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

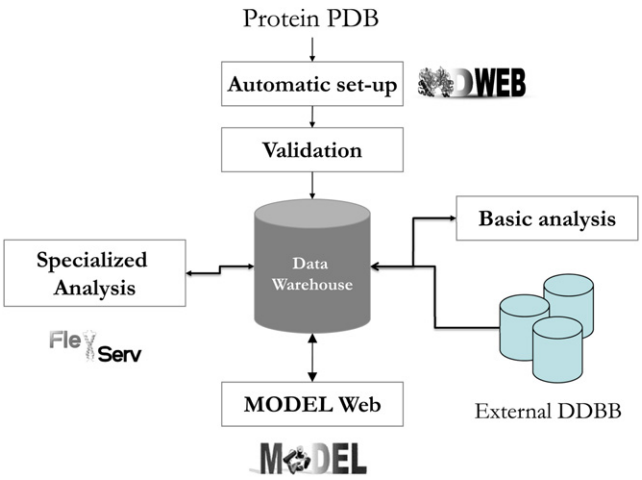

Figure 1. General Flowchart of the MoDEL Platform
The automatic setup tools prepare and run a trajectory from the structure in PDB format. Before storing the results, the trajectory is validated and later analyzed with our analysis tools. MODEL data are available through our public MODEL web server at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

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

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

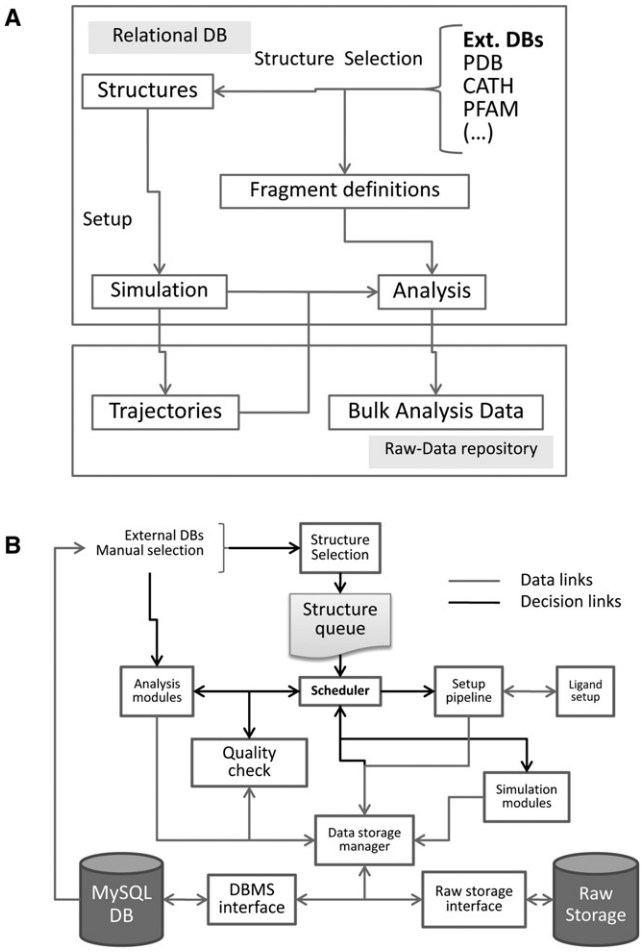

Figure 2. General Structure of the MoDEL Data Warehouse and Management Software

(A) General scheme of MoDEL data warehouse.

(B) Diagram of MoDEL management software.

See also Figures S2–S4, and Table S1.

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),

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

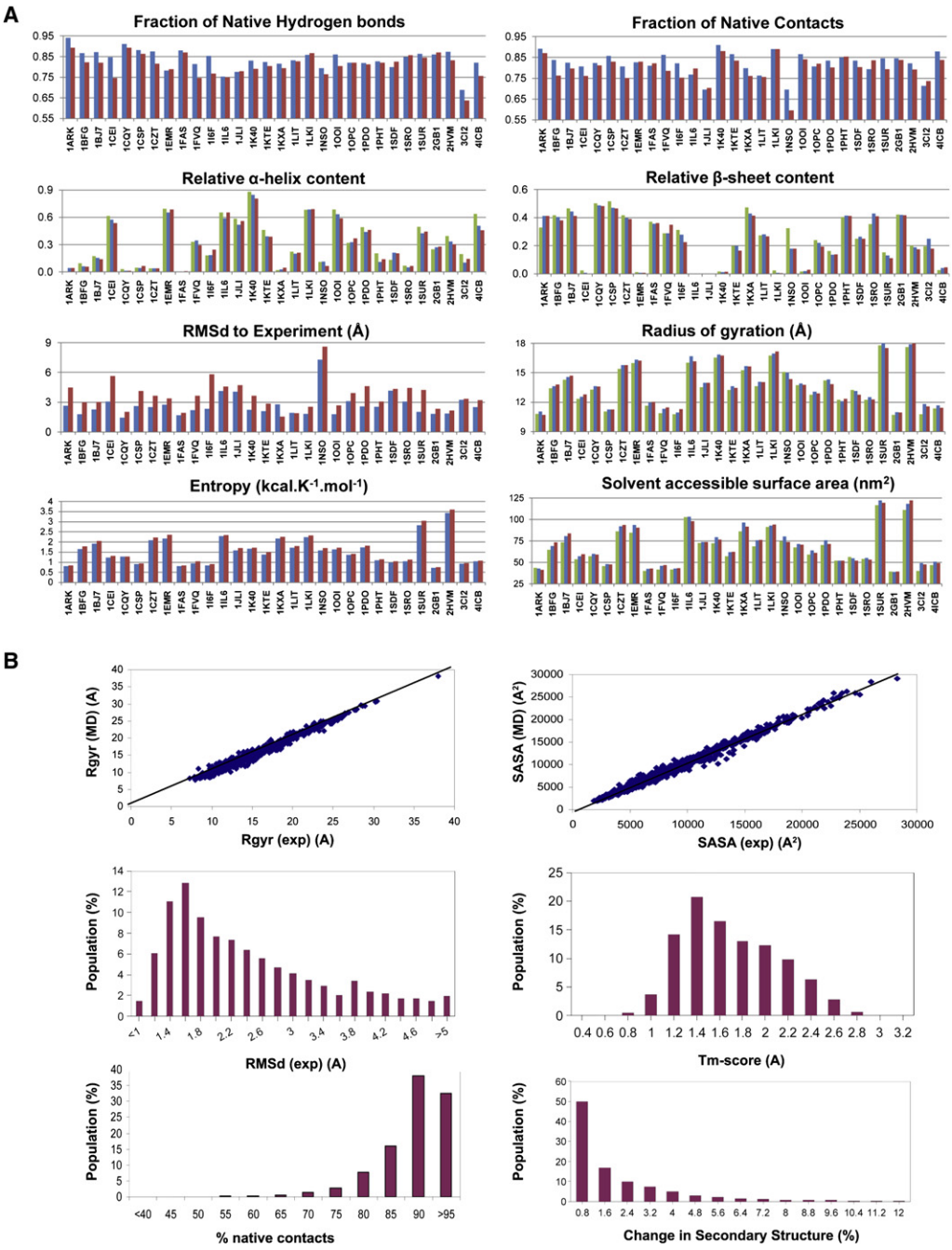

Figure 3. Quality of Simulations in MoDEL

(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

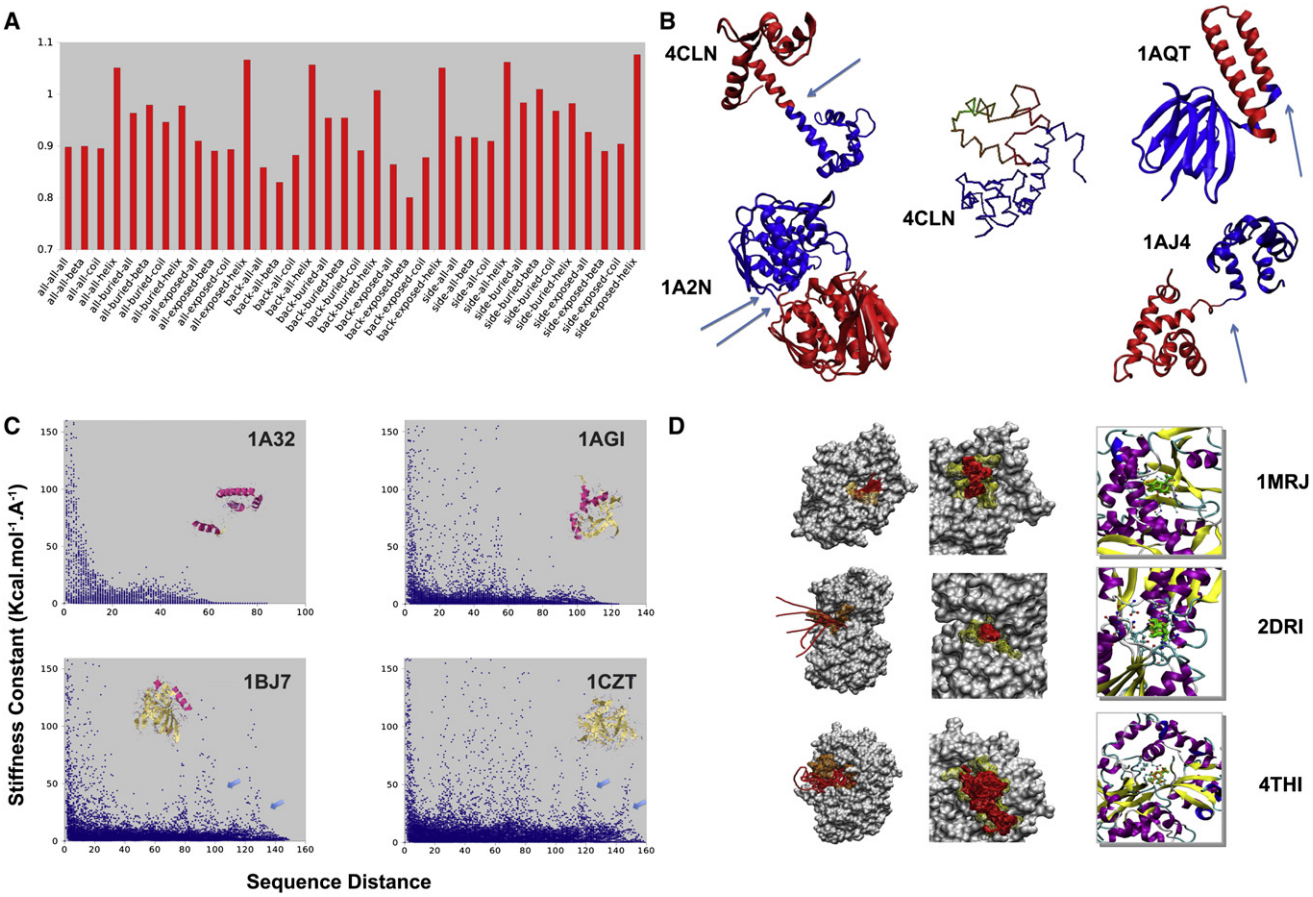

Figure 4. Examples of Data Mining in MoDEL

(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

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

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.

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.