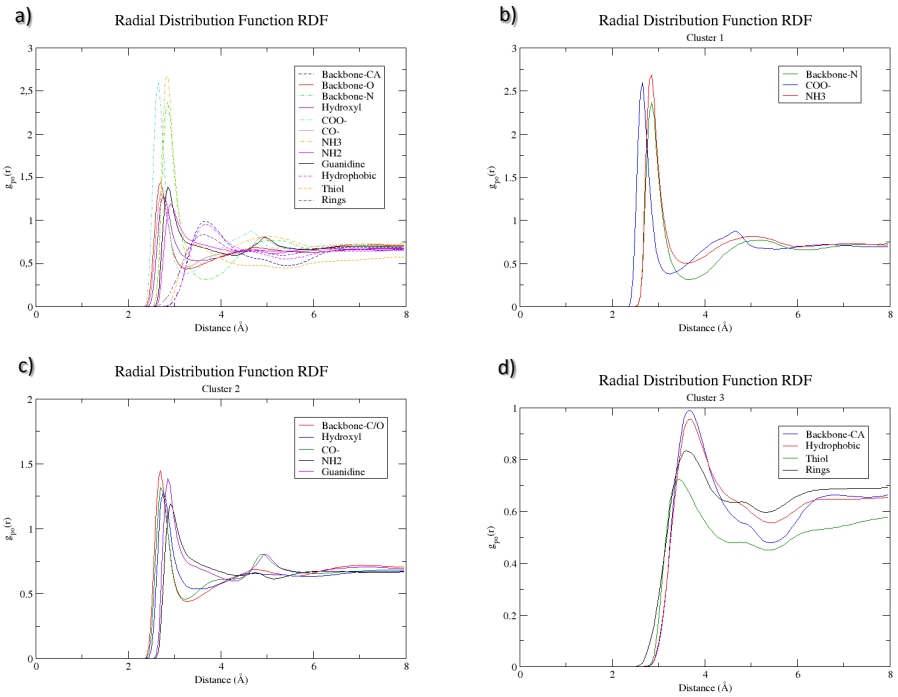


Figure 7

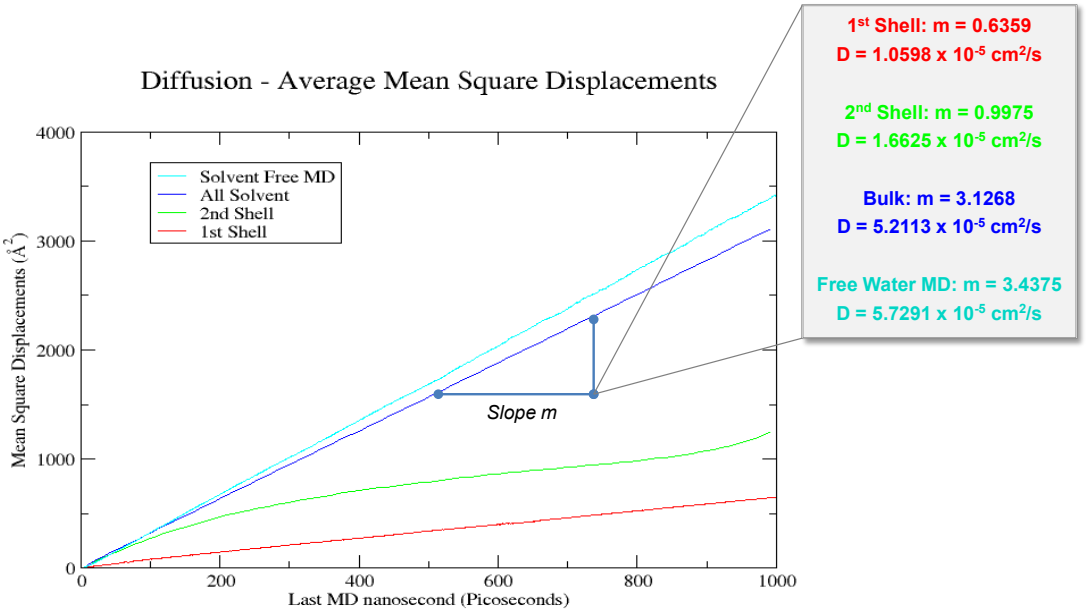


Figure8

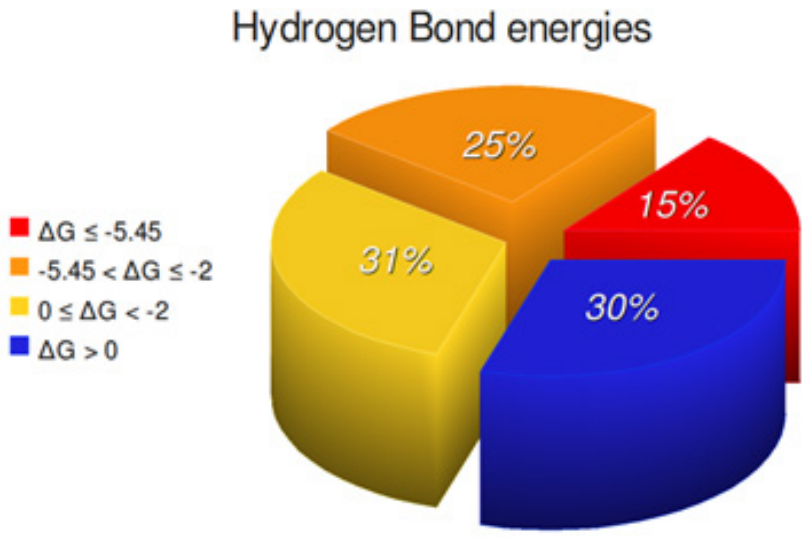


Figure9

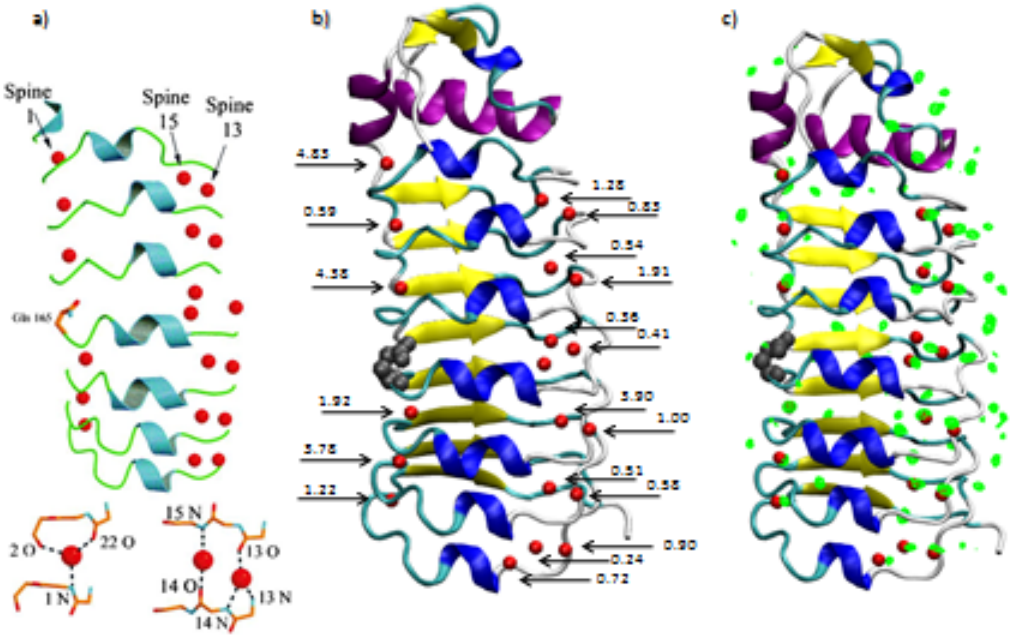
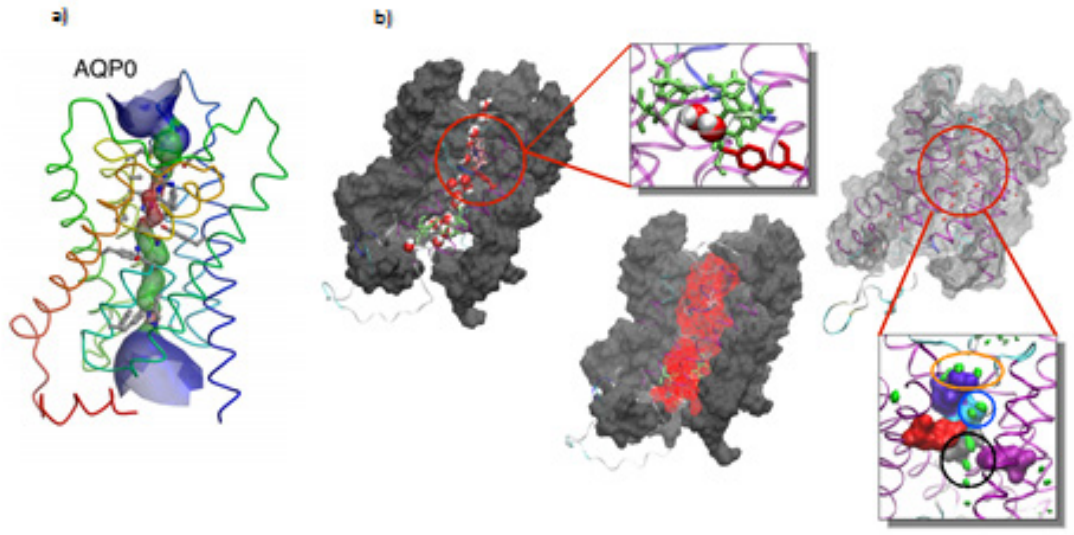


Figure10



Water-omics: high throughput analysis of protein-solvent interaction from molecular dynamics simulations

Supplementary Material

Adam Hospital[1,2,3,4], Modesto Orozco[1,2,3,4,6*] and Josep Lluís Gelpí[1,2,3,5,6*]

[1] Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain

[2] Joint IRB-BSC Program on Computational Biology, Barcelona, Spain

[3] Barcelona Supercomputing Center, Barcelona , Spain[4] Structural Bioinformatics Node, National Institute of Bioinformatics, Barcelona, Spain

[5] Computational Bioinformatics Node, National Institute of Bioinformatics, Barcelona, Spain

[6] Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain

* Correspondence to M.Orozco modesto.orozco@irbbarcelona.org or Josep Lluis Gelpí gelpi@ub.edu

Materials & Methods

- *MoDEL Database* - http://mmb.irbbarcelona.org/MoDEL

Molecular Dynamics Extended Library (MoDEL) is a biosimulation database consisting on 2082 molecular dynamic simulations. The structures were selected from the Protein Data Bank (PDB) (1), getting monomeric soluble structures and selecting one representative protein for every cluster 90 family (i.e. considering only those proteins sharing less than 90% of sequence identity). MD simulations were produced in the NPT ensemble (Temperature: 300K, Pressure: 1atm), using SHAKE (2) to eliminate the vibration of chemical bonds involving hydrogen atoms, thereby allowing the use of a 2-fs integration step. Periodic boundary conditions and Particle Mesh Ewald (3) were used to account for long-range interactions. Simulation lengths were run up to 10ns, with 1ps spacing. The subset of simulations used in this study consists of 1595 proteins simulated with AMBER package (4), PARM99 force field (5) and TIP3P water model (6). Structural water molecules (2% of water displaying the strongest contacts with the protein) were added by iterative CMIP calculations (7), while bulk water was added from pre-equilibrated TIP3P water boxes, filling truncated octahedral simulation cells with a minimum solute-box distance of 12Å.

- *Protein functional groups*

A set of protein functional groups were defined to be used in some of the analysis:

  - Backbone-CA: All backbone alpha carbons.
  - Backbone-C/O: All backbone C and O atoms.
  - Backbone-N: All backbone N atoms.
  - Hydroxyl: Hydroxyl groups from Ser, Thr and Tyr.
  - COO-: Carboxyl groups from Asp, Glu and C-Terminal.
  - CO: Carbonyl groups from Asn and Gln.
  - NH3: Ammonia groups from Lys and N-Terminal.
  - NH2: Amine groups from Asn and Gln.
  - Guanidine: Guanidine group from Arg.
  - Hydrophobic: All hydrophobic atoms.
  - Thiol: Thiol groups from Cys.
  - Rings: Phenyl, side chain rings from Phe, Tyr, and Trp, and His.

- *Water molecules RMSd*

The dynamics of water molecules was studied with an in-house algorithm based on previous studies of linear response theory (LRT) coupled to molecular dynamics simulations (8), that allows to follow all water molecules individually during a MD simulation. The method reports the nearest protein atom (and distance) for each water

molecule at each step of the simulation, as well as water RMSd (Eq.1) and average position in a given trajectory window:

$$RMSd_{wat}(w) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_0)^2}{n}}, \quad (1)$$

being x_i the position of water molecule w in the snapshot i and x_0 the averaged (or initial) position of the water molecule w during the whole simulation.

Both oxygen atom as well as the pair of hydrogen atoms were taken into account in the RMSd calculation, although results considering just oxygen atoms are almost indistinguishable (data not shown).

Solvent shells, hydrogen bond distances or mean residence times can also be extracted from the above indicated LRT algorithm. Combining this information with protein structural properties such as internal/external cavities, solvent accessibility, secondary structure or presence of ligands we could obtain a valuable picture of the hydration dynamics.

- *Hydration Water Analysis*

In the next sections, the set of protein-water hydration analysis done in the study is presented.

  - *Atom/Residue Type Hydration and Water Preference*

In order to study particular preferences for water molecules to certain atom or residue types, we follow each particular water molecule (getting results from a water molecule point of view), tracking its atom/residue preferences:

$$Pref_{resType} = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m} numSnapshotsBound_j}{ASA_i}\right)}{N * n}, \quad (2)$$

where i=*1..n* are all the residues of type *resType* (for example Arg) for all the proteins in the dataset; *j=1..m* are all the atoms of a given residue; ASA_i is the Accessible Surface Area (absolute value computed with NACCESS (9) program) of the corresponding residue side chain or backbone, Note that we only consider exposed residues (*ASA* > 10Å^2); $numSnapshotsBound_j$ is the number of snapshots where atom *j* is in contact with a particular water molecule (being *j* the nearest protein atom for a specific water molecule). Note that we only count waters attached to the same protein residue for at least 10 snapshots; Finally, *N* is the total number of snapshots. Final units are *contacts/ps*Å*^{2}.

Similarly, atom type preferences from a water molecule point of view are calculated from:

$$Pref_{atType} = \frac{\sum_{i=1}^{n}\left(\frac{numSnapshotsBound_i}{ASA_i}\right)}{N * n}, \quad (3)$$

where *i=1..n* are all the atoms of type atType (for example N_Lys) for all the proteins in the dataset; ASA_i is the Accessible Surface Area (absolute value) of the corresponding atom. Note that we only consider exposed atoms ($ASA > 10Å^2$); $numSnapshotsBound_i$ is the number of snapshots where the corresponding atom *j* is in contact with a particular water molecule (being *j* the nearest protein atom for a specific water molecule). Note that we only count waters attached to the same protein atom a minimum of 10 snapshots; Finally, *N* is the total number of snapshots. Units are *contacts/ps*Å²*.

- *MRT: Mean Residence Time*

Water mean residence time is defined as the number of consecutive snapshots where a particular water molecule stays attached ($d_{pw} < 3.5Å$) to the same protein atom. However, a range of MRT variants can be also computed:

- Number of consecutive snapshots where a particular water molecule is attached to the same residue, allowing the water molecule to change contacts between different residue atoms.
- Total number of snapshots where a particular water molecule is attached to the same protein atom, allowing the water molecule to escape for a while and return to the original contact.
- Total number of snapshots where a particular water molecule is attached to the same protein residue, allowing the water molecule to escape for a while and return to the original contact/s.

- *Internal Cavities*

Internal cavities for proteins simulated in MoDEL were computed using CMIP program (7). From the 1595 proteins studied, only 114 were found having stable internal cavities during the dynamics. The pipeline used consist of 200 different runs for simulation (one each 50 ps), from which we keep those trajectories with internal cavities appearing in at least 50% of the time. From them, only those having internal cavities with average volume greater than 100 $Å^3$ were further analyzed.

Water molecules placed inside internal cavities were analyzed in the first and last snapshots of the 10ns-length trajectory, identifying those actually trapped in the cavity, as well as water molecules able to escape from the cavity and external water molecules able to reach this internal cavity (Suppl. Table 1).

- *RDF: Radial Distribution Function*

Radial distribution functions represent water density as a function of distance from the protein:

$$g_{PS}(r) = \frac{1}{\rho\ 4\pi r^2 dr}\sum_S \langle \delta(r - |r_S - r_P|) \rangle \quad , \qquad (5)$$

being ρ the average water density. The summatory corresponds to the number of solvent molecules within a spherical shell of radii between r and r + dr, measured from the reference site *P* in the protein. Index *S* denotes either the water hydrogen or the water oxygen atoms. (10)

RDFs for this study were computed using cpptraj program included in Ambertools package (11,12). Reference sites when computing RDFs for not the entire protein are defined as atoms belonging to one of our protein functional groups (see above). Please note that area occluded by the protein in this case is not taken into account, thus obtaining a density for long range distances lower than 1. Normalization is done by the number of atoms included in the study as well as the number of frames.

- *Water Diffusion*

Water diffusion coefficients can be obtained from molecular dynamics simulations by calculating the time-dependent mean square displacements (MSD). The translational self-diffusion coefficient *D* can be calculated from:

$$6Dt = \langle |\vec{r}(t) - \vec{r}(0)|^2 \rangle \quad t \to \infty, \qquad (6)$$

where $\vec{r}(t)$ is the position vector of the solvent molecule at time t, and the brackets indicate that the average is taken over both the time origins and solvent molecules.

MSDs for this study were computed using ptraj program included in Ambertools package (11,12). It should be noted that due to the simulation conditions, diffusion values reported here have been determined relative to the protein, what may lead to a slight underestimation of true diffusion coefficients.

- *Protein – Solvent Hydrogen Bond dynamics*

Water molecules are known to be an important factor in the formation/destruction of Hydrogen Bonds between exposed protein atom pairs. They can act either as a competitor for the formation of these bonds or as an enhancer, when forming water bridges.

The dynamics of these processes can be obtained studying the surrounding solvent molecules of all the protein atom pairs involved in a hydrogen bond interaction. Information such as number of water molecules attached (d_{pw} < 3.5Å) to the first atom/residue, number of water molecules attached to the other atom/residue HB pair, number of water molecules that can be acting as a water bridge (d_{pw} < 3.5Å for both protein atom/residues involved in the HB), together with dynamic information about the HB itself extracted from the molecular dynamics trajectory give us a complete picture

of the protein-solvent HB dynamics, and allow us to build complex graph systems for a posterior deeper analysis (Suppl. Fig. 8).

HBs are considered to have the form "*acceptor atom – hydrogen donor atom*". Acceptor and donor atoms are specified in Suppl. Table 2. Atom pairs are defined to be forming a HB when the distance between the heavy atoms is below a cutoff of 3.5Å and the angle between the acceptor hydrogen and the donor atom is below a cutoff of 120º. A HB is considered to be formed if it appears just once in the 10000 different snapshots used for the analysis. The program used for finding HBs was ptraj from Ambertools package (11,12).

- *Crystallographic water molecules.*

Crystallographic water molecules have been a field of controversy for many years now (13,14). A significant number of proteins from our MoDEL library have been crystallized together with hydration water molecules (846 out of 1595 from this study). In order to compare MD hydration molecules with water molecules included in X-ray PDB files, we extracted from our MoDEL library a set of simulations corresponding to these proteins, rejecting those having an X-ray resolution worse than 2 Å (resulting in a dataset of 567 proteins). These structures were superimposed with the original PDB file with standard Kabsch algorithm (15). Water molecules with lower RMSd values in our study are expected to be located in those regions where structural water molecules were reported from the electron-density map. Extensive comparison between crystallographic water molecules and MD hydration sites has been performed for the resulting filtered set. A 2Å distance criterion has been used to determine matching between X-ray and MD water molecules.

Suppl. References

(1) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Research*, 28:235-242.

(2) Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J.Comput. Phys.* 23, 327-341.

(3) Darden, T., York, D., Pedersen, L. (1993) Particle Mesh Ewald – An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089.

(4) Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz, Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R. (2005) The amber biomolecular simulation programs. *Computat. Chem.* 26, 1668-1688.

(5) Cornell, W.D., Cieplak, P., Bayli, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117, 5179-5197.

(6) Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R..W., Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926.

(7) Gelpí, J.L, Kalko, S.G., Barril, X., Cirera, J., de la Cruz, X., Luque, F.J. and Orozco, M. (2001) Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins,* 45(4), 428-437.

(8) Morreale, A., de la Cruz, X., Meyer, T., Gelpí, J.L., Luque, F.J., Orozco, M. (2005) Partition of protein solvation into group contributions from molecular dynamics simulations. *Proteins*, 58(1), 101-109.

(9) Hubbard, S.J. & Thornton, J.M. (1993) 'NACCESS' Computer Program, Department of Biochemistry and Molecular Biology, University College London.

(10) Schröder, C., Rudas, T., Boresch, S., Steinhauser, O. (2006) Simulation studies of the protein-water interface I. Properties at the molecular resolution. *J. Chem. Phys.* 124, 234907, 1-18.

(11) Case, D.A., Babin, V., Berryman, J.T., Betz, R.M., Cai, Q., Cerutti, D.S., Cheatham, III, T.E., Darden, T.A., Duke, R.E., Gohlke, H., Goetz, A.W, Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T.S., Legrand, S., Luchko, T., Luo, R., Madej, B., Merz, K.M., Paesani, F., Roe, D.R., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C.L., Smith, W., Swails, J., Walker, R.C., Wang, J., Wolf, R.M., Wu X., and Kollman P.A. (2014), AMBER 14, *University of California, San Francisco.*

(12) Roe, D. and Cheatham III, T.E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput., 9 (7), 3084-3095.*

(13) Halle, B. (2004) Protein Hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc. London B,* 359, 1207-1224.

(14) Nucci, N. V., Pometum, M. S., Wand, A.J. (2011) Site-resolved measurement of water-protein interactions by solution NMR. *Nature structural & Molecular Biology,* 18, 245-249.

(15) Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*. 32, 922-923.

Suppl. Table 1. Water molecules in internal cavities. Cavities have been selected with an average volume > 100Å^3.

PDB Code	Ligand	Cavity Volume (Å^3)	Wats Ini	Wats End	Trapped Wats
1akd	HEM	606.13	2	2	2
1as0	GSP	200.63	8	7	4
1cgt	-	141.63	9	15	5
1dlh	Influenza virus peptide	151.25	2	19	0
1e3y	-	171.00	1	10	0
1f3l	SAH	214.88	2	2	0
1g62	-	117.38	5	5	5
1g7n	-	403.5	6	24	3
1j2m	-	117.13	4	9	0
1kqw	RTL	291.75	0	0	0
1lst	Lys	152.13	3	10	1
1mu0	PHK (no sim)	247.75	5	22	2
1sqc	LDA	377.13	2	19	1
1tia	-	184.38	5	18	0
2sni	-	120.25	8	10	0

Suppl. Table 2. Hydrogen Bond atom acceptors and donors considered in HB analyses (AMBER residue/atom nomenclature).

Acceptor		Donor	
Residue	**Atom(s)**	**Residue**	**Atom(s)**
BCK	N-H	BCK	O
NTER	N-H1 N-H2 N-H3	CTER	OXT
LYS	NZ-HZ1 NZ-HZ2 NZ-HZ3	ASP	OD1 OD2
ARG	NH1-HH11 NH1-HH12 NH2-HH21 NH2-HH22 NE-HE	GLU	OE1 OE2
SER	OG-HG	SER	OG
TYR	OH-HH	TYR	OH
GLN	NE2-HE21 NE2-HE22	GLN	OE1 NE2
ASN	ND2-HD21 ND2-HD22	ASN	OD1 ND2
THR	OG1-HG1	THR	OG1
HIE	NE2-HE2	HIE	ND1
HID	ND1-HD1	HID	NE2
HIP	NE2-HE2 ND1-HD1		
TRP	NE1-HE1		

Legends to Suppl. Figures

Suppl. Fig1. Summary of water molecule's statistics in MoDEL database: a) MoDEL simulation's radii box distribution, b) MoDEL simulation's box volume distribution, c) MoDEL simulation's number of water molecules distribution, and d) correlation between number of water molecules and box volume in MoDEL simulations.

Suppl. Fig2. Water molecule's RMSd distributions computed in different time windows, from 100ps to 5ns. Progressive disappearance of the second shell can be clearly observed.

Suppl. Fig3. a) Percentage of water molecules visiting the 1st shell during different time windows (1, 2, 5 and 10 ns). b) Water density distribution (g/cm^3) for the whole set of water molecules belonging to the 1st shell.

Suppl. Fig4. Mean residence times (MRT) of water molecules attached to a particular protein **atom**. a) MRT for 1st shell water molecules in a 1ns time-window. b) MRT for 1st shell water molecules in a 5ns time-window (only 3ns are shown for clarity). c) MRT for water molecules trapped in internal cavities. d) MRT for 2nd shell water molecules in a 1ns time-window.

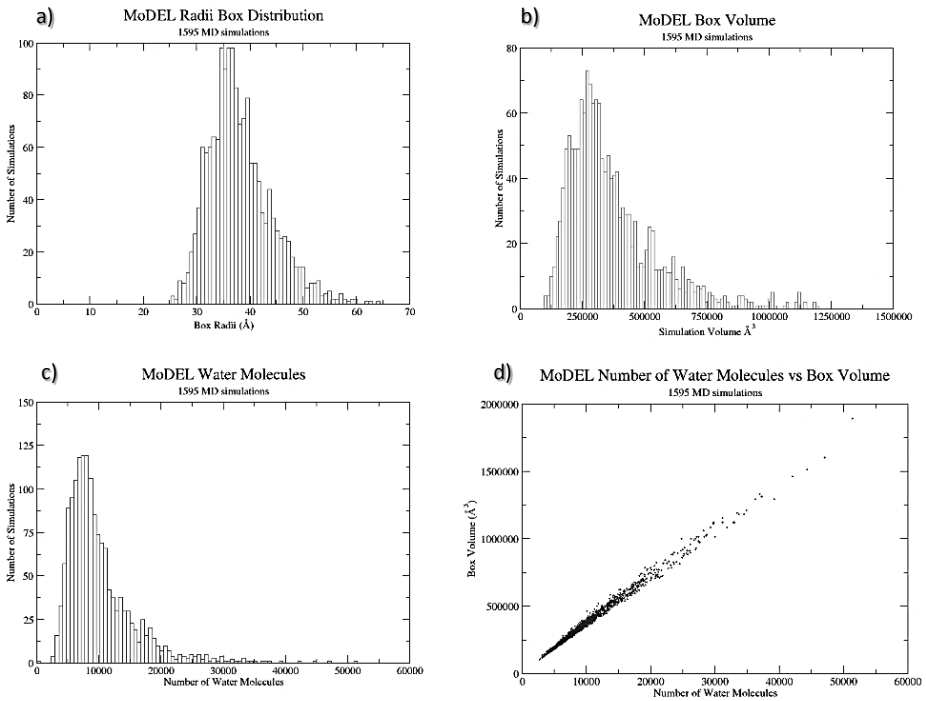
Suppl. Fig5. Radial Distribution Function (RDF) for the complete MoDEL database.

Suppl. Fig6. Solvent diffusion studies computed in the last nanosecond of MoDEL simulations: Velocities for water molecules belonging to the 1st (red), 2nd (green) shell and bulk (dark-blue) solvent. Velocity for solvent in a MD simulation with only water molecules is plotted as a reference (light-blue).

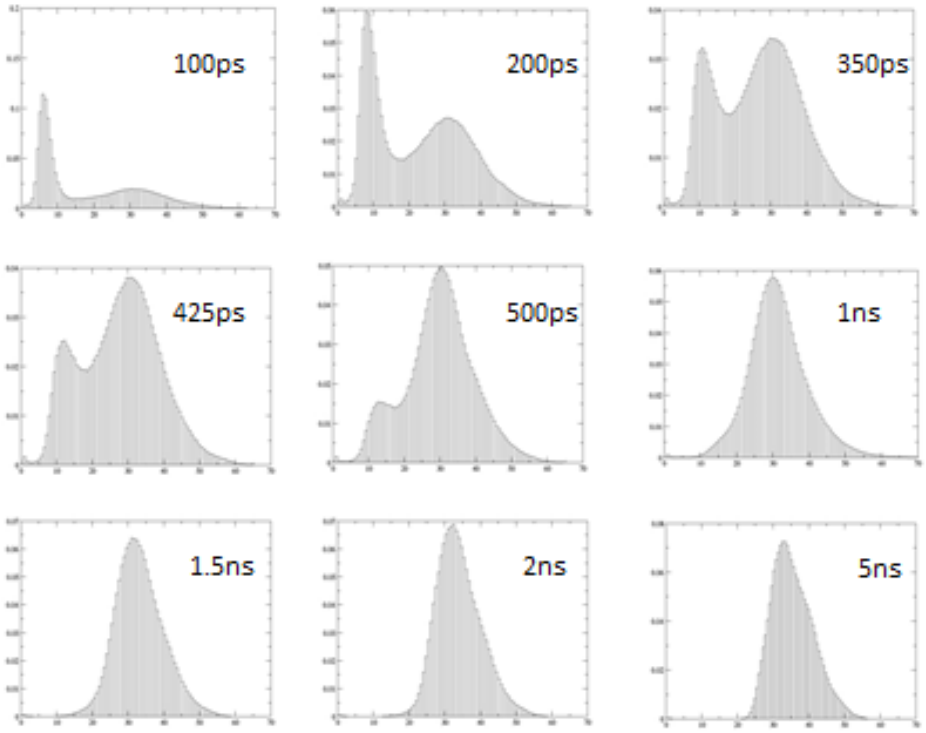
Suppl. Fig7. Analysis of the MSD behavior for some representative water molecules, shown by their velocity profile. The arrows represent the time window in which the water molecule is considered to be in the 1st shell. Inline plots representing average values for 100ps window-lengths are shown for clarity.

Suppl. Fig8. HB graph example, representing states as nodes and transition between states as edges. This particular example shows a 2-bit fingerprint graph, where the first bit represents presence or absence of the hydrogen bond and the second bit represents presence or absence of a water bridge. Size of nodes represents the number of occurrences of the state and edges represent probability of transition.

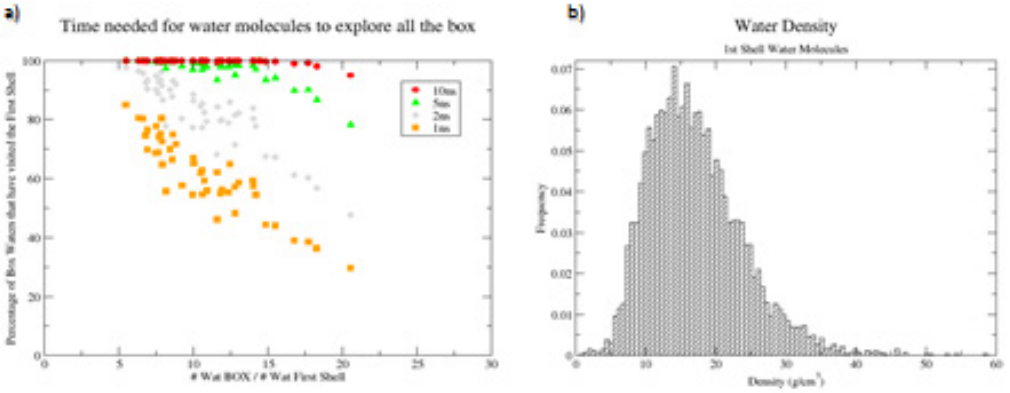
Suppl. Figure 1



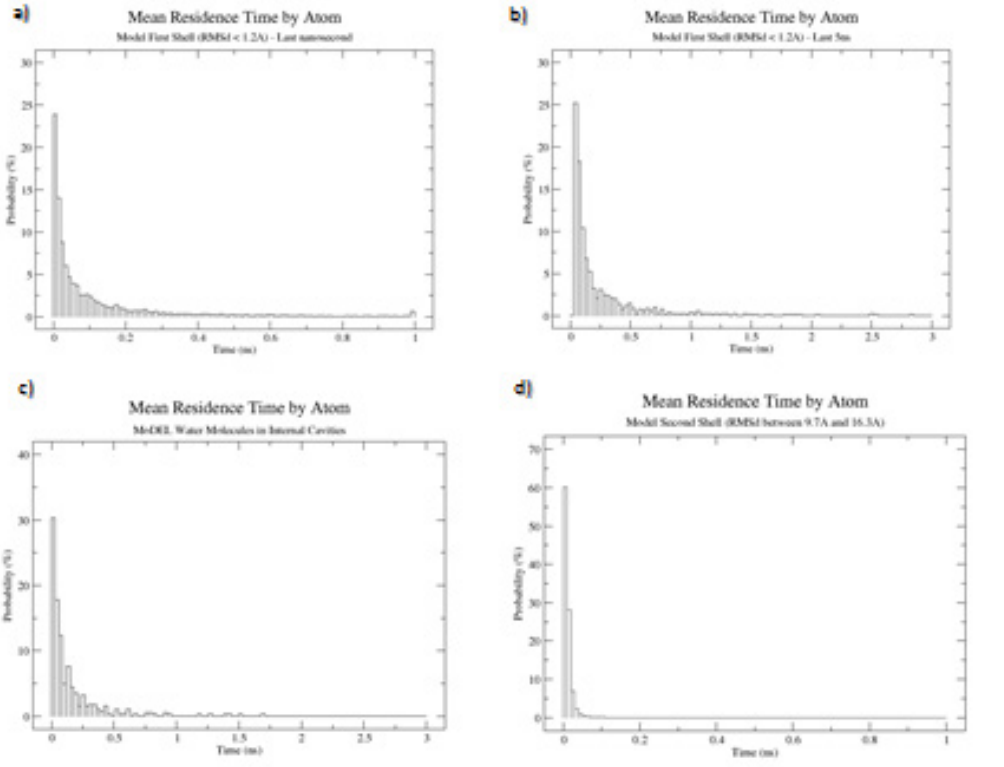
Suppl. Figure 2



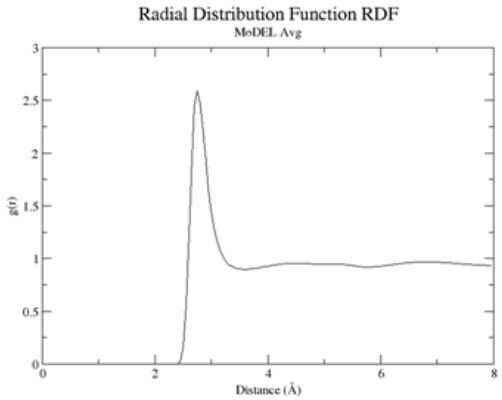
Suppl. Figure 3



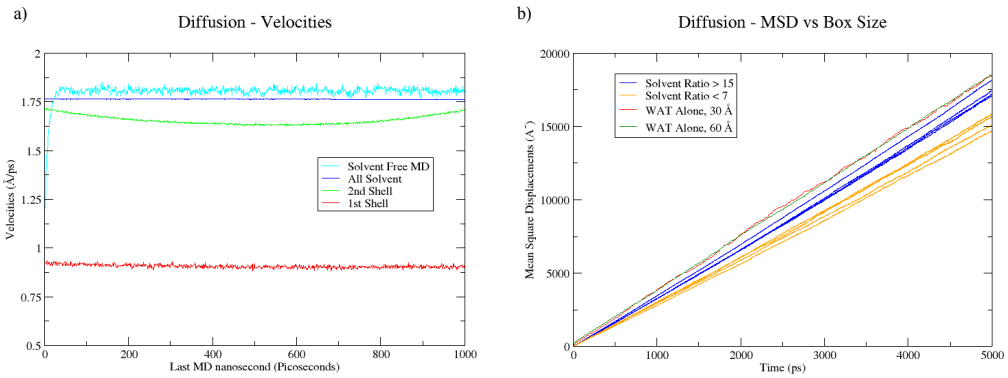
Suppl. Figure 4



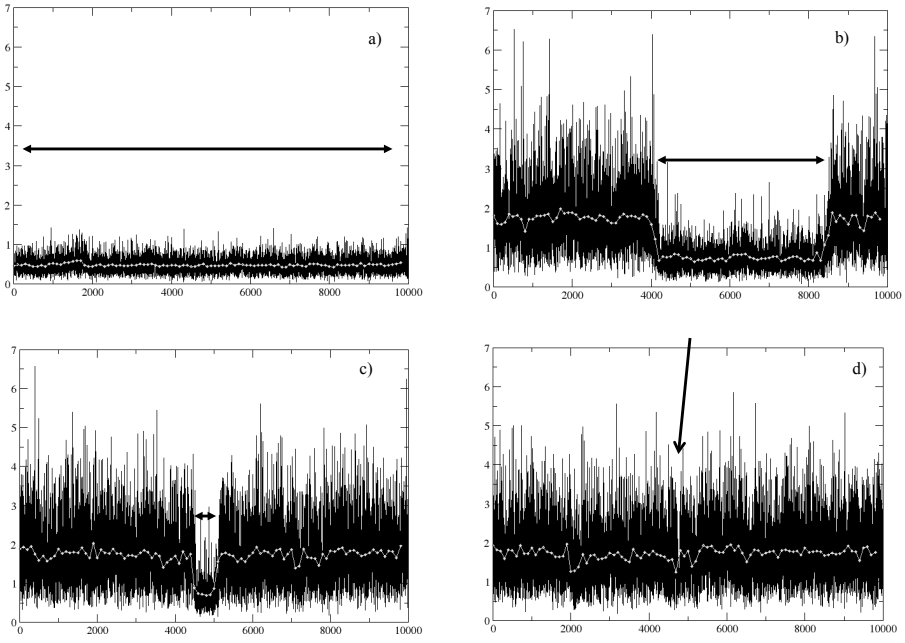
Suppl. Figure5



Suppl. Figure 6



Suppl. Figure 7



5.4 Development of data types and informatics tools to port macromolecular dynamics methods to the high-throughput regime.

5.4.1 Synopsis

Building a platform able to run different MD programs with a large variety of input parameters and output formats requires the generation of a well-designed computational infrastructure. We chose to work with semantic web services coupled with a biological ontology, where we added our own set of MD data types. The set of tools implemented as web services, ontology types and workflows was called MDMoby, and it allowed us the automation of the setup process, and thus, the possibility to port MD simulations to the HT regime.

An additional set of complex analysis tools, mainly based on essential dynamics methods, together with the ones actually running the simulations (atomistic MD as well as CG simulations) was added to complete the platform.

The whole package was compacted and presented as a couple of web portals: MDWeb (http://mmb.irbbarcelona.org/MDWeb) and FlexServ (http://mmb.irbbarcelona.org/FlexServ), the first one is aimed to the setup, running, and analysis of MD and CG simulations whereas the second one is dedicated to advanced protein flexibility analysis (although it also has CG methods integrated for the sake of independency).

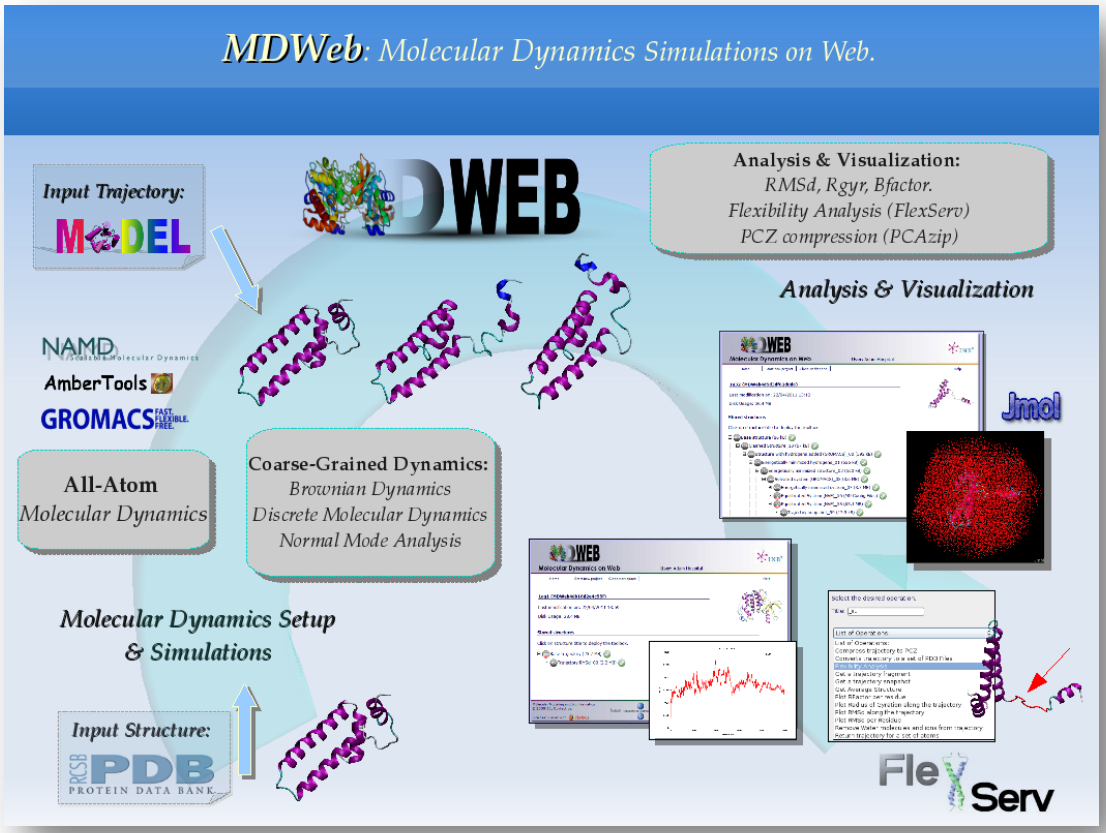
I have been responsible of the development of MDMoby ontology types, web services and workflows. My work in MDWeb server was focused in the internal engine interacting with all the MDMoby workflows and services, design and programming of part of the web code, and writing of the help pages and tutorials. In the FlexServ project, I implemented different analysis tools. I was also involved in the porting of all the analysis integrated in FlexServ into the INB's BioMoby platform, building a set of web services using the newly implemented MDMoby ontology objects.

5.4.2 Paper 3

MDWeb & MDMoby an integrated web-based platform for molecular dynamics simulations.

Adam Hospital, Pau Andrio, Carles Fenollosa, Damjan Cicin-Sain, Modesto Orozco and Josep Lluís Gelpí.

Bioinformatics (2012), 28(9), 1278-1279.



BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 9 2012, pages 1278–1279
doi:10.1093/bioinformatics/bts139

Structural bioinformatics
Advance Access publication March 21, 2012

MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations

Adam Hospital[1,2], Pau Andrio[1,3], Carles Fenollosa[1,3], Damjan Cicin-Sain[1,2], Modesto Orozco[1,2,4,*] and Josep Lluís Gelpí[1,3,4,*]

[1]Joint IRB-BSC Program on Computational Biology, Institute of Research in Biomedicine, Barcelona Science Park, Josep Samitier 1-5, Barcelona 08028, and Barcelona Supercomputing Center, Jordi Girona 31, Barcelona 08034, [2]Structural Bioinformatics Node, National Institute of Bioinformatics, Josep Samitier 1-5, Barcelona 08028, [3]Computatinal Bioinformatics Node, National Institute of Bioinformatics, Jordi Girona 31, Barcelona 08034 and [4]Department of Biochemistry and Molecular Biology, University of Barcelona, Av. Diagonal 643, Barcelona 08028, Spain

Associate Editor: Anna Tramontano

ABSTRACT

Summary: MDWeb and MDMoby constitute a web-based platform to help access to molecular dynamics (MD) in the standard and high-throughput regime. The platform provides tools to prepare systems from PDB structures mimicking the procedures followed by human experts. It provides inputs and can send simulations for three of the most popular MD packages (Amber, NAMD and Gromacs). Tools for analysis of trajectories, either provided by the user or retrieved from our MoDEL database (http://mmb.pcb.ub.es/MoDEL) are also incorporated. The platform has two ways of access, a set of web-services based on the BioMoby framework (MDMoby), programmatically accessible and a web portal (MDWeb).

Availability: http://mmb.irbbarcelona.org/MDWeb; additional information and methodology details can be found at the web site (http://mmb.irbbarcelona.org/MDWeb/help.php)

Contact: gelpi@ub.edu; modesto.orozco@irbbarcelona.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 19, 2011; revised on February 27, 2012; accepted on March 18, 2012

1 INTRODUCTION

Molecular dynamics (MD) has experienced a long evolution since its origins (McCammon *et al.*, 1977), being now a mature technique that allows to obtain accurate pictures of the dynamics of proteins and nucleic acids. Unfortunately, the practical use of MD is limited by three factors: (i) force-fields uncertainties; (ii) the need of large computational resources; and (iii) the high-level expertise needed to use efficiently the technique. Indeed, setting up a system for simulation requires a large series of operations, and a number of decisions that demand a significant degree of expertise and large amounts of human time, in most cases similar to that of computing the trajectory. The end result is that newcomers to the field face a stiff learning curve. These problems reach a maximum in high-throughput projects (Meyer *et al.*, 2010; Rueda *et al.*, 2007; Van der Kamp *et al.*, 2010), where thousands of trajectories need to be launched and supercomputer resources are required.

We present here a tool designed to help naïve users to completely prepare systems for simulation, and that allows expert users to use MD in the high-throughput regime. The tool is presented in two different versions: a web-services-oriented software platform (MDMoby), and a web portal MDWeb. The technology has been adapted to be accessed as web-services following BioMoby (www.biomoby.org; BioMoby Consortium, 2008). MDMoby can be run with usual web-services clients and also programmatically through suitable APIs (http://inb.bsc.es). In turn, the web portal MDWeb provides a friendly environment to setup new systems, run test simulations and perform analysis within a guided interface. Setup files can be prepared, at present time, for Amber (Case *et al.*, 2010), NAMD (Phillips *et al.*, 2005) and Gromacs (Hess *et al.*, 2008), and analyses can be carried out from trajectories written in the most usual formats. Additionally, the platform is interfaced to our flexibility analysis software FlexServ, so providing coarse-grained simulation, and advanced analysis tools (Camps *et al.*, 2009; Emperador *et al.*, 2008), and also with our MoDEL database (Meyer *et al.*, 2010).

2 IMPLEMENTATION

2.1 MDMoby

MDMoby services have been developed under the BioMoby framework (www.biomoby.org; BioMoby Consortium, 2008). To this end, a complete new set of data types was designed (see MDWeb help section for a detailed information about MDMoby ontology). Among them, the most relevant are MD_Topology, MD_Structure and MD_Trajectory, that contain information on topologies, structures and trajectory files, respectively. Those data types have been defined in an abstract form, and specialized in the inherited objects. Hence, data formats are inferred from object name and handled without user intervention. Setup services cover operations like structure repair, titration of ionizable residues and relevant water molecules, neutralization of the system, addition of salt and solvent, energy minimization, thermalization and system equilibration. A limited set of parameterized ligands are available to

*To whom correspondence should be addressed.

be included in the setup process. Full automatic setup procedures in standard conditions for the software packages covered are offered as pre-packed workflows. Underlying software is based in the Ambertools (Case *et al.* 2010) and VMD (Humphrey *et al.* 1996) packages, combined with in-house and publicly available programs (Supplementary Table S1). The tool allows launching test simulations in our servers, and provides the necessary scripts for running production simulations at the user side. Scripts provided contain the necessary instructions to be customized to match the users' configuration. Basic analysis results are provided from Ambertools and Gromacs tools, complemented with in-house software. Tools for conversion between trajectory formats are available to facilitate trajectory sharing, extension and analysis. Supplementary Tables S2 and S3, and MDWeb site's help show the full offer of web-services and operations available.

MDMoby is accessible through Perl and Java APIs provided by the Spanish National Institute of Bioinformatics (http://inb.bsc.es), and, in a limited extent, using clients like Taverna (Hull *et al.*, 2006) or Jorca (Martín-Requena *et al.*, 2010). Supplementary Figure S1 shows a sample Perl script to prepare of a protein system using MDMoby.

2.2 MDWeb

MDWeb is a web portal implemented in PHP and MySQL that provides a GUI to MDMoby. It provides users with a personal workspace where intermediate data, trajectories and analysis results can be stored. Registration is free but necessary to maintain a permanent workspace. The primary entry is a structure (uploaded or obtained from PDB) for setup or a trajectory for analysis. The input structure or trajectory acts as the root of a tree (see Supplementary Figure S2 for some screenshots) to that new sets of data are added according to the operations performed. Every new component of the tree is identified by its BioMoby's data-type, and can be visualized (with JMol, http://www.jmol.org), or downloaded. For every component the appropriate choice of operations is presented. To guide the non-experts, pre-packed workflows including recommended options are also available. However, all web-services in MDMoby are available separately, allowing more experienced users to finely tune the procedures. Results of trajectory analysis are presented through alphanumerical values, 2D plots or Jmol-based 3D visualizations, as appropriate. Analysis results can also be downloaded for further processing. Trajectories from the MoDEL library (Meyer *et al.*, 2010) can also be loaded in MDWeb and analyzed. MDWeb handles efficiently potentially slow operations, which are derived to a batch queue to be executed. Results become available in the workspace as soon as calculations are finished and can be recovered at a later time.

3 CONCLUSIONS

The software platform formed by MDMoby and its portal MDWeb provides a step forward in the current offer of software to help in the use of MD. It is a consequence of the automation required to develop the MoDEL project and integrates the expertise accumulated over the years on massive setup of systems and analysis of MD trajectories. The modular nature of the web-service paradigm in which the platform is created assures that the system can grow to incorporate new operations without a significant change of the interface or even to incorporate MDMoby services to user's codes. The platform provides non-experts users of pre-packed tools allowing to do complete MD analysis without a deep knowledge of the details involved. At the same time, the platform is flexible enough as to allow expert users to perform finely tuned simulations. The platform is not tied to a specific software package, therefore increasing the number of potential users and helping them in code-migration and re-use of trajectories. MDMoby and MDWeb constitute a growing platform where eventually new operations and scenarios will be included in the future.

ACKNOWLEDGEMENT

We are indebted all the beta-testers of the platform.

Funding: Spanish Ministry of Education and Science [grant numbers: BIO2006-01602, CTQ2005-09365, CONSOLIDER E-Science Project in Supercomputation]; The EU Commission [contract INFSO-RI-261523, ScalaLife project]; the Instituto de Salud Carlos III [National Institute of Bioinformatics and the COMBIOMED network]; the ERC for an Advanced Grant to M.O., and the Fundación Marcelino Botín.

Conflict of Interest: none declared.

REFERENCES

BioMoby Consortium. (2008) Interoperability with Moby 1.0 – It's better than sharing your toothbrush. *Brief. Bioinform.*, **1**, 1–12.

Camps,J. *et al.* (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, **25**, 1709–1710.

Case,D.A. *et al.* (2010) *AMBER 11*, University of California, San Francisco.

Emperador,A. *et al.* (2008) Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.*, **95**, 2127–2138.

Hess,B. *et al.* (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

Hull,D. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, 729–732. (Web Server Issue)

Humphrey,W. *et al.* (1996) VMD: visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.

Martín-Requena,V. *et al.* (2010) jORCA: easily integrating bioinformatics Web Services. *Bioinformatics*, **26**, 553–559.

McCammon,J.A. *et al.* (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.

Meyer,T. *et al.* (2010) MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, **18**, 1399–1409.

Phillips,J.C. *et al.* (2005) Scalable molecular dynamics with NAMD. *J. Comp. Chem.*. **26**, 1781–1802.

Rueda,M. *et al.* (2007) A consensous view to protein dynamics. *Proc. Natl Acad. Sci. USA*, **104**, 796–801.

van der Kamp,M.W. *et al.* (2010) Dynameomics: a comprehensive database of protein dynamics. *Structure*, **18**, 423–435.

Supplementary material

MDWeb and MDMoby: An integrated web-based platform for molecular dynamics simulations

Adam Hospital[1,2], Pau Andrio[1,3], Carles Fenollosa[1,3], Damjan Cicin-Sain[1,2], Modesto Orozco[1,2,4,*] and Josep Lluís Gelpí[1,3,4,*]

[1] Joint IRB-BSC Program on Computational Biology, Institute of Research in Biomedicine, Barcelona Science Park, Josep Samitier 1-5, Barcelona 08028, Spain and Barcelona Supercomputing Center, Jordi Girona 31, Barcelona 08034, Spain. [2] National Institute of Bioinformatics, Structural Bioinformatics, Josep Samitier 1-5, Barcelona 08028, Spain. [3] National Institute of Bioinformatics, Computational Bioinformatics, Jordi Girona 31, Barcelona 08034, Spain. [4] Department of Biochemistry and Molecular Biology, University of Barcelona, Av. Diagonal 643, Barcelona 08028, Spain.

Figure S1. Sample script for Molecular Dynamics Setup using MDMoby through the MobyLite Perl API

Script covers the preparation and equilibration of a protein system in standard conditions, starting from a structure obtained from the PDB. Web services, defined as Perl Functions by the API, are highlighted in bold. Error handling code has been deleted from all service invocations but the first one, for the sake of simplicity.

```
my $pdbId = $ARGV[0]  # Input PDB id.

# 1. Obtaining input structure from PDB

my $object  = Object->new($pdbId,'PDB');
my $pdbStruct = getStructureFromPDB (
   'structure' => $object
) -> {'structure'};

# Simple Error handling

if ($@) {
   print "Service execution failed:\n$@";
   exit;
}

# 2. Strip ligand molecules, and hydrogen atoms from the structure

my $cleaned = cleanPDB (
   'structure' => $pdbStruct,
   'waters' => 'false',       # Keeping Waters
   'hydrogens' => 'true',     # Removing Hydrogens
   'ligands' => 'true'        # Removing Ligands
) -> {'structure'};

# 3. Fix missing atoms on side chains

my $fixed = fixSideChains (
   'structure' => $pdbStruct
) -> {'structure'};

# 4. Add hydrogen atoms to the structure

my $hyd =  addHydrogensFromPDBText (
   'structure' => $fixed
) -> {'structure'};

# 5. Energy minimization of added hydrogen atoms

my $minH =
optimizeStructureFromAMBER_MD_Structure (
   'structure' => $hyd,
   'minimize' =>   100# 100 minimization steps
) -> {'structure'};

# 6. Energy minimization of protein atoms with restrains

my $minProt =
optimizeStructureFromAMBER_MD_Structure (
   'structure' => $minH,
   'minimize' =>   200,       # 200steps.
   'restrained_atoms' =>  'protein-h',# heavy
   'restraint' =>   50  # 50 Kcal/mol fcte.
) -> {'structure'};

# 7. Solvates the structure for AMBER

my $solv =
solvateStructureFromAMBER_MD_Structure (
   'structure' => $minProt,
   'ions' => 'true',          # counterions.
   'ionic_concentration' => 0.05,    # 50 mM
   'boxtype' => 'oct',        # Octahedric Box.
    'boxsize' => 15,          # 15 A spacing
) -> {'structure'};

# 8. Energy minimization of added solvent

my $minSolv =
optimizeStructureFromAMBER_MD_Structure  (
   'structure' => $solv,
   'minimize' =>   500,       # 500 steps.
   'restrained_atoms' => 'protein',
   'restraint' =>   50,  # 50 Kcal/mol fcte.
) -> {'structure'};

# 9. System 5 ps Equilibration

my $eq =  runMDFromAMBER_MD_Structure  (
   'structure' => $minSolv,
   'time' =>   5,             # 5 ps.
   'temperature' =>   300,    # 300 K.
) -> {'structure'};

# Equilibrated $eq object is a MD_TrajectoryDCD
# Getting output info and saving

my $struct = $eq->struct; # AMBER_MD_Structure

open PDB ">setup.pdb";
print PDB  $struct->content;  # PDB
close PDB

my $top = $struct->top;       # PRMTOP_Text

open TOP ">setup.top";
print TOP $top->content;      # Topology
close TOP
```

Figure S2. Screenshots of MDWeb Graphical Interface.

Top left: Structure Checking. **Top Right:** Structure/Trajectory Visualization using Jmol. **Bottom left:** Analysis Visualization using Gnuplot. **Bottom right:** Flexibility Analysis using FlexServ.

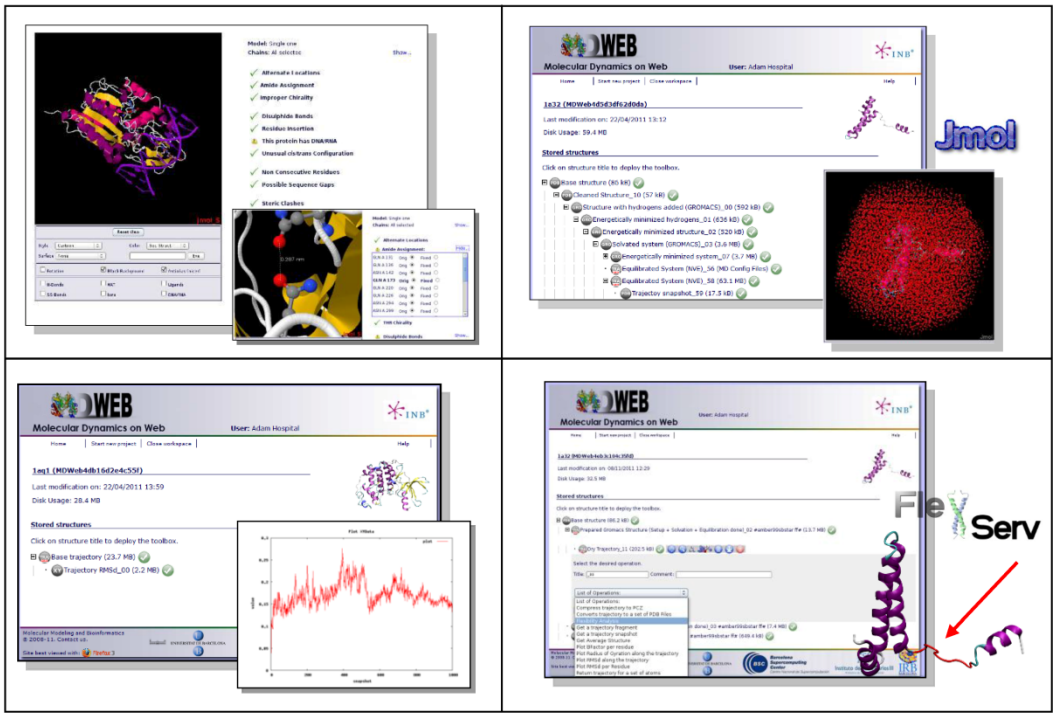


Table S1. Software used in MDWeb & MDMoby.
An updated version of the software list can be found on-line at http://mmb.irbbarcelona.org/MDWeb/help.php?id=software

Program	Description	Package & Version
BioMoby	Biomoby web-services framework	BioMoby 1.0
BLAST	Basic Local Alignment Search Tool	BLAST 2.2.17
CMIP	Classical Molecular Interaction Potential	CMIP 2.5.4
GnuPlot	Plotting tool	Gnuplot 4.2 patchlevel 2
Grace	Plotting tool	Grace 5.1.21
GROMACS	Molecular Dynamics Simulator	GROMACS 4.0.2
Jmol	Molecular Graphics Viewer	JMol 10.00.46
MobyLite PerlAPI	BioMoby Perl API	MobyLite PerlAPI 1.0
MySQL	Relational database manager	MySQL 5.0.51
NAMD	Molecular Dynamics Simulator	NAMD 2.7
PCAsuite	Trajectory compression tool	PCAsuite 1.1
Propka	Prediction of protein pKa values	Propka 2.0
Ptraj	Structure and dynamic analysis of trajectories	Ambertools 1.2
Sun Grid Engine	Batch queue manager	SGE 6.1u2
Tgatoppm, pnmcrop, pnmtopng	Image management	ImageMagick 6.3.7
Tleap	MD preparation program	Ambertools 1.2
VMD	Molecular Graphics Viewer	VMD 1.8.5

Table S2. List of available Molecular Dynamics WebServices (MDMoby).

The following table covers the complete list of available web-services included in MDMoby. Additionally, MDWeb full list of operations, linked to these services, can be found at http://mmb.irbbarcelona.org/MDWeb/help.php?id=ops.
An updated list of available web-services with the specific details of use and parameters is kept at http://mmb.irbbarcelona.org/WebServices/.

Web Service	Description
Molecular Dynamics Analysis > Trajectory Conversion	
fromMD_TrajectoryToMD_TrajectoryBINPOS	Converting MD trajectory formats, from any format (compatible with ptraj program) to Binpos (Scripps Binpos format).
fromMD_TrajectoryToMD_TrajectoryCRD	Converting MD trajectory formats, from any format (compatible with ptraj program) to CRD (Amber text format).
fromMD_TrajectoryToMD_TrajectoryDCD	Converting MD trajectory formats, from any format (compatible with ptraj program) to DCD (Charmm DCD Binary format).
fromMD_TrajectoryToMD_TrajectoryNetCDF	Converting MD trajectory formats, from any format (compatible with ptraj program) to NetCDF (Amber NetCDF format).
fromMD_TrajectoryToPDB_Collection	Converting an MD_Trajectory to a collection of PDBs.
fromPDB_CollectionToMD_TrajectoryNetCDF	Converting a PDB Collection to a NetCDF trajectory (Amber NetCDF format).
Molecular Dynamics Analysis > PDB Output	
getAverageStructureFromMD_Trajectory	Computes an average structure.
getSnapshotFromMD_Trajectory	Gets a single Snapshot from an MD trajectory
getStructureFromMD_Trajectory	Gets an MD_structure object from MD_Trajectory.
Molecular Dynamics Analysis > Trajectory Output	
getDryTrajectoryFromMD_Trajectory	Produces a Dry Trajectory (removing water molecules and/or Ions) from an MD trajectory.
getMaskTrajectoryFromMD_Trajectory	Applies a filter to select specific atoms or residues from an MD trajectory.
getSliceFromMD_Trajectory	Produces a slice (a set of snapshots) from an MD trajectory.
Molecular Dynamics Analysis > Trajectory Analysis	
getTrajectoryRms	Computes the Root Mean Square deviation (RMSd) from a collection of PDBs
getTrajectoryRmsFromMD_Trajectory	Computes the Root Mean Square deviation (RMSd) along an MD_Trajectory.
getTrajectoryRadiusOfGyrationFromMD_Trajectory	Computes the Radius of Gyration along an MD_Trajectory.
Molecular Dynamics Analysis > Analysis per Residue	
getTrajectoryBfactorPerResidue	Computes Bfactor x Residue in a collection of PDBs.
getTrajectoryBfactorPerResidueFromMD_Trajectory	Computes Bfactor x Residue in an MD_Trajectory.
getTrajectoryRmsPerResidue	Computes a Root Mean Square deviation (RMSd) per Residue in a

	collection of PDBs.
getTrajectoryRmsPerResidueFromMD_Trajectory	Computes a Root Mean Square deviation (RMSd) per Residue in an MD_Trajectory.
Molecular Dynamics Analysis > Trajectory Compression	
zipTrajectory	Compress MD Trajectories with PCAZIP tool.
unzipTrajectory	Uncompress protein trajectories previously compressed with PCAZIP tool.
Molecular Dynamics Setup/Simulation > GROMACS	
addLigandToGROMACS_MD_Structure	Add ligand (top+gro) to a GROMACS_MD_Structure
getGROMACS_MD_StructureFromPDBText	Obtains a GROMACS_MD_Structure (pdb + gro + topology) from a PDB-Text.
optimizeStructureFromGROMACS_MD_Structure	Runs an Energy Minimization of the Structure (Using GROMACS package).
runMDFromGROMACS_MD_Structure	Run a Simple Molecular Dynamic Simulation with Protein-Solvent system (Using GROMACS package).
runMDFromMD_TrajectoryXTC	Run a Simple Molecular Dynamic Simulation with Protein-Solvent system (Using GROMACS package).
solvateStructureFromGROMACS_MD_Structure	Solvates a Structure (Protein/Nucleic).
Molecular Dynamics Setup > Ambertools (Leap)	
fixSideChainsFromPDBText	Fix Side Chain Problems, adding missing heavy atoms (where possible).
getAMBER_MD_StructureFromPDBText	Obtains an AMBER_MD_Structure (pdb + topology) from a PDB-Text.
runLeapFromAMBER_MD_Structure	Runs Leap program from AmberTools package (from an AMBER_MD_Structure). Returns an AMBER_MD_Structure (pdb + topology).
runLeapFromPDBText	Runs Leap program from AmberTools package (from a PDB-Text). Returns an AMBER_MD_Structure (pdb + topology).
runLeapWithLigandsFromPDBText	Runs Leap program from AmberTools package (from a PDB-Text). Reads ligand libraries and frcmod files from input. Returns an AMBER_MD_Structure (pdb + topology).
solvateStructureFromAMBER_MD_Structure	Solvates a Structure (Protein/Nucleic).
solvateStructureWithLigandsFromAMBER_MD_Structure	Solvates a Structure (Protein/Nucleic). Reads Ligand libraries and frcmod files from input.
solvateStructureWithLigandsFromPDBText	Solvates a Structure (Protein/Nucleic). Runs Leap program from AmberTools package (from a PDB-Text). Reads Ligand libraries and frcmod files from input. Returns an AMBER_MD_Structure (pdb + topology).
Molecular Dynamics Setup/Simulation > NAMD	
addDisulphideBondsFromPDBText	Adds Disulfide Bonds with a distance criteria.
addHydrogensFromPDBText	Adds Hydrogens in the PDB.
mutateResidueFromPDBText	Mutates a Residue.
cleanStructureFromPDBText	"Cleans" a PDB structure: Removes Crystal Water molecules and/or

	Crystal Hydrogen atoms and/or Non-Parameterized Ligands.
getNAMD_MD_StructureFromPDBText	Obtains a NAMD_Structure (pdb + psf) from a PDB-Text.
solvateStructureFromNAMD_MD_Structure	Solvates a structure (Protein/Nucleic).
optimizeStructureFromAMBER_MD_Structure	Runs an Energy Minimization of an AMBER structure (Using NAMD program).
optimizeStructureFromNAMD_MD_Structure	Runs an Energy Minimization of a NAMD structure (Using NAMD program).
runMDFromAMBER_MD_Structure	Runs a Simple Molecular Dynamics Simulation with Protein-Solvent system (Using NAMD program).
runMDFromMD_TrajectoryDCD	Extends a Molecular Dynamics Simulation with Protein-Solvent system (Using NAMD program).
runMDFromNAMD_MD_Structure	Runs a Simple Molecular Dynamics Simulation with Protein-Solvent system (Using NAMD program).
Molecular Dynamics Setup > CMIP	
protonateHistidinesFromPDBText	Computes correct ionization state for histidine residues.
protonateIonizableResiduesFromPDBText	Computes correct ionization state for all ionizable residues.
titrateStructureFromPDBText	Titrates a Structure, adding structural water molecules and Na/Cl ions at the energetically most favourable positions.
Miscelanea Services > Parsers and Information Services	
getAminoAcidSequenceFromPDB	Gets the amino acid sequence that corresponds to a PDB entry.
getHeaderFromPDB	Retrieves headers from a PDB entry in PDB format.
getStructureFromPDB	Retrieves a structure (PDB format) from a PDB Id. Allows to specify a single chain using the format XXXX_X
parseAminoAcidSequenceFromPDBText	Extracts amino acid sequence from an input structure in PDB format.
plotFeatureAASequence	Returns a 2D graphic from a FeatureAASequence (PNG image format)
Molecular Dynamics > Coarse-Grained Simulations (Flexserv)	
runBrownianMDFromPDBText	Runs a Brownian Molecular Dynamics Simulation.
runDiscreteMDFromPDBText	Runs a Discrete Molecular Dynamics Simulation.
runNormalModeAnalysisFromPDBText	Runs a Normal Mode Analysis.

Table S3. List of available Predefined Molecular Dynamics Workflows (MDMoby).
Nested calls to other workflows are highlighted in bold.
An updated list of available workflows is kept at http://mmb.irbbarcelona.org/MDWeb/help.php?id=workflows

AMBER
Generate Topology for AMBER
ForceField: Parm99SB.* *Program: Leap from AmberTools Package.* 1. Remove crystallographic water molecules. 2. Add hydrogen atoms and missing side chain atoms as appropriate.
AMBER MD Setup. Structure Setup for AMBER forcefield
ForceField: Parm99SB.* *Programs: namd2 from NAMD Package, leap from AmberTools package, protpKa and CMIP.* 1. **Generate topology for AMBER.** 2. Protonate Histidine residues according to protpKa program algorithm. 3. Add 20 water molecules at the energetically best favourable positions of the structure surface using CMIP program. 4. Energy minimize hydrogen atoms for 500 steps of conjugate gradients, while the rest of the structure is kept fixed. 5. Energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 50 Kcal/mol to their initial positions.
AMBER MD Setup with Solvation. Structure Setup + Solvation for AMBER forcefield
ForceField: Parm99SB.* *Programs: namd2 from NAMD Package, leap from AmberTools package, protpKa and CMIP.* 1. **AMBER MD Setup.** 2. Set a truncated Octahedron box of TIP3P water molecules with a spacing distance of 15 Å around the system. 3. Add Cl^- and/or Na^+ ions necessary to neutralize the system. Then, add the appropriate amount of ions up to a concentration of 50 mM. 4. Further energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 50Kcal/mol to their initial positions.
AMBER Advanced Equilibration. System equilibration
Equilibration steps done in NPT ensemble with Periodic Boundary Conditions. *Particle Mesh Ewald (PME) for full-system periodic electrostatics.* *Constant temperature dynamics via Langevin Dynamics.* *Constant pressure dynamics via Nose-Hoover Langevin piston.* *SHAKE is used to maintain all bonds involving hydrogen atoms at their equilibrium values.* 1. Heat solvent to 300K. Solute atoms restrained (F. Const. 10 Kcal/mol). Length 5 ps. 2. Reduce Force constant to 5 Kcal/mol. Length 1 ps. Reduced Timestep 1fs. 3. Reduce Force constant to 2.5 Kcal/mol and limit restrain to backbone atoms. Length 1ps. 4. Reduce Force constant to 1 Kcal/mol. Length 1 ps. 5. Simulation without restraints. Length 1 ps.
AMBER FULL MD Setup. Complete Setup for AMBER forcefield (Structure Setup + Solvation + Equilibration).

ForceField: Parm99SB.*
Programs: namd2 from NAMD Package, leap from AmberTools package, protpKa and CMIP.
Equilibration steps done in NPT ensemble with Periodic Boundary Conditions.
Particle Mesh Ewald (PME) for full-system periodic electrostatics.
Constant temperature dynamics via Langevin Dynamics.
Constant pressure dynamics via Nose-Hoover Langevin piston.
SHAKE is used to maintain all bonds involving hydrogen atoms at their equilibrium values.

1. **AMBER MD Setup with Solvation.**
2. **AMBER Advanced Equilibration.**

NAMD

Generate Topology for NAMD. Generate PSF Topology for Charmm forcefield

ForceField: Charmm-27.
Program: psfgen from NAMD Package.
Warning: Ligands not allowed.

1. Remove crystallographic water molecules.
2. Add hydrogen atoms and missing side chain atoms as appropriate.

NAMD MD Setup. Structure Setup for Charmm Forcefield

ForceField: Charmm-27.
Programs: psfgen, vmd (solvate and autoionize plugins) and namd2 from NAMD Package, protpKa and CMIP.

1. **Generate Topology for NAMD.**
2. Protonate Histidine residues according to protpKa program algorithm.
3. Add 20 water molecules at the energetically best favourable positions of the structure surface using CMIP program.
4. Energy minimize hydrogen atoms for 500 steps of conjugate gradients, while the rest of the structure is kept fixed.
5. Energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 50 Kcal/mol to their initial positions.

NAMD MD Setup with Solvation. Structure Setup + Solvation for Charmm forcefield

ForceField: Charmm-27.
Programs: psfgen, vmd (solvate and autoionize plugins) and namd2 from NAMD Package, protpKa and CMIP.

1. **NAMD MD Setup.**
2. Set a Cubic box of TIP3P water molecules with a spacing distance of 15 Å.
3. Add Cl^- and/or Na^+ ions necessary to neutralize the system. Then, add the appropriate amount of ions up to a concentration of 50 mM.
4. Further energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 50 Kcal/mol to their initial positions.

NAMD Advanced Equilibration. System equilibration

Equilibration steps done in NPT ensemble with Periodic Boundary Conditions.
Particle Mesh Ewald (PME) for full-system periodic electrostatics.
Constant temperature dynamics via Langevin Dynamics.
Constant pressure dynamics via Nose-Hoover Langevin piston.
SHAKE is used to maintain all bonds involving hydrogen atoms at their equilibrium values.

1. Heat solvent to 300K. Solute atoms restrained (F. Const. 10 Kcal/mol). Length 5 ps.
2. Reduce Force constant to 5 Kcal/mol. Length 1 ps. Reduced Timestep 1fs.
3. Reduce Force constant to 2.5 Kcal/mol and limit restrain to backbone atoms. Length 1ps.
4. Reduce Force constant to 1 Kcal/mol. Length 1 ps.
5. Simulation without restraints. Length 1 ps.

NAMD FULL MD Setup. Complete Setup for Charmm force-field (Structure Setup + Solvation + Equilibration).

ForceField: Charmm-27.
Programs: psfgen, vmd (solvate and autoionize plugins) and namd2 from NAMD Package, protpKa and CMIP.
Equilibration steps done in NPT ensemble with Periodic Boundary Conditions.
Particle Mesh Ewald (PME) for full-system periodic electrostatics.
Constant temperature dynamics via Langevin Dynamics.
Constant pressure dynamics via Nose-Hoover Langevin piston.
SHAKE is used to maintain all bonds involving hydrogen atoms at their equilibrium values.

1. **NAMD MD Setup with Solvation**
2. **NAMD Advanced Equilibration.**

GROMACS

Generate Topology for GROMACS. Generate top and itp Topology Files for Gromacs.

Programs: pdb2gmx from Gromacs Package, AmberTools

1. Remove crystallographic water molecules.
2. Add side chain missing atoms using Leap from AmberTools package.
3. Add hydrogen atoms using pdb2gmx.

GROMACS MD Setup. Structure Setup for Gromacs

Programs: pdb2gmx, grompp, editconf, trjconv, make_ndx and mdrun from Gromacs Package, leap from AmberTools package, protpKa and CMIP.

1. **Generate Topology for GROMACS.**
2. Protonate Histidine residues according to protpKa program algorithm.
3. Add 10 water molecules at the energetically best favourable positions of the structure surface using CMIP program.
4. Energy minimize hydrogen atoms for 500 steps of conjugate gradients, while the rest of the structure is kept fixed.
5. Energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 500 KJ mol^{-1} nm^{-2} to their initial positions.

GROMACS MD Setup with Solvation. Structure Setup + Solvation for Gromacs

Programs: pdb2gmx, grompp, editconf, trjconv, make_ndx, genbox, genion and mdrun from Gromacs Package, leap from AmberTools package, protpKa and CMIP.

1. **GROMACS MD Setup.**
2. Set a truncated Octahedron box of TIP3P water molecules (Amber FF) or SPC water molecules (other FF's) with a spacing distance of 10 Å around the molecule.
3. Add Cl^- and/or Na^+ ions necessary to neutralize the system. Then, add the appropriate amount of ions up to a concentration of 50 mM.
4. Further energy minimize the structure for 500 steps of conjugate gradients, restraining heavy atoms with a force constant of 500 KJ mol^{-1} nm^{-2} to their initial positions.

GROMACS Advanced Equilibration. System equilibration

Equilibration steps done in NPT ensemble with Periodic Boundary Conditions.
Particle Mesh Ewald (PME) for full-system periodic electrostatics.
Constant temperature dynamics via Velocity-rescale algorithm.
Constant pressure dynamics via Parrinello-Rahman algorithm.
LINCS Linear Constraint Solver is used to maintain all bonds at their equilibrium values.

1. Heat solvent to 300K. Solute atoms restrained (F. Const. 400 KJ mol^{-1} nm^{-2}). Length 5 ps. Reduced Timestep 1fs.

2. Reduce Force constant to 300 KJ mol^{-1} nm^{-2}. Length 1 ps.
3. Reduce Force constant to 200 KJ mol^{-1} nm^{-2} and limit restrain to backbone atoms. Length 1ps.
4. Reduce Force constant to 100 KJ mol^{-1} nm^{-2}. Length 1 ps.
5. Simulation without restraints. Length 1 ps.

GROMACS FULL MD Setup. Complete Setup for Gromacs Package (Structure Setup + Solvation + Equilibration).

Programs: pdb2gmx, grompp, editconf, trjconv, make_ndx, genbox, genion and mdrun from Gromacs Package, leap from AmberTools package, protpKa and CMIP.
Equilibration steps done in NPT ensemble with Periodic Boundary Conditions.
Particle Mesh Ewald (PME) for full-system periodic electrostatics.
Constant temperature dynamics via Velocity-rescale algorithm.
Constant pressure dynamics via Parrinello-Rahman algorithm.
LINCS Linear Constraint Solver is used to maintain all bonds at their equilibrium values.

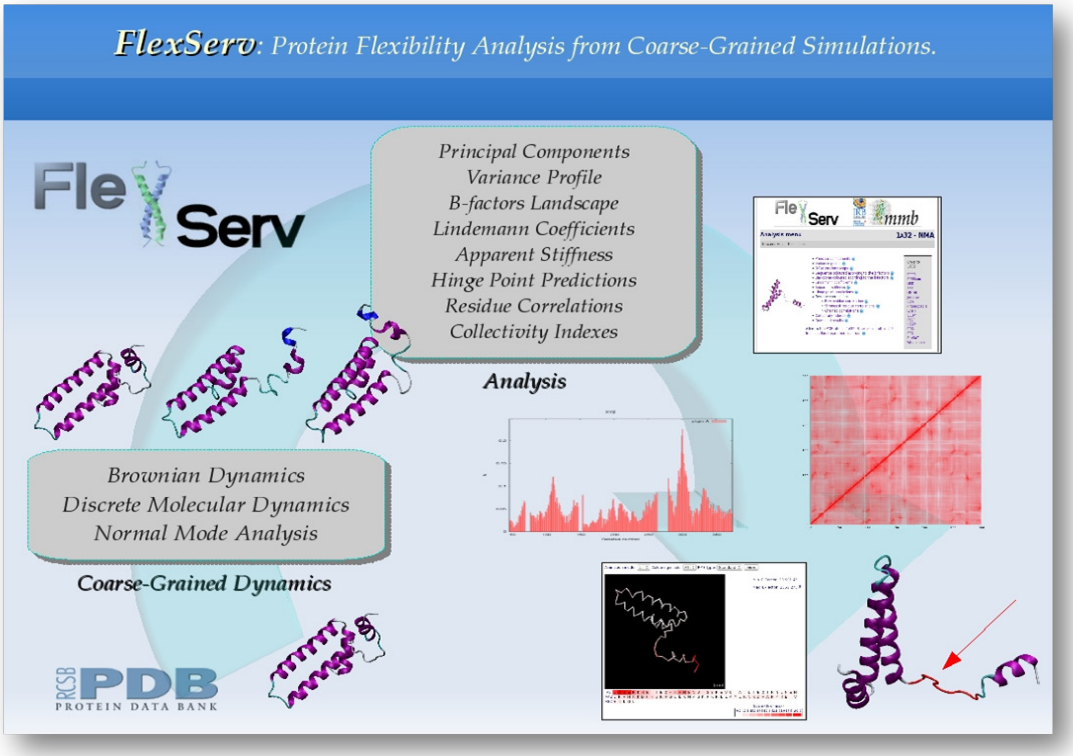
1. **GROMACS MD Setup with Solvation.**
2. **GROMACS Advanced Equilibration.**

5.4.3 Paper 4

FlexServ: An integrated tool for the analysis of protein flexibility

Jordi Camps, Oliver Carrillo, Agustí Emperador, Laura Orellana, **Adam Hospital**, Manuel Rueda, Damjan Cicin Sain, Marco D'Abramo, Josep Lluís Gelpí, Modesto Orozco.

Bioinformatics (2009), 25, 1709-1710.



BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 13 2009, pages 1709–1710
doi:10.1093/bioinformatics/btp304

Structural bioinformatics

FlexServ: an integrated tool for the analysis of protein flexibility

Jordi Camps[1,2], Oliver Carrillo[1], Agustí Emperador[1], Laura Orellana[1], Adam Hospital[1,3], Manuel Rueda[1], Damjan Cicin-Sain[1,3], Marco D'Abramo[1,4], Josep Lluís Gelpí[1,2,4,*] and Modesto Orozco[1,3,4,*]

[1]Joint IRB-BSC Program on Computational Biology, Institute of Research in Biomedicine, Barcelona Science Park, Josep Samitier 1-5, Barcelona 08028 and Barcelona Supercomputing Centre, [2]National Institute of Bioinformatics, Computational Bioinformatics, Jordi Girona 31, Barcelona 08034, Spain, [3]National Institute of Bioinformatics, Structural Bioinformatics, Josep Samitier 1-5 and [4]Department of Biochemistry and Molecular Biology, University of Barcelona, Av Diagonal 645, Barcelona 08028, Spain

Received on December 28, 2008; revised on April 23, 2009; accepted on May 4, 2009

Advance Access publication May 7, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: FlexServ is a web-based tool for the analysis of protein flexibility. The server incorporates powerful protocols for the coarse-grained determination of protein dynamics using different versions of *Normal Mode Analysis* (NMA), *Brownian dynamics* (BD) and *Discrete Dynamics* (DMD). It can also analyze user provided trajectories. The server allows a complete analysis of flexibility using a large variety of metrics, including basic geometrical analysis, B-factors, essential dynamics, stiffness analysis, collectivity measures, Lindemann's indexes, residue correlation, chain-correlations, dynamic domain determination, hinge point detections, etc. Data is presented through a web interface as plain text, 2D and 3D graphics.

Availability: http://mmb.pcb.ub.es/FlexServ; http://www.inab.org

Contact: modesto@mmb.pcb.ub.es or gelpi@mmb.pcb.ub.es

Supplementary information: Additional information and methodology details can be found at http://mmb.pcb.ub.es/FlexServ/help.

1 INTRODUCTION

Proteins have evolved to display not only optimal structures, but also functionally optimal deformability properties (Henzler-Wildman *et al.*, 2007). Many evidences show that evolution carefully conserved deformation patterns as a mechanism to maintain function (Falke, 2002).

Despite recent advances in experimental techniques, the study of flexibility is mostly a task for theoretical methods. The most powerful of them is atomistic molecular dynamics (MD) (McCammon *et al.*, 1977), a rigorous method which provides accurate representations of protein flexibility under physiological-like environments. Unfortunately, MD is complex, computationally expensive and their use requires a certain degree of expertise (Rueda *et al.*, 2007b). Coarse-grained methods coupled to simple potentials (see Dokholyan *et al.*, 1998; Tirion, 1996) are a cheap alternative to MD. By using them we lose atomic detail to gain formal and computational simplicity in the representation of near-native state flexibility of proteins (Emperador *et al.*, 2008; Ma, 2005). Unfortunately, despite its power, the practical use of coarse-grained methods is still limited, due mostly to the lack of standardized protocols for analysis and the existence of a myriad of different algorithms distributed in different websites.

We present here FlexServ, an integrated tool that links most important coarse-grained based methods, structural databases, and atomistic models with a powerful analysis platform. FlexServ incorporates three coarse-grained algorithms for the representation of protein flexibility: (i) discrete dynamics (DMD), (ii) normal mode analysis (NMA) and (iii) Brownian dynamics (BD). DMD assumes that residue-residue interactions are controlled by infinite square wells centered at equilibrium distance with width fitted to reproduce atomistic flexibility (Emperador *et al.*, 2008). Within this approach a particle is either moving at constant velocity or colliding with a wall, which allows the derivation of trajectories using simple ballistic equations (for discussion, see Emperador *et al.*, 2008). On the other hand, NMA and BD assume that inter-residue interactions are controlled by a harmonic-like potential energy expressed as:

$$E_{ij} = \Gamma_{ij} K_{ij}(r_{ij}^0)(r_{ij} - r_{ij}^0)^2 \quad (1)$$

where r_{ij} is the distance between residues i and j, and r_{ij}^0 the equilibrium value, Γ_{ij} the Kirchhoff connectivity matrix, and $K_{ij}(r_{ij}^0)$ the stiffness force constant. In a pure harmonic model $K_{ij}(r_{ij}^0)$ takes a single value K_{ij} and $\Gamma_{ij} = -1$ for atoms within a given cutoff and 0 otherwise. In Kovacs' pseudo-harmonic model (Kovacs *et al.*, 2004), $K_{ij}(r_{ij}^0)$ has a $(1/r_{ij}^0)^6$ dependence and no cutoff is applied. FlexServ incorporates both Hamiltonian definitions as well as a new hybrid one (Orellana *et al.*, unpublished results) which treats differently covalently bonded and non-bonded residues and that is fitted to reproduce atomistic MD results.

NMA [within the Anisotropic Network Model formalism (Altigan *et al.*, 2001)] uses the energy function to define the Hessian whose diagonalization provides the eigenvectors and eigenvalues characterizing the harmonic deformability of the protein. NMA based pseudo-trajectories are generated by activating movements along the eigenvectors (Rueda, *et al.*, 2007a). Within the BD approach the energy functional is used to compute forces, from

*To whom correspondence should be addressed.

which trajectories are derived by using a Brownian algorithm (Emperador *et al.*, 2008).

FlexServ incorporates a large variety of methods to characterize flexibility. Thus, the server performs basic analysis like structural oscillation using either standard (RMSd) or Gaussian (gRMSd) root mean square deviation. gRMSd (Damm and Carlson, 2006) fits the molecules reinforcing the alignment of rigid moieties, localizing more clearly the movements into flexible regions. Essential dynamics routines (Amadei *et al.*, 1993) are used to characterize the most important deformation modes, which are obtained by diagonalization of the trajectory covariance matrix. Essential deformation movements (after either fitting) are ranked by importance, and can be visualized and processed to obtain information (Meyer *et al.*, 2006). Analysis include B-Factor profiles, the 'collectivity' index (a measure of the collective nature of protein motions), the variance profile, the dimensionality (the number of movements defining a percentage of variance), or the size of the essential space (i.e. the number of relevant modes). Lindemann's indexes are computed to evaluate the liquid/solid nature of the entire or partial regions of the protein (Rueda *et al.*, 2007a).

Advanced capabilities of FlexServ include calculation of the apparent stiffness between interacting residues (obtained by inverting the inter-residue covariance (Rueda *et al.*, 2007a) and the determination of residue to residue correlations. Determination of dynamic domains and hinge points is implemented using a variety of techniques: (i) exploration of the B-factor landscape after fitting with the gRMSd method, (ii) analysis of the force-constant profile (Sacquin-Mora and Lavery, 2006) and (iii) clustering by inter-residue correlation (Navizet *et al.*, 2004). Calculations are performed using different sliding windows to reduce noise and false positives.

2 IMPLEMENTATION

FlexServ is written in PHP and implemented as a web-based interface, acting as a front-end of a series of simulation and analysis modules (see the full workflow diagram at http://mmb.pcb.ub.es/FlexServ/img/diagram.png). Results can be obtained through the web or retrieved later using a unique key. The modules are also available as web services in the Spanish National Institute of Bioinformatics platform (http://www.inab.org).

2.1 Input data

The user can upload a coordinate file, retrieve the structure using just the PDB code (Berman *et al.*, 2000), or upload his/her trajectory as a PCZ compressed file [PCAZIP (Meyer *et al.*, 2006); the software is available at the server]. Users can also upload the protein sequence, either in FASTA format or using a UniprotKB code (The Uniprot Consortium, 2008). In this case, analysis is performed on the closest homologue available in the PDB, identified by a standard BLAST analysis (Altschul *et al.,* 1990). Additionally, the program allows the user to analyze already available atomistic MD trajectories included in the MoDEL database (Rueda *et al.*, 2007b).

2.2 Workflow

Once the reference PDB structure is loaded, the program generates a C_α-model, computes several protein descriptors and provides basic visualization. FlexServ offers three engines to generate protein trajectories: (i) NMA, (ii) BD and (iii) DMD. In an elastic NMA calculation, the Hessian is computed and diagonalized. The obtained eigenvectors and eigenvalues are then used to derive a Monte-Carlo pseudo-trajectory in the NMA space. As for BD or DMD, trajectories are obtained directly. Trajectories are then processed with PCAZIP: RMS fits are applied to the data, and trajectory is finally compressed. Uploaded and MoDEL trajectories are processed directly. At this point, the compressed format contains the eigenvectors/eigenvalues data which can be used to animate the movement of the protein along essential dynamics modes as well as to derive the analysis offered by FlexServ (see above).

2.3 Visualization

Data is presented as plain text or 2D plots. When appropriated, 3D data is presented using a JMol applet (http://www.jmol.org). The produced trajectories and raw analysis data can be downloaded from the server for further off-line processing. Besides, a full snapshot of the results can also be downloaded to repeat the analysis at any later time.

Funding: Spanish Ministry of Education and Science [grant numbers: BIO2006-01602, CTQ2005-09365, CONSOLIDER Project in Supercomputation]; the Spanish Ministry of Health [COMBIOMED network]; the Fundación Marcelino Botín; and the National Institute of Bioinformatics.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Amadei,A. *et al.* (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.
Atilgan,A.R. *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.
Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acid Res.*, **28**, 235–242.
Damm,K.L. and Carlson,H.A (2006). Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys. J.*, **90**, 45558–45573.
Dokholyan,N.V. *et al.* (1998) Discrete molecular dynamics studies of the folding of a protein-like model, *Folding Des.*, **3**, 577–587.
Emperador,A. *et al.* (2008) Exploring the suitability of coarse-grained techniques for the representation of protein dynamics, *Biophys. J.*, **95**, 2127–2138.
Falke,J.J. (2002) Enzymology—a moving story. *Science*, **295**, 1480–1481.
Henzler-Wildman,K.A. *et al.* (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, **450**, 913–916.
Kovacs,J.A. *et al.* (2004) Predictions of protein flexibility: first-order measures. *Proteins*, **56**, 661–668.
Ma,J. (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, **13**, 373–380.
McCammon,J.A. *et al.* (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.
Meyer,T. *et al.* (2006) Essential dynamics: a tool for efficient trajectory compression and management. *J. Chem. Theory Comput.*, **2**, 251–258.
Navizet,I. *et al.* (2004) Probing protein mechanics: residue-level properties and their use in defining domains. *Biophys. J.*, **87**, 1426–1435.
Rueda,M. *et al.* (2007a) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**, 565–575.
Rueda,M. *et al.* (2007b) A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA*, **104**, 796–801.
Sacquin-Mora,S. and Lavery,R. (2006) Investigating the local flexibility of functional residues in hemoproteins. *Biophys. J.*, **90**, 2706–2717.
The Uniprot Consortium (2008) *Nucleic .Acid Res.*, **36**, D190–D195.
Tirion,M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.

5.5 Development of a tool for structure generation, dynamic simulation and trajectory analysis of nucleic acids.

5.5.1 Synopsis

As NA simulations require a very specific set of tools for analysis, we extended the presented MDMoby and MDWeb platforms, obtaining a framework (NAFlex) designed specifically for the analysis of nucleic acids flexibility. NAFlex gives the possibility to build structures from nucleotide sequences, using a library of different nucleotide types and base pair conformations, or to work directly with user-provided structures. Besides the traditional atomistic MD simulations, two additional methods (*Elastic Mesoscopic Model* and *Worm-Like Chain Model*) for the simulation of nucleic acids at low resolution are also available. Finally, a complete set of nucleic acid-specific analyses are integrated in the server, including physical properties and NMR observables.

The package is presented as a web server (http://mmb.irbbarcelona.org/NAFlex/) with an intuitive graphical interface, enclosing complete manuals and step by step tutorials.

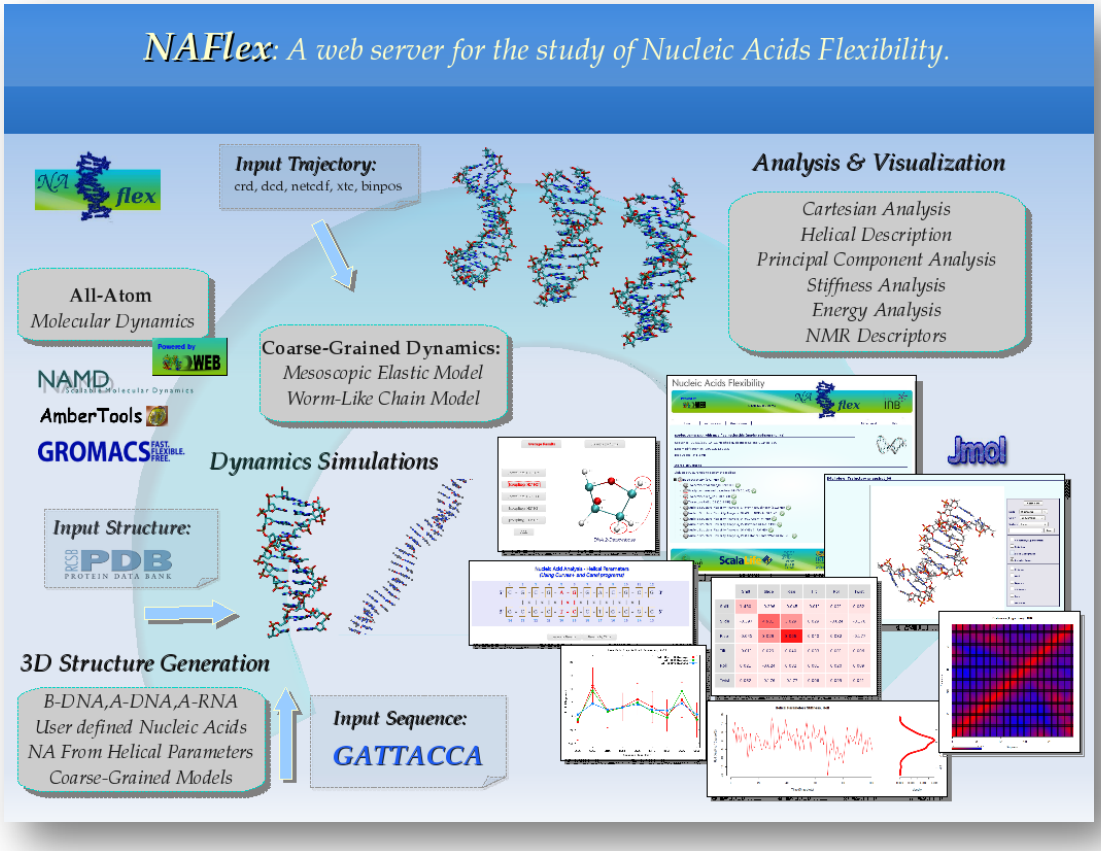
I am the main contributor to the NAFlex server, including the design and implementation of the complete web server, design and programming of the interface between the web application and the scripts developed by different co-authors of the work, as well as the writing of a complete help section containing manuals and tutorials.

5.5.2 Paper 5

NAFlex: A web server for the study of nucleic acids flexibility.

Adam Hospital, Ignacio Faustino, Rosana Collepardo-Guevara, Carlos González, Josep Lluís Gelpí, Modesto Orozco.

Nucleic Acids Research (2013), 41(W1), W47-W55,
NAR Featured Article June 2013.



Nucleic Acids Research Advance Access published May 17, 2013

Nucleic Acids Research, 2013, 1–9
doi:10.1093/nar/gkt378

NAFlex: a web server for the study of nucleic acid flexibility

Adam Hospital[1,2,3,4], Ignacio Faustino[1,2,3], Rosana Collepardo-Guevara[1,2,3], Carlos González[5], Josep Lluis Gelpí[1,2,3,6,7,*] and Modesto Orozco[1,2,3,4,7,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), Molecular Modelling and Bioinformatics (MMB) Department, 08028, Barcelona, Spain, [2]Joint IRB-BSC Programme on Computational Biology, Molecular Modelling and Bioinformatics (MMB) Department, 08028, Barcelona, Spain, [3]Barcelona Supercomputing Center, Life-Science Department, 08034, Barcelona, Spain, [4]National Institute of Bioinformatics, Structural Bioinformatics Node Division, 08028, Barcelona, Spain, [5]Instituto de Química Física 'Rocasolano' C.S.I.C., Biological Physical Chemistry Department, 28006, Madrid, Spain, [6]National Institute of Bioinformatics, Computational Bioinformatics Node Division, 08034, Barcelona, Spain and [7]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain

Received January 30, 2013; Revised April 16, 2013; Accepted April 18, 2013

ABSTRACT

We present NAFlex, a new web tool to study the flexibility of nucleic acids, either isolated or bound to other molecules. The server allows the user to incorporate structures from protein data banks, completing gaps and removing structural inconsistencies. It is also possible to define canonical (average or sequence-adapted) nucleic acid structures using a variety of predefined internal libraries, as well to create specific nucleic acid conformations from the sequence. The server offers a variety of methods to explore nucleic acid flexibility, such as a colorless wormlike-chain model, a base-pair resolution mesoscopic model and atomistic molecular dynamics simulations with a wide variety of protocols and force fields. The trajectories obtained by simulations, or imported externally, can be visualized and analyzed using a large number of tools, including standard Cartesian analysis, essential dynamics, helical analysis, local and global stiffness, energy decomposition, principal components and *in silico* NMR spectra. The server is accessible free of charge from the mmb.irbbarcelona.org/NAFlex webpage.

INTRODUCTION

Nucleic acids are polymorphic molecules whose conformations depend not only on the sequence but also on external factors, such as temperature, solvent, ionic environment and presence of ligands. Sequence and environment-dependent polymorphisms make the experimental description of nucleic acids extremely difficult. Our coverage of the nucleic acid structural space in the near future is not likely to approach that already available for proteins (1–3). The experimental information regarding flexibility is even more scarce, being partial and very much limited to the B-type DNA duplex (1–3). Only recently, NMR spectroscopy has started to provide dynamical information at the atomistic level for some model nucleic acid systems (4–6). Although the results are of impressive quality and impact, the technique is clearly still unable to provide a complete description of the general flexibility of nucleic acids.

In the absence of an experimental approach, simulation techniques are now widely accepted tools with which to describe nucleic acid structure and flexibility (7–11). According to PUBMED, in 2012, nearly 900 articles were published quoting the words 'DNA', 'RNA' or 'nucleic acids' in combination with 'simulation' or 'molecular dynamics'. Given the improvements in the simulation methods and the accessibility to increasingly faster computers, greater popularization of the field is expected in the near future (9).

Theoretical approaches for the study of nucleic acids are diverse, but can be classified on the basis of two basic parameters: (i) the nature of the Hamiltonian used; and (ii) the level of resolution (7,8,11,12). The simplest simulation approaches, such as the wormlike-chain (WLC, 13) model, assume simple Hamiltonians that describe global helical properties of the DNA fiber. Such methods were developed to examine general properties of very long fibers of canonical B-DNA with great computational efficiency. Atomistic molecular dynamics (MD) takes a

*To whom correspondence should be addressed. Tel: +34 934037155; Fax: +34 934037157; Email: modesto.orozco@irbbarcelona.org. Correspondence may also be addressed to Josep Lluis Gelpí. Tel: +34 934037155; Fax: +34 934037157; Email: gelpi@ub.edu

completely different approach. It combines a complex physical Hamiltonian with a rigorous method based on the resolution of Newton's equations of motion to obtain trajectories. MD allows an accurate atomistic description of any nucleic acid (8,9), but at the expense of a very large computational cost. The applicability of MD for very large systems or for the analysis of processes happening over large time scales is very limited. Between WLC and MD lie a wide variety of mesoscopic simulation methods. These were designed for the study of medium-to-large DNA fibers at the base-pair or even pseudo-base level of resolution using Hamiltonians of intermediate complexity (7,8,11–16).

For many years, the simulation of nucleic acids was performed by a small number of expert groups, often the developers of the software, force fields or algorithms. As the popularity of the simulation techniques has increased, groups with limited knowledge of the theoretical tools and often of the physics behind them have become users of these approaches. This development has generated significant confusion, as the potential user now faces a plethora of simulation packages based on a wide range of physical models and dealing with different levels of resolution. Furthermore, in general, a nonexpert user can find little information on how to setup the systems and encounters significant problems when attempting to perform meaningful analysis of the trajectories collected.

Here, we present NAFlex, a web tool designed to facilitate the use of nucleic acid simulation tools for newcomers to the field. The server allows the introduction of nucleic acid structures from diverse sources, as well as automatic structure generation from the sequences. The structural models can be subjected to several simulations, these are based on: (i) a colorless WLC model (13); (ii) a base-pair resolution mesoscopic model (14,15); or (iii) atomistic MD simulations. In the latter case, MDWeb technology (17) is used to help the user during all the setup and equilibration steps, providing all the input files required to launch the simulations. Finally, the trajectory obtained (or uploaded) can be visualized and analyzed using a large set of nucleic acid-specific tools. The server is freely accessible from the mmb.irbbarcelona.org/NAFlex webpage.

MATERIALS AND METHODS

NAFlex is divided (see Figure 1) into three main blocks: (i) *Input*; (ii) *Simulation engines*; and (iii) *Analysis*. The webserver has been designed to obtain a maximum coverage of potential user needs and as such is extremely flexible at the three levels (*input*, *simulation and analysis*).

Input

The user can introduce nucleic acid information in many ways depending on the nature of the problem and the calculation planned (see Figure 1). Thus, it is possible to upload a Protein Data Bank (PDB) structure (18) (of either isolated or complexed nucleic acids), or any user-derived structural model. The user can also upload a trajectory (all usual formats are accepted), which is then sent directly to the analysis modules of the server. When the user plans to work with canonical DNA and/or RNA duplexes, NAFlex allows the automatic generation of double helices using the Nucleic Acid Builder (nab) program from the Ambertools package (19). Structure generation can use parameters from various sources: (i) fiber diffraction data (20,21); (ii) sequence-dependent average X-ray structural information (1); (iii) tetramer-dependent average MD-simulation results (22–24); and (iv) user-defined helical values. The structures generated can be relaxed later to remove any distortion that may occur during model generation before being used in mesoscopic or atomistic simulations. In the case of WLC calculations, B-DNA geometry is assumed, and only sequence information is required as input.

After input, the server checks the structure to correct for potential gaps or missing atoms. Before launching the simulation, the setup procedures and a series of quality controls (see server help) are applied to the model. The server's check list includes, among others, alternate atom/residue locations, unusual distances between consecutive bases, steric clashes and the presence of metal ions or modified nucleotides/ligands. The server warns the user about potential structural errors (not trivial to correct).

Simulation engines

NAFlex offers the nonexpert user a selected variety of simulation tools that differ in complexity and resolution and that are designed to make their use as simple as possible.

WLC model (13) was introduced using the formalism of Jian *et al.* (25,26), implemented into a Monte Carlo sampling procedure. Accordingly, the DNA is represented as a set of N beads, each comprising M base-pairs (typically M = 10 base-pairs, but the exact bead resolution is selected by the user), and the potential energy of the DNA is defined by a simple Hamiltonian (see the help section for additional details):

$$E = E_S + E_B + E_T + E_{ele} \quad (1)$$

where the stretching energy (E_S) is computed as:

$$E_S = 0.5 K_S \sum_{i=1}^{N-1} (l_i - l_0)^2 \quad (2)$$

where K_s is the stretching constant, l_i is the actual distance between beads and l_0 is the optimum bead–bead distance. Based on (25,26), the value of K_s was set at 100 $k_B T/l_0^2$ (as it reproduces the correct DNA bond variance), with k_B the Boltzmann constant and T the temperature. The bending energy is determined as:

$$E_B = 0.5 K_B \sum_{i=1}^{N-2} \beta_i^2 \quad (3)$$

where K_B is the bending constant and β_i is the Euler bending angle between the local coordinate systems

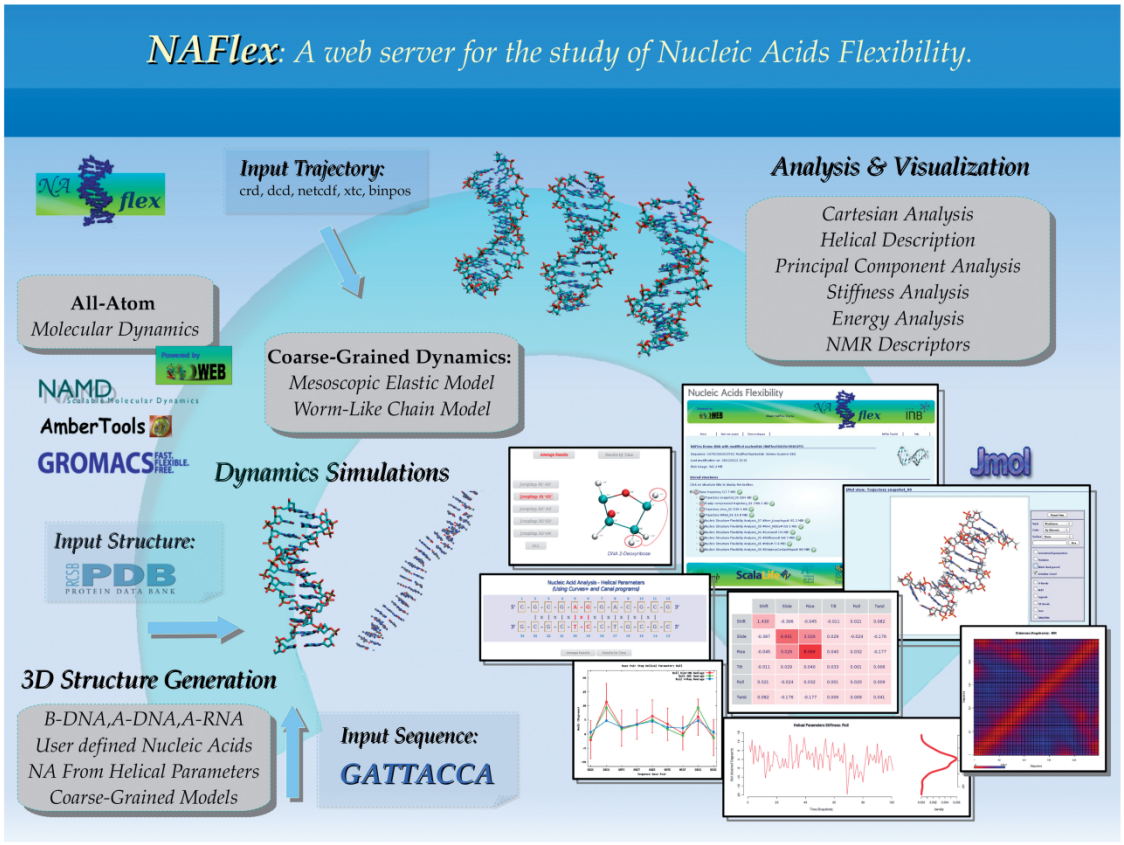


Figure 1. Representation of the NAFlex web server overview showing its three main blocks: **(i)** the inputs accepted: Sequence, structure and trajectory; **(ii)** the simulation engines implemented: molecular dynamics, mesoscopic simulations and WLC calculations; and **(iii)** the set of nucleic acid-specific flexibility analysis and visualization tools offered.

of two consecutive beads; and the torsion potential is defined by:

$$E_T = 0.5K_T \sum_{i=1}^{N-1} (\alpha_i + \gamma_i - \phi_0)^2 \quad (4)$$

where K_T stands for the torsional rigidity, and the sum of α and γ Euler angles defines the torsion between the local coordinate systems of two consecutive beads. ϕ is a parameter that gives the mean DNA twist, and it depends on the bead–bead equilibrium distance and the helical repeat. The values of the bending and torsional rigidity constants are related to the DNA bending (P) and twist (C) persistence lengths, respectively, and the bead–bead equilibrium distance:

$$K_B = \frac{Pk_BT}{l_0} \quad (5)$$

and

$$K_T = \frac{Ck_BT}{l_0} \quad (25,26) \quad (6)$$

The electrostatic repulsion energy (E_{ele}), the only nonlocal term in the Hamiltonian, is determined using Debye–Hückel potential:

$$E_{ele} = \frac{v^2 l_0^2}{D} \sum_{i,j} \frac{e^{-\kappa r_{ij}}}{r_{ij}} \quad (7)$$

where v is the salt-dependent Stigter's effective DNA linear charge density (27), D is the dielectric constant of water, κ is the inverse Debye length and r_{ij} is the distance between two beads. Alternatively, the user can define a value for the effective DNA-bead charge, q_{DNA}, and account for the electrostatic potential as follows:

$$E_{DH} = \frac{qDNA^2}{4\pi\varepsilon\varepsilon_0} \sum_i \sum_{j \neq i} \frac{e^{-Kr_{ij}}}{r_{ij}} \quad (8)$$

where ε_0 is the electric permittivity of vacuum, and ε is the dielectric constant (set to 80).

The elastic mesoscopic model provides an intermediate level of resolution and potential energy complexity (7,8,11,12,14–16). It assumes that DNA deformations can be approximated as the addition of harmonic

distortions of equilibrium base-pair step geometries. Three rotational (twist, roll and tilt) and three translational (slide, shift and rise) degrees of freedom are considered, thereby allowing us to define the Hamiltonian as:

$$E = \Xi\,(\Delta X)^2$$

$$with\ \Xi = \begin{pmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{st} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{st} & k_s & k_{ls} & k_{lf} \\ k_{wl} & k_{rl} & k_{tl} & k_{ls} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{lf} & k_{lf} & k_f \end{pmatrix} = k_B T \subset^{-1} \quad (9)$$

where k_B is the Boltzman constant, T is the absolute temperature, E is the energy associated with the deformation ΔX and k_{XY} stands for the different stiffness constants defined by the 36 elements of the stiffness matrix (Ξ) [twist (w), roll (r), tilt (t), rise (s), slide (l) and shift (f)]. The Ξ can be calculated (see equation 9) by inversion of the covariance matrix obtained from either analysis of MD trajectories (at dinucleotide or tetranucleotide level, 22–24) or from the analysis of dinucleotide step variability in the crystal structures of DNAs and DNA–protein complexes (1,15). In all the cases, sampling is obtained using Monte Carlo simulations in the helical coordinate space following the DNALive (28) protocol.

Atomistic molecular dynamics (MD) is the most flexible, universal and accurate simulation engine incorporated to NAFlex. Unfortunately, it is also the most expensive, thus limiting its practical use to medium-sized nucleic acid structures (7–10). After the quality control check, and before running a MD simulation, the system is neutralized, immersed in the desired solution, minimized, thermalized and equilibrated in a few steps (Supplementary Figure S1). Once the equilibration is finished, the server provides NAMD (29), AMBER (30) or GROMACS (31) adapted input files, which can be then used to perform the production run on the user's local systems (see server help). Output trajectories can be uploaded back to the server for analysis.

Analysis tools

NAFlex integrates a variety of analysis packages for mining nucleic acid trajectories. Analysis results are presented in several numerical and graphical formats. Available analyses include the following: standard Cartesian and helical analysis (17,32); principal component analysis (PCA) (33,34) to determine the essential deformation movements and their associated stiffness; and helical stiffness analysis, performed following Olson–Lankas's approach (14,15). A basic energy analysis is conducted to determine stacking and hydrogen bonding interaction energies along the trajectory. Finally, NAFlex also provides estimates of a number of NMR observables (see help for details), which can be used for either trajectory validation or for *ab initio* spectral predictions.

SERVER IMPLEMENTATION AND USAGE

NAFlex is a web portal implemented in PHP and MySQL. It provides users with a personal workspace where intermediate data, trajectories and results of analysis can be stored. Access to the server is free; however, registration is required to maintain a permanent workspace. Nonregistered users can use the full functionality of the server, although they need to download the results at the end of the session. In the present implementation, registered users are provided with 2 Gb of storage space. Storage for specific projects can be increased on demand. NAFlex is powered by the MDWeb platform, which was designed for automated MD simulations of proteins (17). MDWeb provides the necessary modules for data management, structure checking, MD simulation setup and standard Cartesian trajectory analysis (see http://mmb.irbbarcelona.org/MDWeb/). NaFlex incorporates a variety of tools specifically oriented for nucleic acids, including input facilities, simulation engines and analysis tools (see 'Materials and Methods' section).

Workspace

The starting points for the analysis could be single structures, uploaded as PDB files or derived from a given sequence, or previously simulated trajectories (see above). The input structure or trajectory is used as initial step to initiate a project in the personal workspace. There is no limit to the number of initiated projects while the permitted storage capacity is not exceeded. For each project, the workspace holds all intermediate structures and results, organized as a tree view (Supplementary Figure S2), thereby allowing the user to track the history of operations performed. For each entry of the tree, a number of operations, selected on the basis of the type of data, are offered to the user. For all intermediate results, a download tool and a series of JMol-powered (http://www.jmol.org) visualization tools are available. Complete projects can also be downloaded for local storage. Downloaded projects can be restored in NAFlex again at a later time, thus recovering the original workspace.

Simulations

Coarse-grained simulations from initial structures can be performed within the NAFlex environment. When atomistic MD simulations are done, a series of preparation tools is required. Specific setup procedures adapted to nucleic acids are available, although experienced users may design and incorporate their own protocols into the workflow. Running short MD simulations for testing purposes is allowed on the server; however, production simulations should be run locally in the user's own facilities using NAFlex-generated input files (see above). The user can then upload trajectories into his/her workspace for analysis.

Analysis tools

NAFlex provides a wide repertoire of specific tools for the analysis of nucleic acids, some of them local, others

incorporated from external sources. All the tools are offered in a common interface, which requires no additional expertise for its use. The user has complete control over the level of resolution in the analysis through an interactive duplex viewer (Figure 2a). He/she can also analyze time course or average values and study in detail the resulting structural models (see Figure 2b and c). When available, standard values obtained from the literature are included for comparison purposes (Figure 2b). Plots can be opened in separate windows to compare the structural or mechanical behavior of different regions of the molecules. In all cases, raw data can be downloaded for local analysis. It should be noted that full analyses are done upon the initial request. This allows the user to browse the results even for a large structure without noticeable delay. Trajectories can be manipulated to extract individual snapshots for further inspection (Figure 2d).

Please note that trajectory upload is limited by network bandwidth. The current limit for NAFlex-uploaded trajectories is 100 Mb (as indicated in the corresponding help pages). However, in our experience, once the trajectory has been stripped of solvent and ions, and taking a representative ensemble of snapshots, the analysis of the resulting trajectory with NAFlex can provide useful information about the general flexibility of the molecule. See Example 4 in the next section as an illustration of a real analysis.

EXAMPLES OF USE

NAFlex is an extremely flexible server that offers a large number of options, and it does not require deep knowledge of either simulation engines or the physics of nucleic acids. A few examples of use are included here as references. These examples can be also accessed online on the NAFlex server using the 'demo' account.

Example 1: Atomistic MD from a nucleotide sequence

The files required to run an atomistic MD simulation can be obtained following a robust protocol. First of all, the user should define the type of input (Supplementary Figure S3a), in this case, *DNA/RNA Simulation From sequence*. The only inputs required are the nucleotide sequence (either typed or from a FASTA-formatted file) and the desired set of DNA/RNA helical parameters (Supplementary Figure S3b). For instance, to generate a DNA molecule with a higher value for twist than the canonical B-DNA, *User-defined DNA* should be chosen at *DNA/RNA type* selector and the *Twist* value should be modified accordingly (in the example from 36.0° to 38.0°). The structure will then be generated using the Nab program from the Ambertools package (19). A new window will appear showing the newly created structure in a JMol applet, together with the results of a set of structural checks (Supplementary Figure S3c and d). After validation of the original structure, a new workspace for the project is prepared (see Supplementary Figure S3e). The interface allows the user to select the required operation, in this case a complete setup for an AMBER MD simulation (see Supplementary Figure S3e) using the ff99SB (35) and PARMBSC0 (36) force fields with recent corrections [ildn (37) and OL3 (38)]. These are the recommended settings for a wide set of structures, including protein–DNA complexes and RNA; however, users have the choice of other force-field combinations. The evolution of the running workflow can be visualized in real time (Supplementary Figure S3f), and the final files can be downloaded for local production runs (Supplementary Figure S3g).

Example 2: Mesoscopic simulation of DNA

This example illustrates a mesoscopic simulation of a long protein-free DNA molecule at the base-pair level of resolution. As in the previous example, the first steps are the definition of the type of input (*DNA/RNA Simulation From sequence*), the selection of a title for the project and the introduction of the sequence (see Supplementary Figure S4a and b). In this case, *DNA/RNA type* should be set to *Coarse-Grained DNA Model (Base Step Level)*, to select the desired coarse-grained resolution level. The prepared structure can be visualized using the JMol applet (Supplementary Figure S4c). To launch the simulation, the user should select the desired number of ensemble snapshots (500 in this example), while the type of mesoscopic Hamiltonian (*Coarse-grained DNA Elastic Mesoscopic Model*) is already defined by the resolution level (see Supplementary Figure S4d). After the Monte Carlo ensemble has been collected, the snapshots can be further analyzed to trace the flexibility of the fiber being studied. In this particular example, the probability of cross-talk between distant segments of the DNA is analyzed. This can be done by simple inspection of the ensemble (Supplementary Figure S4e) and from the averaged nucleotide-contact maps (see Supplementary Figure S4f). The appearance of short contacts (red) off-diagonal is indicative of long-range contacts of potential biological relevance.

Example 3: Nucleic acid flexibility analysis

The example illustrates the analysis of a 100-ns MD simulation of a DNA dodecamer (dCGCGAGGACGCG). The trajectory in this case was uploaded from an external source. Analysis is initiated by selecting the *Nucleic Acid Flexibility Analysis* operation in the *Analysis Tools* box (Supplementary Figure S5a). The first results to be inspected are the helical characteristics of the duplex, obtained from a CURVES (32) analysis. Results (Supplementary Figure S5b and c) show the presence of a subtle *Twist* bimodality (Time-course *Twist* plot, Supplementary Figure S5b) and a marked *Roll* increase (taking X-ray as reference) in the first CpG step (Averaged *Roll* plot, Supplementary Figure S5c). This observation suggests that this step can be especially flexible and can display spontaneous curvature. Additional analysis shows that the DNA is highly robust in terms of helical structure, with strong hydrogen bonds and stacking interactions (see Supplementary Figure S5d); however, both ends show fraying effects (see, Supplementary Figure S5e), which are expected also to impact Nuclear Overhauser Effects (NOEs) (see Supplementary Figure S5f).

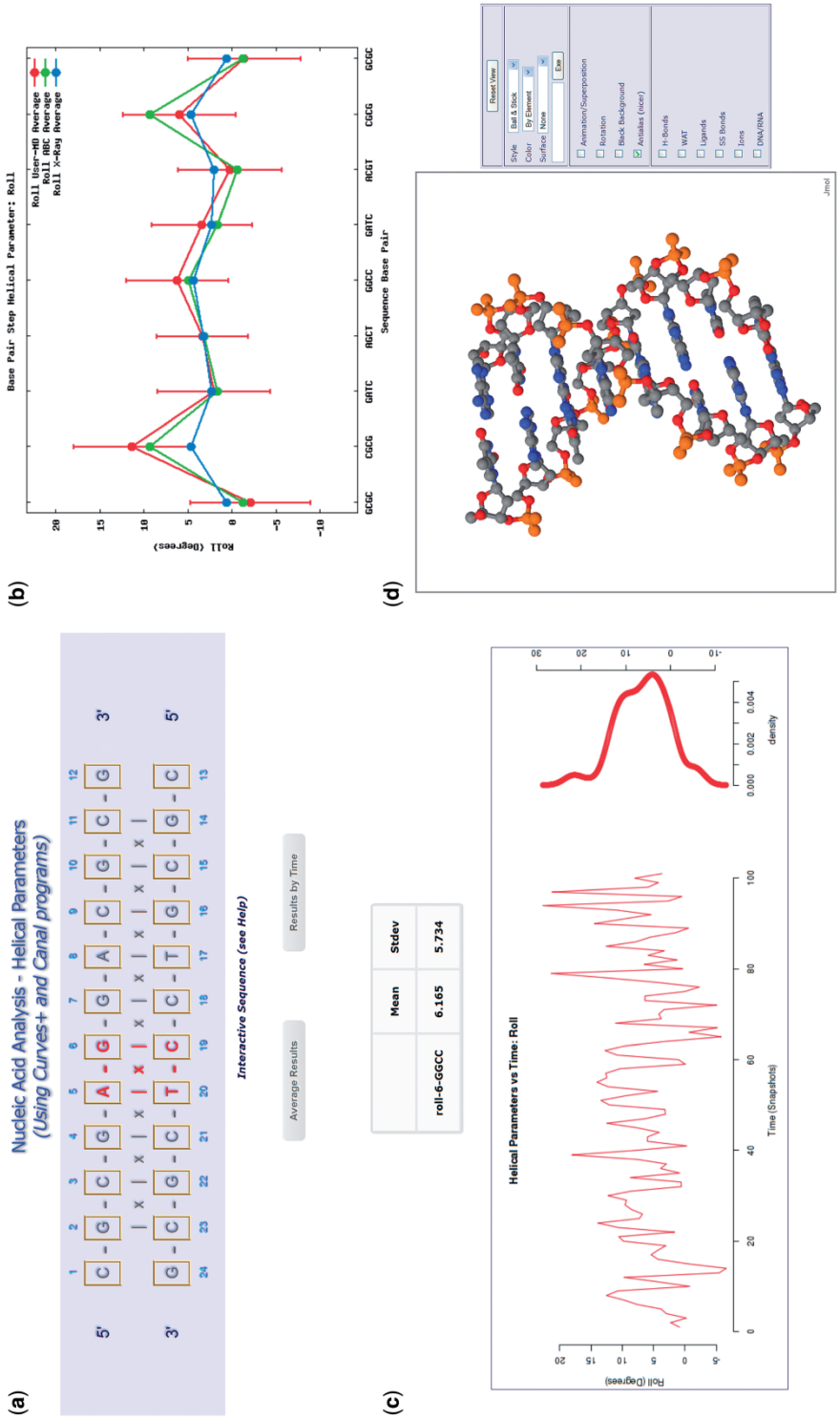


Figure 2. Screenshots of the analysis results interface. (**a**) Interactive duplex viewer allowing the level of resolution where the analysis is focussed: base-pairs (*|*), base-pair steps (*X*) and nucleotide pair steps (*-*). (**b**) Sequence distribution of time-averaged *Roll*, compared with MD- and X-ray averaged values. (**c**) Time course of the *Roll* at *6-GGCC* base-pair step, with an accessory histogram (mean and standard deviations are also displayed). (**d**) JMol interactive representation of a nucleic acid structure. Graphical visualization is offered for structures and also for trajectories.

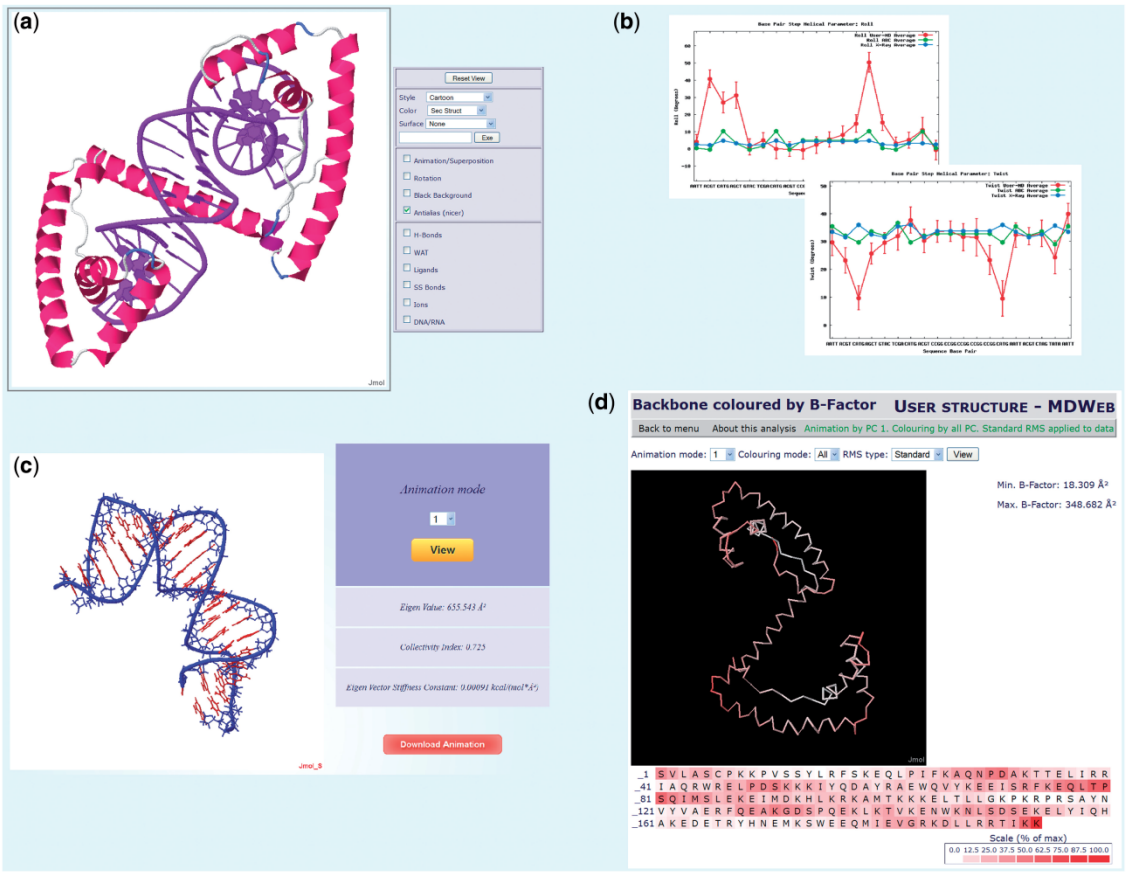


Figure 3. Protein–DNA complex TFAM–LSP as a case study. (**a**) JMol representation of the TFAM–LSP complex. To facilitate inspection, the protein part is represented by a dark pink cartoon, while the nucleic part is shown by a purple cartoon. (**b**) Average inter-base-pair helical parameters *Roll* and *Twist* computed on the LSP oligonucleotide plotted together with standard values obtained from the literature for comparison purposes. (**c**) Snapshot of the animation of the first mode of the DNA obtained from PCA. (**d**) B-factor analysis of the TFAM protein.

Example 4: Protein–DNA complex, human mitochondrial transcription factor A as a case study

The example shows the analysis of an important protein–DNA complex. The structural experimental information available (39–41) suggests that flexibility is crucial for the functionality of the complex. Analysis is initiated from a user-provided 100-ns trajectory obtained with GROMACS 4 (31). In the example, a 1-ns time window is retained for analysis.

A JMol visualizer provides initial structural views of the trajectory. Thus, in Figure 3a, the human mitochondrial transcription factor A (TFAM) protein can be clearly identified in dark pink cartoon representation, with the two high mobility group (HMG) protein domains joined by the inter-domain linker helix. The 22-base-pair oligonucleotide mitochondrial light strand promoter (LSP) attached to the TFAM protein is shown in purple. Once uploaded, the trajectory is automatically split into protein and DNA, allowing separate analysis. An extensive flexibility analysis of the protein counterpart (TFAM) can be performed using FlexServ platform (42), also available through the NAFlex workspace.

Basic helical analysis (see Figure 3b) shows that *roll* and *twist* angles are completely distorted along the trajectory in the regions where the protein residues intercalate into DNA base-pairs (Figure 3a). PCA was used to obtain a clear picture of the essential movements of the oligonucleotide in the complex. In this example, the first three eigenvectors explain most of the variance (Figure 3c and interactive JMol online). The first two modes involve the kinking of the complete LSP along the direction of the linker helix, bringing both DNA ends closer together, thus adopting a U-turn. The third mode shows the twisting of the nucleic acid caused by the motion of the two TFAM domain boxes HMG1 and HMG2. Domain motions can be visualized using the available JMol Applet. In accordance with the hypothesis that the HMG domains induce an overall DNA U-turn stabilized by the helix linker (40,41), FlexServ analysis of the protein moiety (Figure 3d) reveals the presence of flexible

segments near the domain boxes, while the linker helix appears to be relatively rigid.

CONCLUSIONS

The field of nucleic acid simulation has reached maturity. It has moved away from the times when only 'proof of concept' simulations were done by a very small number of highly specialized groups toward a 'full production' situation, where nucleic acid simulation tools are used by many groups, often not experts in the physics of nucleic acids or the theory behind the simulation package. NAFlex is a bioinformatics tool created to facilitate the use of simulation tools for nonexpert users interested in gaining insight into the dynamics of nucleic acid systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

We thank Drs G. Portella, F. Battistini, A. Alibés and P. Dans for their extensive testing of the platform, P. Sfriso for his help in NOE intensity computation, P. Andrio for his help in the design of the graphical interface and J. Alcántara for the design and maintenance of the computers used by the server. We are also indebted to the developers of the CURVES+, CANAL, GROMACS, AMBERTOOLS, NAMD, VMD and JMol software, which are being used in the server.

FUNDING

Funding for open access charge: *Spanish Ministerio de Economía y Competitividad* [BIO2012-3286]; European Research Council Advanced Grant (ERC); *Instituto Nacional de Bioinformática* (INB); Consolider E-Science Project; EU-Scalalife project; Fundación Marcelino Botín; European Union's Seventh Framework Programme (FP7/2007–2013) [275096 to R.C.-G. and M.O.].

Conflict of interest statement. None declared.

REFERENCES

1. Dans,P., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
2. Pérez,A., Noy,A., Lankas,F., Luque,F.J. and Orozco,M. (2004) The relative flexibility of DNA and RNA: database analysis. *Nucleic Acids Res.*, **32**, 6144–6151.
3. Zheng,G., Colasanti,A.V., Lu,X.J. and Olson,W.K. (2010) 3DNA Landscapes: a database for exploring the conformational features of DNA. *Nucleic Acids Res.*, **38**, D267–D274.
4. Nikolova,E.N., Bascom,G.D., Andricioaei,I. and Al-Hashimi,H.M. (2012) Probing sequence-specific DNA flexibility in A-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13C relaxation and molecular dynamics simulations. *Biochemistry*, **51**, 8654–8664.
5. Bothe,J.R., Lowenhaupt,K. and Al-Hashimi,H.M. (2011) Sequence-specific B-DNA flexibility modulates Z-DNA formation. *J. Am. Chem. Soc.*, **133**, 2016–2018.
6. Nikolova,E.N., Kim,E., Wise,A.A., O'Brien,P.J., Andricioaei,I. and Al-Hashimi,H.M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, **470**, 498–502.
7. Orozco,M., Pérez,A., Noy,A. and Luque,F.J. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, **32**, 350–364.
8. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acids flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.
9. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.*, **45**, 196–205.
10. Laughton,C.A. and Harris,S.A. (2011) The atomistic simulation of DNA. *WIREs Comput. Mol. Sci.*, **1**, 590–600.
11. Drsata,T. and Lankas,F. (2013) Theoretical Models of DNA Flexibility. *WIREs Comput. Mol. Sci.*, 14 Feb (doi:10.1002/wcms.1144; epub ahead of print).
12. Lankas,F. (2012) Modelling nucleic acid structure and flexibility: from atomic to mesoscopic scale. In: Schlick,T. (ed.), *Innovations in Biomolecular Modeling and Simulations*, Vol. 2. Royal Society of Chemistry, London, pp. 3–32.
13. Allison,S.A. (1986) Brownian dynamics simulation of wormlike chains. Fluorescence depolarization and depolarized light scattering. *Macromolecules*, **19**, 118.
14. Lankas,F., Sponer,J. and Langowski,J. (2000) Sequence-dependent elastic properties of DNA. *J. Mol. Biol.*, **299**, 695–709.
15. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
16. Olson,W.K. (1996) Simulating DNA at low resolution. *Curr. Opin. Struct. Biol.*, **6**, 242–256.
17. Hospital,A., Andrio,P., Fenollosa,C., Cicin-Sain,D., Orozco,M. and Gelpí,J.L. (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics*, **28**, 1278–1279.
18. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
19. Case,D.A., Darden,T.A., Cheatham,T. III, Simmerling,C.L., Wang,J., Duke,R.E., Luo,R., Walker,R.C., Zhang,W., Merz,K.M. *et al.* (2012), AMBER 12. San Francisco: University of California.
20. Arnott,S., Hukins,D.W.L., Dover,S.D., Fuller,W. and Hodgson,A.R. (1973) Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations. X-ray diffraction analyses of the molecular conformations of polyadenylic acid—polyuridylic acid and polyinosinic acid—polycytidylic acid. *J. Mol. Biol.*, **81**, 107–122.
21. Lakshiminarayanan,A.V. and Sasisekharan,V. (1970) Stereochemistry of nucleic acids and polynucleotides II. Allowed conformations of the monomer unit for different ribose puckerings. *Biochim. Biophys. Acta*, **204**, 49–59.
22. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–2394.
23. Faustino,I., Pérez,A. and Orozco,M. (2010) Towards a Consensus view of duplex RNA flexibility? *Biophys. J.*, **99**, 1876–1885.
24. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D., Cheatham,T. III, Dixit,S., Jayaram,B., Lankas,F., Laughton,C.A. *et al.* (2010) A systematic molecular dynamics study of the nearest-neighbor effects on base pair and base step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–311.
25. Jian,H., Vologodskii,A.V. and Schlick,T. (1997) A combined wormlike-chain and bead model for dynamic simulations of long linear DNA. *J. Comp. Phys.*, **136**, 168–179.
26. Jian,H., Schlick,T. and Vologodskii,A.V. (1998) Internal motions of supercoiled DNA: brownian dynamics simulation of site juxtaposition. *J. Mol. Biol.*, **284**, 287–296.
27. Stigter,D. (1977) Interactions of highly charged colloidal cylinders with applications to double-stranded DNA. *Biopolymers*, **16**, 1435–1448.

28. Goñi,J.R., Fenollosa,C., Pérez,A., Torrents,D. and Orozco,M. (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.
29. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L. and Schulten,K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
30. Case,D.A., Cheatham,T.E. III, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
31. Hess,B., Kutzner,C., van der Spoel,D. and Lindahl,E. (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.
32. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
33. Amadei,A., Linssen,A.B. and Berendsen,H.J. (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.
34. Noy,A., Meyer,T., Rueda,M., Ferrer,C., Valencia,A., Pérez,A., de la Cruz,X., López-Bes,J.M., Luque,F.J. and Orozco,M. (2006) Datamining of molecular dynamics trajectories of nucleic acids. *J. Biomol. Struct. Dyn.*, **23**, 447–455.
35. Hornak,V., Abel,R., Okur,A., Strockbine,B., Roitberg,A. and Simmerling,C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, **65**, 712–725.
36. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force-field for nucleic acid simulations. Improving the representation of α/β conformations. *Biophys. J.*, **92**, 3817–3829.
37. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. and Shaw,D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950–1958.
38. Zgarbova,M., Otyepka,M., Sponer,J., Mladek,A., Banas,P., Cheatham,T.E. and Jurecka,P.J. (2011) Refinement of the Cornell et al. nucleic acid force field based on reference quantum chemical calculations of torsion profiles of the glycosidic torsions. *J. Chem. Theory Comput.*, **7**, 2886–2902.
39. Kaufman,B.A., Durisic,N., Mativetsky,J.M., Costantino,S., Hancock,M.A., Grutter,P. and Shoubridge,E.A. (2007) The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. *Mol. Biol. Cell.*, **18**, 3225–3236.
40. Rubio-Cosials,A., Sidow,J.F., Jiménez-Menéndez,N., Fernández-Millán,P., Montoya,J., Jacobs,H.T., Coll,M., Bernadó,P. and Solà,M. (2011) Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. *Nat. Struct. Mol. Biol.*, **18**, 1281–1289.
41. Rubio-Cosials,A. and Solà,M. (2013) U-turn DNA bending by human mitochondrial transcription factor A. *Curr. Opin. Struct. Biol.*, **23**, 116–124.
42. Camps,J., Carrillo,O., Emperador,A., Orellana,L., Hospital,A., Rueda,M., Cicin-Sain,D., D'Abramo,M., Gelpí,J.L. and Orozco,M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, **25**, 1709–1710.

LEGENDS TO SUPPLEMENTARY FIGURES

Suppl. Figure S1: Setup workflow. Steps involve cleaning the structure (removing for example, undesired crystallographic waters and/or ligands), fixing side chains (adding for example missing atoms), solvate and neutralize the system, minimization, thermalization and equilibration of the system using a standard procedure: G.Shields, C.Laughton and M.Orozco. *J.Am.Chem.Soc* , 119, 7463-7469 (1997)

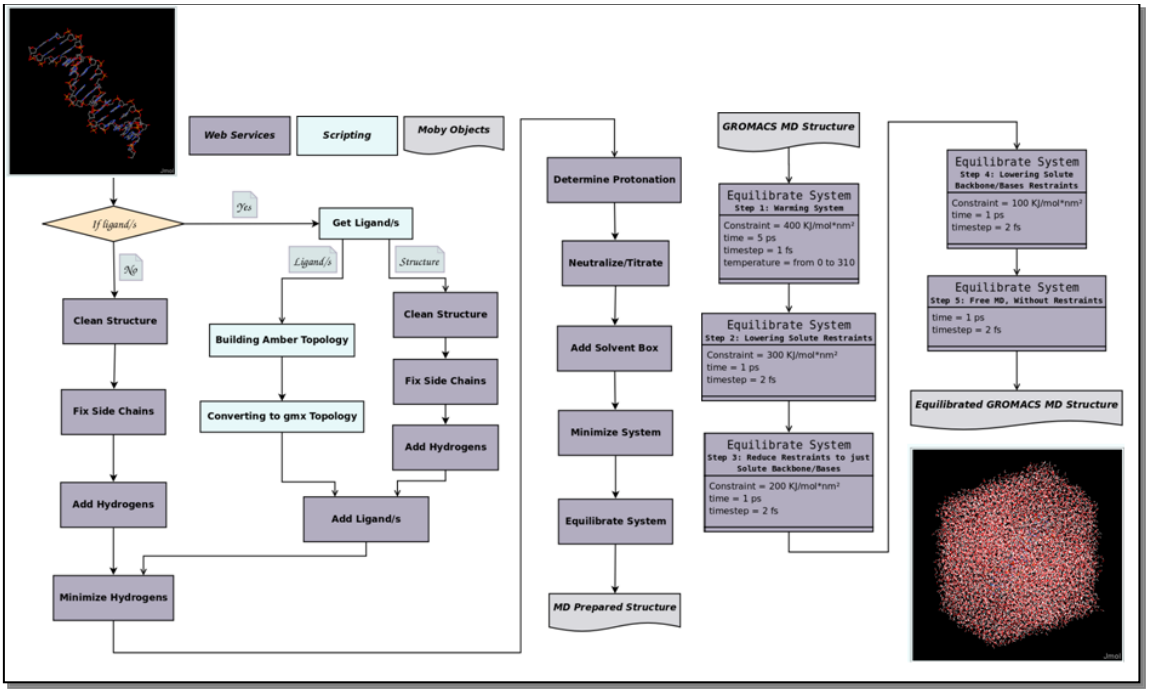
Suppl. Figure S2: NAFlex workspace screenshots. For every project the workspace holds all intermediate structures and results, organized as a tree view, allowing the user to track the history of operations performed. For every entry of the tree, a number of operations are offered to the user, according to the type of data. Moreover, download tools and Jmol visualizations are also available.

Suppl. Figure S3: Example 1: Atomistic Molecular Dynamics from a nucleotide sequence. a) Title and input selection. b) DNA/RNA nucleotide sequence input. c) Structure quality check. d) Summary of the final structure selections and integrity checks done. e) Operation selection: Complete Setup for running a molecular dynamics simulation with Amber package. f) Workflow progress in real-time. g) Downloading the final results.

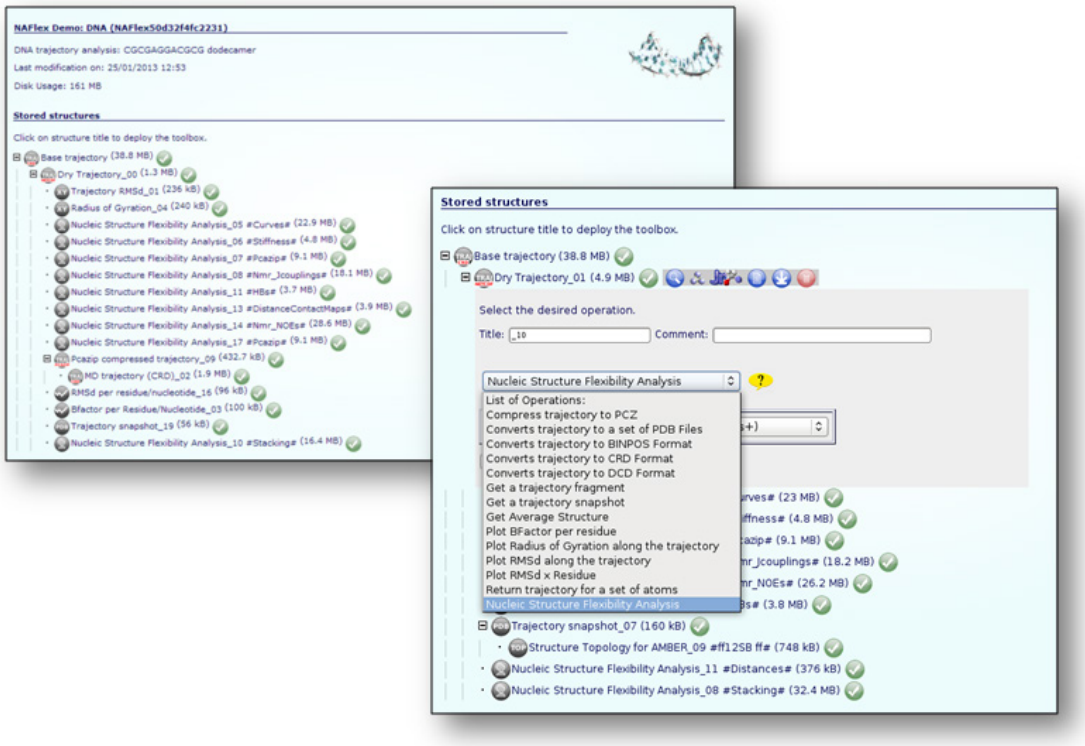
Suppl. Figure S4: Example 2: Mesoscopic simulations from a nucleotide sequence. a) Title and input selection. b) DNA/RNA sequence. c) Inspection of the coarse-grained nucleic acid used as starting conformation. d) Operation selection: Coarse-grained DNA Elastic Mesoscopic Model. e) Jmol visualization of the collected ensemble. f) Flexibility analysis of the Coarse-grained simulation: distance contact map.

Suppl. Figure S5: Example 3: Nucleic Acids flexibility analysis. a) First step, Nucleic acids flexibility operation selection from an uploaded trajectory. b) Time course (and histogram) plot of the *Twist* at the first CpG step. c) Time averaged representation of the *Roll* for all the sequence overlapping tetramers, with X-Ray and MD-reference values for other oligos. d) Nucleobase-Nucleobase hydrogen bond and stacking energies; High values in the top-left to bottom-right diagonal show HB interaction energies for base pairs, while top-right to bottom-left diagonal values show stacking interaction energies. e) Plot of the canonical hydrogen Bond distances; a grey-shadowed region indicates the expected values. f) Time-dependent analysis of a NOE along the trajectory.

Suppl. Figure S1

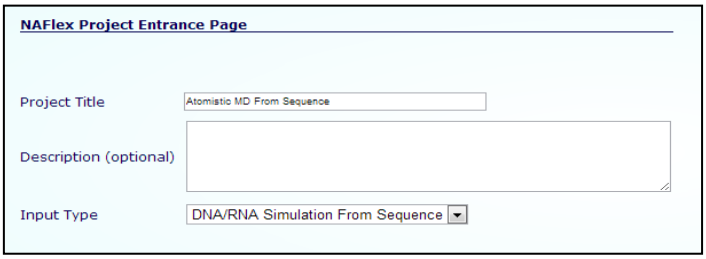


Suppl. Figure S2

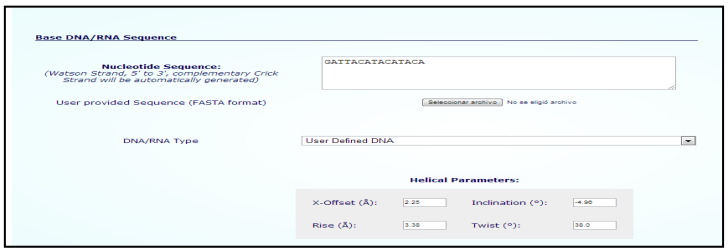


Suppl. Figure S3

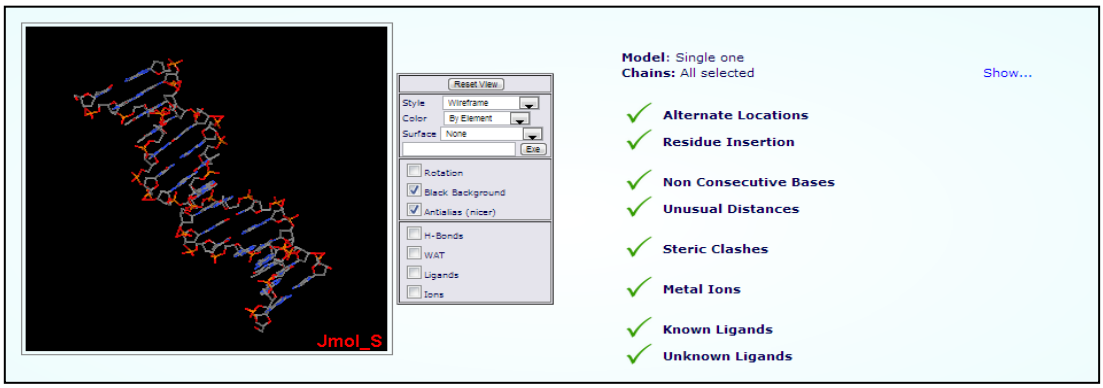
(a)



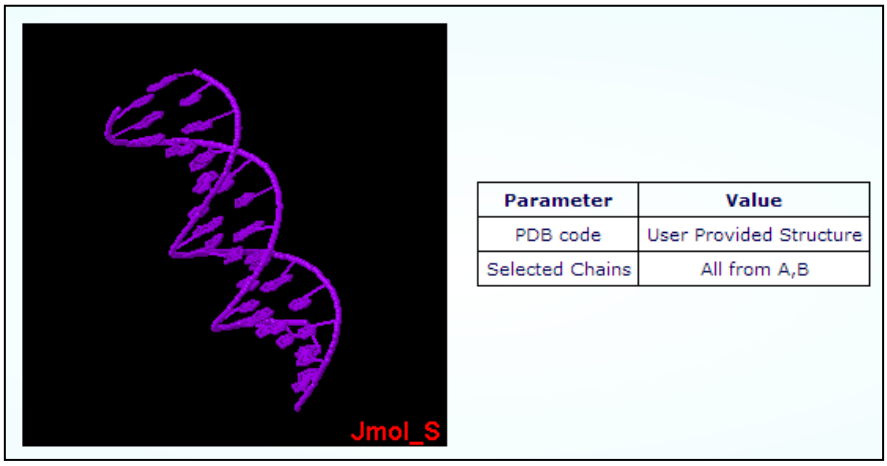
(b)



(c)

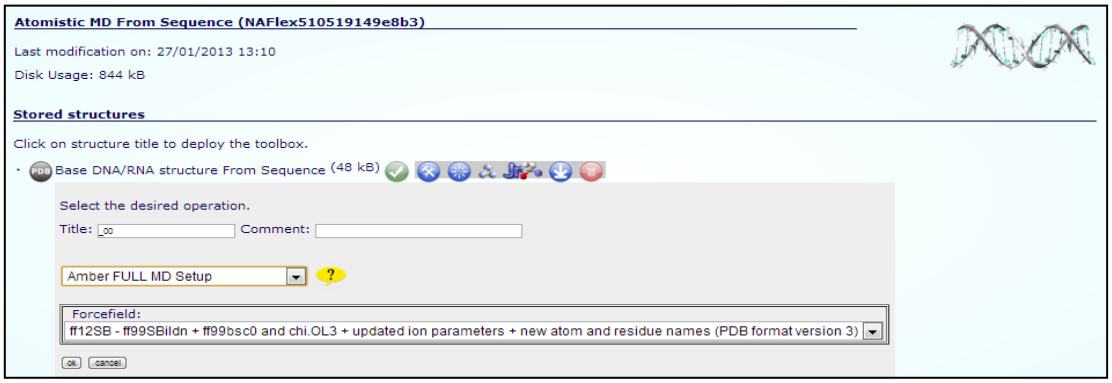


(d)

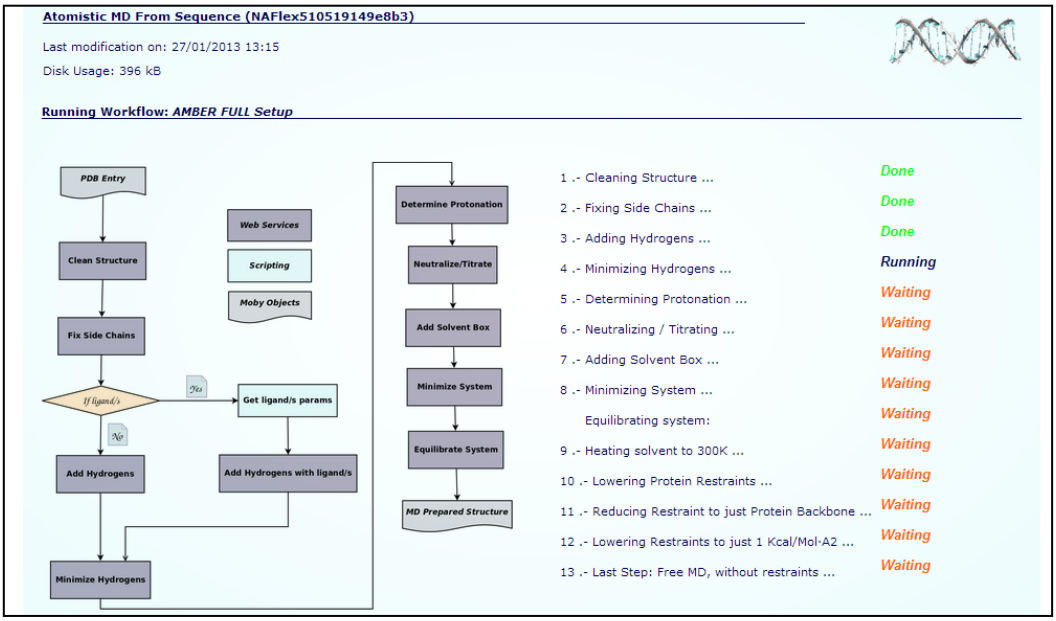


Suppl. Figure S3 (cont.)

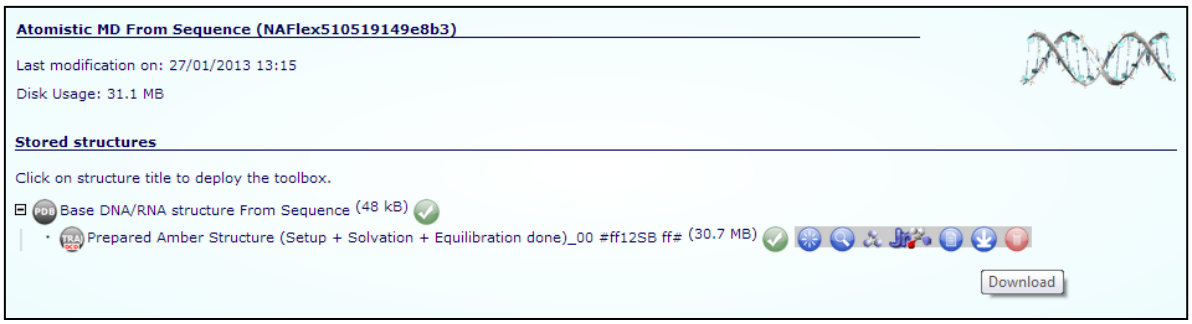
(e)



(f)

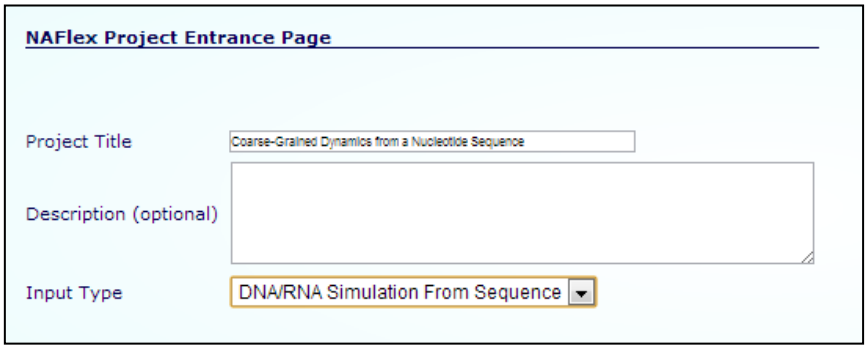


(g)

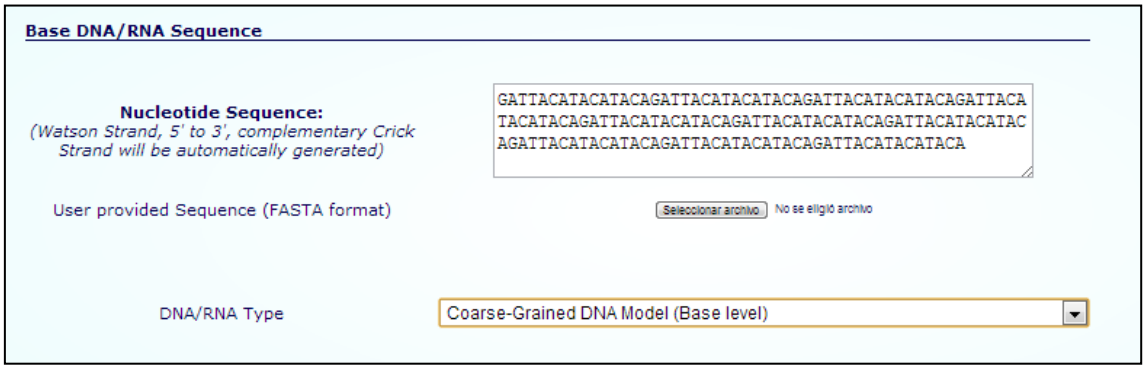


Suppl. Figure S4

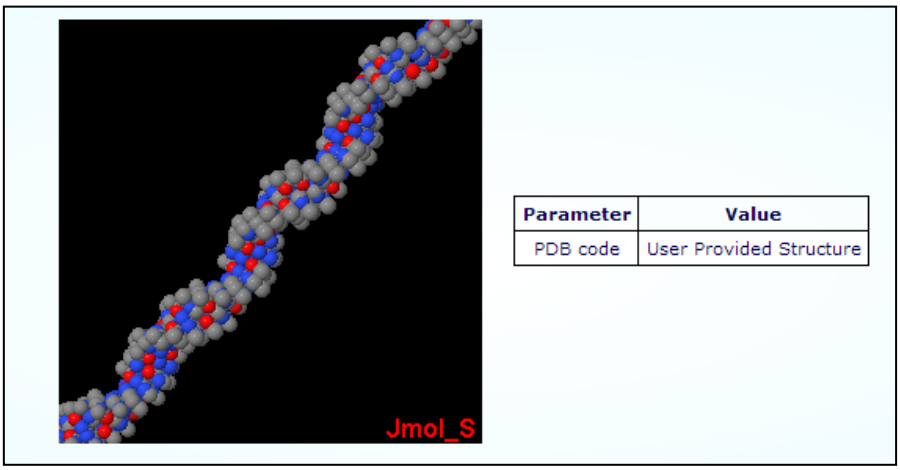
(a)



(b)

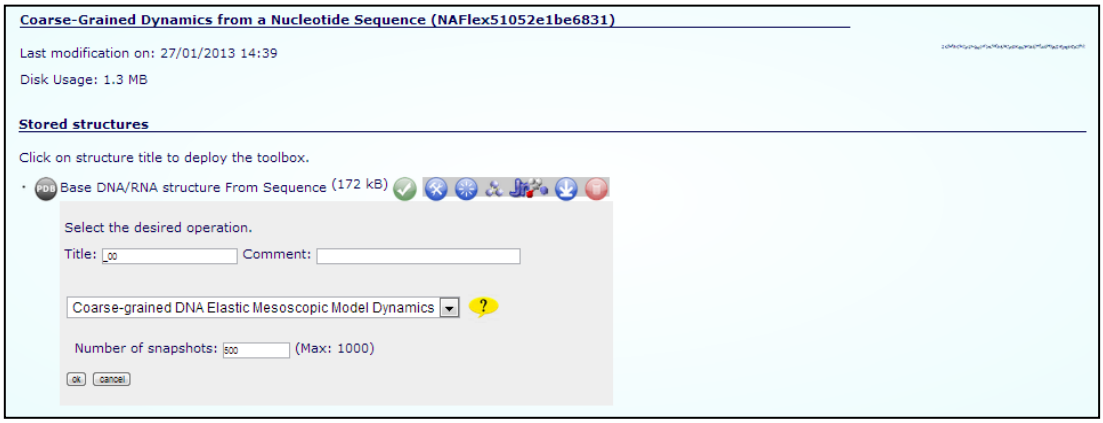


(c)

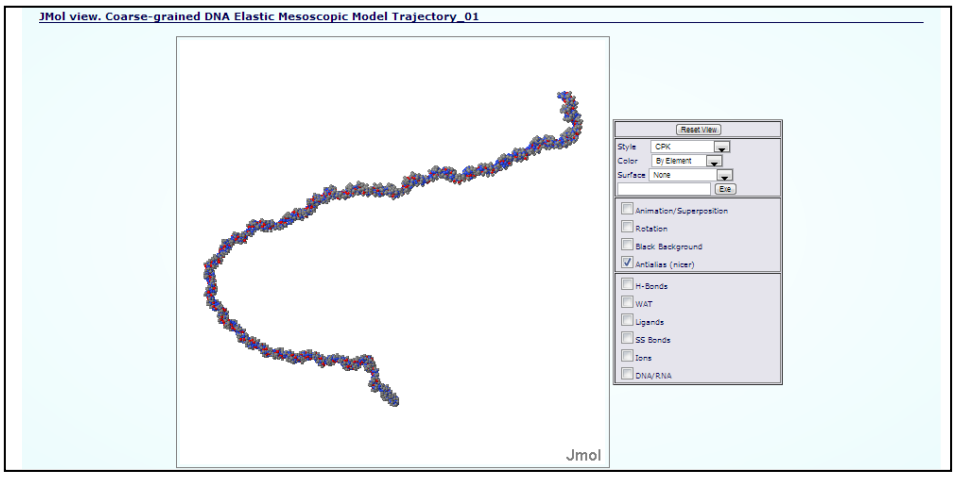


Suppl. Figure S4 (cont.)

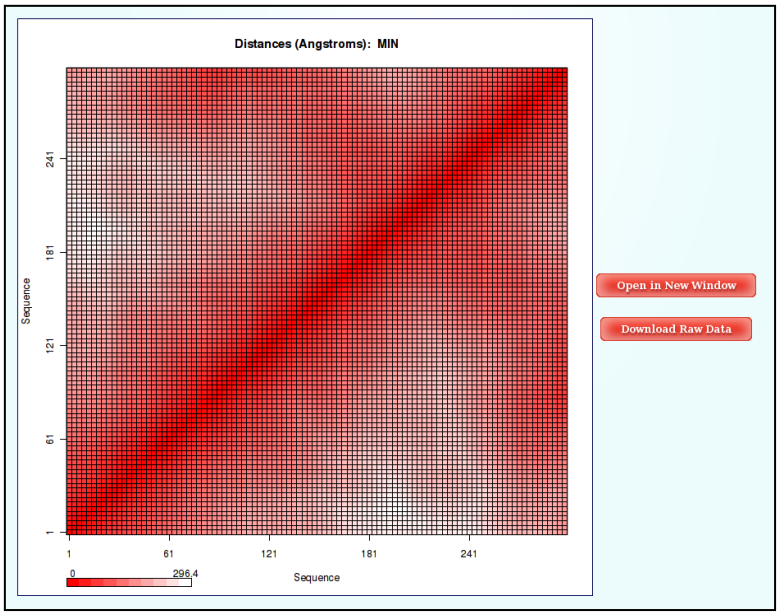
(d)



(e)

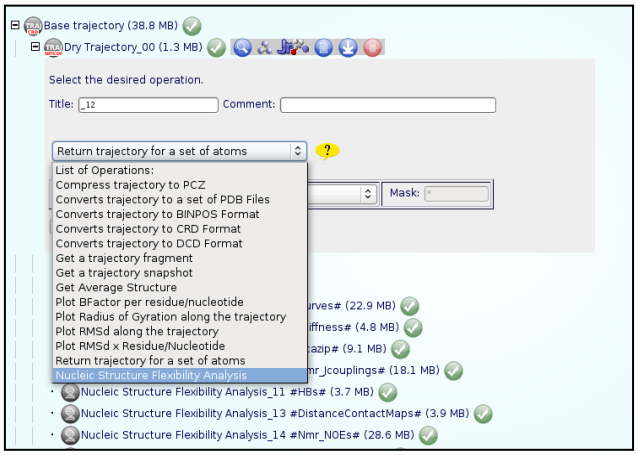


(f)

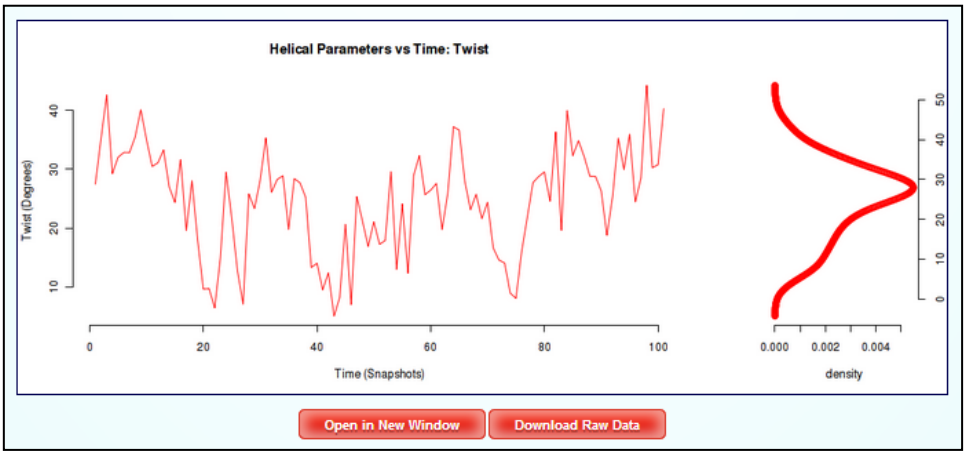


Suppl. Figure S5

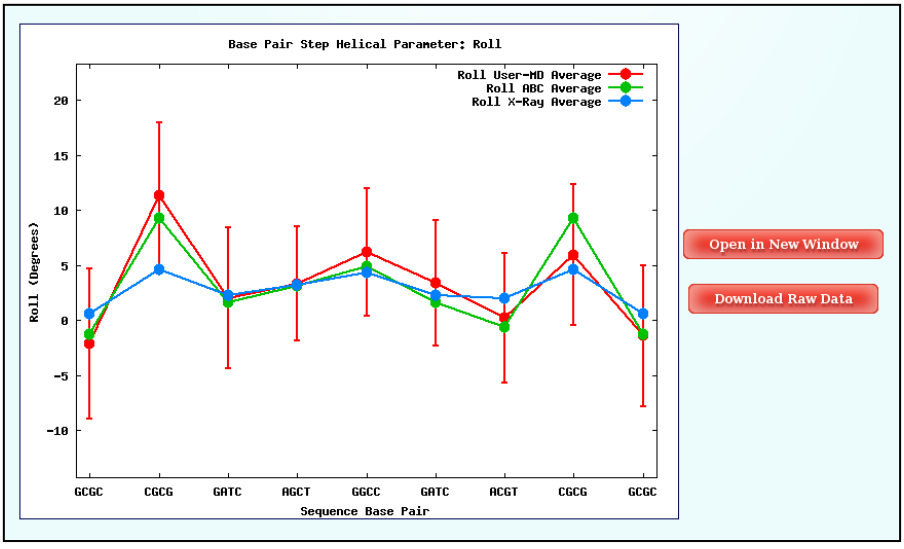
(a)



(b)

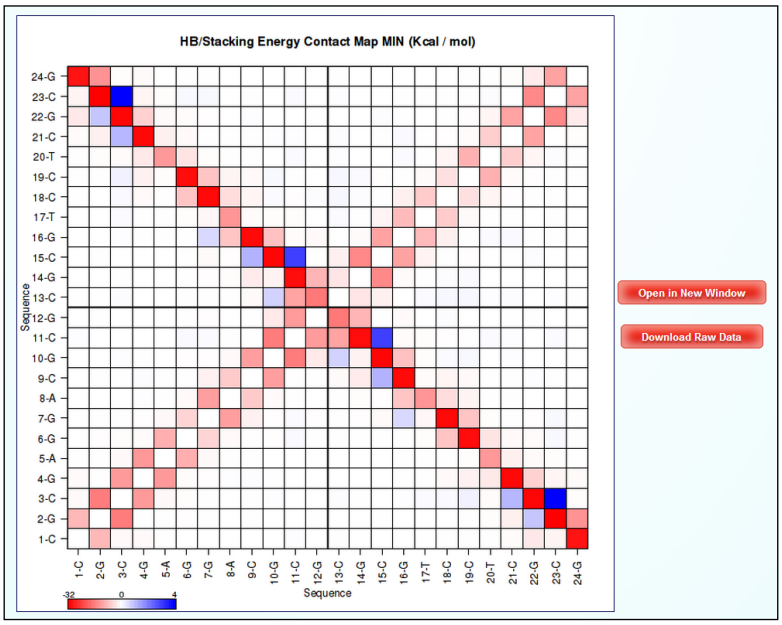


(c)

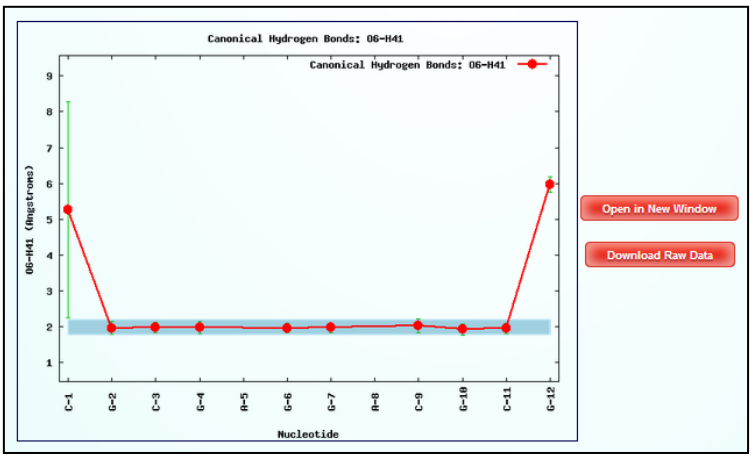


Suppl. Figure S5 (cont.)

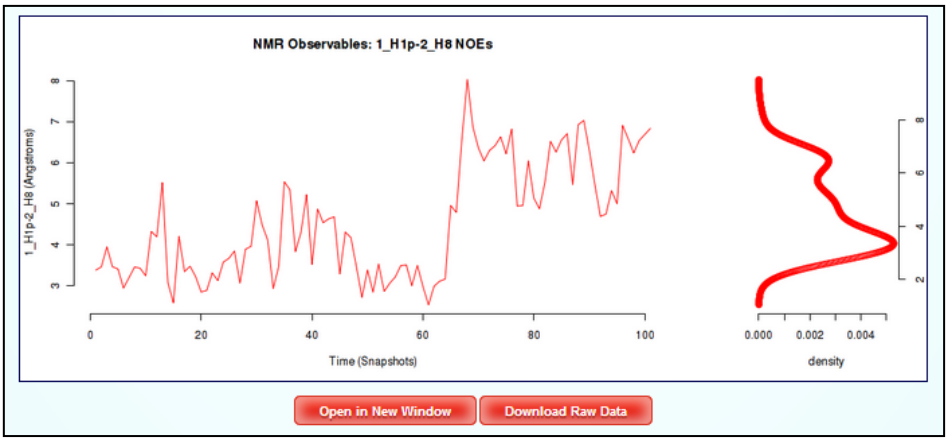
(d)



(e)



(f)



5.6 Section Bibliographic References

[1] R. A. Laskowski, "PDBSum new things," *Nucleic Acids Res.,* vol. 37, pp. D355-D359, 2009.

[2] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Durmousseau, M. Feuermann, U. Hinz, C. Jandraslts, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard and H. Hermjakob, "The IntAct molecular interaction database in 2012," *Nucleic Acids Research,* vol. 40, no. D1, pp. D841-D846, 2011.

[3] G. Lopez, A. Valencia and M. Tress, "FireDB - a database of functionally important residues from proteins of known structure," *Nucleic Acids Research,* vol. 35, no. 1, pp. D219-D223, 2006.

[4] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, "CATH - a hierarchic classification of protein domain structures," *Structure,* vol. 5, no. 8, pp. 1093-1109, 1997.

[5] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, "SCOP: A structural classification of proteins databaase for the investigation of sequences and structures," *J. Mol. Biol.,* vol. 247, pp. 536-540, 1995.

[6] Y. Lu, R. Wang, C. Y. Yang and S. Wang, "Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes," *J. Chem. Inf. Model.,* vol. 47, no. 2, pp. 668-675, 2007.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.,* vol. 215, pp. 403-410, 1990.

[8] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics,* vol. 22, pp. 1658-1659, 2006.

[9] C. Lipinski, F. Lombardo, B. Dominy and P. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.,* vol. 23, pp. 3-25, 1997.

6. Global Discussion

Breakthroughs such as the sequencing of the whole human genome (followed by that of a large list of species, and by structural genomics projects) have led to a spectacular growth in sequence, structure and function-related data of crucial importance for the scientific community. For an efficient access to these varied types of information data, they must be stored not only in a very structured way, but also using interconnection methods that allow us to easily joining distant-related data. In this field, relational databases have played a key role in bioinformatics.

Databases gathering together basic information such as protein sequences, structures or domains are nowadays essential in bioinformatics studies. Biological DBs are present in the majority of the recent bioinformatics research projects thanks to their ability to easily connect this basic information with complex data such as disease-related information or function. That has given rise to integrated initiatives such as the EMBL-EBI web portal (http://www.ebi.ac.uk/services), where users can interrogate the whole range of freely available and up-to-date molecular databases stored (DNA/RNA, gene expression, protein, chemical biology, ontologies, and more) [1]. It is clear that in the days coming, databases will continue to be a crucial part of the bioinformatics world, maybe changing the paradigm (going to a cloud approach), but still following the initial idea of storing and cross-referencing biological data. European initiatives like Elixir (http://www.elixir-europe.org/) seek to provide permanent infrastructure to maintain "*core*" databases as the necessary basis of bioinformatics data.

Together with the important usage of databases in the biological field, a growing need to make data available to the scientific community appeared. That need led to the first on-line servers, offering information usually related to a single database. With the evolution of computer power, data storage capabilities, and internet bandwidth, bioinformatics on-line servers proliferated, becoming more and more complex, offering not only direct access to stored databases but also the possibility to remotely run bioinformatics algorithms and tools. This need is becoming very evident in the field of molecular simulation, where the improvement in algorithm and in computational resources has raised the possibility to perform the so-called *High-Throughput* (*HT*) studies, where usually thousands of calculations are run at the same time either in a GRID or in a single machine. This new type of work, and the fact that the use of simulation tools has extended out of a limited group of experts, has generated important informatics problems that need to be solved.

In this thesis, we have developed computational tools to facilitate the study of one of the most important and still not well understood feature of macromolecules, their dynamics and flexibility.

Structural Databases

The current exponential growth in available structural data thanks to initiatives such as *Structural Genomics* and the *Protein Structure Initiative* [2, 3] generates an increasing need for secondary databases dedicated to analyse, process and present this large amount of information in a usable form.

Relational databases for storage of structural data (Section 5.1) and subsequent efficient retrieval were designed to confront appealing issues studied in protein structure field. Apart from the metadata directly extracted from the PDB files, we focus our interest in protein active sites and protein-protein recognition regions. Our *Active Site* DB (section 5.1.2, Fig. 5.3) followed the same approach used by a well-known database, *FireDB* [4], focusing special attention to biological and pharmacologically relevant ligands. Information such as the one stored in *FireDB* and in our own DB is still of main importance in the field, and updated publications (*FireDB* [5]) and even new projects following very similar approaches (*BioLiP* [6], *ProBiS* [7]) are still appearing. Our *Structure Important Parts* DB was designed to deal with the need of an automatic way to obtain datasets of already known protein-protein interactions, to work in protein-protein docking or protein-protein interface prediction. At the time of the development of this DB (2007), the main dataset used in the field came from a published benchmark by the group of Zhiping Weng (Protein-protein docking benchmark, currently in version 4.0 [8]), but it was not updated regularly enough in comparison with the exponential growth of new solved structures stored in the PDB. With our DB, weekly updates allowed us to obtain not only larger sets, but also information regarding real atomic contacts in an automatic way. The last version of the protein-protein benchmark is from 2010, and this methodology is nowadays replaced for new HT approaches obtaining protein–protein information from complete interactomes [9, 10].

The developed infrastructure has allowed us to quickly and easily obtain the protein datasets to be used in different projects. Cross-relations of distinct structural data (protein parts, domains, active sites, etc.) offered us the possibility to apply a set of filters to the whole PDB in order to select just the most appropriate structures (or even structure domains/regions). An automatic procedure to update all the databases (coupled to the PDB mirror synchronisation) ensures an up-to-date information storage facility. They were used for the selection and filtering of the protein datasets employed in the published studies 1, 3, 5 of the list of publications [11, 12, 13]. It is used in all the web servers presented in this thesis, as well as the web tools accompanying publications number 8 and 11 from the list of publications [14, 15]. Examples of real queries that were performed using these databases are:

- Datasets to be used in protein-protein docking prediction studies. Examples could be: obtain a benchmark dataset of hetero-dimers from the complete PDB, together with atoms involved in the protein-protein interface, or automatically obtain the set of known PDB structures having a certain monomer (90% sequence similarity) crystallized in complex with one or more protein domains. Work in collaboration with the *Protein interactions and docking group* from *Barcelona Supercomputing Center (BSC)* life science department.

- Obtain a complete dataset of non-redundant monomeric proteins covering the whole PDB, minimizing the possible problems for use them in MD studies: retrieve structures only with a desired resolution, removing for example,

membrane proteins, proteins with gaps in the structure (missing residues), proteins with non-standard residues and proteins containing polymeric or non-constitutive ligands difficult to parameterize by automatic procedure (works nº 1 and 5 from the list of publications [11, 13]).

- Obtain a complete set of homologous protein pairs in the whole PDB being one of them a thermophilic protein and the other a mesophilic one, to be used in a comparison of dynamic properties between them. Work in collaboration with the *Theoretical and Computational Chemistry Group* from *Institute of Advanced Chemistry of Catalonia (iQAC),* paper in preparation.

- Obtain the set of PDB proteins crystallized together with water molecules, with the number of water molecules per protein and their coordinates, to compare with structural water molecules predicted by programs used in the setup of proteins done before running a MD simulation (work nº 14 from the list of publications).

- Obtain a set of ligand active sites in the PDB having a crystallographic water molecule in the vicinity, to be used in studies about the importance of solvent in protein-ligand docking approaches (work nº 14 from the list of publications).

Molecular Dynamics Library

Using a high throughput approach to setup, run and analyse MD simulations, we were able to implement an ambitious project to obtain a protein dynamics library (MoDEL, Section 5.2). A non-redundant set of 1,595 monomeric proteins (at the time of publication), covering a 38% of the whole PDB and a 30% of DrugBank [16] (including homologous structures) were chosen and automatically prepared to run atomistic MD. This allowed us to move from the original PDB structure to a completely equilibrated system formed by the protein surrounded by a box of water molecules and counterions. Simulations to generate MoDEL database were run in the *MareNostrum* supercomputer at BSC. Finally, a complete set of flexibility analysis implemented in our group was applied to the whole library.

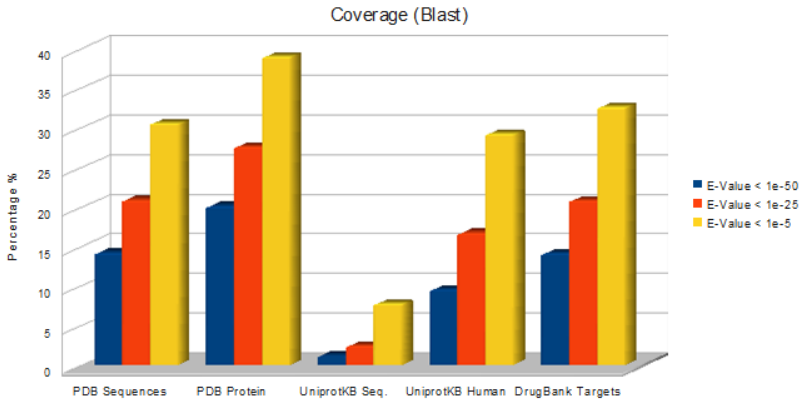


Fig 6.1.- MoDEL Coverage. *MoDEL database covered at the time of publication up to 38% of the entire PDB and 30% of DrugBank (removing redundancy).*

The difficulties in large MD projects related to computational power and storage needed had resulted in a couple of unsuccessful projects, *BioSimGrid* [17] and *P-Found* [18]. Both initiatives aimed to work as a repository (a "dynamic" PDB) were scientific community would deposit their own trajectories. Unfortunately, at the time of these projects (2004-2006), computer power, data storage and bandwidth was still far behind their needs. A few years after, a couple of new approaches appeared, our own *MoDEL* database, and a similar project called *Dynameomics* [19] (see introduction section 1.5). In these cases, not only a big storage infrastructure was built, but also populated with already computed simulations. The main difference between both projects relies on the central objective. While our *MoDEL* approach aims to maximize the PDB coverage, *Dynameomics* wants to study protein stability issues, folding and unfolding, through simulations at different temperatures. Later to MoDEL a new database has been generated by Grubmuller's group: Dynasome [20], with which they try to classify proteins according to their internal mobility patterns, and obtain correlations between structure, dynamics and function.

At the technology level, for an efficient storage of massive MD resulting data, we decided to follow a dual approach, working with the described analysis and metadata database, but also keeping trajectory raw files in common disk raids. This approximation differs from the one used in *Dynameomics* project, where trajectory coordinates were loaded into a multidimensional database (*MOLAP*) instead of using common flat files, and queries were directly addressed to this DB. Although this infrastructure was shown to be efficient, it was also inconvenient when working with the original trajectory files was recurrent, as it needs a previous format conversion.

More recently, *iBIOMES* (*integrated BIOmolEcular Simulations*) project redefined this dual approach, implementing a distributed solution to data storage and sharing across research laboratories, based on the new cloud technology [21]. *iBIOMES* system provides a command-line interface along with a web interface for extra visualization components. Apart from the metadata database, iBIOMES uses a completely distributed file system approach (*iRODS, integrated Rule Oriented File System* [22]), containing another internal database, (*iCAT, iCatalog*) that helps in efficient searches within the distributed files, mimicking the internal engines used in the new *NoSQL databases*. Nowadays, with the fast evolution of these so-called *NoSQL databases* based on distributed systems, new frameworks like the one used in iBIOMES with storage of large files in distributed and replicated infrastructures are starting to appear [23], and are expected to become a standard approach for the years to come. MoDEL database is now being adapted to such approach using *Cassandra* (http://cassandra.apache.org/) as file system manager.

Current MoDEL database store raw trajectory files in the original format (*netCDF* binary files, see Materials and Methods section 3.2.2), with basically two different versions: original raw trajectory, with solvent and counter ions, and "dry" trajectory, with just the information of the protein (and ligands if present). The whole set of trajectories takes up nearly 20TB of disk space. Storing of these files in the original trajectory format allows the direct execution of new analysis, without any previous conversion needed. A hierarchical folder strategy reaching a number of 141,089 nested directories was designed to obtain a greater input/output efficiency. Finally, in order to make the complete simulation trajectories available to the scientific community, we created compressed trajectories for each of the proteins with our *PCAZip* package [24], which, using an algorithm based on *Essential Dynamics*, is able to reduce trajectory size

with up to a 10-fold compression. That is really useful in projects like MoDEL, where downloading a set of MD trajectories would have been very slow and tiresome. We know, both from the web server statistics and from the scientific community feedback, that these compressed trajectories are downloaded and used regularly. Moreover, this work, together with the work published using a representative set of proteins from *MoDEL* (*μMoDEL*) comparing MD results using different force fields (nº 3 of the list of publications) have a non-negligible number of 176 citations (June 2014), indicating the impact of the project. *MoDEL* is an on-going project, it was designed and prepared to be extended with new analysis, and new simulations are being added every day, including trajectories in unusual environments (gas phase [25], urea [26]).

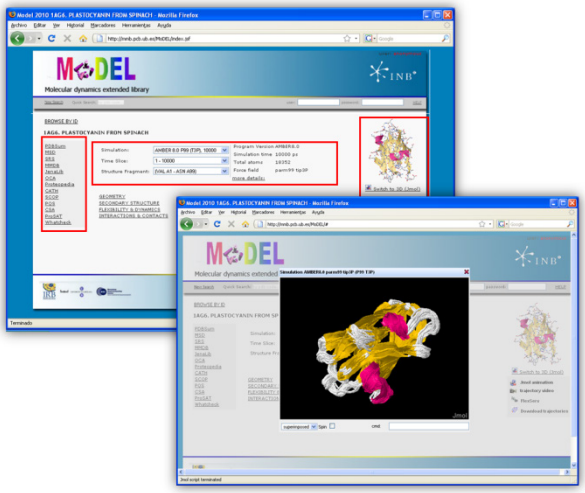


Fig 6.2.- MoDEL web server. *Web server for retrieval of flexibility information from our Molecular Dynamics Extended Library.*

Having a large simulation database is not very useful, unless analysis methods are also moved to the high-throughput regime. In this thesis, as a proof of concept of the information power contained in *MoDEL*, a high throughput study of solvent environment was performed. This massive study (information on more than 16 million water molecules (Section 5.3), provided a comprehensive picture of the dynamics of protein hydration. The study revealed a much more dynamic behaviour of the hydration process than expected with even water molecules trapped in protein internal cavities being very mobile. Water molecules registered as structural waters are still continuously interchanging with other solvent molecules, reaching mean residence times of just a few nanoseconds, as can be seen in the InlB example of section 5.3. These results change the usual notion of structural waters. Also, a significant inconsistency between MD usually defined Hydrogen Bonds and that with real favourable free energy was found. Indeed, the importance of water molecules in the dynamic processes of protein hydrogen bonds and the large amount of data provided by *MoDEL* trajectories for this analysis has resulted in an independent project that is now submitted for its publication.

Macromolecular flexibility for everybody

One of the main problems in molecular dynamics simulation is the difficulty in the preparation of the structure. System setup is a major challenge for non-expert users, to the limit of giving up the study due to the frustration in the very first steps. The result is either going to another method or use default procedures without knowing the meaning of the different steps done, with a high probability of obtaining artifactual trajectories (not easily distinguishable from the correct ones).

The most widely-used programs in the MD field, *CHARMM* [27], *NAMD* [28], *AMBER* [29] and *GROMACS* [30], come with a set of accompanying programs to help in the system setup process. A number of web-tools have appeared, allowing the user to perform a structure setup for a specific MD package running the corresponding programs through a graphical and user-friendly interface. Specific initiatives for all of the above-mentioned programs exist: *CHARMM-Gui* [31] and *CHARMMing* [32] for *CHARMM* MD package, *VMD* [33], with a set of system setup plug-ins and *ClickMD* [34] for *NAMD* package, *Glycam-Web* (*Glycam Biomolecule Builder*) [35] and *AMPS-NMR* [36] for *AMBER* package and *Guimacs* [37], *Gromita* [38], *jSimMacs* [39] and a *PyMOL GUI* plugin [40] for *GROMACS* package. The majority of these tools are typical standalone GUI tied to a programming language (e.g. Python [40], Java [37, 39], Perl/tcl-tk [33, 38]), that usually need an installation process, are typically linked to a given program, and sometimes can only be used in particular operatives systems (UNIX vs Windows). Web-based tools, on the other hand, are more easily accessible due to the lack of downloading and installation [31, 32, 34, 36].

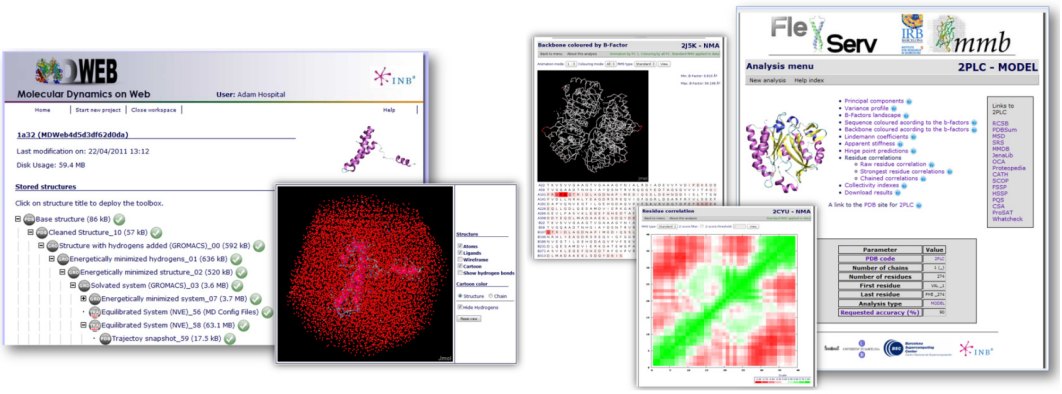


Fig 6.3.- MDWeb & FlexServ web servers. *Web servers for setup, run and analyse Molecular Dynamics and Coarse-Grained Dynamics simulations.*

The difficulty of the setup process is even more complicated for HT MD studies, since a non-negligible number of issues must be solved before obtaining a prepared system to run a simulation (see Fig. S8 of section 5.2). These includes, among others, identification of missing regions in the structure, unparameterized ligands, and even errors in the interpretation of experimental data (e.g. improper amide or chirality assignments). A typical MD project commonly requires a manual setup, with a considerable effort in solving all possible structural issues, but this cannot be done in HT studies. Both problems require the automation of simulation setup. For the non-expert user this implies the need for the generation of a set of pre-configured workflows, with a complete explanation about all what is done within, programs used,

methods applied, etc. For the expert users, this will imply the generation of a robust software suite with which they will be able to prepare a large set of unrelated macromolecular structures to be simulated automatically using a variety of program packages.

MD produces huge amount of data, which is currently provided in software-dependent format. Recently we and others in the European project Scalalife (*Scalable Software Services for Life Sciences*; http://www.scalalife.eu) [41] have tried to develop standards for the representation of molecular simulation data, we hope that such standards will be accepted in the field, since it will largely facilitate generation of new runs, and the analysis of trajectories. Such analysis requires efficient methods for storing and transmitting trajectory data, which is growing in size with the ability to perform larger and bigger simulations mainly due to the new HPC facilities. As for the system setup, each MD package offers a set of programs to analyze the resulting trajectories, but in this case the availability of web-tools to perform such analysis in a user-friendly and graphical way is scarce. As a side product of the MoDEL project we developed a set of data types and informatics tools (web services and workflows, Section 5.4) related to Molecular and Coarse-Grained Dynamics simulations in order to obtain a platform to port these studies to the high-throughput regime. The flexibility of the web services infrastructure helps in building personal workflows. As each of the WS encodes a small step of the simulation pipeline, they can be joined as desired to build specific workflows. The same flexibility allows the design and implementation of new WS to be added to the MDMoby group. By now, MDMoby contains services to setup, run and analyse dynamic trajectories obtained either from atomistic MD simulations as from low-resolution CG simulations (see section 5.4.2 suppl. material), and are part of the whole set of WS offered by the INB platform (http://www.inab.org).

Our work was not restricted to protein, but we also adapted our tools to the representation of nucleic acids. The developed tools allowed us to obtain dynamic information, from low-resolution coarse-grained to atomistic molecular dynamics simulations. For the sake of completeness, we also add methods to build NA structures from just the nucleotide sequence, allowing then the full pipeline, from sequence to structure, dynamics and flexibility. The developed infrastructure (NAFlex, section 5.5; Fig. 6.4) is flexible enough as to be useful both for highly expert users and newcomers in the field.

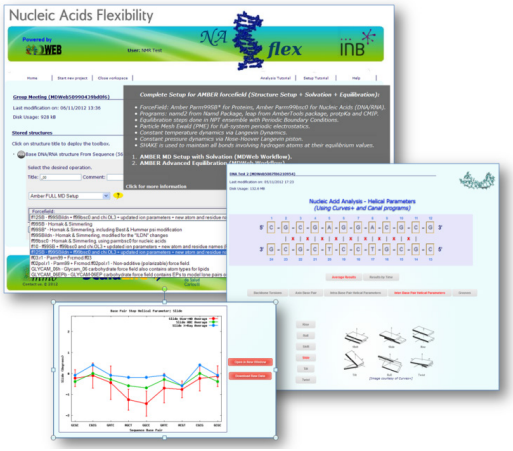


Fig 6.4.- NAFlex Web Server. *Web server for the study of nucleic acid flexibility.*

An important utility provided by our web interfaces for dynamic analysis of macromolecules are the tools for the quality of input structure as a first step of the simulation process (Fig. 6.5). The user can select which parts of the structure are going to be used, is warned about possible structural problems, such as atom clashes or missing residues, and sometimes even has the possibility to correct some of these dangerous problems as can be the case of wrong amide assignments. The set of checks are integrated in a very intuitive design, with an interactive Jmol applet showing all the resulting checks on the structure, zoomed on the atoms involved. Structural quality is automatically computed every time the user selects one of the possible options. Implementation of this server part was made by Pau Andrio, co-author of the work.

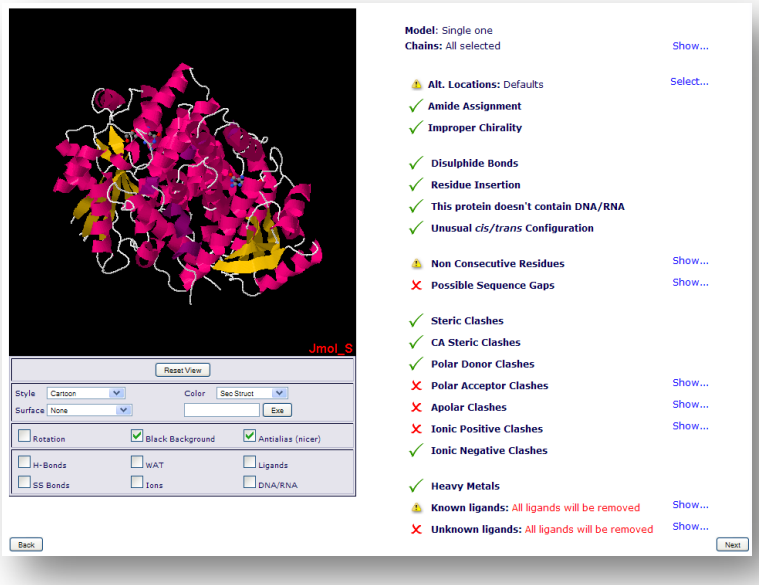


Fig 6.5.- MDWeb & NAFlex structure checking.
Possible structure problems are checked as a first step in our MD server.

Taking advantage of the power offered by the Web-Service technology, we designed and implemented pre-configured workflows to run the entire setup pipeline for the most used MD packages *NAMD*, *AMBER* and *GROMACS*, with the possibility to use different existing force-fields. The complete list of available pre-configured workflows included in MDWeb and NAFlex can be found in the supplementary material of paper 3 (section 5.4.2). These workflows can be really complex, and thanks to the web servers, users can run them with just one mouse click, easing the usage of MD methods to novices. That workflow approach had already been used in automatic platforms like *Guimacs* [37] with its *AutoMACS* option or *CHARMM-GUI* [31], with its *Quick MD Simulator*, but as far as we know, this was the first time that a single platform offered a complete set of different workflows for the most well-known existent MD packages and force-fields.

In spite of its complexity, our flexibility platform has also the possibility to be easily moved to other hardware, thanks to its internal configuration (Fig. 6.6). Virtualization (e.g. Xen virtualization project [42]) together with the layer-distributed configuration (web interface, internal engine, computing cluster and storage facilities) allows a flexible transport to other machines, through a simple ISO image. *Sun Grid Engine* (*SGE*) queue manager [43] is the responsible for the processes distribution, and is the only part that needs to be configured, depending on the number of available

machines. That gives the opportunity to install the complete pipeline in other groups, to be used in HT projects or through the integrated graphical web interfaces.

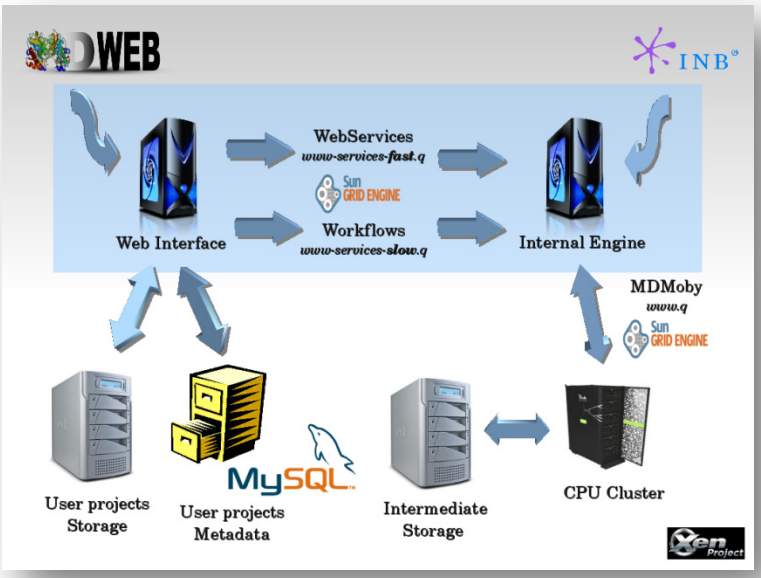


Fig 6.6.- MDWeb & NAFlex internal configuration. *Virtualization and SGE queue manager ensures an easy installation of the complete flexibility package in any type of hardware.*

The increasing number of tools for the automatic preparation and run of MD simulations has provoked the appearance of a skeptical group of scientists pointing out that these software packages raise the risk for MD field to be pervaded with unreliable results, due to the low expertise needed to use them [44]. It is clear that the probability to obtain misleading results is higher for novice users, but an immense effort is put in accompanying on-line servers with complete manuals and step-by-step tutorials, that sometimes includes also MD basic information helping them in understanding the basis, while running a real simulation. The structure quality step incorporated in our MD servers is also aimed to tackle this point, reducing the possibilities of failure due to a lack of experience.

The impact of platforms like those developed here is demonstrated by the usage statistics of a platform like MDWeb (Fig. 6.7). We have already reached 1,000 registered users, keeping a monthly average number of visits higher than 1,000 since last year, with a 50% of returning visitors. Continuous feedback with registered members is helping us identifying possible new methodologies to offer, as well as real end-users interests and opinions about the platform. The maintenance of this kind of servers is an issue to be taken into account. As an example, a couple of the published tools presented, *ClickMD* [34] and *Glycam-Web* [35], are not available as automatic MD platforms anymore, and the majority of the other software and servers presented had stacked in the first release, using software versions considered obsolete as of today. We are very aware of that, and for this reason we are continuously working in adapting our tools to the fast evolution of MD algorithms, although a thoroughly set of test cases must be performed before any new version should be released.

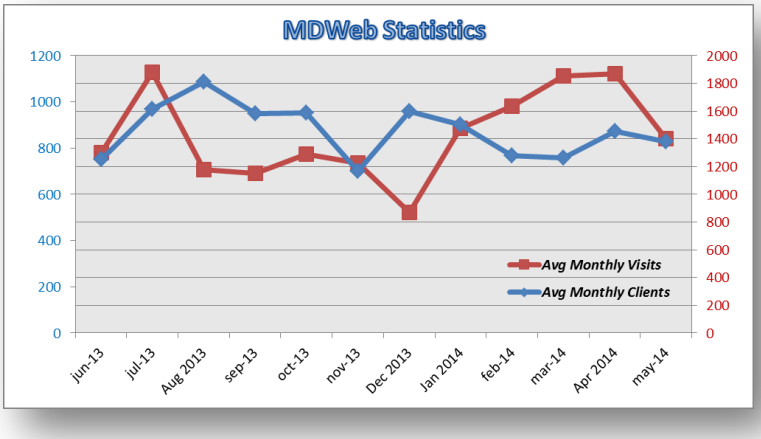


Fig 6.7.- MDWeb usage statistics. *MDWeb server has been used this past year for an average number of nearly 900 users per month, with a monthly average of 1,500 visits.*

The infrastructure developed here has been used also in projects developed in the group. New students joining the group launch their first simulations using these platforms, and even entire projects have been dealt with using them. For example, MDMoby workflows were the main tool for the mutations and setups of the large set of EGFR MD simulations in publication nº 12, and a new study about the different flexibility accounted by thermophilic vs mesophilic proteins that is currently in preparation was entirely performed using MDWeb server.

The next step in this long-term project will be the integration of all web servers presented in this thesis, together with other structural flexibility methods and web pages developed in the group in a single web portal: *FlexPortal, INB Integrated platform for macromolecular flexibility* (Fig. 6.8). That future MMB-INB portal will provide a private workspace (similar to that found in MDWeb and NAFlex servers), where an interested user can store all his/her calculations, regardless of the tool used to generate it. The platform will give the possibility to cross-reference all the macromolecular flexibility tools presented in this thesis with new methods such as conformational transition dynamic analysis (list of publications number 8 and 11 [14, 15]). As an example, a user would be able to enter a FASTA sequence which with the portal will obtain a family of related structures, offer a simulated MD trajectory if one of them was included in the MoDEL library or, alternatively, prepare the structure with MDWeb to obtain such MD trajectory, that, in turn, could be analysed using FlexServ. All this information will be stored in the user's private workspace. We hope that FlexPortal, already in construction, will become the common-site for people aiming to understand the dynamics of macromolecules.

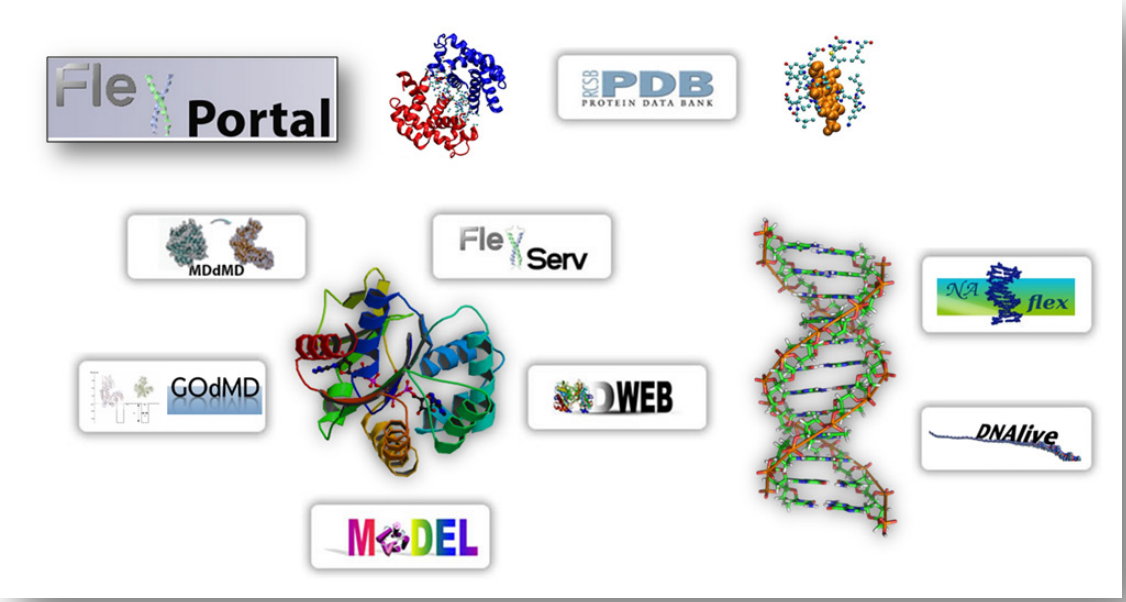


Fig 6.8.- FlexPortal: INB Integrated platform for macromolecular flexibility. *Integration of all tools presented in this thesis, together with other flexibility related servers, for proteins and nucleic acids.*

Summary of On-line tools

In order to make our tools publicly available to the scientific community, we followed the common bioinformatics approach of presenting all our platforms through on-line web servers. Thus, our contribution to the field of macromolecular flexibility is presented in a set of web pages that are being briefly summarized in the next sections.

- **Structural database infrastructure**
    - http://mmb.irbbarcelona.org/pdb

Molecular Modelling and Bioinformatics (MMB) PDB mirror web server is an on-line platform to query the in-house relational databases storing structural information. It offers a set of searching possibilities, by PDB code, Uniprot accession number, resolution, compound type, experimental type and keyword. Information presented is all PDB metadata (classification, type, deposition date, title, source, authors, resolution, experimental type, etc.) as well as ligand information (if any), sequences by chain and links to cluster families (50, 70, 90, 95 and 100). Molecule can be visualized with Jmol or Chime software, and downloaded from the local mirror repository. Links to other databases are provided for deeper analysis.

- **MDMoby/MDWeb (Molecular Dynamics on Web)**:
    - http://mmb.irbbarcelona.org/MDWeb

MDWeb server is a web graphical interface to facilitate the use of the powerful MDMoby molecular dynamics web services and workflows, designed to run MD and Coarse Grained dynamics simulations with different program packages and different inputs. It is divided in three main sections: structure setup, simulation run (limited to external users to test trajectories of max. 0.5ns) and trajectory analysis. A quality checking test is run if working with an input structure, which gives the user a graphical overview of the possible problems that can later affect the dynamic simulation. Extensive manual and step-by-step tutorials for using the platform are offered, either on-line as well as with pdf files.

- **FlexServ (Flexibility Server)**:
    - http://mmb.irbbarcelona.org/FlexServ

FlexServ server is a compilation of protein flexibility analysis joined together in a graphical web page. To facilitate the extraction of dynamic information, three different Coarse-Grained algorithms (BD, DMD and NMA, see Materials & Methods) are integrated so that user may just enter the desired PDB code or structure to get a complete flexibility picture of the protein. All the analyses are represented graphically with very intuitive plots and an interactive Jmol applet. The whole flexibility analysis results can be downloaded together in a single compressed tar file.

- **MoDEL (Molecular Dynamics Extended Library):**
    - http://mmb.irbbarcelona.org/MoDEL

MoDEL server is a graphical interface connected to our library of MD protein simulations. Information about the different simulations together with the stored analysis is offered in an easy and intuitive web page. It is divided in three main sections:

- Links to external databases.
- Simulation information and analysis.
- Visualization and download.

MoDEL simulations can be accessed from just a PDB code or FASTA sequence (or file), but it also offers the possibility to browse all the available simulations by alphabetical/program order, or by *CATH* fold. Trajectories can be visualized with an interactive Jmol applet, or directly with a previously computed video. They are freely downloadable in a compressed PCZ format, with Cα, backbone and heavy atoms resolutions. A set of varied protein flexibility analysis based on FlexServ platform is offered, accompanied by graphical plots.

- **NAFlex (Nucleic Acids Flexibility)**:
    - http://mmb.irbbarcelona.org/NAFlex

NAFlex server is a platform to study nucleic acid-specific flexibility properties. It is powered by MDWeb server, taking advantage of its interface and atomistic MD tools. Apart from all the possibilities inherited from MDWeb/MDMoby, it adds:

- Run simulations: new Coarse-Grained methods, nucleic acid-specific, from very low-resolution *Colorless Wormlike-Chain* (*WLC*) model, designed for very large sequences to an almost atomistic resolution base-pair mesoscopic model designed to quickly obtain dynamic information for short sequences.

- Trajectory Analysis: a complete set of nucleic acid-specific flexibility analysis have been integrated in the server:

- Possibility to enter just a sequence and automatically built a desired nucleic acid structure type (DNA/RNA), with several kinds of pre-configured structures (right/left handed B-DNA, right handed A/A'-RNA) or alternatively built taking average experimental or theoretical helical parameter values.

Future Perspective

- **FlexPortal (Flexibility Portal)**:
    - Under construction at http://mmb.irbbarcelona.org/FlexPortal

In summary, along the years of this thesis, we have developed a set of computational tools to study macromolecular structure flexibility, covering the fields of protein structures, nucleic acid structures and even protein-nucleic complexes. All the platforms were made available through on-line servers, offering our results and new developed tools to the scientific community through the *Spanish National Institute of Bioinformatics* (INB).

Hardware & Software evolution

The fast evolution of hardware and high-throughput studies is increasing the rate of software growth in the bioinformatics world. A typical example of this trend can be seen in the MD arena, where new algorithms to work with parallel processors and GPUs have appeared, and where force-fields are in continuous evolution. This amazingly fast evolution applied to the macromolecular flexibility demands an exclusive dedication to have up-to-date platforms. In our case, a revision of MDWeb server is already prepared to be launched, with extremely important updates in MD programs used and force-fields (and bug fixes).

In the other hand, evolution of new computers now enables the previously unavoidable generation of a comprehensive view of the dynamic properties of macromolecules. Usage of new supercomputers and new efficient software, together with the already available HT setup and analysis tools would be enough to obtain a first approximation to a "dynamic PDB". Of course, simulations should be extended in length, as the currently state of the art simulations are reaching the microsecond time scale. An extended overview about high-throughput molecular dynamics simulations can be found in the annexed paper: *High-throughput molecular dynamics simulations: toward a dynamic view of macromolecular structure* (Annex I).

Different approaches to store the huge amount of data generated by HT projects are tested and compared in this thesis, with a slight tendency to the dual system: relational database to store the simulation metadata and flat files to store the raw trajectories. Transfer of trajectory files is another hot topic, that can be solved using our compressed PCZ files (see Materials & Methods, section 3.2.2) that, although loosing part of the original information, are able to reduce the file size to just MBytes. Ideas such the one in *BioSimGrid*, where the library was opened to the scientific community to deposit their own trajectories, working as a repository like the *Protein Data Bank*, is currently hindered by the lack of standards in Molecular Dynamics. Almost every MD package work with its own file formats for storing topology information, trajectories and simulation restart data. As noted above, initiatives such as the ontology in MDMoby and the *European Scalalife* initiative wanted to address this important issue, clue in a future world of trajectory sharing [41, 45].

Section References

[1] C. Brooksbank, M. Bergman, R. Apweiler, E. Birney and J. Thornton, "The European Bioinformatics Institute's data resources 2014," *Nucleic Acids Res.,* vol. 42 , no. (Database Issue), pp. D18-25, 2014.

[2] K. Khafizov, C. Madrid-Aliste, S. Almo and A. Fiser, "Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative," *Proc. Natl. Acad. Sci. USA,* vol. 111, no. 10, pp. 3733-3738, 2014.

[3] M. Levitt, "Growth of novel protein structural data," *Proc. Natl. Acad. Sci. USA ,* vol. 104, pp. 3138-3188, 2007.

[4] G. Lopez, A. Valencia and M. Tress, "FireDB-a database of functionally important residues from proteins of known structure.," *Nucleic Acids Res.,* vol. 35, no. Database Issue, pp. D219-D223, 2007.

[5] P. Maietta, G. Lopez, A. Carro, B. Pingilley, L. Leon, A. Valencia and M. Tress, "FireDB: a compendium of biological and pharmacologically relevant ligands," *Nucleic Acids Research ,* vol. 42, no. Database Issue, pp. D267-D272, 2014.

[6] J. Yang, A. Roy and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Research,* vol. 41, no. Database Issue, pp. D1096-D1103, 2013.

[7] J. Konc, T. Cesnik, J. Trykowska Konc, M. Penca and D. Janezic, "ProBiS-Database: Precalculated Binding Site Similarities and Local Pairwise Alignments of PDB Structures," *J. Chem. Inf. Model.,* vol. 52, no. 2, pp. 604-612, 2012.

[8] H. Hwang, T. Vreven, J. Janin and Z. Weng, "Protein-protein docking benchmark version 4.0," *Proteins,* vol. 78, pp. 3111-3114, 2010.

[9] R. Mosca, A. Céol and P. Aloy, "Interactome 3D: adding structural details to protein networks," *Nature Methods,* vol. 10, no. 1, pp. 47-56, 2013.

[10] R. Mosca, A. Céol, A. Stein, R. Olivella and P. Aloy, "3did: a catalog of domain-based interactions of known three-dimensional structure," *Nucleic Acids Research ,* vol. 42, no. Database Issue, pp. D374-D379, 2014.

[11] D. Talavera, A. Morreale, A. Hospital, C. Ferrer-Costa, J. Gelpí, X. de la Cruz, T. Meyer, R. Soliva, F. Luque and M. Orozco, "A fast method for the determination of fractional contributions to solvation in proteins," *Protein Science,* vol. 15, pp. 2525-2533, 2006.

[12] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, A. Hospital, J. Gelpí and M. Orozco, "A consensus view of protein dynamics," *Proc. Natl. Acad. Sci.,* vol. 104, pp. 796-801, 2007.

[13] T. Meyer, M. D'Abramo, A. Hospital, M. Rueda, C. Ferrer-Costa, A. Pérez, O. Carrillo, J. Camps, C. Fenollosa, D. Repchevsky, J. Gelpí and M. Orozco, "MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories," *Structure,* vol. 18, no. 11, pp. 1399-1409, 2010.

[14] P. Sfriso, A. Emperador, L. Orellana, A. Hospital, J. Gelpí and M. Orozco, "Finding conformational transition pathways from discrete molecular dynamics simulations," *Journal of Chemical Theory and Computation - JCTC,* vol. 8, no. 11, pp. 4707-4718, 2012.

[15] P. Sfriso, A. Hospital, A. Emperador and M. Orozco, "Exploration of conformational transitions pathways from Coarse-Grained simulations," *Bioinformatics,* vol. 29, no. 16, pp. 1980-1986, 2013.

[16] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. Chi Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. Dame, B. Han, Y. Zhou and D. Wishart, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research,* vol. 42, no. Database Issue, pp. D1091-D1097, 2014.

[17] K. Tai, S. Murdock, B. Wu, M. Ng, S. Johnston, H. Fangohr, S. Cox, P. Jeffreys, J. Essex and M. Sansom, "BioSimGrid: towards a worldwide repository for biomolecular simulations," *Org. Biomol. Chem.,* vol. 2, pp. 3219-3221, 2004.

[18] C. Silva, V. Ostropytskyy, N. Loureiro-Ferreira, D. Berrar, W. Dubitzky and R. Brito, "P-found: The protein folding and unfolding simulation repository," *Computational Intelligence and Bioinformatics and Computational Biology,* vol. 1, no. 8, pp. 28-29, 2006.

[19] A. M. Simms, R. D. Toofanny, C. Kehl, N. C. Benson and V. Daggett, "Dynameomics: design of a computational lab workflow and scientifc data repository for protein simulations," *Protein Eng. Des. Sel.,* vol. 21, no. 6, pp. 369-377, 2008.

[20] U. Hensen, T. Meyer, J. Haas, R. Rex, G. Vriend and H. Grubmüller, "Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function," *PLOS one,* vol. 7, p. e33931, 2012.

[21] J. C. Thibault, J. C. Facelli and T. E. Cheatham III, "iBIOMES: Managing and Sharing Biomolecular Simulation Data in a Distributed Environment," *J. Chem. Inf. Model.,* p. Ahead of print, 2013.

[22] A. Rajasekar, R. Moore, C. Hou, C. Lee, R. Marciano, A. de Torcy, M. Wan, W. Schroeder, S. Chen and L. Gilbert, "iRODS Primer: Integrated Rule-Oriented Data System.," in *Faceted Search (Synthesis Lectures on Information Concepts, Retrieval and Services)*, San Francisco, CA, Morgan and Claypool Publishers., 2010, pp. Vol2, 1-143.

[23] G. Manyam, M. Payton, J. Roth, L. Abruzzo and K. Coombes, "Relax with CouchDB - Into the non-relational DBMS era of Bioinformatics," *Genomics,* vol. 100, no. 1, pp. 1-7, 2012.

[24] T. Meyer, C. Ferrer-Costa, A. Perez, M. Rueda, A. Bidon-Chanal, F. J. Luque, C. A. Laughton and M. Orozco, "Essential dynamics: a tool for efficient trajectory compression and management," *J. Chem. Theory Comput.,* vol. 2, pp. 251-258, 2006.

[25] T. Meyer, X. de la Cruz and M. Orozco, "An atomistic view to the gas phase proteome," *Structure,* vol. 17, no. 1, pp. 88-95, 2009.

[26] M. Candotti, S. Esteban-Martin, X. Salvatella and M. Orozco, "Toward an atomistic description of the urea-denatured state of proteins," *Proc. Natl. Acad. Sci. USA,* vol. 110, no. 15, pp. 5933-5938, 2013.

[27] B. R. Brooks, C. L. 3. Brooks, A. D. J. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels and S. Boresch, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.,* vol. 30, pp. 1545-1614, 2009.

[28] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kale, R. D. Skeel and K. Schulten, "NAMD: a parallel, object oriented molecular dynamics program.," *Int. J. Supercomput. Appl. High Perf. Comput.,* vol. 10, pp. 251-268, 1996.

[29] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang and K. M. Merz, "AMBER 12," *San Francisco, California: University of California.,* 2012.

[30] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.,* vol. 4, pp. 435-447, 2008.

[31] S. Jo, T. Kim, V. G. Iyer and W. Im, "CHARMM-GUI: a web-based graphical user interface for CHARMM.," *J. Comput. Chem.,* vol. 29, pp. 1859-1865, 2008.

[32] B. T. Miller, R. P. Singh, J. B. Klauda, M. Hodoscek, B. R. Brooks and H. L. 3. Woodcock, "CHARMMing: a new, flexible web portal for CHARMM.," *J. Chem. Inf. Model.,* vol. 48, pp. 1920-1929, 2008.

[33] W. Humphrey, A. Dalke and K. Schulten, "VMD - Visual Molecular Dynamics," *J. of Molec. Graphics,* vol. 14, pp. 33-38, 1996.

[34] A. Cristiani, N. Brisotto, F. Chatwin Cedrati, M. Floris, L. Scapozza and S. Moro, "ClickMD: an intuitive web-oriented molecular dynamics platform," *Future Med. Chem. ,* vol. 3, no. 8, pp. 923-931, 2011.

[35] R. J. Woods, "http://glycam.org," [Online]. Available: http://glycam.org. [Accessed 05 2014].

[36] I. Bertini, D. A. Case, L. Ferella, A. Giachetti and A. Rosato, "A Grid-enabled web portal for NMR structure refinement with AMBER," *Bioinformatics,* vol. 27, no. 17, pp. 2384-2390, 2011.

[37] P. Kota, "GUIMACS-a Java based front end for GROMACS," *In Silico Biology,* vol. 7, pp. 95-99, 2007.

[38] D. Sellis, D. Vlachakis and M. Vlassi, "Gromita: a fully integrated graphical user interface to gromacs 4," *Bioinf. Biol. Insights,* vol. 3, pp. 99-102, 2009.

[39] S. Roopra, B. Knapp, U. Omasits and W. Schreiner, "jSim-Macs for GROMACS: a Java application for advanced molecular dynamics simulations with remote access capability," *J. Chem. Inf. Model.,* vol. 49, pp. 2412-2417, 2009.

[40] T. Makarewicz and R. Kazmierkiewicz, "Molecular Dynamics Simulation by GROMACS using GUI plugin for PyMOL," *J. Chem. Inf. Model,* vol. 53, pp. 1229-1234, 2013.

[41] R. Apostolov, L. Axner, H. Agren, E. Ayugade, M. Duta, J. L. Gelpi, J. Gimenez, R. Goñi, B. Hess and F. Jamitzky, "ScalaLife: scalable software services for Life Science," *Proceedings of the 9th HealthGrid Conference, Bristol, UK,* 2011.

[42] Xen Project, "Xen Project," [Online]. Available: http://www.xenproject.org/. [Accessed 05 2014].

[43] Sun Grid Engine, "Sun Grid Engine," Oracle, [Online]. Available: http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html. [Accessed 05 2014].

[44] B. Knapp and W. Schreiner, "Graphical user interfaces for Molecular dynamics - Quo vadis?," *Bioinformatics and Biology Insights,* vol. 3, pp. 103-107, 2009.

[45] J. Thibault, D. Roe, J. Facelli and T. Cheatham III, "Data model, dictionaries, and desiderata for biomolecular simulation data indexing and sharing," *J. Cheminform.,* vol. 6, no. 1, p. 4, 2014.

7. Conclusions

1. A methodology for the automatic extraction and classification of information from available structural data system has been developed. It allows a fast and easy obtaining of molecular datasets, from atomic to macromolecular resolution.

2. A library containing a picture of macromolecular structure flexibility (through Molecular Dynamics (MD) simulations) has been built. It contains dynamics for 1,595 proteins covering a 38% of the total PDB content (2010).

3. A dual approach strategy (a relational database containing simulation metadata and trajectory raw files in common disk raids) for the storage of data generated by the MD library has been developed, and has proved to be successful for the growth of content (simulations and analysis) as well as for the efficient retrieval of metadata information.

4. High-throughput study of solvent water molecules in *MoDEL* database has been performed. The analysis shows that water molecules have a much more dynamic behaviour than expected, with very few structural water molecules and mean residence times generally in the range of picoseconds, only overtaken by water molecules placed in concave protein pockets or internal cavities. Hydration water molecules have an active role in the dynamics of intra-protein hydrogen bonds.

5. A set of bioinformatics data types and tools for automatic setup, simulation and analysis of macromolecules using MD and Coarse-Grained (CG) methods have been designed and implemented. Web-Service technology used in the development of these tools allows the generation of personal workflows and their usage in HT projects.

6. Access to the before mentioned infrastructure is made available to non-expert users through user-friendly web interfaces (*MDWeb*, *NAFlex*, *FlexServ*). Addition of pre-configured workflows, integration of macromolecular flexibility analyses and visualization possibilities enhance the value of the final project. The set of web server applications designed and implemented in this thesis is publicly accessible for the scientific community, forming an integrated macromolecular flexibility portal, which can be reached directly or through INB-portals.

8.- Summary (Spanish)

Introducción

El descubrimiento de la doble hélice de ADN por James Watson y Francis Crick en el año 1953 [1] y la resolución de la estructura tridimensional de la mioglobina por Max Perutz y John Kendrew en el año 1958 [2] abrieron las puertas al estudio de las estructuras tridimensionales de las macromoléculas y mostraron el camino para explicar el funcionamiento de los sistemas biomoleculares a partir de principios físicos básicos.

Otro gran avance científico, la completa secuenciación del genoma humano en el año 2001 [3, 4], incrementó el interés en el estudio de la estructura de macromoléculas, y en particular de las proteínas; una vez desentrañada la composición del genoma, el siguiente y obvio (aunque no trivial) paso debía ser la obtención del proteoma y su definición estructural. Este interés desembocó en la aparición de varias iniciativas (conocidas como *Structural Genomics*) enfocadas a determinar estructuras de proteínas de manera masiva [5, 6]. El resultado ha sido un crecimiento exponencial en el número de estructuras disponibles en la base de datos *Protein Data Bank* (PDB) [7]. Menos se ha avanzado en el estudio dinámico de las macromoléculas, fundamentalmente, porque las técnicas experimentales muestran grandes problemas en representar la promiscuidad estructural inherente a la dinámica.

La tremenda dificultad del estudio experimental de las propiedades dinámicas de las macromoléculas ha hecho aparecer un conjunto de técnicas teóricas con las que obtener una simulación de su movimiento tanto en situación de equilibrio como cuando son sometidas a cierta tensión. Una de las técnicas teóricas más conocidas y usadas por la comunidad científica es la dinámica molecular (MD) [8, 9]. Los algoritmos de simulación de dinámicas moleculares trabajan con una representación atomística del sistema (cada átomo de la molécula se representa como una partícula del sistema), empleando representaciones clásicas de la energía de interacción inter-partícula (el force-field o campo de fuerzas), a partir de las cuales se derivan fuerzas, de las que por integración de las ecuaciones de Newton se derivan las trayectorias de los átomos de las macromoléculas a lo largo del tiempo. Aunque la MD es actualmente la técnica teórica más usada para obtener información dinámica de macromoléculas, el método tiene unas claras limitaciones, tal vez las más evidente su gran coste computacional que limita las escalas de tiempo accesibles (típicamente de nanosegundos a microsegundos), insuficientes para simular ciertos procesos de gran importancia biológica.

Durante los últimos años, nuevas técnicas teóricas, más simples computacionalmente, han ido ganando relevancia: las simulaciones *Coarse-Grained* (CG) o de baja resolución [10, 11]. Con las técnicas de CG podemos llegar a simular sistemas muy grandes en escalas de tiempo mayores gracias a trabajar con modelos de baja resolución de la molécula, típicamente comprimiendo varios átomos de un residuo en una única partícula. La importancia de los métodos CG ha llevado a la aparición de un gran repertorio de aproximaciones, que permiten abordar desde la simulación mesoscópica de largas cadenas nucleotídicas hasta la simulación casi atomística de pequeñas proteínas en una escala de milisegundos.

Los métodos teóricos de simulación de dinámicas moleculares van inexorablemente unidos a los continuos avances en el campo de la computación. La reciente revolución en el campo de los supercomputadores (conocida como *High*

Performance Computing - *HPC*, computación de alto rendimiento), con la aparición de potentes máquinas formadas por miles de procesadores trabajando en paralelo, ha desencadenado una gran evolución en los algoritmos de dinámica molecular. La posibilidad de dividir el sistema en trozos y enviar cada uno de ellos a un procesador distinto, ha permitido alcanzar escalas de tiempo impensables hace tan solo 5 años.

Dentro de este nuevo entorno de computación de alto rendimiento, cabe destacar un sistema diseñado en primera instancia para ser usado en el campo de los juegos por ordenador, las tarjetas gráficas con procesador integrado, las *Graphical Processing Units* (GPUs). Este particular tipo de procesador ha resultado ser ideal para cálculos de dinámica molecular, debido a su implementación interna similar a la de los procesadores vectoriales, cuyo diseño favorece el proceso de código independiente en paralelo. Aunque para adaptar los algoritmos existentes a estos nuevos procesadores ha sido necesario un rediseño casi total del código, el resultado ha sido espectacular, llegando a obtener un rendimiento hasta 10 veces mayor que el obtenido usando CPUs de última generación.

El nivel máximo de interrelación entre la simulación y la química computacional lo representa el diseño e implementación de un supercomputador completamente dedicado a calcular simulaciones moleculares: *Anton* (D.E. Shaw Research) [12]. El resultado es una máquina capaz de ofrecer un rendimiento 100 veces mayor que cualquier supercomputador, llegando a escalas de tiempo nunca antes exploradas con este tipo de métodos teóricos.

Finalmente, el nuevo paradigma de cálculo y almacenamiento en la nube (*Cloud Computing*) se está convirtiendo también en una posible alternativa a los supercomputadores. Con los cálculos en la nube, también conocidos como computación distribuida o computación en red, podemos usar recursos computacionales conectados a una red en paralelo como si de un supercomputador se tratara. De esta forma, se puede aprovechar el tiempo en que un procesador está en espera para este tipo de cálculos. Aproximaciones como *Folding at home* (http://folding.stanford.edu) o los más recientes *GPUGRID* [13] y *PS3GRID* [14] han usado este paradigma con gran éxito.

Los grandes y rápidos avances tanto en la computación como en los estudios teóricos de flexibilidad de macromoléculas han abierto la posibilidad de llevar a cabo estudios masivos de alto rendimiento (*High throughput*). Sin embargo, para lograr realizar este tipo de estudios, no solo se requieren algoritmos potentes y poder computacional, sino también una automatización de los distintos pasos necesarios en el proceso de cálculo de trayectorias así como de su posterior análisis. Casi tan importante como los cálculos, es necesario un sistema de almacenamiento que permita tanto guardar como consultar de manera eficiente una cantidad enorme de datos generados por el estudio masivo.

En esta tesis, se han estudiado, diseñado e implementado diferentes sistemas de automatización *high throughput* de cálculos de dinámica molecular, tanto atomística como de baja resolución, así como herramientas para su posterior análisis. Un ejemplo del potencial aportado por estos estudios masivos está reflejado en un caso concreto de análisis de la dinámica del solvente (moléculas de agua alrededor de una macromolécula) en simulaciones moleculares atomísticas, trabajando con una cantidad de información nunca antes disponible: más de 16 millones de moléculas de agua (Resultados, sección 3) y 20 Terabytes de datos.

Objetivos

Los objetivos de esta tesis se pueden dividir en cuatro apartados:

1. El primer objetivo es la extracción automática de información de interés a partir de las estructuras experimentales de macromoléculas, en particular el diseño e implementación de un conjunto de bases de datos relacionales para el almacenamiento de información estructural de interés y su posterior consulta de manera eficiente.

2. Obtener una visión dinámica de las estructuras macromoleculares, a partir de la generación de una librería de dinámicas moleculares, para ello nos propusimos:

    - Seleccionar un conjunto representativo de estructuras de proteínas, cubriendo la mayor parte posible del número total de estructuras disponibles en el PDB.

    - Obtener las propiedades dinámicas de las mismas a partir de cálculos de dinámica molecular.

    - Calcular un conjunto de análisis de flexibilidad sobre las trayectorias generadas.

    - Diseñar una base de datos relacional para el almacenamiento de los análisis de flexibilidad, así como de datos importantes relacionados con las distintas simulaciones.

    - Diseñar e implementar un servidor web para ofrecer toda la información obtenida al conjunto de la comunidad científica.

    - Finalmente, aprovechando la librería de simulaciones moleculares atomísticas generada, y como prueba de concepto, investigar la dinámica del solvente en disoluciones de proteínas.

3. Desarrollar un conjunto de tipos de datos y herramientas bioinformáticas para una fácil implementación de estos métodos en estudios de alto rendimiento. Esto incluye:

    - Diseñar un conjunto de tipos de datos para métodos de dinámica molecular, válidos para los paquetes más conocidos y usados a nivel mundial: AMBER, NAMD y GROMACS.

    - Desarrollar un conjunto de herramientas interconectadas para una total automatización de los procesos de preparación, simulación y análisis de dinámicas macromoleculares.

    - Recopilar, integrar y visualizar un amplio conjunto de análisis de flexibilidad para trayectorias generadas con dinámicas moleculares o de

baja resolución.

4. Desarrollar una herramienta para la generación de estructuras, simulación dinámica y análisis de trayectorias específica para ácidos nucleicos. La herramienta debía permitir:

   - Generar estructuras tridimensionales a partir de secuencias nucleotídicas.

   - Simular dinámicas de ácidos nucleicos, a partir de diferentes técnicas teóricas y diferentes resoluciones, desde métodos mesoscópicos de baja resolución hasta dinámicas moleculares atomísticas.

   - Recopilar, integrar y visualizar un conjunto de análisis de flexibilidad específico para ácidos nucleicos, con especial atención puesta en las propiedades físicas.

Resultados

1. Extracción automática de información de interés a partir de estructuras macromoleculares.

Los estudios de alto rendimiento (*High throughput*) necesitan un acceso a los datos lo más eficiente posible. En el campo de la estructura macromolecular, esto significa acceso a los ficheros en formato PDB que contienen la información de la estructura tridimensional (coordenadas 3D de todos los átomos que forman una molécula). El repositorio *Protein Data Bank* contiene el conjunto de estructuras resueltas experimentalmente, disponibles públicamente [7]. El primer paso en nuestro proyecto fue la construcción de una copia (*mirror*) del conjunto total de estructuras en discos locales. Esta copia se actualiza automáticamente cada semana, con la seguridad de tener siempre los últimos datos disponibles. Una vez disponible toda la información en discos locales, queríamos extraer información que considerábamos interesante e insertarla en una base de datos relacional. Esta base de datos nos permitiría, una vez creada, consultar información de manera mucho más eficiente.

Diseñamos un conjunto de tablas para almacenar información relacionada con cada una de las estructuras como por ejemplo: código PDB, secuencia aminoacídica, resolución o número de cadenas. En otro conjunto de tablas almacenamos información derivada de las distintas partes que contiene una estructura. Partes proteicas, no proteicas o moléculas de agua, con su respectivo tamaño, composición, etc. Una de las informaciones más interesantes almacenada es la relativa a los átomos situados en la interacción entre distintas partes de la molécula. Familias de proteínas compartiendo un determinado porcentaje de identidad de secuencia se guardó en diferentes tablas llamadas *clusters*. Finalmente, decidimos almacenar también información relativa a centros activos. Para esto, diseñamos un algoritmo donde se identificaba un centro activo como un conjunto de residuos situados dentro de una distancia umbral con respecto a un ligando, y a partir de estos residuos se obtenían todos los correspondientes centros activos de la familia de la proteína en cuestión, aplicando un mapeo estructural.

El conjunto de bases de datos implementado (PDB, partes, familias y centros activos) se integró en el proceso de actualización automática. Así, cada vez que actualizamos la copia del PDB, se lanzan los scripts necesarios para actualizar las correspondientes tablas de las bases de datos. Este diseño nos permite un acceso eficiente para consultas puntuales a información relacionada con estructuras macromoleculares. Además, es muy útil en procesos de selección de conjuntos de estructuras para ser estudiadas en un determinado proyecto. Consultas complejas interrelacionando conceptos pueden ser implementadas en minutos, obteniendo información tan variada como estos ejemplos reales:

- Conjunto de hetero-dímeros presente en todo el PDB, junto con los átomos situados en la interfaces entre los dos monómeros, para ser usado en validación de estudios de predicción de interfaces de interacción proteína-proteína.

- Obtención de un conjunto no redundante de estructuras de proteínas monoméricas cubriendo la mayor parte posible del PDB, con una resolución menor que un determinado valor límite, descartando proteínas de membrana, o con zonas de la estructura no determinadas, residuos no estándar, o proteínas unidas a ligandos de difícil parametrización para su uso en dinámica molecular.

- Obtención de un conjunto de pares de proteínas homólogas en el PDB donde una de ellas proviene de un organismo termófilo mientras que la otra proviene de un organismo mesófilo, para su uso en un estudio de comparación de propiedades dinámicas entre ellas.

- Obtención de un conjunto de proteínas cristalizadas junto a moléculas de agua, recuperando las coordenadas 3D para cada una de las aguas, para su uso posterior.

- Obtención de un conjunto de centros activos en todo el PDB con una molécula de agua (cristalográfica) o más alrededor, para su uso en estudios sobre el papel del solvente en el reconocimiento ligando-proteína.

2. Generación de una librería de dinámicas moleculares de proteínas.

A medida que el conocimiento estructural sobre macromoléculas iba creciendo, más y más indicios surgían apuntando que el estudio de las estructuras estáticas no era suficiente para entender su funcionamiento. Las macromoléculas no existen como estructuras fijas, bien al contrario, son máquinas biológicas en continuo movimiento, adaptando configuraciones diferentes. Hoy en día se cree que es precisamente esta flexibilidad la que determina la función de la mayoría de las proteínas y que es un parámetro cuidadosamente refinado por la evolución. Desafortunadamente, son pocas las técnicas experimentales que logran extraer información dinámica de las macromoléculas. Este hecho ha popularizado el uso de diferentes técnicas teóricas para el estudio de flexibilidad en estructuras de macromoléculas. Uno de los métodos computacionales más conocidos y usados en la actualidad es la dinámica molecular. En ella, los átomos son modelados como partículas en movimiento, logrando las siguientes posiciones en el tiempo aplicando las ecuaciones de Newton.

Aunque la dinámica molecular es una técnica actualmente usada por un gran número de grupos, su uso es todavía demasiado complejo para ser integrada en un proyecto de alto rendimiento, debido básicamente a tres obstáculos:

- Gran coste computacional.
- Dificultad en la preparación previa de la estructura a simular.
- Problemas para el posterior almacenamiento y análisis de los resultados.

Los grandes avances en el mundo de la computación, junto con la evolución de los algoritmos de dinámica molecular han facilitado la popularización de la simulación molecular, lo que ha llevado en paralelo a un problema importante en el manejo de la ingente cantidad de datos producida. No podemos olvidar que las trayectorias obtenidas con métodos de dinámica molecular requieren de un generoso espacio de disco, y la tendencia claramente apunta a necesitar más y más, con la posibilidad de simulaciones más largas realizadas sobre sistemas cada vez más grandes. Para un eficiente almacenamiento masivo de trayectorias y su posterior análisis, se requiere una buena planificación previa.

Reuniendo el conocimiento obtenido después de años trabajando con dinámica molecular, se decidió generar una librería de trayectorias de dinámica molecular, que permitiera analizar la flexibilidad de las proteínas a nivel del proteoma. El proyecto se inicia con una selección de un conjunto representativo de proteínas del PDB, debidamente filtradas para ser usadas en estudios de dinámica molecular, su simulación para obtener una trayectoria, y el posterior análisis y almacenamiento de los datos. Para ello, se diseñó un flujo de trabajo, usando nuestras bases de datos estructurales junto con unos *workflows* automatizados de preparación y simulación de MD, lanzando los cálculos en el supercomputador *MareNostrum (BSC)*. Se utilizaron entre 64 y 192 procesadores en paralelo por cada cálculo, de los 4812 procesadores *IBM PowerPC* con los que en ese momento contaba *MareNostrum*, con el código en paralelo del programa AMBER [15], y el campo de fuerzas considerado de última generación en ese momento: *parm99* [16], con una longitud final de 10ns. Finalmente, se analizaron las trayectorias resultantes en las máquinas del laboratorio y se almacenaron los datos obtenidos. La base de datos se extendió posteriormente con simulaciones

más extensas para algunas proteínas representativas, con cálculos realizados con otros campos de fuerza más recientes, y con trayectorias obtenidas en medios no acuosos.

El resultado final es una base de datos de simulaciones de dinámica molecular para 1,595 proteínas monoméricas (con más de 1,700 trayectorias generadas), representando un 38% del total de proteínas disponibles en el PDB, con al menos 10 nanosegundos por simulación. Para cada simulación, se calculó un conjunto de análisis de flexibilidad obteniendo así una visión dinámica del PDB, estos análisis y las trayectorias se almacenaron, juntamente con toda la información asociada a las simulaciones (programa, campo de fuerza, longitud, etc.). Para ello se usó una aproximación dual: las trayectorias se almacenaron en ficheros planos, mientras que para los análisis y la información de simulación se diseñó una base de datos específica. Esta base de datos es totalmente flexible, permitiendo añadir en cualquier momento nuevas simulaciones y análisis.

La mayoría de los datos obtenidos en el proyecto están disponibles públicamente a partir de una interfaz web interconectada con la base de datos. El servidor web de MoDEL permite diferentes búsquedas de proteínas, a través de código PDB o palabras clave o bien navegando por el conjunto total de simulaciones disponibles. Para cada simulación, todos los análisis de flexibilidad calculados se muestran en tablas y gráficos, en una interfaz totalmente fácil e intuitiva para el usuario. La página tiene integrado un visualizador interactivo de trayectorias (*Jmol: an open-source Java viewer for chemical structures in 3D*; http://www.jmol.org/), además de videos previamente generados del movimiento de la proteína en el tiempo. Las trayectorias comprimidas en formato PCZ [17] se encuentran también disponibles desde la aplicación web. La librería MoDEL tuvo en el último año (Julio 2013 – Mayo 2014) una media de 1,344 visitas al mes (con 31,573 accesos), correspondientes a unas 45 visitas al día (1052 accesos).

MoDEL es una librería de dinámicas moleculares viva y que evoluciona con el tiempo, ampliándose a medida que nuevas dinámicas se van generando o nuevos análisis aparecen.

Artículo 1:

MoDEL (Molecular Dynamics Extended Library): una base de datos de trayectorias generadas con dinámica molecular atomística.

Tim Meyer*, Marco D'Abramo*, **Adam Hospital***, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluís Gelpí, Modesto Orozco. (* *Autores con la misma aportación al trabajo)*

Structure. (2010) 18(11), 1399-1409.

Sinopsis: Se han generado más de 1700 trayectorias de proteínas seleccionadas como representantes del conjunto de estructuras solubles y monoméricas de todo el PDB a partir de métodos de dinámica molecular atomística de última generación, con simulaciones en condiciones próximas a las fisiológicas. Las

trayectorias y sus respectivos análisis se encuentran almacenados en un enorme repositorio donde pueden ser consultadas para obtener información sobre la dinámica de las respectivas proteínas, incluyendo interacciones. En este artículo, describimos el proyecto y la estructura y contenido de nuestra base de datos, junto con ejemplos sobre cómo podemos usarla para describir propiedades globales de flexibilidad en las proteínas. Análisis básicos y trayectorias sin solvente comprimidas con una resolución reducida están disponibles en nuestro servidor.

Link: http://mmb.irbbarcelona.org/MoDEL

3. Estudio de la dinámica del solvente y su efecto sobre las proteínas en simulaciones moleculares atomísticas.

En condiciones fisiológicas las proteínas se encuentran completamente recubiertas de solvente (mayoritariamente moléculas de agua). Esta capa acuosa es precisamente la responsable de la estructura globular de las proteínas y en muchas ocasiones ha demostrado ser crucial en la definición de su función. Desafortunadamente, a pesar de su importancia, la mayoría de estudios se limitan a describir la dinámica de las proteínas dejando a un lado el efecto causado por el solvente sobre estas moléculas. Actualmente, la falta de datos experimentales sobre la dinámica del agua es total. Los estudios teóricos, por otro lado, son también escasos, además de estar siempre centrados en proteínas específicas. En este trabajo, queremos aprovechar la gran cantidad de información generada en el proyecto de alto rendimiento MoDEL, diseñando e implementando un estudio teórico a gran escala sobre el conjunto final de moléculas de agua incluidas como solvente en las cerca de 1800 simulaciones disponibles en la librería, con el objetivo de obtener una visión global de su dinámica y de cómo esta afecta a las proteínas. El conjunto de datos analizado contiene más de 16 millones de moléculas de agua, siendo el trabajo teórico sobre propiedades del solvente más grande realizado hasta el momento.

La dinámica general de las moléculas de agua se estudió en una primera etapa a partir del análisis de las desviaciones en el espacio observadas respecto a su punto de partida (RMSd). Calculando estas desviaciones para una ventana de tiempo de 500ps, se obtuvieron tres capas de hidratación bien definidas:

- Una primera capa de hidratación, formada por las moléculas de agua con una desviación inferior a 1.2Å.
- Una segunda capa de hidratación, con una dinámica mayor que la primera capa, con desviaciones entre 9.7Å y 16.3Å.
- Finalmente una tercera capa, correspondiente a la mayoría de las moléculas de agua que forman la llamada caja de simulación, con una desviación mayor de 16.3Å.

Las moléculas de agua pertenecientes a la primera capa de hidratación son una pequeña minoría del conjunto analizado, y presentan una dinámica muy limitada comparada con el resto. Mayoritariamente se encuentran en cavidades internas de las proteínas o bien en regiones cóncavas de su superficie, con tiempos de residencia que pueden llegar hasta 5ns. Estas moléculas de agua, aún pudiéndose considerar estructurales, no están completamente estáticas. Después de un cierto espacio de tiempo, se intercambian con otras, que pasan a ocupar su lugar, incluso en el caso de cavidades que podrían considerarse cerradas desde el punto de vista de estructura estática.

Las moléculas de agua en la primera capa de hidratación prefieren formar puentes de hidrógeno como dadoras y no como aceptoras, al contrario de lo que se había descrito. Este hecho puede deberse a nuestra definición de primera capa, restringida a moléculas de agua altamente estructurales, con altos tiempos de

residencia y poca dinámica, mientras que la definición general de primera capa de hidratación sólo tiene en cuenta un criterio de distancia respecto a la superficie de la proteína; de hecho, en el análisis de la segunda capa de hidratación, sí que encontramos una preferencia hacia actuar como aceptoras de puente de hidrógeno. En esta segunda capa, encontramos moléculas de agua situadas en puntos de la superficie de las proteínas con alta densidad de agua, pero que no tienen una sola molécula atrapada, sino que hay un intercambio rápido entre diferentes moléculas (del orden de picosegundos).

No se observa una clara diferencia en la densidad de agua alrededor de diferentes estructuras secundarias, aunque sí se pueden diferenciar unos perfiles característicos para densidades alrededor de ciertos grupos atómicos, como pueden ser átomos cargados, hidrofílicos o hidrofóbicos. La principal diferencia se observa en los grupos hidrofílicos, con un pico acentuado de densidad a distancia 2.8Å de la proteína, indicando una fuerte interacción mediante puentes de hidrógeno. Un segundo pico de densidad aparece también en los grupos hidrofílicos, a distancias entre 4Å y 6Å de la proteína, acorde con lo esperado. Interferencias causadas por grupos polares cercanos a los grupos estudiados aparecen en ciertas categorías, generando un perfil muy característico, con un primer pico pronunciado seguido de un decaimiento progresivo de la densidad, que puede llegar hasta distancias cercanas a 3Å (grupos guanidino de Arg y amino de Asn y Gln), e incluso generando un nuevo pico a 5Å (grupos guanidino de Arg y carbonil de Asn y Gln). Los grupos hidrofóbicos, a su vez, muestran también un primer pico de densidad, pero a distancia cercana a un posible contacto de Van der Waals (3.6Å).

El estudio de la dinámica de las moléculas de agua individualmente resulta complejo pero informativo. Las moléculas pertenecientes a la primera capa de hidratación suelen tener una velocidad baja (en comparación con la velocidad media de las moléculas de agua libres) durante la mayor parte de la simulación, resultado de su fuerte interacción con la proteína. En cambio, las pertenecientes a la segunda capa exhiben una dinámica en general muy rápida, frenándose sólo al aproximarse a la superficie de la proteína. Esta dinámica se refleja en los coeficientes de difusión de las distintas capas de hidratación, donde primera y segunda capa ven altamente reducido su valor promedio (1.05 y 1.66 $x10^{-5}$ cm^2/s, respectivamente) en comparación con el resto de moléculas de agua incluidas en la simulación (5.21 $x10^{-5}$ cm^2/s), que ven sólo ligeramente modificada su movilidad respecto al agua libre (5.73 $x10^{-5}$ cm^2/s).

Una clara demostración de la compleja red dinámica generada por las moléculas de agua se obtiene al comprobar cómo todas y cada una de las moléculas pertenecientes a la caja de un sistema son capaces de visitar como mínimo una vez la superficie de la proteína en las simulaciones. Este análisis nos sirve también para demostrar que la longitud del conjunto de simulaciones de la librería MoDEL utilizadas en este estudio, si bien corta para estudiar la dinámica total de la proteína, puede considerarse suficiente para asegurar que la caja de moléculas de agua rodeando a la proteína está debidamente equilibrada.

Uno de los efectos de más interés producido por las moléculas de agua en contacto con la superficie de las proteínas es su capacidad para romper o favorecer

puentes de hidrógeno entre átomos expuestos. En algunos casos, el solvente compite con estos átomos de la proteína para formar interacciones, mientras que en otros casos una molécula de agua puede situarse justo en medio de dos átomos expuestos, interaccionando con los dos a la vez, actuando así como un puente (*water bridge*). El estudio de más de 800.000 puentes de hidrógeno intra-proteína identificados en nuestra librería denota una estabilidad muy grande en los puentes situados en estructuras secundarias, así como un importante porcentaje de puentes estables generados entre átomos de los grupos amida de la cadena principal (*backbone*) y grupos hidroxilo, carbonilo y carboxilo. Aún así, la dinámica de creación y destrucción de estos puentes sigue siendo muy elevada y el agua juega un papel clave en la misma. Independientemente de los átomos involucrados en las interacciones, el estudio indica que en un 20-50% de los casos de destrucción de puente de hidrógeno, existe una molécula de agua interaccionando con cada uno de los átomos de la proteína, mientras que en casi el total del resto de puentes perdidos, como mínimo uno de los átomos está interaccionando con una molécula de agua. Únicamente en un 5% de los casos estudiados, el puente de hidrógeno desaparece sin presencia de moléculas de agua en la cercanía. La creación de nuevos puentes mediante *water bridges* se reduce a un 10% de los casos, con el grupo carboxilo de los ácidos Asp y Glu mostrando una mayor preferencia a formarlos.

El cálculo de energías (ΔG) a partir de datos termodinámicos extraídos directamente de las trayectorias (mediante porcentajes de población de estados) permite clasificar los puentes de hidrógeno según su estabilidad. El resultado del análisis deja entrever que la clásica definición de los puentes de hidrógeno a partir de dinámicas moleculares, típicamente a partir de criterios geométricos, difiere de su definición mediante energías de interacción: un 60% de los puentes de hidrógeno analizados muestran una energía negativa pero no muy favorable ($\Delta G > -2$ Kcal/mol), mientras que sólo el 40% restante serían puentes de hidrógeno entre átomos de la proteína realmente estables a juzgar por sus energías. Este resultado sugiere un posible cambio de metodología en lo que respecta al cálculo de puentes de hidrógeno a partir de datos de simulación molecular.

En congruencia con lo ya conocido, los puentes formados por átomos no expuestos (según su superficie accesible al solvente), así como los situados en el *backbone* de la proteína son los que muestran una mejor estabilidad. En general, la estabilidad de los puentes de hidrógeno que se ven destruidos por causa de moléculas de agua es baja, con energías no muy favorables ($\Delta G > -2$ Kcal/mol), mostrando una cierta tendencia a estar situados en regiones desestructuradas de la proteína.

El proceso de destrucción de puentes de hidrógeno aparece como un simple paso, sin involucrar un complejo camino de aparición/desaparición de interacciones. La llegada de una molécula de agua favorece la división del puente de hidrógeno original, simultáneamente. Las moléculas de agua mediando interacciones (*water bridge*) aparecen en caminos más complejos, menos probables, pero todavía posibles, con un porcentaje mayor en grupos carboxilo (Glu, Asp) y amonio (Lys), y en general en puentes con menor energía, menos estables.

Los resultados del estudio sugieren un comportamiento de las moléculas de agua mucho más dinámico de lo esperado, con escasas moléculas consideradas estructurales, completamente estáticas. El intercambio entre moléculas de agua en la capa de hidratación de la proteína es constante. El seguimiento de más de 18 millones de moléculas de agua muestra un rápido descenso de la velocidad en las moléculas que pasan a interaccionar con la superficie de la proteína, dónde generalmente no pasan más de unos picosegundos unidas. Uno de los resultados de esta gran movilidad lo encontramos en la comparación con las aguas cristalográficas (moléculas de agua identificadas con técnicas de cristalografía de rayos X). Sólo un 18% de éstas se encuentran representadas por moléculas de agua con alto tiempo de residencia en las dinámicas moleculares, mientras que el resto coincide con puntos de alta densidad. Por otro lado, el papel determinante otorgado a las moléculas de agua en los procesos de destrucción de puentes de hidrógeno en la superficie de la proteína se refleja claramente en nuestros resultados, con la gran mayoría (>90%) de los átomos involucrados en puentes de hidrógeno interaccionando con moléculas de agua durante el proceso de ruptura del puente. La simple definición de puentes de hidrógeno a partir de criterios geométricos se pone en duda en la interpretación de los resultados de este estudio, señalando como desfavorables hasta un 60% de los puentes de hidrógeno analizados, lo que coincide con la idea generalizada de que los puentes de hidrógeno no son los que están guiando el plegamiento de la proteína.

Artículo 2:

Water-omics: Análisis de alta disponibilidad sobre interacciones proteína-solvente a partir de simulaciones de dinámica molecular.

Adam Hospital, Modesto Orozco, Josep Lluís Gelpí.

Enviado (2014)

Sinopsis: El agua que rodea a las proteínas juega un papel crucial sobre su estructura, dinámica y función. Desafortunadamente, el análisis experimental de la hidratación en proteínas es, a día de hoy, difícil de realizar y todavía más difícil de interpretar, lo que ha popularizado el uso de herramientas de simulación. En este campo, la dinámica molecular se ha erigido como la técnica escogida para el estudio de las propiedades del agua en las proximidades de la superficie de las proteínas. La gran potencia de cálculo disponible en la actualidad nos ha permitido realizar un estudio a nivel de proteoma de la flexibilidad de las moléculas de agua situadas en el entorno de las proteínas. En este estudio, analizamos las moléculas de agua que forman el solvente de las aproximadamente 1800 simulaciones de dinámica molecular de proteínas con solvente explícito disponibles en la base de datos MoDEL, extrayendo información relativa a capas de hidratación, tiempos de residencia, átomos/residuos preferentes, funciones de distribución radial (RDFs), difusión, puentes de hidrógeno, y aguas en cavidades internas y externas. Los resultados sugieren un comportamiento mucho más dinámico de lo esperado, observando muy pocas moléculas de agua estructurales, y muchas móviles que afectan la pauta de interacciones de puente de hidrógeno intra-proteína. El estudio ofrece una metodología sistemática para el estudio de interacciones entre proteína y solvente, que puede ser fácilmente exportable a otros conjuntos de datos.

4. Desarrollo de un conjunto de tipos de datos y herramientas bioinformáticas para la implementación de dinámicas macromoleculares en estudios de alto rendimiento.

La automatización de cálculos de dinámica molecular es un requisito al uso de la dinámica molecular en régimen masivo. A su vez, esta automatización simplifica el uso de la técnica a los usuarios no expertos. Para ello, decidimos diseñar un conjunto de tipos de datos bioinformáticos que nos ayudaran a estandarizar los múltiples formatos de ficheros usados en dinámica molecular. Así, generamos un conjunto de objetos integrados en una ontología dentro del proyecto *BioMOBY* [18], con definiciones de estructura, topología, trayectoria y *restart* (información para el reinicio de una simulación).

Una vez hecho esto, generamos un conjunto de servicios web semánticos para cada uno de los pasos necesarios para la preparación de una dinámica molecular, para la simulación, y para un conjunto de análisis básicos de la trayectoria resultante. Implementamos servicios para diferentes paquetes de dinámica molecular (*AMBER* [15], *GROMACS* [19] *y NAMD* [20]) y para diferentes campos de fuerza, además de métodos de simulación de baja resolución (*Coarse-Grained*) que permitían realizar estudios de flexibilidad de manera más rápida. Estos servicios web, gracias a la ontología previamente construida, permiten la construcción de complejos *workflows* (proceso dividido en pequeños pasos) con su simple concatenación. El resultado es un conjunto de procesos pre-configurados para preparación de estructuras desde su obtención en la base de datos PDB hasta el sistema final generado con la molécula rodeada de una caja de aguas preparado para la simulación, con diferentes programas y campos de fuerza. El conjunto final de tipos de datos, servicios web y *workflows* se nombró *MDMoby* en referencia a la librería usada.

MDMoby nos permite lanzar preparaciones, simulaciones e incluso análisis de manera masiva, pero por otro lado sigue siendo un complejo sistema para un usuario inexperto. Para estos usuarios, diseñamos el servidor *MDWeb*. *MDWeb* permite hacer uso de toda la potencia de *MDMoby*, desde una interfaz gráfica diseñada para el usuario no experto. Finalmente, integramos un conjunto de análisis de flexibilidad de proteínas en una sola herramienta, para generar a partir de una trayectoria un exhaustivo estudio de propiedades dinámicas: *FlexServ*. Para incrementar el poder de *FlexServ* en casos en los que no se dispusiera de una dinámica atomística, integramos tres métodos de baja resolución (dinámica Browniana, dinámica discreta y análisis de modos normales).

Tanto *MDWeb* como *FlexServ* son servidores web con una interfaz gráfica, ofreciendo los resultados en formato de diagramas 2D y 3D, y con el visualizador interactivo de estructuras Jmol integrado, para una fácil comprensión de los análisis de flexibilidad. El número de usuarios registrados en el servidor *MDWeb* es ya superior a 1,000, y sus estadísticas del último año (Julio 2013 – Mayo 2014) son de 1,508 visitas al mes (con 53,124 accesos), correspondientes a unas 50 visitas al día (1,770 accesos). *FlexServ* por su parte es nuestro servidor más popular, con una media de 1,917 visitas al mes (con 436,170 accesos), correspondientes a unas 64 visitas al día (14,539 accesos).

Artículo 3:

MDWeb & MDMoby, una plataforma web para simulaciones de dinámica molecular.

Adam Hospital, Pau Andrio, Carles Fenollosa, Damjan Cicin-Sain, Modesto Orozco and Josep Lluís Gelpí.

Bioinformatics (2012), 28(9), 1278-1279.

Sinopsis: MDWeb y MDMoby constituyen una plataforma web para facilitar el acceso a métodos de dinámica molecular tanto en sistemas estándar como en sistemas de alto rendimiento. La plataforma ofrece herramientas para preparar sistemas a partir de una estructura PDB de manera similar al procedimiento seguido por expertos en el campo. También ofrece ficheros de descarga para enviar la simulación de manera local, y es capaz de lanzar simulaciones para tres de los paquetes de dinámica molecular más populares (AMBER, NAMD y GROMACS). El programa incorpora a su vez herramientas para análisis de trayectorias, tanto proporcionadas por el usuario como directamente obtenidas desde nuestra base de datos MoDEL (http://mmb.irbbarcelona.org/MoDEL). La plataforma ofrece dos accesos diferentes, un conjunto de servicios web basados en la librería BioMoby (MDMoby) de acceso programático, y un portal web (MDWeb).

Link: http://mmb.irbbarcelona.org/MDWeb

Artículo 4:

FlexServ: una herramienta de integración de análisis de flexibilidad en proteínas.

Jordi Camps, Oliver Carrillo, Agustí Emperador, Laura Orellana, **Adam Hospital**, Manuel Rueda, Damjan Cicin Sain, Marco D'Abramo, Josep Lluís Gelpí, Modesto Orozco.

Bioinformatics (2009), 25, 1709-1710.

Sinopsis: FlexServ es una herramienta basada en web para el análisis de flexibilidad en proteínas. El servidor incorpora potentes protocolos de baja resolución para la obtención de dinámicas de proteínas a partir de diferentes aproximaciones: Análisis de modos normales (NMA), dinámica Browniana (BD) y dinámica discreta (DMD). Las trayectorias pueden ser de origen local o proporcionadas por el usuario. El servidor ofrece un completo análisis de flexibilidad a partir de una gran variedad de métricas, incluyendo análisis geométrico básico, factores de temperatura (*B-factors*), dinámica esencial, análisis de rigidez (*stiffness*), medidas de colectividad, índices de Lindemann, correlación entre residuos, correlaciones encadenadas, determinación de dominios dinámicos, detección de puntos de bisagra (*hinge points*), etc. Los resultados aparecen en la página web en forma de texto plano y gráficos 2D y 3D.

Link: http://mmb.irbbarcelona.org/FlexServ

5. Desarrollo de una herramienta para la generación de estructuras, simulación dinámica y análisis de trayectorias de ácidos nucleicos.

Los ácidos nucleicos son posiblemente las macromoléculas biológicas más flexibles, pero su simulación es compleja, más que la de las proteínas, y normalmente requiere un grado de experiencia superior por parte del usuario. Decidimos por ello generar una herramienta específica para el estudio de flexibilidad en ácidos nucleicos, que empleaba la estructura general de *MDWeb*, pero la adaptaba a las particularidades de estas moléculas. La herramienta consta de tres partes importantes:

- Generación de estructuras de ácidos nucleicos a partir de secuencias nucleotídicas.
- Simulación dinámica de estructuras de ácidos nucleicos, con técnicas a diferentes resoluciones, desde métodos mesoscópicos a baja resolución hasta métodos atomísticos (dinámica molecular).
- Recopilación, integración y visualización de análisis de flexibilidad específicos para ácidos nucleicos, con una especial atención a las propiedades físicas.

El resultado es un servidor web llamado *NAFlex*. *NAFlex* permite trabajar a partir de: i) secuencia, generando el tipo de estructura de ácido nucleico deseado a partir de librerías de tipos de estructuras nucleicas; ii) estructura, ya sea a partir de un simple código PDB o a partir de una estructura proporcionada por el usuario; y iii) a partir de una trayectoria previamente generada.

El servidor integra tres tipos diferentes de métodos de simulación dinámica, principalmente diferenciados por la resolución con la que trabajan. El algoritmo de *Colorless Wormlike-Chain (WLC)* [21] utiliza modelos donde una partícula representa N pares de bases (típicamente N=10), y es apropiado para simular cadenas nucleotídicas muy largas. El algoritmo mesoscópico a nivel de pares de bases [22] utiliza propiedades elásticas para generar un conjunto de conformaciones, adecuado para largas secuencias. Y finalmente, el método más preciso implementado en *NAFlex*, pero también el más costoso computacionalmente, es la dinámica molecular atomística.

Una de las partes más importantes de *NAFlex* es la recopilación de análisis de flexibilidad aplicados a las trayectorias obtenidas. Desde análisis cartesianos básicos (desviaciones cuadráticas medias, radios de giro, factores de temperatura, etc.) hasta complejos estudios sobre parámetros helicoidales o estimaciones de observables obtenidos por resonancia magnética nuclear. El conjunto de herramientas implementado se puede encontrar en el servidor con una interfaz gráfica heredada del proyecto *MDWeb*. Los resultados se presentan de forma gráfica siempre que sea posible, con gráficos 2D y 3D y visualizaciones interactivas con *Jmol*. Completos tutoriales y manuales complementan el servidor, con el que un usuario inexperto puede obtener fácilmente, siguiendo los simples pasos indicados, un conjunto de propiedades de flexibilidad de su ácido nucleico de interés. El servidor *NAFlex* suma ya una media de 23 visitas diarias en su primer año de vida.

Artículo 5:

NAFlex: un servidor web para el estudio de flexibilidad en ácidos nucleicos.

Adam Hospital, Ignacio Faustino, Rosana Collepardo-Guevara, Carlos González, Josep Lluís Gelpí, Modesto Orozco.

Nucleic Acids Research (2013), 41(W1), W47-W55,
NAR Featured Article June 2013.

Sinopsis: Presentamos NAFlex, una nueva herramienta web para el estudio de flexibilidad en ácidos nucleicos, tanto aislados como unidos a otras moléculas. El servidor permite al usuario incorporar estructuras desde el PDB, rellenando espacios vacíos en la estructura y arreglando posibles inconsistencias estructurales. El servidor también ofrece la posibilidad de generar estructuras canónicas de ácidos nucleicos (promedio o adaptadas a la secuencia), a partir de un conjunto de librerías internas, así como crear conformaciones específicas de ácidos nucleicos a partir de su secuencia. El servidor integra una variedad de métodos para explorar la flexibilidad en ácidos nucleicos, como por ejemplo un modelo *Colorless wormlike-chain*, un modelo mesoscópico con resolución sobre pares de bases e incluso simulaciones de dinámica molecular atomística, con una amplia variedad de protocolos y campos de fuerza. Las trayectorias generadas con las simulaciones, o bien proporcionadas desde el exterior, pueden ser visualizadas y analizadas con un gran número de herramientas, incluyendo análisis Cartesianos estándar, dinámica esencial, análisis helicoidal, rigidez local y global, descomposición de energías, componentes principales y espectros de resonancia magnética nuclear. El servidor es accesible en http://mmb.irbbarcelona.org/NAFlex.

Conclusiones

Los proyectos de genómica estructural han producido una explosión en el número de estructuras resueltas experimentalmente. Actualmente, el repositorio PDB recoge una cifra superior a las 100.000 estructuras (Mayo 2014). El gran volumen de información experimental no tiene, desafortunadamente, un paralelismo en nuestro conocimiento sobre la dinámica de las macromoléculas, lo que limita nuestra capacidad de racionalizar el funcionamiento de las mismas. La falta de datos experimentales sobre propiedades dinámicas de las macromoléculas ha hecho aflorar una serie de métodos teóricos para su cálculo, destacando entre ellos los cálculos de dinámica molecular atomística (MD). La popularización del uso de las técnicas de MD, ligado a las mejoras en los programas y a la constante evolución de la computación de alto rendimiento ha generado nuevos retos informáticos, que han sido abordados a lo largo de la presente tesis. Hemos construido herramientas para realizar simulaciones masivas, creando herramientas que simulan las decisiones y el procedimiento de trabajo de un experto, que verifican la calidad de una simulación, la analizan, ponen la información en contexto y la guardan de una manera que facilita su uso posterior. Hemos también implementado en nuestros portales web herramientas de resolución múltiple, que acercan metodologías complejas a usuarios no expertos. Todas estas herramientas se han implementado en entornos gráficos, que directamente, o vía el portal del Instituto Nacional de Bioinformática (INB), permiten su uso por una amplia comunidad científica.

Futuro

El ambicioso proyecto de estudio computacional de flexibilidad en estructuras macromoleculares no acaba aquí. El siguiente paso será la integración de todas las plataformas web diseñadas para ofrecer públicamente nuestras herramientas y datos obtenidos en esta tesis (junto con más herramientas ofrecidas por nuestro grupo de investigación) en un portal llamado *FlexPortal*. Este portal, ya en construcción, va a contener un espacio privado para cada usuario, donde se almacenarán todas las estructuras, cálculos y análisis del mismo. De esta forma, los usuarios tendrán acceso a todas las herramientas generadas para el estudio de flexibilidad de estructuras en un solo espacio. Además, y como en todos nuestros proyectos, el diseño será totalmente flexible para adaptar nuevos servidores web que puedan desarrollarse en un futuro.

La continua evolución en el campo de la dinámica molecular, tanto atomística como de baja resolución, unido a la nueva revolución de la computación de alto rendimiento, favorece que los métodos mejoren y se actualicen a un ritmo nunca antes visto. Por eso, las herramientas implementadas en esta tesis deben ser revisadas y actualizadas regularmente. El claro ejemplo lo encontramos en *MDWeb*, donde los usuarios (1,068 usuarios registrados actualmente) piden la inclusión de determinados campos de fuerza publicados durante los últimos años, así como ejemplos de cálculos en GPUs, consultas que han impulsado una segunda versión del proyecto ya en preparación. También prevemos la incorporación de nuevos modelos de baja resolución, o incluso la de métodos híbridos de frontera abierta.

La suma de la rápida evolución de las técnicas de simulación y el incremento en la potencia de cálculo, en las capacidades de almacenaje y en el ancho de banda de las comunicaciones informáticas apuntan a un futuro próximo donde las simulaciones de dinámicas de macromoléculas van a dejar de ser esporádicas y se van a convertir en básicas para el estudio de su función. Con el trabajo presentado en esta tesis esperamos haber contribuido dando el primer paso hacia el uso extensivo de estos métodos.

Referencias

[1] J. D. Watson and F. H. Crick, "The structure of DNA," *Cold Spring Harb Symp Quant Biol,* vol. 18, pp. 123-131, 1953.

[2] J. Kendrew, G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff and D. Phillips, "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis," *Nature,* vol. 181, pp. 662-666, 1958.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. S. Granger and H. O. Smith, "The sequence of the human genome," *Science,* vol. 291, no. 5507, pp. 1304-1351, 2001.

[4] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, 2001.

[5] T. C. Terwilliger, D. Stuart and S. Yokoyama, "Lessons from Structural Genomics," *Annu. Rev. Biophys.,* no. 38, p. 371–383, 2009.

[6] G. T. Montelione, "The Protein Structure Initiative: achievements and visions for the future," *F1000 Biology Reports,* vol. 4, p. 7, 2012.

[7] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne and H. M. Berman, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, pp. 235-242, 2000.

[8] J. A. McCammon, B. R. Gelin y M. Karplus, «Dynamics of folded proteins,» *Nature,* vol. 267, pp. 585-590, 1977.

[9] C. L. Brooks, M. Karplus and B. M. Pettitt, "Proteins: A theoretical perspective of dynamics, structure and thermodynamics.," Cambridge: Cambridge University Press, 1987.

[10] D. A. Potoyan, A. Savelyev and G. A. Papoian, "Recent successes in coarse-grained modeling of DNA," *WIREs Comput. Mol. Sci.,* vol. 3, pp. 69-83, 2013.

[11] V. Tozzini, "Coarse-grained models for proteins," *Current Opinion in Structural Biology,* vol. 15, pp. 144-150, 2005.

[12] R. O. Dror, R. M. Dirks, J. P. Grossman, X. Huafeng and D. E. Shaw, "Biomolecular simulation: a computational microscope for molecular biology," *Annu. Rev. Biophys.,* vol. 41, pp. 429-452, 2012.

[13] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson and G. De Fabritis, "High-throughput All-Atom molecular dynamics simulations using distributed computing," *J. Chem. Inf. Model,* vol. 50, pp. 397-403, 2010.

[14] M. Harvey, G. Giupponi, J. Villà-Freixa and G. De Fabritiis, "PS3GRID.NET: Building a distributed supercomputer using the PlayStation 3.," *Distributed and Grid Computing-Science Made Transparent for Everyone. Principles, Applications and Supporting Communities.,* 2007.

[15] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang and K. M. Merz, "AMBER 12," *San Francisco, California: University of California.,* 2012.

[16] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell and P. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.," *J. Am. Chem. Soc. ,* vol. 117, no. 19, pp. 5179-5197, 1995.

[17] T. Meyer, C. Ferrer-Costa, A. Pérez, M. Rueda, A. Bidon-Chanal, F. J. Luque, C. A. Laughton and M. Orozco, "Essential Dynamics: A tool for efficient trajectory compression and management," *J. Chem. Theory Comp.,* vol. 2, pp. 251-258, 2006.

[18] The BioMoby Consortium, "Interoperability with Moby 1.0 - It's better than sharing your toothbrush!," *Briefings in Bioinformatics,* vol. 9, no. 3, pp. 220-231, 2008.

[19] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation.," *J. Chem. Theory Comput.,* vol. 4, pp. 435-447, 2008.

[20] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kale, R. D. Skeel and K. Schulten, "NAMD: a parallel, object oriented molecular dynamics program.," *Int. J. Supercomput. Appl. High Perf. Comput.,* vol. 10, pp. 251-268, 1996.

[21] S. A. Allison, "Brownian dynamics simulation of wormlike chains. Fluorescence depolarization and depolarized light scattering.," *Macromolecules,* vol. 19, p. 118, 1986.

[22] J. Goñi, C. Fenollosa, A. Pérez, D. Torrents and M. Orozco, "DNAlive: a tool for the physical analysis of DNA at the genomic scale," *Bioinformatics,* vol. 24, no. 15, pp. 1731-1732, 2008.

9. List of publications

(** Part of this thesis, * work using tools or methodology developed in this thesis)

1. (*) David Talavera, Antonio Morreale, Tim Meyer, **Adam Hospital**, Carles Ferrer-Costa, Josep Lluís Gelpí, Xavier de la Cruz, Robert Soliva, F. Javier Luque, Modesto Orozco.
   A fast method for the determination of fractional contributions to solvation in proteins.
   Protein Science (2006), 15, 2525-2533.

2. David Talavera, **Adam Hospital**, Modesto Orozco, Xavier de la Cruz.
   A procedure for identifying homologous alternative splicing events.
   BMC Bioinformatics (2007), 8, 1-11.

3. (*) Manuel Rueda, Carles Ferrer-Costa, Tim Meyer, Alberto Pérez, Jordi Camps, **Adam Hospital**, Josep Lluís Gelpí, Modesto Orozco.
   A consensus view of protein dynamics.
   Proc. Natl. Acad. Sci. (2007) 104, 796-801.

4. (**) Jordi Camps, Oliver Carrillo, Agustí Emperador, Laura Orellana, **Adam Hospital**, Manuel Rueda, Damjan Cicin Sain, Marco D'Abramo, Josep Lluís Gelpí, Modesto Orozco.
   FlexServ: An integrated tool for the analysis of protein flexibility
   Bioinformatics (2009), 25, 1709-1710.

5. (**) Tim Meyer*, Marco D'Abramo*, **Adam Hospital***, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Lluís Gelpí, Modesto Orozco.
   MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories.
   Structure. (2010) 18(11), 1399-1409.

6. (*) Modesto Orozco Laura Orellana, **Adam Hospital**, Athi N. Naganathan, Agustí Emperador, Oliver Carrillo, Josep Lluís Gelpí.
   Coarse Grained Representation of Protein Flexibility. Foundations, successes and shortcomings.
   Computational Chemistry Methods in Structural Biology. Advances in Protein Chemistry and Structural Biology, Ed. C.Christov. Elsevier. (2011) 85, 183-215.

7. **(**) Adam Hospital**, Pau Andrio, Carles Fenollosa, Damjan Cicin-Sain, Modesto Orozco and Josep Lluís Gelpí.
   MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations.
   Bioinformatics. (2012) 28(9), 1278-1279.

8. (*) Pedro Sfriso, Agustí Emperador, Laura Orellana, **Adam Hospital**, Josep Lluís Gelpí, Modesto Orozco.
   Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations.
   Journal of Chemical Theory and Computation – JCTC (2012) 8(11), 4707-4718.

9. **(**) Adam Hospital** and Josep Lluís Gelpí.
   High-throughput molecular dynamics simulations. Towards a dynamic view of macromolecular structure.
   Wiley Interdisciplinary Reviews: Computational Molecular Science (2013), 3 (4), 364–377.

10. **(**) Adam Hospital**, Ignacio Faustino, Rosana Collepardo-Guevara, Carlos González, Josep Lluís Gelpí, Modesto Orozco.
    NAFlex: A web server for the study of nucleic acids flexibility.
    Nucleic Acids Research (2013), 41(W1), W47-W55, NAR Featured Article June 2013.

11. (*) Pedro Sfriso, **Adam Hospital**, Agustí Emperador and Modesto Orozco.
    Exploration of conformational transitions pathways from Coarse-Grained simulations.
    Bioinformatics (2013), 29 (16), 1980-1986.

12. (*) Laura Orellana, **Adam Hospital** and Modesto Orozco.
    Oncogenic mutations of the EGF-Receptor ectodomain clustered at interdomain regions reveal a mechanism for ligand-independent activation.
    Submitted (2014).

13. (*) Montse Barbany, Tim Meyer, **Adam Hospital**, Ignacio Faustino, Marco D'Abramo, Jordi Morata, Modesto Orozco and Xavier de la Cruz.
    Naturally existing cavity couplings in proteins and the origins of allostery.
    Submitted (2014).

14. **(**) Adam Hospital**, Modesto Orozco and Josep Lluís Gelpí.
    Water-omics: High throughput analysis of protein-solvent interactions from MD simulations.
    In preparation (2014).

ANNEX I

High-throughput molecular dynamics simulations. Towards a dynamic view of macromolecular structure.

Adam Hospital and Josep Lluís Gelpí.

Wiley Interdisciplinary Reviews: Computational Molecular Science (2013), 3 (4), 364–377.

Overview

High-throughput molecular dynamics simulations: toward a dynamic view of macromolecular structure

Adam Hospital[1] and Josep Ll Gelpi[2,3]*

Molecular dynamics (MD) simulation is nowadays the eligible theoretical technique to account for macromolecular flexibility. Molecular simulations, from *ab initio* to coarse-grained representation levels, allow having a direct visualization of macromolecular dynamic behavior and its influence in molecular recognition. In the postgenomic era, where most bioinformatics studies should be performed genome-wide, molecular simulations should not be an exception. However, moving MD simulations to the high-throughput regime is not a trivial issue. High-performance computing systems (highly parallel supercomputers or graphical processing units-based systems) allow performing large and complex simulations at a significantly reduced time. Besides, data storage strategies have also been largely improved. However, system preparation and trajectory analysis are still performed almost manually and become highly limited by the need of human intervention. A number of projects, like Dynameomics, BioSimGrid, or MoDEL have addressed the issue of high-throughput MD. Such initiatives have released tools that have meant a significant progress in the field. The extension of structural studies, mainly molecular simulations, to the high-throughput regime, will allow matching genomic-wide studies now performed.

How to cite this article:
WIREs Comput Mol Sci 2013. doi: 10.1002/wcms.1142

INTRODUCTION

Seeking understanding of the structure to function relationship in biological macromolecules has been and still constitutes the main object of study in structural biology. In the postgenomic era, the amount of biological information available is in constant growth.[1–4] Even though structural information increases at a much slower rate than sequence data, there is now a significant probability that the three-dimensional (3D) structure of a given macromolecular system is already available. Besides experimental structures, comparative modeling or similar strategies[5] contribute to extent the number of structures available for simulation. Protein Data Bank (PDB)[1] is the main source of data in structural biology. However, information contained there is restricted to a static view of 3D structures. It is a well-known fact that the dynamic properties of macromolecules play a key role in molecular recognition. The understanding of enzyme mechanisms, small molecules recognition, or protein–DNA interactions requires acquiring a proper view of the flexibility of the involved macromolecules. Irrespective to the experimental technique used, flexible or moving regions do not accumulate enough signal intensity to give a

*Correspondence to: gelpi@ub.edu

[1]Institute of Research in Biomedicine, Joint IRB-BSC Computational Biology Programme, Structural Bioinformatics, National Institute of Bioinformatics, Barcelona, Spain

[2]Barcelona Supercomputing Center, Joint Program IRB-BSC for Computational Biology, Computational Bioinformatics, National Institute of Bioinformatics, Barcelona, Spain

[3]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain

DOI: 10.1002/wcms.1142

clear picture of the atomic positions. Hence, flexible regions lead to low information regions in the obtained structures. For instance in the interpretation of X-ray diffraction data, B-factors account partially for this lack of resolution but, most often, large flexible regions are simply missing from the PDB record. Only theoretical techniques can really give an accurate picture of flexible regions in macromolecules. Molecular dynamics (MD) is nowadays the most accepted technique to reveal the dynamic properties of biological systems. The advances achieved in simulation algorithms and force fields, together with the availability of large supercomputing facilities, allow envisioning the availability of a 'dynamic PDB' where not only structures but also structures in motion would be stored. Fields like drug discovery or protein–protein docking, for instance, would take a large benefit of having such comprehensive view of conformational variation. On the other hand, the production of sequence data, since the availability of next-generation sequencing techniques, makes possible to obtain high quality genomic data of individuals in a short period of time. Information about individual genetic variation is being already used in drug discovery and in more ambitious projects centered in 'personalized medicine'. The structural point of view is still missing in this kind of studies. Again, high-throughput MD appears as the election technique to better understand the effects of genetic variations on the biological function of macromolecules.

Although MD is a mature technique, the requirements to move it to the high-throughput regime are not trivial. Simulation itself can now be extended to the microsecond range, closer to the biologically relevant time scales. Also, fairly large, biologically meaningful complexes can be simulated. At the same time, techniques like replica exchange MD, Markov state models, and others can take benefit of large computer systems by launching a series of parallel simulations instead of a single long one. However, the preparation of a macromolecular system for simulation is still a fairly manual process, often limited by the need of human intervention. In turn, the analysis of large trajectories requires new methods to store, and to extract data in an efficient way. Analysis methods themselves are constantly evolving, and methods to handle trajectory data must take this into account.

The present paper will give an overview of the technology available to perform MD simulations in the high-throughput regime, and the possibilities and limitations of present approaches.

MACROMOLECULAR FLEXIBILITY, A KEY ISSUE

Flexibility is a key feature to understand biological function. Biological macromolecules are large and flexible entities that perform their function through the recognition of other molecular entities. Such recognition process, intuitively linked to the classical Fischer's 'lock and key' hypothesis, is only possible after the interacting partners have undergone a process of mutual adjustment. Changes in structure before binding not only make more difficult the prediction of the structure of complexes, but also have a significant role in the energetic profile of the process. Binding of transcription factors to DNA, for example, is not only dependent on DNA sequence recognition but a direct consequence of the ability of the DNA molecule to adapt to the protein surface.[6,7] Proteins have the intrinsic ability to undergo functionally relevant conformational changes under native state conditions.[8–10] Allosteric regulation of enzymes is based on their ability to exist in two or more alternative conformations.[10,11] A great number of enzyme mechanisms rely in steps of structural rearrangement.[8] Loop or domain closures contribute to isolate the active site from solvent and, in so doing, alter the chemical environment around substrates, triggering the catalytic event.[11–13] Additionally, some features of protein function can only be understood when dynamic properties are taken into account. For instance, diffusion of small substrates through heme-dependent enzyme molecules requires the transient appearance of channels in the protein structure.[14–19] It becomes clear that the correct interpretation of macromolecular behavior require considering structure ensembles rather than single 3D structures taken from the PDB.

From a practical point of view, structural bioinformatics relies on experimental structures as the main source of data. Any prediction made on structures, such as the effect of point mutations, either made for engineering purposes, or consequence of a disease-related DNA point mutation; the binding modes of protein complexes; or drug–receptor docking poses, uses static structures. The size of the present PDB makes possible to find several copies of the same molecule, or very close related molecules. Taking these 'experimental ensembles', a partial view of the macromolecule flexibility can be obtained, although PDB composition is necessarily biased. Indeed, the study of PDB as source for molecular flexibility has been exploited in some extent.[20,21] Perhaps the most clearly negative effect of the lack of flexibility

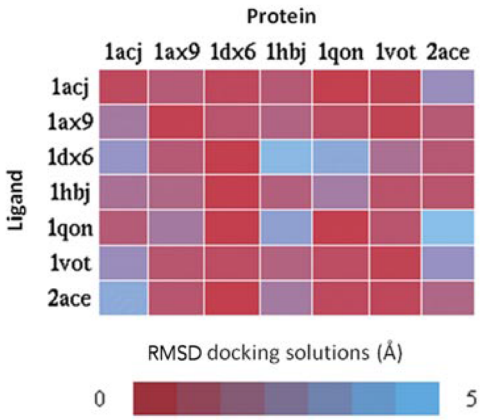


FIGURE 1 | Molecular flexibility influence on docking. Cross-docking experiment using acetyl cholinesterase structures obtained from the Protein Data Bank. Ligands were docked back in the available structures. RMSD from the best docking solution is presented.

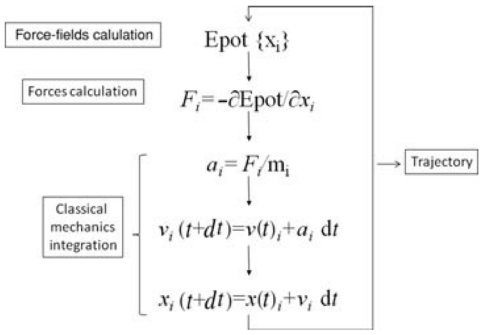


FIGURE 2 | Simplified schema for the molecular dynamics algorithm.

information appears in docking predictions. Figure 1 shows a typical cross-docking study made with some acetyl cholinesterase structures found at PDB. Ligands extracted from PDB structures systematically fail to dock in 3D structures made from the same protein. This illustrates that natural conformational variations are far larger that what can be accounted with standard docking procedures. Selection of receptor structures in large docking studies or in virtual screening is a key point to define the success of the prediction.

Theoretical techniques like MD simulations appear as the most convenient way to obtain a picture of macromolecular dynamic properties.

MD SIMULATIONS, GOING HIGH THROUGHPUT

MD simulation, first developed in the late 70s,[22] has advanced from simulating several hundreds of atoms to systems with biological relevance including entire proteins in solution with explicit solvent representations, membrane-embedded proteins, or large macromolecular complexes like nucleosomes[23] or ribosomes.[24,25] Simulation of systems having about 50,000–100,000 atoms are now a routine, and simulations around 500,000 atoms are common when the appropriate computer facilities are available. This remarkable improvement is in a large part a consequence of the use of high-performance computing (HPC), and the simplicity of the basic MD algorithm (Figure 2). An initial model of the system is obtained from either experimental structures or comparative modeling data. The simulated system could be represented at different levels of detail. Atomistic representation is the one that allows the best reproduction of the actual systems. However, coarse-grained representations are becoming very popular when large systems or long simulations are required (see Ref 26 for a review of such strategies). Solvent representation is a key issue in system definition. Several approaches have been assayed[27,28] but, again, the most used is the simplest one, the explicit representation of solvent molecules, although at the expense of increasing the size of the simulated systems. Once the system is built, forces acting on every atom are obtained by deriving equations, the force fields, where potential energy is deduced from molecular structure.[29–34] Force fields are complex equations but they are easy to calculate. The simplicity of the force-field representation of molecular features: springs for bond length and angles, periodic functions for bond rotations, and Lennard-Jones potentials and the Coulomb's law for Vdw and electrostatic interactions, respectively, assures that energy and force calculations are extremely fast even for large systems. Several force fields are currently used in atomistic molecular simulations, which differ in the way they are parameterized. Parameters are not necessary interchangeable, and not all force fields allow to represent all molecule types, but simulations conducted using either modern force field are normally equivalent.[35,36] Once the forces acting on individual atoms are obtained, classical Newton's law of motion is used to calculate accelerations and velocities and to update the atom positions. As integration of movement is done numerically, a time step shorter of the fastest movements in the molecule should be used.

This ranks normally between 1 and 2 femtoseconds for atomistic simulations, and is the major bottleneck of the simulation procedure. Microsecond long simulations, barely scratching the time scales of biological processes, require iterating over this calculation cycle 10^9 times. Algorithmic advances, which include fine tuning of energy calculations, parallelization, or the use of graphical processing units (GPUs), have largely improved the performance of MD simulations. Microsecond long simulations are now possible, and nanosecond long simulations routine.

HPC and MD

The present generation of computers takes benefit of parallelism and accelerators to speed up the process. The most popular simulation codes (AMBER,[37] CHARMM,[38] GROMACS,[39] or NAMD[40]) have been long ago compatible with the Messaging Passing Interface (MPI). MPI is a protocol for computer-to-computer communication that allows sharing work between processors. When a large number of computer cores can be used simultaneously, MPI can greatly reduce the computation time. To benefit the locality of interactions, the general strategy is to distribute the system to simulate among processors. The strategy is called spatial decomposition. Only a small fragment of the system has to be simulated in each processor. The most efficient division is not based in the list of particles, but in their position in space. Each processor deals with a region of space irrespective of which particles are present there. Communication between processors is also reduced, as only those simulating neighboring regions have to share information. A number of improvements in the selection of such 'neighboring' regions further reduce communications (see Ref 41 for a review). Efforts to improve the parallelism efficiently extend to most steps of the calculation (for instance, implicit solvent models,[42] numerical integrators,[43] and nonbonded interactions).[44,45] The use of accelerators, mainly GPUs, has become a major breakthrough in simulation codes. GPUs are hardware components, which were originally designed to accelerate computer graphics. GPUs have evolved into general-purpose, fully programmable, high-performance processors, and represent a major technical improvement to perform atomistic MD. They can deliver over an order of magnitude more floating-point operations per second than classical central processing units (CPUs). Last generation of large supercomputers incorporates GPUs (Titan, USA and Tianhe-1A, China; first, and eight in the Top 500 computer ranking http://www.top500.org). To take full advantage of GPU power, a major redesign of the computer codes is necessary. However, most major MD codes are prepared already for GPUs, and even MD codes written specifically to be used on GPUs have been developed (ACEMD).[46] Simulation on GPUs alone or combined with MPI is, at present, the default strategy for high-throughput MD simulations. Although it is not of general use, hardware specifically tailored to MD calculations has also been developed. An examples of such is MDGRAPE-3 by Yonezawa and coworkers,[47] which uses special hardware pipelines to implement fast particle mesh Ewald calculations. More importantly, Anton, a whole computer system, has been designed to run MD simulations. Anton core is constituted by a significant number of application-specific integrated circuits designed to improve force-field evaluation, together with a high-speed 3D torus network interconnection. Anton is by far, the fastest alternative available, reaching the millisecond time scale.[48] Unfortunately, Anton is only available for public research at the National Resource for Biomedical Supercomputing at the Pittsburgh Supercomputing Center.

MD codes have largely improved, and nanosecond up to microsecond are the routine time scales for present MD simulations. However, computers have grown much further. Power consumption limitations have lead hardware vendors to increase computer power by increasing the number of processors, instead of increasing clock frequency. It is no longer difficult to find supercomputers with hundreds of thousands of cores or even more (first computer on the Top500 list holds 560,640 cores, second 1,572,864 cores, and even 10th in the list still holds 63,360 cores; www.top500.org; November, 2012 classification). Present parallelization strategies are unable to use efficiently supercomputers with more than about 1000–5000 cores in a single simulation. The obvious change of approach consists in launching several simulations simultaneously, instead of a single long one. Simulations can be run in parallel. Calculations are independent, or have only low frequency communication steps among them. Such strategy shows a much larger scalability (that should be ideal for truly independent simulations), and uses efficiently large supercomputers. The limit of such approach is not using supercomputers at all, but distributed computing strategies.[49–51] The concept can be exploited in two different levels. First, a series of parallel simulations is an efficient way to built comprehensive ensembles of structures. Second, available computer power could be used to obtain a comprehensive view of macromolecular flexibility, by generating a database of such ensembles.

Ensembles and Replicated Simulations: Using Large Supercomputers

The naïve use of simulations is to generate movies depicting conformational events. However, the goal of most real simulations is the production of ensembles of structures. Ensembles can be analyzed to derive thermodynamic properties of the system, like entropy or free energy. If properly built, ensembles can also be used to reconstruct complex conformational transitions or even folding events. Because of the stochastic nature of MD, simulation trajectories obtained with the same system should not be identical. Although differences are usually small, depending on the system and the initial structure used, trajectories can explore completely independent regions of the conformational landscape. Nonexperts tend to consider such MD simulations as nonreproducible results, but in fact they are just partial samplings of the same ensemble of states. The quality of the representation of such ensemble increases as the number of simulations increases. The traditional approach, deriving the ensemble from a single trajectory, is the least efficient way. Single MD trajectories are biased by their initial conformations. The only way to improve sampling is extending the length of the simulation. Tricks like metadynamics,[52–54] for instance, where already visited conformations are penalized, help to improve the sampling. However, even with a perfect sampling, MD simulations cannot surmount barriers in the energy landscape higher than the total energy added to the system. The obtained ensemble with a single simulation is limited to those states that are accessible at the simulation temperature. Combining several trajectories starting from a representative set of structures would produce a more complete sampling. But such set is not normally available from experiment. Alternatively, replica exchange methods can been used to improve sampling.[55,56] Such methods launch parallel simulations in different conditions. The easiest variation is simulation temperature. The sampling ability of the simulation increases with temperature. Higher temperature simulations can surmount energy barriers and explore new regions of the ensemble. Periodically, energies of the different simulations are compared and structures are swapped according to its energy rank. The resulting simulation has sampled a larger conformational space, due to high temperature simulations, and retains the ability to represent the low temperature states of the system. The idea has been extended to other simulation strategies. Most remarkably, replicas based in differences in the Hamiltonian (Hamiltonian Replica Exchange), including alchemical free energy calculations[57] or constant-pH simulations,[58,59] have become popular.

Simulation Databases, toward a Dynamic PDB

Large databases are common in bioinformatics, but not in the simulation world. Two initial projects were reported, BioSimGrid[60] and P-Found.[61] Both initiatives provided software infrastructure to build biosimulation databases. The philosophy behind was that having a central repository of trajectories would allow to obtain a comprehensive view of biomolecular structure. At the time that these projects were started, computer power was still limited, and hence, both projects were conceived as open databases where trajectories should be provided by their users. Unfortunately, this approach was less effective than expected. This philosophy has been recovered lately in the European project Scalalife (Scalable Software Services for Life Sciences, http://www.scalalife.eu)[62] in the conception of its Competence Center. Two additional projects reported a major release in 2010: Dynameomics[63] and MoDEL.[64] In these projects, besides of the software infrastructure, a significant amount of trajectories for proteins were reported. The general aim of these projects does not differ from the previous ones, but in these cases a significant effort in terms of computer power was invested in populating the database. In fact, none of them is actually open for the community to upload trajectories. Table 1 shows some comparative data between Dynameomics and MoDEL. Although both databases contain an impressive number of trajectories, they only represent a fraction of the kingdom of known protein structures. Dynameomics is more oriented to folding–unfolding studies, and selection of targets has been done from a controlled library of folds.[65] Dynameomics uses simulation strategies linked to protein stability studies (several simulation temperatures, influence of point mutations, and general absence of nonessential ligands). On the contrary, MoDEL aims to maximize PDB coverage, and uses Cluster 90 representatives. MoDEL started from the simplest structures (monomeric, without ligands), and progressively incorporated larger structures with a rich collection of parameterized ligand molecules.[64]

No equivalent initiatives in the nucleic acid simulation have been reported, although the ABC-Consortium created a simulation database oriented to characterize DNA flexibility.[69] The database provides 50–100 nanoseconds trajectories of 39 different DNA

TABLE 1 | Comparison of Main Characteristics of Major Protein Trajectory Databases

	Dynameomics	MoDEL
Number of simulations	7263 (100 available)	1875 (all available in compressed format)[66]
Number of proteins	996 + 230 variants	1596
Selection criteria	Fold classification	PDB cluster
PDB coverage	N.R.	38% including homologous structures
Type of simulation	Several temperatures, explicit solvent	NPT, explicit solvent
Typical simulation length	31 nanoseconds	10 nanoseconds
Software	*In lucem* MM[67]	AMBER,[37] NAMD,[42] GROMACS[39]
Storage strategy	Relational DB + MOLAP[68]	Relational DB + native trajectory files

Data obtained from the 2010 releases of Dynameomics[63] and MoDEL.[64]

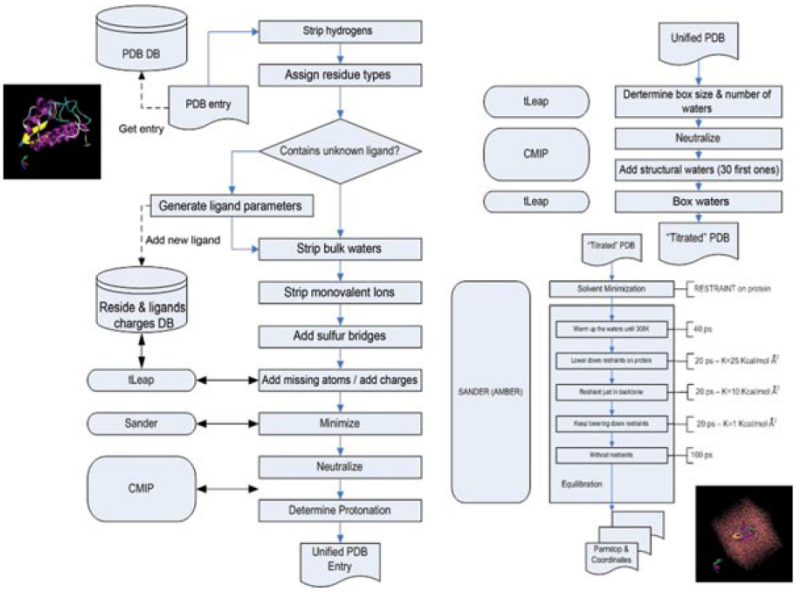


FIGURE 3 | Molecular dynamics setup used in MoDEL project based on Ambertools and in-house software.

oligomers, which cover all possible nearest neighbor sequence environments for the 10 possible DNA base steps.

Automated Tools for System Preparation

The availability of databases, like Dynameomics or MoDEL, allows envisioning a time when dynamics data for most known 3D structures, and even for nonstructured proteins, will be available. In that scenario, obtention of MD trajectories would be just an additional step in the preparation of the structure. However, such situation is still far from being feasible.

Preparation for large-scale trajectory databases implies to set up a huge variety of protein systems. Expected issues include nonstructured or missing regions or residues, nonstandard ligands, or even structures bearing errors in the interpretation of experimental data. When a single system is simulated, all the effort in the preparation of the system is worth, as it assures the quality of the simulation result. Such setup is usually done manually, with a considerable human effort. Figure 3 shows the preparation workflow used for MoDEL simulations.[64] It implies a number of well-known procedures: fixing structure errors; ionization of titrable amino acids; addition of structural water molecules, counter ions, and solvent; and energy

minimization and equilibration of the system at the desired temperature. An expert modeler normally carries out these procedures using a set of helper programs. Such expert has the necessary knowledge to surmount specific problems that may arise. However, although the quality of the results is still desirable, when preparing thousands of protein structures for simulation, there is no place for human intervention. Workflow shown in Figure 3 was programmed to run automatically, but a nonnegligible fraction of the proteins prepared failed at some point of the process.

On a completely different point of view, system setup is a major challenge for nonexpert users. For newcomers to MD simulation, even a single simulation could represent an unaffordable problem. Even worse, nonexpert users tend to blindly use default procedures leading easily to artifactual trajectories, which are hard to distinguish from the correct ones. This strongly contributes to the lack of popularity of biomolecular simulations among the bioinformatics or the biochemical community.

Both problems require the automatization of simulation setup. We would be looking for a clever black box for the nonexperts, but also a robust software suite that can account for a large set of unrelated protein structures. All major MD codes come with a set of accompanying programs, which perform most steps of the preparation. A number of initiatives, combining those tools with a user-friendly interface, have come into the scene to cover this problem. CHARMM-Gui[70] or CHARMMing[71] provides a web-based user interface to generate the necessary input files to be simulated with CHARMM MD package. Guimacs,[72] Gromita,[73] or jSimMacs[74] provide a similar functionality for the GROMACS package. VMD,[75] a popular structure and trajectory visualizer, provides a number of plug-ins that allow to launch simulations with NAMD. Most of these tools provide a friendly environment to prepare systems for simulation without the need of a deep knowledge of the underlying operations. This facilitates the access to the field for the newcomers. Some of them, like CHARMMing, allow also to launch test simulations, or even to control production simulations, through the interactions with most common HPC batch systems. Unfortunately, the lack of a standard for the representation of molecular simulation data has as a consequence that most helper applications are restricted to a single MD package, and data are not easily interchangeable.

These tools do succeed in making MD available to nonexperts. They pack standard procedures and allow a visual inspection of the quality of the setup. However, they do not necessarily provide solutions for the difficult structures. Besides, although most use some kind of embedded scripting language, automatization of procedures is not a straightforward task. Lessons learnt in the preparation of the MoDEL database, in our group, led to the generation of a new set of tools, MDMoby and MDWeb,[76] that try to cover both aspects of the problem. On one hand, MDMoby provides a full set of web-services, based in the BioMoby framework (http://www.biomoby.org),[77] covering all setup, simulation, and analysis operations. The modular nature of such web-services collection allows incorporating them as a tool-kit in a very flexible way to the design of complex setup protocols and to run them programmatically. In turn, MDWeb, a web-based interface run on top of MDMoby, provides a user-friendly bench where user can check for the quality of the input structure (Figure 4), tailor their own setup protocols or use a collection of predefined ones (Figure 5). MDWeb is not linked to a specific simulation package (as MoDEL database contained trajectories from several origins), being able to prepare input files for AMBER, NAMD, and GROMACS.

Trajectory Analysis: Quality Control and Flexibility

Often the focus of high-throughput simulation is directed to setup and HPC, however, the real benefit comes from the analysis of the data. We can generate hundreds of microseconds of trajectory data in a fraction of the time, but getting information out of such data requires efficient methods for storing and transmitting trajectory data, and high-throughput analysis procedures. Although simulation itself and even setup (see above) has been adapted to the high-throughput regime, analysis tools are far from that achievement. A large number of analysis tools exist. GROMACS,[39] Ambertools,[37] or VMD[75] provides a full set of analysis set for the corresponding trajectory formats. For independent examples, CAMDAS[78] is an automated conformation analyzer, FlexServ[79] provides a complete flexibility analysis, and ALADYN[80] performs structure alignment based on dynamics properties, and also the already mentioned MDWeb[76] provides an extended set of analysis procedures.

Quality of the simulations is a key issue and assessing whether a trajectory is a faithful representation of the behavior of the system is a primary concern of the analysis. The usual approach is to compare snapshots of the trajectory with experimental data. No systematic recipe for such assessment is available. The MoDEL project[64] reported its own quality control approach. All trajectories were analyzed for

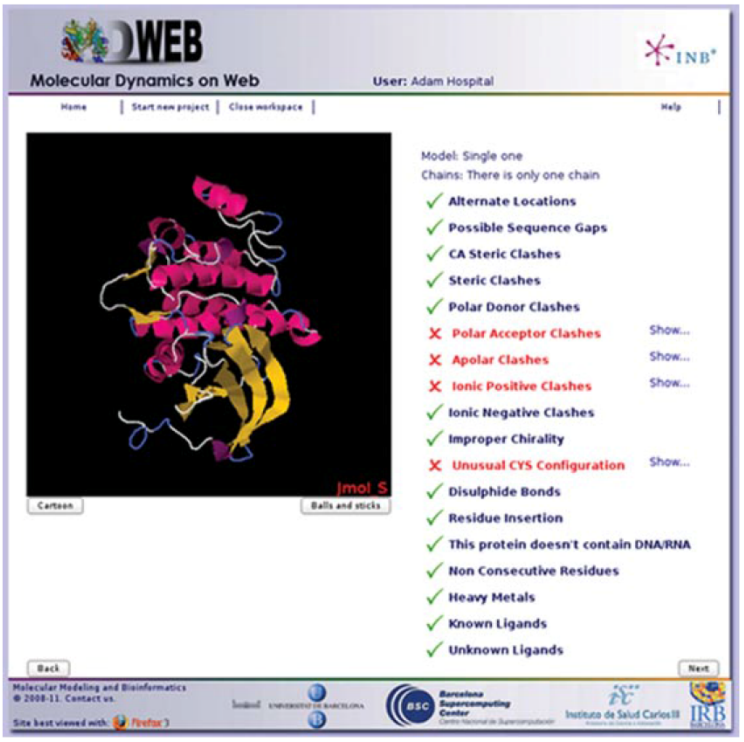


FIGURE 4 | Screenshot from MDWeb. Structure checking module. http://mmb.irbbarcelona.org/MDWeb. Reproduced with permission of Molecular Modelling & Bioinformatics Group.[76]

global descriptors, such as the absolute and relative RMSD, the TM-score$_{rmsd}$,[81] the radii of gyration, and solvent accessible surface. They were also analyzed for local descriptors, the number of native contacts, and the secondary structure. Trajectories were analyzed after the first nanosecond to check for technical problems in the setup (these usually lead to anomalous diffusion or velocities in protein, ligand or solvent), which were rare and were easy to correct in most cases. Experimental data could be also used to assess the quality of ensembles or even to guide their building giving the appropriate weight to the found conformations (see Ref 82 for a review).

Although simulation trajectories allow visualizing biomolecular flexibility, one step further is required to really target quantitatively dynamical properties. The simplest way is to calculate B-factor values, which are formally equivalent to those obtained from X-ray diffraction analysis, but in this case derived from the actual atomic fluctuations. However, a much more powerful way to gathering dynamic properties out of trajectories or ensembles is the use of essential dynamics[83] (ED). Following ED, the orthogonal movements describing the variance of a system are obtained by diagonalization of the covariance matrix derived from the MD simulation. The result of the analysis is the generation of a set of eigenvectors (the modes or the principal components), which describe the nature of the deformation movements of the protein and a set of eigenvalues, which indicate the stiffness associated with every mode. Sorting the eigenvectors by their associated eigenvalues, the largest part of the variance will be explained by the first few eigenvectors. Because the eigenvectors represent a full-basis set, the original Cartesian, trajectory can be always projected into the eigenvectors space, without loss of information. Furthermore, if a restricted set of eigenvectors is used, information is lost, but the level of error introduced in the simplification is always on user control by considering the recovered variance. Inspection of the atomic components of the most important eigenvalues helps to determine

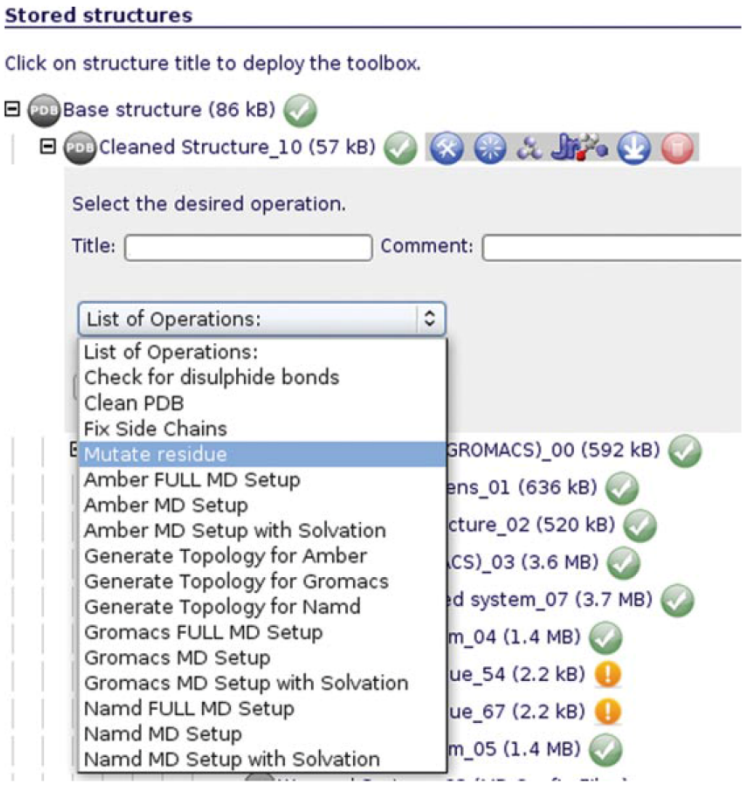


FIGURE 5 | Screenshot of MDWeb. Workspace with selector of setup operations open. http:///mmb.irbbarcelona.org/MDWeb. Reproduced with permission of Molecular Modelling & Bioinformatics Group.[76]

the contribution of different residues to the key essential deformations of the protein. Visualization of the ED modes allows to gain insight into the nature of the most prevalent protein movements. A comprehensive set of tools oriented to ED can be found in PCASuite[66] (http://mmb.irbbarcelona.org/software/pcasuite/) and its web frontend FlexServ.[79]

Data Storage and Transmission

Storage and access to the data is a major issue in high-throughput MD. Problem is not different from what the genomic world is facing with the appearance of new generation sequencing data. Major projects, Dynameomics and MoDEL, use (see Table 1) different strategies to store and eventually distribute data. Both projects have adopted a dual approach. Simulation metadata is stored in relational databases, whereas raw trajectory data are stored in a disk-based structure. The main difference lays in that Dynameomics uses a Multidimensional Online Analytical Processing (MOLAP)[68] to store trajectory data, whereas MoDEL prefers to store data in the original format, trusting in advanced file systems approaches to speed up data access. Both strategies have pros and cons. Advanced approaches like MOLAP allow a large flexibility in the access to data, using less disk space than the equivalent relational approach. However, although tools to access this kind of storage exist, traditional simulation analysis software should be adapted or even rewritten from scratch to get maximum benefit of the strategy. Using original formats for storing trajectory data allow using any already existing analysis or visualization software without modification, at the expense of the poor indexing capabilities of those tools. For routine analyses that imply the use of whole trajectories this is a less important problem. For advanced analysis where, for example, only specific parts of the system are needed or several trajectories should be compared, the efficiency is poor.

Transmission of simulation data is also an unsolved issue. MD packages have assayed several approaches like NetCDF, HDF5, and others to replace the inefficient text-based formats, but the level of compression is small. Trajectories offered by the MoDEL web site for download come in a compressed format based on ED.[66] This method reduces the size of the trajectory by eliminating high frequency modes that are not useful to describe conformational movements of the molecule. The amount of compression is significant, but it implies losing information and only can be applied to unsolvated structures.

CONCLUSION

In an era when classical biochemical studies has to be extended to whole genome studies and a huge amount of data are available on the sequence side, the structural bioinformatics style needs to be updated. Fortunately, during the last years, the development of new and more efficient simulations engines, and the availability of large supercomputer platforms has inspired a number of projects aimed to add a fourth dimension (time) to structural databases. A major change in philosophy is required to move MD to the high-throughput, genomic like, regime. Large-scale parallelization, albeit extremely useful to extend the length of simulations, will be complemented by grid-like approaches, even though they could be implemented in the same supercomputers. A huge number of simulations will be necessary to achieve a comprehensive view of the dynamics properties of macromolecules. Simulations should be extended in length or efficiency of sampling, as much as the improvement of algorithm makes possible, and this is a nonstop task. Projects like Dynameomics and MoDEL represent an encouraging start. They may serve as seed for major initiatives, probably involving large consortia, where large-scale simulation databases will be constructed. The original idea by BioSimGrid involving a community contribution has not to be discarded. Tools for automatic setup and analysis are already available, but a lot of work is needed to standardize data formats, allowing nonrelated researchers to share trajectory data. Improvements are needed in the field of data storage and transmission, in a way that the concept of a distributed MD database can be technically feasible. The systematic study of macromolecular dynamic properties through the use of theoretical methods opens a new perspective in the understanding of structure to function relationships in biology and will facilitate structural studies to cope with the plethora of genomic information available.

URLS RELEVANT TO HIGH-THROUGHPUT MD SIMULATIONS

MD Codes and Helper Applications

ACEMD & ACEMDtk[46]

multiscalelab.org/acemd

AMBER & AMBERTOOLS[37]

ambermd.org

CHARMM[38]

www.charmm.org

CHARMM-GUI[70]

www.charmm-gui.org

CHARMMing: Web portal for CHARMM[71]

www.charmming.org

GROMACS[39]

www.gromacs.org

Gromita. Integrated GUI for Gromacs 4[73]

bio.demokritos.gr/gromita

MDWeb & MDMoby. An integrated web platform for molecular dynamics simulations[76]

mmb.irbbarcelona.org/MDWeb

NAMD[42]

www.ks.uiuc.edu/Research/namd

VMD[75]

www.ks.uiuc.edu/Research/vmd

Trajectory Databases

BioSimGrid, Repository for biomolecular simulations[60]

www.BioSimGrid.org

Dynameomics: A comprehensive database of protein dynamics[63]

www.dynameomics.org

MoDEL. Molecular Dynamics Extended Library[64]

mmb.irbbarcelona.org/MoDEL

P-Found The protein folding and unfolding simulation repository.[61]

www.p-found.org

Analysis Tools

FlexServ. Integrated tool for the analysis of protein flexibility[79]

mmb.irbbarcelona.org/FlexServ

PCAsuite. Compression based on Essential dynamics[66]	mmb.irbbarcelona.org/software/pcasuite/pcasuite.html

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000, 28:235–242.
2. UnitProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* 2010, 38:D142–D148.
3. Magrane M, Consortium U. UniProt knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, 2011:bar009.
4. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2011, 40:D48–D53.
5. Tramontano A. Comparative modelling techniques: where are we? *Comp Funct Genomics* 2003, 4:402–405.
6. Bouvier B, Zakrzewska K, Lavery R. Protein–DNA recognition triggered by a DNA conformational switch. *Angew Chem Int Ed Engl* 2011, 50:6516–6518.
7. Zakrzewska K, Lavery R. Towards a molecular view of transcriptional control. *Curr Opin Struct Biol* 2012, 22:160–167.
8. Kumar S, Ma B, Tsai CJ, Wolfson H, Nussinov R. Folding funnels and conformational transitions via hinge-bending motions. *Cell Biochem Biophys* 1999, 31:141–164.
9. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. *Protein Sci* 1999, 8:1181–1190.
10. Bahar I, Chennubhotla C, Tobi D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr Opin Struct Biol* 2007, 17:633–640.
11. Stevens RC, Lipscomb WN. Allosteric control of quaternary states in *E. coli* aspartate transcarbamylase. *Biochem Biophys Res Commun* 1990, 171:1312–1318.
12. Shinoda T, Arai K, Shigematsu-Iida M, Ishikura Y, Tanaka S, Yamada T, Kimber MS, Pai EF, Fushinobu S, Taguchi H. Distinct conformation-mediated functions of an active site loop in the catalytic reactions of NAD-dependent D-lactate dehydrogenase and formate dehydrogenase. *J Biol Chem* 2005, 280:17068–17075.
13. Karplus M. Role of conformation transitions in adenylate kinase. *Proc Natl Acad Sci USA* 2010, 107:E71; author reply E72.
14. Kalko SG, Gelpi JL, Fita I, Orozco M. Theoretical study of the mechanisms of substrate recognition by catalase. *J Am Chem Soc* 2001, 123:9665–9672.
15. Jakopitsch C, Droghetti E, Schmuckenschlager F, Furtmuller PG, Smulevich G, Obinger C. Role of the main access channel of catalase-peroxidase in catalysis. *J Biol Chem* 2005, 280:42411–42422.
16. Hara I, Ichise N, Kojima K, Kondo H, Ohgiya S, Matsuyama H, Yumoto I. Relationship between the size of the bottleneck 15 A from iron in the main channel and the reactivity of catalase corresponding to the molecular size of substrates. *Biochemistry* 2007, 46:11–22.
17. Bidon-Chanal A, Marti MA, Estrin DA, Luque FJ. Dynamical regulation of ligand migration by a gate-opening molecular switch in truncated hemoglobin-N from *Mycobacterium tuberculosis*. *J Am Chem Soc* 2007, 129:6782–6788.
18. Guallar V, Lu C, Borrelli K, Egawa T, Yeh SR. Ligand migration in the truncated hemoglobin-II from *Mycobacterium tuberculosis*: the role of G8 tryptophan. *J Biol Chem* 2009, 284:3106–3116.
19. Daigle R, Guertin M, Lague P. Structural characterization of the tunnels of *Mycobacterium tuberculosis* truncated hemoglobin N from molecular dynamics simulations. *Proteins* 2009, 75:735–747.
20. Perez A, Noy A, Lankas F, Luque FJ, Orozco M. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res* 2004, 32:6144–6151.
21. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. *Biophys J* 2005, 88:1291–1299.
22. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature* 1977, 267:585–590.
23. Roccatano D, Barthel A, Zacharias M. Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. *Biopolymers* 2007, 85:407–421.
24. Tinoco I, Jr., Wen JD. Simulation and analysis of single-ribosome translation. *Phys Biol* 2009, 6:025006.
25. Brandman R, Brandman Y, Pande VS. A-site residues move independently from P-site residues in all-atom molecular dynamics simulations of the 70S bacterial ribosome. *PLoS One* 2012, 7:e29377.
26. Orozco M, Orellana L, Hospital A, Naganathan AN, Emperador A, Carrillo O, Gelpi JL. Coarse-grained

representation of protein flexibility. Foundations, successes, and shortcomings. *Adv Protein Chem Struct Biol* 2011, 85:183–215.

27. Orozco M, Luque FJ. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem Rev* 2000, 100:4187–4226.
28. Vorobjev YN. Advances in implicit models of water solvent to compute conformational free energy and molecular dynamics of proteins at constant pH. *Adv Protein Chem Struct Biol* 2011, 85:281–322.
29. Hermans J, Berendsen HJC, Vangunsteren WF, Postma JPM. A consistent empirical potential for water–protein interactions. *Biopolymers* 1984, 23:1513–1518.
30. Mackerell AD, Wiorkiewiczkuczera J, Karplus M. An all-atom empirical energy function for the simulation of nucleic-acids. *J Am Chem Soc* 1995, 117:11946–11975.
31. Ott KH, Meyer B. Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. *J Comput Chem* 1996, 17:1068–1084.
32. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998, 102:3586–3616.
33. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J Am Chem Soc* 1995, 117:5179–5197.
34. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001, 105:6474–6487.
35. Rueda M, Ferrer-Costa C, Meyer T, Perez A, Camps J, Hospital A, Gelpi JL, Orozco M. A consensus view of protein dynamics. *Proc Natl Acad Sci USA* 2007, 104:796–801.
36. Perez A, Lankas F, Luque FJ, Orozco M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res* 2008, 36:2379–2394.
37. Case DA, Darden TA, Cheatham ITE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, et al. *AMBER 12*. San Francisco, California: University of California, San Francisco; 2012.
38. Brooks BR, Brooks CL 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. CHARMM: the biomolecular simulation program. *J Comput Chem* 2009, 30:1545–1614.
39. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 2008, 4:435–447.
40. Nelson MT, Humphrey W, Gursoy A, Dalke A, Kale LV, Skeel RD, Schulten K. NAMD: a parallel, object oriented molecular dynamics program. *Int J Supercomput Appl High Perf Comput* 1996, 10:251–268.
41. Larsson P, Hess B, Lindahl E. Algorithm improvements for molecular dynamics simulations. *WIREs Comput Mol Sci* 2010, 1:93–108.
42. Tanner DE, Chan KY, Phillips JC, Schulten K. Parallel generalized born implicit solvent calculations with NAMD. *J Chem Theory Comput* 2011, 7:3635–3642.
43. Predescu C, Lippert RA, Eastwood MP, Ierardi D, Xu H, Jensen MO, Bowers KJ, Gullingsrud J, Rendleman CA, Dror RO, et al. Computationally efficient molecular dynamics integrators with improved sampling accuracy. *Mol Phys* 2012, 110:967–983.
44. Bowers KJ, Dror RO, Shaw DE. The midpoint method for parallelization of particle simulations. *J Chem Phys* 2006, 124:184109.
45. Shaw DE. A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. *J Comput Chem* 2005, 26:1318–1328.
46. Harvey M, Giupponi G, De Fabritiis G. ACEMD: accelerated molecular dynamics simulations in the microseconds timescale. *J Chem Theory Comput* 2009, 5:1632–1639.
47. Kikugawa G, Apostolov R, Kamiya N, Taiji M, Himeno R, Nakamura H, Yonezawa Y. Application of MDGRAPE-3, a special purpose board for molecular dynamics simulations, to periodic biomolecular systems. *J Comput Chem* 2009, 30:110–118.
48. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Commun ACM* 2008, 51:91–97.
49. Shirts M, Pande VS. Computing: screen savers of the world unite! *Science* 2000, 290:1903–1904.
50. Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G. High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model* 2010, 50:397–403.
51. Bertini I, Case DA, Ferella L, Giachetti A, Rosato A. A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* 2011, 27:2384–2390.
52. Micheletti C, Laio A, Parrinello M. Reconstructing the density of states by history-dependent metadynamics. *Phys Rev Lett* 2004, 92:170601.
53. Laio A, Rodriguez-Fortea A, Gervasio FL, Ceccarelli M, Parrinello M. Assessing the accuracy of metadynamics. *J Phys Chem B* 2005, 109:6714–6721.
54. Raiteri P, Laio A, Gervasio FL, Micheletti C, Parrinello M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J Phys Chem B* 2006, 110:3533–3539.

55. Nymeyer H, Gnanakaran S, Garcia AE. Atomic simulations of protein folding, using the replica exchange algorithm. *Methods Enzymol* 2004, 383:119–149.
56. Kubitzki MB, de Groot BL. Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. *Biophys J* 2007, 92:4262–4270.
57. Meng Y, Dashti DS, Roitberg AE. Computing alchemical free energy differences with Hamiltonian replica exchange molecular dynamics (H-REMD) simulations. *J Chem Theory Comput* 2011, 7:2721–2727.
58. Burgi R, Kollman PA, Van Gunsteren WF. Simulating proteins at constant pH: an approach combining molecular dynamics and Monte Carlo simulation. *Proteins* 2002, 47:469–480.
59. Stern HA. Molecular simulation with variable protonation states at constant pH. *J Chem Phys* 2007, 126:164112.
60. Tai K, Murdock S, Wu B, Ng MH, Johnston S, Fangohr H, Cox SJ, Jeffreys P, Essex JW, Sansom MS. BioSimGrid: towards a worldwide repository for biomolecular simulations. *Org Biomol Chem* 2004, 2:3219–3221.
61. Silva CG, Ostropytskyy V, Loureiro-Ferreira N, Berrar D, Dubitzky W, Brito RMM. *P-found: The protein Folding and Unfolding Simulation Repository*. New York: IEEE; 2006.
62. Apostolov R, Axner L, Agren H, Ayugade E, Duta M, Gelpi JL, Gimenez J, Goni R, Hess B, Jamitzky F, et al. ScalaLife: scalable software services for Life Science. *Proceedings of the 9th HealthGrid Conference*. Bristol, UK; 2011.
63. van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, et al. Dynameomics: a comprehensive database of protein dynamics. *Structure* 2010, 18:423–435.
64. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Perez A, Carrillo O, Camps J, Fenollosa C, Repchevsky D, et al. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure* 2010, 18:1399–1409.
65. Day R, Beck DA, Armen RS, Daggett V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* 2003, 12:2150–2160.
66. Meyer T, Ferrer-Costa C, Perez A, Rueda M, Bidon-Chanal A, Luque FJ, Laughton CA, Orozco M. Essential dynamics: a tool for efficient trajectory compression and management. *J Chem Theory Comput* 2006, 2:251–258.
67. Beck DA, Alonso DOV, Daggett V. *In Lucem molecular Mechanics*. Seattle, Washington: University of Washington; 2000–2010.
68. Simms AM, Toofanny RD, Kehl C, Benson NC, Daggett V. Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. *Protein Eng Des Sel* 2008, 21:369–377.
69. Lavery R, Zakrzewska K, Beveridge D, Bishop TC, Case DA, Cheatham T, 3rd, Dixit S, Jayaram B, Lankas F, Laughton C, et al. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res* 2009, 38:299–313.
70. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* 2008, 29:1859–1865.
71. Miller BT, Singh RP, Klauda JB, Hodoscek M, Brooks BR, Woodcock HL, 3rd. CHARMMing: a new, flexible web portal for CHARMM. *J Chem Inf Model* 2008, 48:1920–1929.
72. Kota P. GUIMACS—a Java based front end for GROMACS. *In Silico Biol* 2007, 7:95–99.
73. Sellis D, Vlachakis D, Vlassi M. Gromita: a fully integrated graphical user interface to gromacs 4. *Bioinform Biol Insights* 2009, 3:99–102.
74. Roopra S, Knapp B, Omasits U, Schreiner W. jSimMacs for GROMACS: a Java application for advanced molecular dynamics simulations with remote access capability. *J Chem Inf Model* 2009, 49:2412–2417.
75. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996, 14:33–38, 27–38.
76. Hospital A, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M, Gelpi JL. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics* 2012, 28:1278–1279.
77. Wilkinson MD, Senger M, Kawas E, Bruskiewich R, Gouzy J, Noirot C, Bardou P, Ng A, Haase D, Saiz EdA, et al. Interoperability with Moby 1.0—it's better than sharing your toothbrush. *Brief Bioinfor* 2008, 9:220–231.
78. Tsujishita H, Hirono S. CAMDAS: an automated conformational analysis system using molecular dynamics. Conformational analyzer with molecular dynamics and sampling. *J Comput Aided Mol Des* 1997, 11:305–315.
79. Camps J, Carrillo O, Emperador A, Orellana L, Hospital A, Rueda M, Cicin-Sain D, D'Abramo M, Gelpi JL, Orozco M. FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* 2009, 25:1709–1710.
80. Potestio R, Aleksiev T, Pontiggia F, Cozzini S, Micheletti C. ALADYN: a web server for aligning proteins by matching their large-scale motion. *Nucleic Acids Res* 2010, 38:W41–W45.
81. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004, 57:702–710.
82. Fenwick R, Esteban-Martín S, Salvatella X. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J* 2012, 40:1339–1355.
83. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. *Proteins* 1993, 17:412–425.

FURTHER READING/RESOURCES

Harvey MJ, Giupponi G, De Fabritiis G. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* 2009, 5:1632–1639.

Perez A, Luque FJ, Orozco M. Frontiers in molecular dynamics simulations of DNA. *Acc Chem Res* 2011, 45:196–205.

Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *BMC Biol* 2011, 9:71.

Harvey MJ, De Fabritiis G. High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug Discov Today* 2012, 19–20:1059–1062.

Borhani DW, Shaw DE. The future of molecular dynamics simulations in drug discovery. *J Comput Aided Mol Des* 2012, 26:15–26.

ANNEX II

Semantic Web and Web Services Definitions

General

- **W3C (World Wide Web Consortium):** The W3C is the main international standards organization for the World Wide Web. W3C's primary activity is to develop protocols and guidelines that ensure long-term growth for the Web. W3C's standards define key parts of what makes the World Wide Web work.

- **Semantic Web:** The Semantic Web is the extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data".
  http://www.w3.org/standards/semanticweb/

- **Ontology:** Ontologies are considered one of the pillars of the Semantic Web, although they do not have a universally accepted definition. A (Semantic Web) vocabulary can be considered as a special form of (usually light-weight) ontology, or sometimes also merely as a collection of URIs with an (usually informal) described meaning.
  http://www.w3.org/standards/semanticweb/ontology

Web Services

- **WS (*Web Services*):** WS is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Two major classes of Web services exist: *REST-compliant Web Services* and *Arbitrary Web Services* (SOAP). http://www.w3.org/standards/webofservices/

- **SOAP (*Simple Object Access Protocol*):** SOAP is a protocol specification for exchanging structured information in the implementation of Web Services. It uses XML for the message format, but it can be used with many different transport protocols (HTTP, FTP, TCP, UDP, etc.).

  SOAP based services strictly define the format of messages passed back and forth. A SOAP message contains the data, the action to perform on it, the headers, and the error details in case of failure. Security is provided by WS-Security standards and is end-to-end. It supports identity through intermediaries, not just point to point (SSL).

  SOAP provides a mechanism for services to describe themselves to clients (WSDL), and to advertise their existence (UDDI). SOAP also provides reliable messaging (WS-ReliableMessaging) that has successful/retry logic built in and provides end-to-end reliability through SOAP intermediaries.
  http://www.w3.org/standards/techs/soap

- **REST (*Representation State Transfer*):** REST is a software architecture style for designing networked applications; it describes a set of architectural principles by which data can be transmitted over a standardized interface (HTTP). REST does not contain an additional messaging layer and focuses on design rules for creating stateless services (i.e. state of the service doesn't persist between subsequent requests and response). A client can access the resource using the unique URI and a representation of the resource is returned.

  REST recognizes everything as a resource (e.g. Protein, Sequence, etc.) and each resource implements a standard uniform interface (typically HTTP interface), resources have name and addresses (URIs), each resource has one or more representation (like JSON, XML or CSV) and resource representations move across the network usually over HTTP.

  With each new resource representation, the client is said to transfer state. While accessing RESTful resources with HTTP protocol, the URL of the resource serves as the resource identifier and GET, PUT, DELETE, POST and HEAD are the standard HTTP 1.1 operations to be performed on that resource.

  REST provides a lighter weight alternative. Instead of using XML to make a request, REST relies on a simple URL in many cases. As REST web services do not require the amount of information included in SOAP XML envelope, and also because of the ability to transfer binary data, they can be considerably more lightweight.

- **RESTful:** Used to refer to web services implementing REST software architecture.

- **WSDL (*Web Services Description Language*):** WSDL is a document used to describe Web Services and how to access them, written in XML language. http://www.w3.org/standards/techs/wsdl - http://www.w3.org/TR/wsdl/

- **WSDL2 (*Web Services Description Language, v2.0*):** WSDL version 2.0 tries to solve interoperability issues found with the previous versions of WSDL specifications. One of the major updates is the possibility to describe RESTful web services. Another remarkable improvement is the introduction of Internationalized Resource Identifiers (IRIs) for its described components. The ability to unambiguously identify every WSDL2 object as a resource allows the use of these identifiers within ontologies, which in turn can be expressed in Web Ontology Language (OWL).
  http://www.w3.org/standards/techs/wsdl - http://www.w3.org/TR/wsdl20/

- **SAWSDL (*Semantic Annotations for WSDL*):** SAWSDL defines mechanisms using which semantic annotations can be added to WSDL components. SAWSDL does not specify a language for representing the semantic models, e.g. ontologies. Instead, it provides mechanisms by which concepts from the semantic models that are defined either within or outside the WSDL document

can be referenced from within WSDL components as annotations. These semantics when expressed in formal languages can help disambiguate the description of WSs during their automatic discovery and composition. http://www.w3.org/standards/techs/sawsdl - http://www.w3.org/TR/sawsdl/

- **SA-REST (*Semantic Annotation for REST*):** SA-REST defines mechanisms to include semantic annotations in HTML/XHTML documents, typically embedded ontological meta-data, allowing a capable processor to gain extra information about the content of the document.
  http://www.w3.org/Submission/SA-REST/

- **WADL (*Web Application Description Language*):** WADL is a machine-readable XML description of HTTP-based web applications (typically REST web services). WADL models the resources provided by a service and the relationships between them. WADL is intended to simplify the reuse of web services that are based on the existing HTTP architecture of the Web. It is platform and language independent and aims to promote reuse of applications beyond the basic use in a web browser.

  WADL was submitted to the World Wide Web Consortium by Sun Microsystems on 31 August 2009, but the consortium has no current plans to standardize it and it is not yet widely supported. WADL is the REST equivalent of SOAP's Web Services Description Language (WSDL), which can also be used to describe REST web services. http://www.w3.org/Submission/wadl/

Web languages and Protocols

- **HTTP (*Hypertext Transfer Protocol*):** The Hypertext Transfer Protocol is an application protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web. http://www.w3.org/Protocols/

- **HTML (*Hypertext Markup Language*):** HTML is the publishing language of the World Wide Web since 1990s. It is used to create web pages and other information that can be displayed in a web browser. HTML defines a series of tags enclosed in angle brackets (<HTML>), which are read and interpreted by web browsers to finally compose them into visible or audible web pages.
  http://www.w3.org/html/

- **XHTML (*eXtensible Hypertext Markup Language*):** XHTML is a "reformulated" HTML, to conform to the XML standard. It takes profit of both, HTML and XML languages, to gain in power and flexibility. XHTML has much stricter language syntax than HTML, still being backward compatible with older, non-XHTML compliant web browsers. http://www.w3.org/TR/xhtml1/

- **HTML5 (*Hypertext Markup Language, v.5*):** HTML5 is the latest standard for HTML. As the Internet has changed significantly since the definition of the first HTML version (1999), a new version was required, adapted to the hardware revolution that we are living. HTML5 is specially designed to deliver rich

content without the need for additional plugins. It is able to deliver everything from animation to graphics, or music to movies, and can also be used to build complicated web applications. HTML5 is also cross-platform, is designed to work whether you are using a PC, or a Tablet, a Smartphone, or a Smart TV. http://www.w3.org/TR/html5/

- **XML (*eXtensible Markup Language*):** XML is a markup language that defines a set of rules for encoding document in a format that is both human-readable and machine-readable. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all free open standards. http://www.w3.org/XML/

- **XSD (*XML Schema Definition*):** An XML Schema Definition describes the structure of an XML document. XSD can be used to express a set of rules to which an XML document must conform in order to be considered *valid* according to that schema. http://www.w3.org/standards/techs/xmlschema

- **JSON (*Java Script Object Notation*):** JSON is a lightweight data-interchange format, easy for humans to read and write, and easy for machines to parse and generate, consisting on *attribute-value* pairs. Although originally derived from the JavaScript scripting language, JSON is a language-independent data format, and code for parsing and generating JSON data is readily available in a large variety of programming languages.

- **RDF (*Resource Description Framework*):** RDF is the standard for encoding metadata and other knowledge on the Semantic Web. Developed under the guidance of the World Wide Web Consortium, RDF was designed to allow developers to build search engines that rely on the metadata and to allow Internet users to share Web site information more readily. RDF relies on XML as interchange syntax, creating an ontology system for the exchange of information on the Web.
  http://www.w3.org/standards/techs/rdf

- **SPARQL (*SPARQL Protocol and RDF Query Language*):** SPARQL is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format.
  http://www.w3.org/standards/techs/sparql

- **URI (*Uniform Resource Identifier*):** A URI is the way you identify any Internet points of content; Internet space is formed by many points of content: a page of text, a video or sound clip, a still or animated image, a program, etc. A URI typically describes:

  - The mechanism used to access the resource.
  - The specific computer that the resource is housed in.
  - The specific name of the resource on the computer.

The most common form of URI is the Web page address, which is a particular form or subset of URI called a Uniform Resource Locator (URL).
http://www.w3.org/Addressing/

- **URL (*Uniform Resource Locator*):** is the unique address for a file that is accessible on the internet. A common way to get to a web site is to enter the URL of its home page file in your web browser's address line. However, any file within that web site can also be specified with a URL. Such a file can be any web page other than the home page (HTML), an image file (PNG, JPEG, etc.) or even a program (CGI, Java applet, etc.).

  URL is a specialization of URI that defines the network location of a specific representation for a given resource (a URL is a URI, but a URI is not a URL). The URL contains the name of the protocol to be used to access the file resource, a domain name that identifies a specific computer on the Internet, and a pathname, a hierarchical description that specifies the location of a file in that computer.
  http://www.w3.org/Addressing/

- **IRI (*Internationalized Resource Identifier*):** The URI syntax essentially restricts Web addresses to a small number of characters: basically, just upper and lower case letters of the English alphabet, European numerals and a small number of symbols. User's expectations and use of the Internet have moved on since then, and there is now a growing need to enable use of characters from any language in Web addresses. A Web address in your own language and alphabet is easier to create, memorize, transcribe, interpret, guess, and relate to.

  IRIs are a generalization of URIs; while URIs are limited to a subset of the ASCII character set, IRIs may contain characters from the Universal Character Set, including Chinese, Japanese, Cyrillic characters, etc.

  IRIs were included in the WSDL Version 2.0 specifications.
  http://www.w3.org/Addressing/

Ontology related

- **OWL (*Web Ontology Language*):** The Web Ontology Language is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases. The languages are characterized by formal semantics and RDF/XML-based serializations for the Semantic Web.
  http://www.w3.org/standards/techs/owl

- **OBO (*Open Biomedical Ontology*):** OBO is a collaborative experiment involving developers of science-based ontologies. It is concerned with establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain.