Protein Engineering, Design & Selection vol. 21 no. 6 pp. 353–368, 2008
Published online April 14, 2008 doi:10.1093/protein/gzn011

Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations

David A.C. Beck[1], Amanda L. Jonsson[2], R. Dustin Schaeffer[2], Kathryn A. Scott[4], Ryan Day[2,5], Rudesh D. Toofanny[1], Darwin O.V. Alonso[1] and Valerie Daggett[1,2,3,6]

[1]Department of Bioengineering, [2]Biomolecular Structure and Design Program and, [3]Biomedical and Health Informatics, University of Washington, Box 355013, Seattle, WA 98195-5013,USA, [4]Current address: Structural Bioinformatics and Computational Biochemistry Unit, University of Oxford, Oxford OX1 3QU, UK
[5]Current address: Department of Physics, Applied Physics, and Astronomy, Renssalaer Polytechnic Institute, Troy, NY 12181, USA

[6]To whom correspondence should be addressed.
E-mail: daggett@u.washington.edu

The goal of Dynameomics is to perform atomistic molecular dynamics (MD) simulations of representative proteins from all known folds in explicit water in their native state and along their thermal unfolding pathways. Here we present 188-fold representatives and their native state simulations and analyses. These 188 targets represent 67% of all the structures in the Protein Data Bank. The behavior of several specific targets is highlighted to illustrate general properties in the full dataset and to demonstrate the role of MD in understanding protein function and stability. As an example of what can be learned from mining the Dynameomics database, we identified a protein fold with heightened localized dynamics. In one member of this fold family, the motion affects the exposure of its phosphorylation site and acts as an entropy sink to offset another portion of the protein that is relatively immobile in order to present a consistent interface for protein docking. In another member of this family, a polymorphism in the highly mobile region leads to a host of disease phenotypes. We have constructed a web site to provide access to a novel hybrid relational/multidimensional database (described in the succeeding two papers) to view and interrogate simulations of the top 30 targets: http://www.dynameomics.org. The Dynameomics database, currently the largest collection of protein simulations and protein structures in the world, should also be useful for determining the rules governing protein folding and kinetic stability, which should aid in deciphering genomic information and for protein engineering and design.

Keywords: database/Dynameomics/molecular dynamics/protein dynamics/protein folds

Introduction

The Protein Data Bank (PDB) (Berman *et al*., 2000) currently contains over 40 000 protein structures and the effort to expand the PDB continues with renewed vigor in the age of high-throughput structural genomics efforts. The coordinates deposited in the PDB are static, average structures, yet proteins are necessarily dynamic entities that sample an ensemble of conformers in their folded (native) states. The recent improvements in moderate resolution protein structure prediction from amino acid sequences (and by extension genes themselves) have also increased the number of static structures in the public domain (Lattman, 2005). However, the focus on static structures is a known problem for functional annotation, and it is recognized as a severe limitation of drug design methods (Carlson, 2002; Lin *et al*., 2003; Meagher and Carlson, 2004). For example, it may be that binding-competent conformational states have low populations and are only accessed dynamically within the native ensemble. In a commentary entitled 'Not just your average structures', Petsko (1996) described a hypothetical day in New York in which the law of averages is repealed. This leads to chaos as everyone showers at the same time, causing a water shortage. Everyone makes their toast at the same instant, resulting in a massive power outage. Then, all of the commuters try to get through the Holland Tunnel at the same time, leading to a traffic jam that takes hours to clear. His point is that 'We have designed our civilization around averages, but we understand that reality consists of fluctuations about those averages'. While often overlooked, the same is true for proteins. The beautiful static, time and ensemble-average structures from NMR and X-ray crystallography represent only part of the story. Protein function can depend on what we do not readily see: the excursions from the average.

The processes by which proteins convert between states on the folding pathway or substates in a given conformational ensemble are difficult to probe experimentally. Consequently, realistic all-atom explicit solvent molecular dynamics (MD) simulation (Beck and Daggett, 2004) is an attractive approach because of its ability to provide atomic detail of protein dynamics, function, folding and unfolding. MD has been used to study a wide variety of proteins, mini-proteins and peptides for purposes as diverse as protein folding, drug design, nuclear magnetic resonance (NMR) structure determination and protein structure prediction validation. In addition, there are a number of diseases associated with protein instability and unfolding, such as amyloid diseases, and MD has been instrumental in providing structural information in these cases (Daggett, 2006).

Our simulations of cytochrome b_5 provide an example of new functional information that only became apparent through the excursions from the average observed in MD simulations. From a native-state simulation we discovered a novel, dynamic and hydrophobic cleft on the surface of the protein (Storch and Daggett, 1995). This cleft is unobservable in the crystal structure; however, in the MD simulation, it opens and transiently exposes hydrophobic residues and provides access to the heme

group, which is required for function. We proposed that this cleft would provide an ideal position for docking of the protein's electron transfer partners, allowing for protected transfer of the electron through a channel lined with aromatic residues. Subsequently, we performed experiments that verified that the protein moves as predicted from MD (Storch *et al.*, 1999a, b). Additional MD and experimental NMR studies of the cytochrome c—cytochrome b_5 complex indicated that cytochrome c does indeed bind to the cleft identified from our MD simulations, which uniquely explains the experimental data (Hom *et al.*, 2000). Thus, MD simulations have the ability to reveal many interesting biologically relevant phenomena that cannot be seen in static structures.

Here, we combine MD for studying protein dynamics with a high-throughput approach to simulation and annotation of protein fold space. It is our intention to simulate at least one representative of each known protein fold in its native biological state and along its thermal unfolding pathway, a project which we have dubbed 'Dynameomics'. Both experimental and theoretical studies comparing different members of a fold family suggest that for many folds simulating a single representative may be sufficient to describe the general behavior of the entire family (Clarke *et al.*, 1999; Gunasekaran *et al.*, 2001; Gianni *et al.*, 2003). Therefore, we are primarily restricting ourselves to one representative from each fold; however, we are expanding our simulations to include multiple representatives of more highly populated folds.

To date, we have performed over 3000 simulations of more than 400 proteins, resulting in 10^3 times more structures than the PDB. Here we focus on an initial set of 188 native-state simulations, which are the result of working our way down our previously described population-ranked target list of 1130 folds from most- to least-populated (Day *et al.*, 2003). In choosing targets for our fold list, we focused on proteins with well-resolved experimental structures, biomedical relevance and with experimental data available to validate the simulations. Examples of some of the medically relevant proteins in our dataset include: amyloid β-precursor protein involved in Alzheimer's disease; glutathione S-transferase (GST), which is involved in resistance to chemotherapy; HIV-1 protease; MAP30, which is implicated in HIV and cancer, and serum amyloid P component, amyloidosis. Our fold list was also used recently by another group to choose targets for simulations comparing different force fields (Rueda *et al.*, 2007). For the 188 simulations presented here, we include several measures of validation of the trajectories against available experimental data and summary statistics derived from the trajectories. In addition, we present specific simulations in detail to demonstrate general phenomena observed across the dataset. Due to the size and complexity of the database constructed from the simulations and their analyses, we were forced to develop novel strategies for storing, managing and analyzing our multi-terabyte dataset in the accompanying papers (Kehl *et al.*, 2008; Simms *et al.*, 2008). The simulations and analyses of the top 30 folds by population are available at our public web site for researchers to browse, search and download at http://www.dynameomics.org.

Methods

The Dynameomics protocol can be broken down into five steps (Fig. 1). The first step, known as 'Target Selection' is where protein structures are selected as fold representatives from the PDB. The second step involved preparing the protein structure from the PDB for all-atom protein simulations in a bath of explicit water molecules at 298 Kelvin (K), which is the third step. Subsequent to simulation, in the fourth step, a standardized set of 32 analyses are calculated for the trajectory. Finally, in step five, the simulations (i.e. coordinates) and analyses are loaded into a database for annotation, browsing and data-mining. A more complete discussion of the fifth step is available in the accompanying papers (Kehl *et al.*, 2008; Simms *et al.*, 2008).

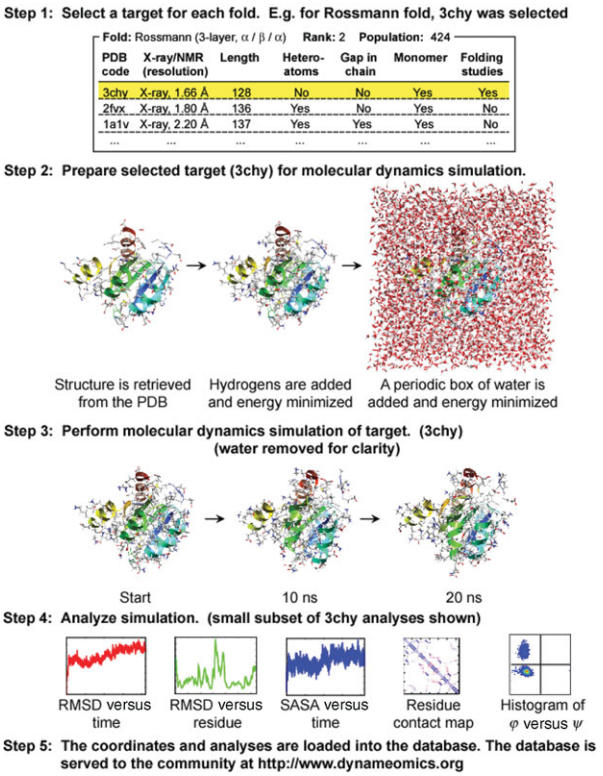

Fig. 1. Dynameomics protocol. For each fold identified by a consensus treatment of CATH, SCOP and Dali, a representative structure (or target) was chosen from the list of fold members (Day *et al.*, 2003). This target selection process, Step 1, involves choosing a fold member's structure from the PDB, giving preference to those with shorter sequences, no gaps in sequence (e.g. missing density), structures of monomeric proteins and proteins of biological, disease or folding interest. Once the target has been selected, in Step 2, the structure is retrieved from the PDB and any missing hydrogen atoms are added. The protein is minimized and placed in a box of water. This water is subjected to minimization and dynamics while the protein is held fixed. This allows the water to adopt a hydration layer around the protein without perturbing it. In Step 3, the solvated protein is simulated in *il*mm (Beck *et al.*, 2000–2008) at 298 Kelvin for at least 21 ns. Here, only the protein atoms of the starting structure, the 10 ns structure and the 20 ns structure are depicted. After the completion of the simulation, the trajectory is analyzed (see Table IV). For that step, we show representative plots of a tiny subset of the analyses performed in Step 4. After analysis, in Step 5, the trajectory and the calculated analytical properties from Step 4 are loaded into a database. The data for the targets selected to represent the top 30 most populated folds are available at our web site.

Step 1: target selection from folds

Protein folds were identified by a consensus approach of three popular fold/domain classification systems: (i) structural classification of proteins (SCOP) v1.65 (Murzin *et al.*, 1995); (ii) class, architecture, topology and homologous

superfamily (CATH) v2.4 (Orengo *et al*., 1997); (iii) Dali Domain Dictionary v3.1β (Dietmann *et al*., 2001) as described previously (Day *et al*., 2003). In that work, a fold was defined as a set of protein chains that shared two or more classification assignments from these three well-known systems. For each fold, individual targets were selected from the structures in the PDB. Selection criteria were such that single chains were preferred over multimers and multi-domain proteins, shorter chains were preferred over longer ones and high-resolution crystal structures were preferred over those of lower resolution and those derived from NMR. For folds where several candidate targets remained after applying these filters, the final target selection was made giving deference to those systems with available experimental data in the forms of NMR observables, folding and unfolding studies, drug design interest, many protein–protein interactions or biomedical relevance.

At this time, we have simulated and analyzed 188 targets. Their PDB codes, chain and segment specifications, fold name, source organism and CATH class codes are shown in Table I. These targets represent 181 folds from among the 450 most populated folds. The 181 subset was chosen from the larger collection of 450 because their targets were the most amenable to simulation. In particular, each target's main chain was contiguous in sequence and was long enough that we expected it to fold autonomously. The 181 folds represent 67% of known protein domains. The discrepancy between the number of targets (188) and folds (181) is a result of two special cases: (i) multiple targets for eight of the folds were simulated where a fold was of special interest (e.g. the Rossmann fold of rank 2, DNA/RNA-binding 3-helix bundles of rank 6, and SH3 barrels of rank 17), and (ii) 2-folds were simulated together in one target. The latter case only applied to the GST target, which contains two specific folds: the N- and C-terminal domains of GST folds (ranks 12 and 27, respectively).

We present a breakdown of the 188 targets by their experimental source (X-ray crystallography, NMR) and how much of the source PDB was simulated in Table II. For the 72% of targets whose structures were solved by X-ray crystallography, the mean resolution was 1.97 ± 0.38 Ångstroms (Å). Table II distinguishes between targets where the entire protein chain(s) in the PDB (61%) was used and those cases where an entire subunit, i.e. a single covalently bound chain (13%) or a partial subunit was used, i.e. the protein chain was cut (26%). The mean target length was 127 ± 77 residues with the smallest target having 29 residues and the largest 411 residues (per monomer).

Step 2: target structure preparation

Before MD simulations can be performed, the structure must be abstracted from the PDB, missing atoms added, energy minimized and then solvated. When the structure was determined by NMR, the first model in the PDB was used. When multiple conformations for specific residues were present in the PDB, the first conformation was chosen. Hydrogen was then added and disulfide bonds were included where indicated in the PDB file by SSBOND records. Then the coordinates of the structure were minimized with steepest descent (SD) minimization for 1000 steps or until the potential energy converged. The minimized structure was then solvated in a box of water extending at least 10 Å from the protein. No waters were permitted to be closer than 1.8 Å to any protein heavy atom. The box size was adjusted slightly such that its density matched the experimental value of: 0.997 g/ml for 298 K (Kell, 1967). Then, the water molecules (only) were subjected to 1000 steps of SD minimization, followed by 1 ps of MD where the temperature of the water was raised to approximately 40 K. The water molecules were again then SD minimized for 500 steps and, finally, the protein was SD minimized for 500 steps.

As a quality control measure, the resulting molecular systems were individually inspected. Problems such as minimization or preparation failures (e.g. failure of the total energy to decrease during minimization, typically a result of omission of a disulfide bond), and large structural changes in the protein were identified and corrected. To ensure that all systems were correct, each prepped structure was examined by hand.

Step 3: molecular dynamics simulation of target

All-atom, explicit solvent MD simulations of the 188 targets were conducted. For each target, a simulation at 298 K was conducted for at least 21 nanoseconds (ns). Structures were saved from the simulation at 1 ps resolution (yielding a trajectory) for later analyses and full restart files with velocities and accelerations were saved once per nanosecond.

The simulations were performed in keeping with our core MD protocols and simulation methods that are published elsewhere (Beck and Daggett, 2004; Beck *et al*., 2005). The simulations used fully flexible representations of protein and solvent in the Levitt *et al*. force field with our standard parameters (Levitt *et al*., 1995, 1997). The flexible three center (F3C) water model was used as it has demonstrated success in reproducing a large number of experimental observables across a range of temperatures (Levitt *et al*., 1997; Beck *et al*., 2003). The microcanonical ensemble was used, where the number of atoms, unit cell volume and total energy were conserved (NVE), as well as linear momentum (NVE*p*). The simulation timestep was 2 femtoseconds (fs) such that each picosecond (ps) of simulation time required 500 steps (cycles of force field evaluation and numerical integration). In turn, a 21 ns simulation requires 1.05×10^7 steps. The unit cell was periodic with minimum image conventions employed (Allen and Tildesley, 1987). For the non-bonded interactions, a force-shifted 10 Å cut-off range was used (Levitt *et al*., 1995). Non-bonded interactions between atoms in charge groups separated by three bonds were included and scaled by 0.4. The non-bonded interaction pairlists were used for no more than 6 fs (i.e. they were updated every three steps).

The simulations were performed with *in lucem* Molecular Mechanics (*il*mm) (Beck *et al*., 2000–2008). A variety of computational resources were employed for the simulations. Most of the simulations were performed on Seaborg: the Department of Energy's National Energy Research Supercomputing Council's 6080 processor supercomputer. Each IBM AIX based supercomputer node (IBM Nighthawk) has 16 POWER3 processors operating at 375 Mhz. In addition, some of the more recent simulations were performed on quad-core Intel Woodcrest 5130 nodes (2.00 Ghz) running Windows Server 2003 R2.

Table I. List of 188 proteins simulated

PDB	Chain	Range	Species	Fold name
1wit		1–93	*Caenorhabditis elegans*	Immunoglobulin-like
3chy		2–129	*Escherichia coli*	Rossman folds I (RF1)
1ypi	A	2–248	*Saccharomyces cerevisiae*	TIM barrels
1sac	A	1–204	*Homo sapiens*	Jelly roll
1ris		1–97	*Thermus thermophilus*	α-βplaits
1enh		3–56	*Drosophila melanogaster*	DNA/RNA-binding 3-helix bundles
1a6n		1–151	*Physeter catodon*	Globins
1hgu		2–191	*Homo sapiens*	Four-helix bundles
1ubq		1–76	*Homo sapiens*	β-grasps
1ev4	C	1–78	*Rattus norvegicus*	Thioredoxin-like
1mjc		2–70	*Escherichia coli*	Oligonucleotide-binding (OB) folds
1e65	A	1–128	*Pseudomonas aeruginosa*	IG-like II
2giw		1–104	*Equus caballus*	Cytochrome C folds
1ght	A	1–105	*Escherichia coli*	Rossmann folds II
1shg		6–62	*Gallus gallus*	SH3 barrels
1shf		84–142	*Homo sapiens*	SH3 barrels
1ebd	A	155–271	*Bacillus stearothermophilus*	FAD/NAD(P) binding domains
1snb		1–64	*Buthus martensii karsch*	Defensin A-like I
2afp		1–129	*Hemitripterus americanus*	C-type lectin-like
1ifc		1–131	*Rattus norvegicus*	Lipocalins
2adr		102–161	*Saccharomyces cerevisiae*	C2H2 and C2HC Zinc finger
1ntn		1–72	*Naja naja oxiana*	Snake toxin-like
1g6l	A	1–99	Human immunodeficiency virus type 1	Acid proteases
2pth		1–193	*Escherichia coli*	Rossmann folds III
1ev4	C	79–207	*Rattus norvegicus*	GST C-terminal folds
1sso		1–62	*Sulfolobus solfataricus*	IL-8 like (OB folds)
1ixa		46–84	*Homo sapiens*	Laminin-like (knottins)
1bp5	A	82–246	*Homo sapiens*	D-maltodextrin binding proteins
1ihb	A	5–160	*Homo sapiens*	α–α superhelices
1vap	A	1–123	*Agkistridon piscivorus piscivorus*	Phospholipase A2-like
1dro		1–122	*Drosophila melanogaster*	PH-domain like
1iad		1–200	*Astacus astacus l.*	Zincin-like
1tmc	A	1–175	*Homo sapiens*	MHC-like
2baa		1–243	*Hordeum vulgare*	Lysozyme-like
1ugi	A	2–84	Bacteriophage pbs2	Cystatin-like
1eqk	A	1–102	*Oryza sativa subsp. japonica*	Cystatin-like
1ceq	A	163–328	*Plasmodium falciparum*	Lactate and malate dehydrogenases
2blt	A	2–360	*Enterobacter cloacae*	β-lactamases
1ab2		1–109	*Homo sapiens*	SH2-like
1d1n	A	143–241	*Bacillus stearothermophilus*	Reductase/isomerase/elongation factor common domains
1uok		480–558	*Bacillus cereus*	α-amylases (IG-like)
1amm		1–85	*Bos taurus*	γ-crystalin-like
2tgi		1–112	*Homo sapiens*	Cystine-knot cytokines
1shp		1–55	*Stichodactyla helianthus*	BPTI-like
1aap	A	1–56	*Homo sapiens*	BPTI-like
1bf0		1–60	*Dendroaspis angusticeps*	BPTI-like
1cvz	A	1–212	*Carica papaya*	Cysteine proteinases
1ril		2–147	*Thermus thermophilus*	RNase H-like
1c0g	A	137–334	*Dictyostelium discoideum* and *tetrahymena*	RNase H-like
1tfb		111–207	*Homo sapiens*	Cyclin-like
1grj		2–79	*Escherichia coli*	Membrane all-α helix hairpins
2lao		91–191	*Salmonella typhimurium*	D-maltodextrin binding proteins
1ge8	A	2–117	*Pyrococcus furiosus*	DNA clamps
11as	A	4–330	*Escherichia coli*	Class II aaRS and biotin synthetases
1ypr	A	1–125	*Saccharomyces cerevisiae*	Profilin-like
1lbd		225–462	*Homo sapiens*	Nuclear receptor ligand-binding domains
1php		177–381	*Bacillus stearothermophilus*	Rossmann folds V
1byl		1–121	*Streptoalloteichus hindustanus*	Dihydroxylbiphenyl dioxygenases
2hnp		5–282	*Homo sapiens*	Protein–tyrosine phosphatases
1cok	A	1–68	*Homo sapiens*	SAM domain-like/DNA polymerase domain
1cem		33–395	*Clostridium thermocellum*	α/α toroids
1agi		1–125	*Bos taurus*	RnaseA-like
3grs		347–478	*Homo sapiens*	Heme-dependent peroxidases
1d0n	A	27–159	*Equus caballus*	Actin depolymerizing proteins
1ezg	A	2–83	*Tenebrio molitor*	Single-stranded right-handed β-helices
1b6b	A	34–201	*Ovis aries*	Acyl-CoA N-acyltransferases
1iyu		1–79	*Azotobacter vinelandii*	Barrel-sandwich hybrid (OB folds)
5hpg		1–84	*Homo sapiens*	Kringle-like
2glt		1–122	*Escherichia coli*	Rossmann folds VI
1cun	A	1–116	*Gallus gallus*	Spectrin repeat-like

Continued

Table I. *Continued*

PDB	Chain	Range	Species	Fold name
1ayi		1–86	*Escherichia coli*	Acyl carrier protein-like
1unk	A	1–87	*Escherichia coli*	Acyl carrier protein-like
1hpt		1–56	*Homo sapiens*	Ovomucoid/PCI-1 like inhibitors/4-Layer sandwiches
1gad	O	147–316	*Escherichia coli*	Glyceraldehyde-3-Phosphate dehydrogenase-like
1fad	A	89–183	*Mus musculus*	DEATH domains
1qts	A	826–938	*Mus musculus*	Clathrin adaptor appendage domains
1f8d	A	1–388	Influenza a virus	6-bladed β-propellors
2sil		2–382	*Salmonella typhimurium*	6-bladed β-propellors
1bw3		1–125	*Hordeum vulgare*	Double psi β-barrels
1uxc		1–50	*Escherichia coli*	λ-repressor like DBD
2trc	B	1–340	*Bos taurus*	7-bladed β-propellor
1iva		1–48	*Agelenopsis aperta*	Omega-Agatoxin V (knottins)
1a2p	A	3–110	*Bacillus amyloliquefaciens*	Microbial ribonucleases
1ai9	A	1–192	*Candida albicans*	Dihydrofolate reductases
1fzt	A	1–211	*Schizosaccharomyces pombe*	Rossmann folds VII
1dyw	A	1–172	*Caenorhabditis elegans*	Cyclophilin-like
1imf		147–276	*Homo sapiens*	Inositol polyphosphate phosphatases
1hcc		1–59	*Homo sapiens*	Complement control module/SCR domains
1jd1	A	3–128	*Saccharomyces cerevisiae*	Bacillus chorismate mutase-like
1dco	A	6–104	*Rattus norvegicus*	α-β plaits II
1d8v	A	1–263	*Momordica charantia*	MHC antigen-recognition domains
1fqn	A	4–261	*Homo sapiens*	Carbonic anyhdrases
1bea		5–120	*Zea mays*	Bifunctional inhibitor/lipid-transfer protein/seed storage
1i11	A	1–70	*Mus musculus*	HMG-boxs
1qlp	A	23–394	*Homo sapiens*	Serpins
1ep0	A	3–185	*Methanobacterium thermoautotrophicum*	Double-stranded β-helix/Jelly Rolls
1bkp		2–279	*Bacillus subtilis*	Thymidylate synthase/dCMP hydroxymethylase
1bhd	A	147–254	*Homo sapiens*	Calponin-homology domains
1cxw	A	1–60	*Homo sapiens*	Kringle-like
1dfa	A	299–415	*Saccharomyces cerevisiae*	Homing endonuclease-like
1b8w	A	1–42	*Ornithorhynchus anatinus*	Enolase N-terminal domains
1c2a	A	4–123	*Hordeum vulgare*	Defensin-like II
1hrd	A	52–187	*Clostridium symbiosum*	Aminoacid dehydrogenase-like, N-terminal domain
1b34	A	2–81	*Homo sapiens*	SM snRNP motifs / SH3 barrels
1ah6		2–214	*Saccharomyces cerevisiae*	Hsp90-like
1cuk		156–203	*Escherichia coli*	RuvA C-terminal domain-like
1e6y	B	2041–2188	*Methanosarcina barkeri*	α-βplaits III
3rub	S	1–123	*Nicotiana tabacum*	Rubisco, small subunit
1dsx	A	33–119	*Rattus norvegicus*	Potassium channel domains
1fkb		1–107	*Homo sapiens*	FKBP-like
1chu	A	244–348	*Escherichia coli*	Succinate dehydrogenase/ fumarate reductase catalytic domains
1evl	A	531–642	*Escherichia coli*	Rossmann folds VIII
1c25		335–495	*Homo sapiens*	Rhodanese/Cell-cycle control phosphatases
1ldl		1–46	*Homo sapiens*	Ligand-binding domain of LDL receptors
1bg2		3–325	*Homo sapiens*	P-loop containing NTP hydrolases
4mat	A	2–272	*Escherichia coli*	Creatinase/aminopeptidase
1esj	A	1–272	*Bacillus subtilis*	Ribokinase-like
1b2p	A	1–119	*Scilla campanulata*	β-prisms II
1huu	A	1–90	*Bacillus stearothermophilus*	IHF-like DNA-binding proteins
1e5 m	A	6–416	*Synechocystis sp.*	Ketoacyl-synthetases
4pga	A	220–335	*Pseudomonas 7a*	Rossmann folds IX
1g5 m	A	3–207	*Homo sapiens*	Toxin membrane translocation domains
1dpb		395–637	*Azotobacter vinelandii*	CoA-dependent acyltransferases
1pkl	A	87–187	*Leishmania mexicana*	PK β-barrel domain-like
1dpt	A	1–117	*Homo sapiens*	Tautomerase/migration inhibitory factor
1d6t	A	1–117	*Staphylococcus aureus*	Ribosomal protein S5 domain 2-like
1eo9	A	4–205	*Acinetobacter calcoaceticus adp1*	Prealbumin-like
1f68	A	730–832	*Homo sapiens*	Bromodomain-like
1d2d	A	1–56	*Cricetulus griseus*	S15/NS1 RNA-binding domain
1erd		1–40	*Euplotes raikovi*	Protozoan pheremone proteins
1crn		1–46	*Crambe abyssinica*	Crambin-like
1egl		1–70	*Hirudo medicinalis*	CI2-like serpins
1mut		1–129	*Escherichia coli*	Nudix-like
1srv	A	192–336	*Thermus thermophilus*	Swivelling β/β/α domains
1akz		82–304	*Homo sapiens*	DNA glycosylase
1tul		7–108	*Autographa californica nuclear*	β-clips
1vmo	A	1–163	*Gallus gallus*	β-prisms
1f60	A	333–441	*Saccharomyces cerevisiae*	EF-Tu/eEF-1α C-terminal domains
1cmz	A	79–206	*Homo sapiens*	Regulator of G-protein signaling
1eai	C	1–61	*Sus scrofa*	Serpins II
1b10	A	125–228	*Mesocricetus auratus*	Prion-like

Continued

Table I. *Continued*

PDB	Chain	Range	Species	Fold name
1knb		396–581	Human adenovirus type 5	Adenovirus fiber protein head domains (knob domain)
1baz	A	5–53	Bacteriophage p22	Met repressor-like
1bx8		5–53	*Hirudo medicinalis*	Antistasin (knottins)
3pro	C	6–84	*Lysobacter enzymogenes*	α-lytic protease prodomain-like
1e1q	A	24–93	*Bos taurus*	α/βsubunits F1 ATPase/thrombin
1edi		1–56	*Staphylococcus aureus*	Bacterial immunoglobulin/albumin-binding domains
1dzf	A	143–215	*Saccharomyces cerevisiae*	RPB5-like RNA polymerase subunit
1cz4		92–185	*Thermoplasma acidophilum*	CDC28 domain 2-like
1bol	A	1–222	*Rhizopus niveus*	Ribonuclease Rh-like
1ops		2–65	*Macrozoarces americanus*	β-clips II
1gpr		4–161	*Bacillus subtilis*	Barrel-sandwich hybrids
1du5	A	1–206	*Zea mays*	Osmotin, thaumatin-like protein
1cip	A	62–180	*Rattus norvegicus*	Transducin (α subunit), insertion domains
1bup	A	68–116	*Bos taurus*	Defensin-A like III
2fus	A	405–459	*Escherichia coli*	RNA reductase protein R1
1jud		18–93	*Pseudomonas*	L-2-haloacid dehydrogenase domains
1a3q	A	142–185	*Homo sapiens*	3-methyladenine DNA Glycosylase II, chain
1fvl		1–70	*Trimeresurus flavoviridis*	Disintegrin/echistasin
1igl		1–67	*Homo sapiens*	Insulin-like
1e09	A	1–159	*Prunus avium*	TBP-like
1a6s		1–87	Rous sarcoma virus	Retroviral matrix proteins/DNA polymerase domain 1
1eo0	A	1–77	*Saccharomyces cerevisiae*	N-cbl like
1bxy	A	1–60	*Thermus aquaticus*	α–β plaits III
1g24	A	41–251	*Clostridium botulinum*	ADP-ribosyltransferases
1tpg		1–46	*Homo sapiens*	Fibronectin type I module/Complement module, domain 1
1ew4	A	1–106	*Escherichia coli*	N domain of copper amine oxidase-like
1a5 m	A	1–100	*Klebsiella aerogenes*	Urease, γ-subunit
2ife	A	90–180	*Escherichia coli*	IF3-like
1aa3		268–330	*Escherichia coli*	Anti-LPS factor/recA domains
1qme	A	635–691	*Streptococcus pneumoniae*	Defensin A-like IV
1c3 g	A	180–258	*Saccharomyces cerevisiae*	HSP70/DnaJ peptide-binding domains
1coo		249–329	*Escherichia coli*	SAM domain-like/RNA polymerase α subunits
1g1e	B	295–383	*Mus musculus*	PAH2 domains
1kjs		1–74	*Homo sapiens*	Anaphylotoxins
1ycq	A	21–108	*Xenopus laevis*	MDM2
1h9f	A	1–57	*Homo sapiens*	LEM domain
1utg		1–70	*Orcytolagus cuniculus*	Uteroglobin-like
1f17	A	200–286	*Homo sapiens*	6-phosphogluconate dehydrogenase C-terminal domain-like
1ba3		440–544	*Photinus pyralis*	GMP synthetases
1c06	A	52–139	*Bacillus stearothermophilus*	α helix orthogonal bundles III
1eth	B	4–90	*Sus scrofa*	Lipases
1pnk	A	150–195	*Escherichia coli*	Helix hairpins II
1xo1	A	180–260	Bacteriophage t5	DNA polymerase domains
1uaa		107–180	*Escherichia coli*	α helix orthogonal bundles IV
1i50	A	72–155	*Saccharomyces cerevisiae*	RPB5-like RNA polymerase subunit
1d1r	A	29–111	*Escherichia coli*	eIF1-like
3gcc		144–206	*Arabidopsis thaliana*	DNA-binding domains
1yge		350–499	*Glycine max*	Inhibitor stefin B (Cystatin B)
1ypi	A,B	2–248, 2–248	*Saccharomyces cerevisiae*	TIM barrels
1g6l	A	1–99, 1001–1099	Human immunodeficiency virus type 1	Acid proteases
1okt	A,B	1–211, 1–211	*Plasmodium falciparum*	Thioredoxin-like; GST C-terminal
1ev4	C,D	2–222, 2–222	*Rattus norvegicus*	Thioredoxin-like; GST C-terminal
2lao		1–238	*Salmonella typhimurium*	D-maltodextrin binding proteins
1php		1–394	*Bacillus stearothermophilus*	Rossmann folds II; Rossmann folds V
1gad	O	1–330	*Escherichia coli*	Rossmann folds I; Glyceraldehyde-3-Phosphate dehydrogenase-like
1hue	A,B	1–90, 1–90	*Bacillus stearothermophilus*	IHF-like DNA-binding proteins
1cip	A	32–347	*Rattus norvegicus*	Rossmann Folds I; Transducin (α subunit), insertion domains
1f17	A	12–304	*Homo sapiens*	6-phosphogluconate dehydrogenase C-terminal domain-like

This table contains the 188 protein target domains simulated and analyzed at the time this manuscript was submitted. The table contains the PDB code of the source structure, the chain identifier (if any) and the residue range. Also included are the source organism and our assigned fold name. Protein targets are ordered by rank from most to least populated.

Step 4: analysis of molecular dynamics trajectories

Each trajectory was subjected to an extensive array of analyses. Broadly, these analyses can be summarized as tools to characterize gross structural changes, assessment of secondary structural changes, calculation of the number of contacts between protein atoms and solvent accessible surface area (SASA). The set of analyses and how they were calculated is shown in Table III. Some analyses could be broken down by side chain or main chain, polar or non-polar, native (i.e. present in the crystal structure) or non-native (i.e. not present in the crystal structure) and various combinations of these could be created, e.g. native non-polar contacts between side-chains. Such combinations bring the total number of properties to 32.

Table II. Breakdown of 188 Dynameomics targets by source

	Entire PDB	Subunit of PDB[a]	Partial chain from PDB[b]	Total
X-ray	67	24	44	135
NMR	47	1	5	53
Total	114	25	49	188

[a]A complete chain from a multi-chain (non-covalently bound) structure.
[b]The chain was cut and a partial chain corresponding to the desired fold was used.

The trajectories were compared to NMR data where the experimental data were available from the BioMagResBank (BMRB) in formats that could be directly parsed or readily converted by the DOCR tools (Doreleijers *et al*., 2003, 2005). For the purposes of comparison to NOEs, an experimental NOE was considered satisfied if the r^{-6} weighted distance of the closest pairs of protons specified by the constraint was <5 Å or the experimentally reported upper bound. Order parameters (S^2), in the form of main-chain amide bond vectors, were calculated from MD trajectories as described previously (Wong and Daggett, 1998). The analyses in Table III were performed in *il*mm with the exception of chemical shifts, which were calculated with SHIFTS (Osapay and Case, 1991).

Residue-based contact preferences were generated from the simulations and the starting structures. We define the contact preference as the negative logarithm of the ratio of the number of observed contacts and the number of expected contacts (from some reference state), multiplied by the Boltzmann constant (k) and temperature (T). A modified version of the method of Godzik *et al*. (1995) in which buried and solvent-exposed residues were considered, was used to calculate the reference state. This method yields 210 pairwise (i.e. Ala to Ala, Ala to Arg) residue contact preferences per dataset. Differences between corresponding pairs in the starting structures and the simulations were examined to identify shifts in contact preferences.

In a number of instances, analyses were compared to data derived from static structures in the PDB. Taken as its whole, the PDB is highly redundant, i.e. there are many structures of SH3 domains. To avoid problems associated with the inherent oversampling of the PDB, the ASTRAL-40 database (v1.65, Brenner *et al*., 2000; Chandonia *et al*., 2002, 2004) was used as a biased reduced set of static structures. This database is a subset of SCOP and is comprised of 5674 structures in the PDB whose sequences have less then 40% sequence identity.

Mining the Dynameomics database for helix motion

In addition to the above analyses, we used the Dynameomics database (Kehl *et al*., 2008; Simms *et al*., 2008) to search for pairs of helices in contact at the start of the simulations and then to calculate the angle between them over the course of the trajectories. We began by using a SQL query against the DSSP analysis stored in the database to identify the location of helices at least 7 residues long in the simulation starting structures. Next, also using a SQL query, we used the coordinates of heavy atoms in the previously identified helices to find pairs of helices in the starting structures that had 10 or more contacts between them (see the contact definition in Table III). Then, Mathematica (Wolfram Research, 2005) was connected to the Dynameomics database to read in the

Table III. Analytical properties calculated from Dynameomics simulations

Property	Subset	Method
Root-mean-squared deviations (RMSD)	Cα atoms from all residues	Kearsley (1989)
	Cα atoms from residues in secondary structure	Kearsley (1989)
$RMSD_{100}$	Cα RMSD normalized by that of a 100 residue protein	Carugo and Pongor (2001)
Structural dissimilarity (CONGENEAL)	Cα atoms from all residues	Yee and Dill (1993)
Solvent accessible surface area (SASA)	Non-polar	Lee and Richards (1971)
	Polar	Lee and Richards (1971)
	Main chain	Lee and Richards (1971)
	Side chain	Lee and Richards (1971)
Heavy atom contacts	Non-polar	X–$X < 5.4$ Å, where X is not charged
	Polar	Y–$Y < 4.6$ Å, where Y is charged
	Other	X–$Y < 4.6$ Å, where X is not charged and Y is charged
	Main chain	Z–Z, where Z is N,CA,C,O
	Side chain	Z–Z, where Z is not N,CA,C,O
	Native	Contacts present in the PDB structure
	Non-native	Contacts not present in the PDB structure
Secondary structure by ϕ/ψ angles	Helix	$-100° < \phi < -30°$, $-80° < \psi < -5°$ (Daggett *et al*., 1991)
	Beta-strand	$-170° < \phi < -50°$, $80° < \psi < -170°$ (Daggett *et al*., 1991)
	Extended	(ϕ, ψ) within 20° of linear
	Other	not helix, beta or extended
Secondary structure by Definition of Secondary Structure of Proteins (DSSP)	Helix	Kabsch and Sander (1983)
	Beta strand	Kabsch and Sander (1983)
	Other/loop	Kabsch and Sander (1983)
	Alpha strand	Alternating: $-122° < \phi < -52°$, $-84° < \psi < -14°$ and $10° < \phi < 80°$, $57° < \psi < 127°$ (Scouras and Daggett, 2008)
Nuclear magnetic resonance (NMR)	Nuclear Overhauser effects (NOE)	Satisfied if the r^{-6} weighted mean distance was less than the experimental upper-bound or 5.0 Å, whichever was larger
	Chemical shifts	Osapay and Case (1991)

list of pairs of helices in contact. Using Mathematica, we computed the angle between the first principal components calculated from the coordinates of each helix's Cα atoms over the course of the trajectory and wrote these data back to the database for subsequent statistical analysis. Pairs of helices with significant fluctuations in this angle were identified by finding those angles with large standard deviations.

Results

The combined simulation time for the native-state simulations of the 188 targets presented here was 6.5 μs. Due to how simulations were packaged to run on the National Energy Resource Supercomputing Center (NERSC) supercomputer, not all simulations were conducted for precisely the same amount of time. However, every simulation was at least 21 ns long with an average length of 29.4 ± 7.1 ns. The resulting coordinate data for the trajectories required 0.5 terabyte (TB) in a binary format of our own design, at approximately 50% compression. The analyses in Table III were applied to the trajectories resulting in an additional 0.33 TB of data.

Computational assessments of simulation quality

As a check on the quality of the simulations, we can monitor the drift in total energy (sum of potential and kinetic at each simulation timestep). This assessment is unique to the microcanonical (NVE) ensemble, as conservation of energy is inherent in the classical equations of motion. However, the numerical integration of these equations with the limited precision inherent to all computer algorithms results in some amount of energy drift. We evaluate this drift in terms of the percentage of total energy change per ns. This value is quite low for all simulations, and ranged from 0.13 to 0.20% per ns. The energy change per ns scaled linearly with the number of atoms in the system, as did the total energy of the system. As temperature is not a fixed quantity in this ensemble, it can fluctuate. Over the last 10 ns of our simulations, the mean temperature was 298 ± 4.6 K.

Assessment of simulation quality by comparison with experiment

MD simulations can be validated by comparison with NMR observables (Table IV) such as Nuclear Overhauser effect crosspeaks (NOEs), which report on hydrogen that is close in space. NOE sets were readily available in a parsable form from BMRB (Doreleijers *et al*., 2003) for 27 of our 188 proteins. The mean percentage of NOEs satisfied across the 27 targets was 92 ± 4% (Table IV). In our experience, more extensive sampling improves the level of agreement. In at least two cases (engrailed homeodomain, rank 6 and ubiquitin, rank 9) where a crystal structure was used and NMR NOE data were available, we found that the crystal structure satisfied fewer NOEs than did the simulation. These cases highlight how MD can improve our understanding of protein behavior by ensemble sampling in solution rather than using a single static structure.

Proton chemical shifts from NMR were also found to be in good agreement with experiment for the 15 proteins for which data were available in a readily processed format from the BMRB. Such chemical shifts reflect the local environment around the atom in question. The correlation coefficient between the predicted proton chemical shifts from the simulations and experiment was 0.96 (Table IV).

Table IV. Comparison of simulations to available NMR observables

	Level of agreement with experiment	Available BMRB data
NOE[a]	92 ± 4%	27 proteins (28 504 NOEs)
Chemical shifts[b]	$R = 0.96$	15 proteins (5778)

[a]Level of agreement is mean per-protein satisfaction. The 27 proteins with available data (by PDB code) are: 1aa3, 1c06, 1d1r, 1gle, 1kjs, 2ife, 3gcc, 1bf0, 1cmz, 1cok, 1cz4, 1d1n, 1d8v, 1enh, 1fad, 1fvl, 1fzt, 1ght, 1i11, 1iyu, 11dl, 1mut, 1sso, 1tfb, 1ubq, 1uxc, 3chy.
[b]Proton chemical shifts from MD structures were calculated with SHIFTS (Osapay and Case, 1991). The 15 proteins with data available (by PDB code): 1mjc, 1hcc, 1ubq, 1baz, 1cz4, 1a2p, 1e65, 1ill, 3chy, 1ght, 1cmz, 1gpr, 1byl, 1fzt, 1b10.

Case study comparison with experiment: ubiquitin

To illustrate the comparison between simulation and different experimental results, we focus on ubiquitin. For ubiquitin [PDB:1ubq], the crystal structure satisfied 94.4% of 2727 NOEs. The 21 ns simulation satisfied 95.2% of the NOEs, showing a net gain in satisfaction of 19 long-range NOEs and a decrease in the mean violation distance from 0.84 to 0.65 Å relative to the crystal structure. While 70 long-range NOEs ($i \leftrightarrow i+5$ or greater) were gained, 58 were lost by the sampling of the simulation. Of the 58 lost NOEs, 41 were violated by <0.5 Å, 11 were violated by <1 Å, 5 were violated by <2 Å and 1 was violated by >2 Å. The experimental proton chemical shifts from experiment were compared with the MD-predicted values, and the correlation coefficient between the two sets of values was 0.98.

Comparison with B-factors is also of interest because they reflect the extent of mobility of the protein in its crystalline state. B-factors contain contributions from internal motion (modeled as root mean square fluctuations about the mean) as well as lattice disorder and phasing. Thus, the comparison with mobility in solution can be problematic. Nevertheless, the B-factors (Vijay-Kumar *et al*., 1987) and Cα RMSF values (Figs 2A and B) both indicate that the tail residues are dynamic and the correlation between the two sets is acceptable ($R = 0.73$), particularly given the different environments. Ubiquitin is attached through its C-terminal tail to proteins to tag them for degradation. The Cα RMSD calculated for all residues and for all residues excluding the tails (residues 2 to 71) were compared (Fig. 2C); the tails contribute an average of ~0.4 Å Cα RMSD throughout the simulation. The S^2 values for NH bond motion from NMR relaxation experiments of ubiquitin (Schneider *et al*., 1992) were also compared to S^2_{MD} values calculated from the simulations (Wong and Daggett, 1998) (Fig. 2D), and the correlation is acceptable ($R = 0.76$).

Summary statistics from the Dynameomics simulations

Summary statistics based on the set of properties calculated for all of our trajectories (see Table III) are included in Tables V and VI. In order to aggregate these statistics in a consistent manner from a set of proteins that is designed to be varied in structure and chain length, we have used two normalization approaches. The first was to normalize the

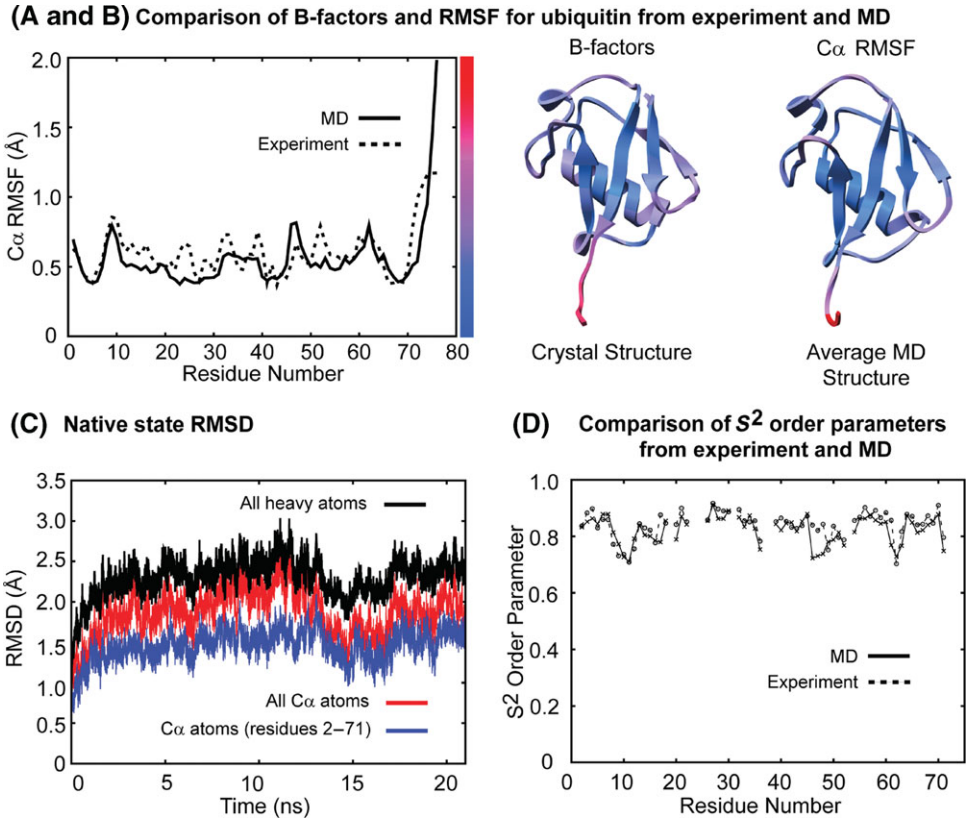

Fig. 2. Comparison of experimental observables to the native-state simulation of ubiquitin. (**A**) Comparison of Cα RMSF and B-factors. B-factors were converted to an RMSF value by RMSF $= \sqrt{3B/8\pi^2}$ where B =B-factor of the Cα atom (correlation, $R = 0.73$). (**B**) Ubiquitin crystal structure colored by experimental B-factors (left) and the average native structure from simulation colored by Cα RMSF values (right). A scale colored from blue to red, see right hand scale bar in (A), representing low to high values respectively is used. (**C**) Native-state RMSD calculated using all heavy atoms, black line; all Cα atoms, red line; Cα atoms excluding tail residues (residues 2–71), blue line. The tail residues are unstructured in the simulation. (**D**) S^2 order parameters from experiment compared to calculated S^2_{MD} values from the native-state simulation of ubiquitin (correlation, $R = 0.76$). MD values are depicted with a solid line and crosses, experimental values are depicted with a dashed line and circles.

Table V. Summary statistics of average properties from simulations (by structural class)

Property	All classes	All α	All β	α+β	α/β
Average properties derived from simulations (average ± Std Dev.)					
Cα RMSD (Å)	2.94 ± 1.12	2.92 ± 1.37	2.83 ± 0.89	2.82 ± 1.04	3.16 ± 1.11
Cα $RMSD_{100}$ (Å)	3.05 ± 1.44	3.16 ± 1.74	2.69 ± 0.82	2.29 ± 1.01	2.83 ± 1.05
Total SASA per residue (Å^2)	64.55 ± 9.15	67.49 ± 9.20	60.27 ± 6.66	56.09 ± 4.85	60.45 ± 6.47
Radius of gyration (Å)	13.77 ± 2.60	13.77 ± 2.95	13.76 ± 2.21	15.78 ± 2.20	14.76 ± 2.27
Fraction of residues in α-helix	0.40 ± 0.20	0.64 ± 0.15	0.17 ± 0.10	0.41 ± 0.09	0.36 ± 0.11
Fraction of residues in β-sheet	0.37 ± 0.16	0.17 ± 0.11	0.55 ± 0.10	0.37 ± 0.09	0.40 ± 0.09
Fractional change in properties relative to thermal equilibration period (0–1 ns) (Average ± Std Dev.)					
Cα RMSD	1.59 ± 0.34	1.46 ± 0.42	1.61 ± 0.35	1.61 ± 0.40	1.62 ± 0.34
Cα $RMSD_{100}$	1.59 ± 0.34	1.46 ± 0.42	1.61 ± 0.35	1.61 ± 0.40	1.62 ± 0.34
Total SASA	1.00 ± 0.04	0.98 ± 0.06	1.02 ± 0.06	1.01 ± 0.06	1.00 ± 0.06
Radius of gyration	1.00 ± 0.03	0.99 ± 0.04	1.00 ± 0.04	1.00 ± 0.05	1.00 ± 0.04
No. of residues in α-helix	1.06 ± 0.18	1.01 ± 0.08	1.16 ± 0.22	1.01 ± 0.10	1.03 ± 0.08
No. of residues in β-sheet	0.94 ± 0.08	0.94 ± 0.12	0.93 ± 0.08	0.96 ± 0.06	0.94 ± 0.06

individual properties by the number of residues in the target of origin, before aggregation. For properties like radius of gyration, Cα RMSD and fraction of residues in a given secondary structure type, this normalization is inherent to the underlying analysis. In other cases, the final units are per residue, e.g. SASA has units of Å^2/residue. The second approach was to normalize by the mean value derived from each simulation's equilibration period (0–1 ns). This essentially converts the quantities into fractional changes in the simulation relative to the equilibration values. Table V breaks down these statistics into protein structural classes based on the targets' folds, i.e. all α, all β, α/β, α+β. Table VI breaks these statistics down into the experimental source for each target, i.e. X-ray crystallography or NMR. Interestingly, comparison of the values reflecting fractional changes in various protein properties over the simulations

Table VI. Summary statistics of average properties (Average ± Std Dev.) from simulations by experimental source of starting structure

Property	Average properties derived from simulations		Fractional change relative to equilibration period (0–1 ns)	
	NMR	X-ray	NMR	X-ray
Cα RMSD	3.45 ± 1.21 Å	2.75 ± 1.04 Å	1.52 ± 0.38	1.62 ± 0.34
Cα $RMSD_{100}$	4.00 ± 1.55 Å	2.69 ± 1.29 Å	1.52 ± 0.38	1.62 ± 0.34
Total SASA	70.44 ± 8.48 $Å^2$/residue	62.27 ± 8.43 $Å^2$/residue	0.97 ± 0.05	1.01 ± 0.04
Radius of gyration	12.47 ± 1.96 Å	14.25 ± 2.67 Å	0.99 ± 0.04	1.00 ± 0.03
Fraction of residues in α-helix	0.41 ± 0.22	0.39 ± 0.20	1.12 ± 0.29	1.04 ± 0.12
Fraction of residues in β-sheet	0.33 ± 0.17	0.38 ± 0.16	0.98 ± 0.11	0.99 ± 0.07

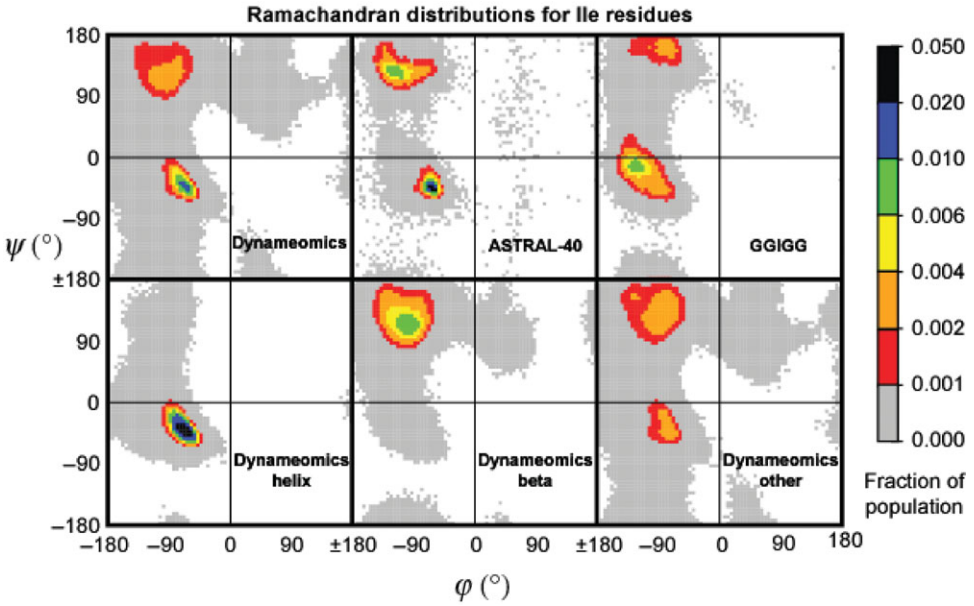

Fig. 3. Ramachandran distributions for Ile residues from Dynameomics simulations, ASTRAL-40 and simulations of GGIGG. (ϕ, ψ) range (−180° to +180°) is divided into 72 bins of 5° widths. Each bin is colored by fractional population (figure left) on a non-linear scale. The top three distributions, from left to right, contain of samples from (i) all time samples Ile residues in the Dynameomics simulation dataset, (ii) all Ile in the ASTRAL-40 dataset (from static PDBs) and (iii) all time samples from simulations of the (acetylated and amidated) penta-peptide Gly-Gly-Ile-Gly-Gly (GGIGG). In the bottom row, again from left to right, are the distributions from: (i) Ile residues in the Dynameomics simulation dataset that began their respective simulations in at least three residues of α-helix (by DSSP), (ii) as (i) but with Ile residues that began in at least three residues of β-sheet (by DSSP) and (iii) all other Ile not included in (i) and (ii).

indicates that the structural properties such as solvent accessibility, secondary structure and radius of gyration change very little even though the C_α RMSD may increase.

Ramachandran distributions

Ramachandran distributions for each residue type were calculated and as an example the data for Ile are shown in Fig. 3. The distributions from Dynameomics are calculated with the (ϕ, ψ) angles for Ile residues at every saved trajectory time point, post-equilibration, resulting in at least 20 000 points for each Ile residue. In addition to the distribution for all Ile residues in the Dynameomics dataset, we depict distributions of selected Ile residues that started the simulations in helical, beta and other (i.e. turns, extended, 'other' or unstructured) conformations. For the purposes of these secondary structure assignments, we required at least three consecutive residues of α-helix or β-sheet to be present. We found it useful to examine the Dynameomics data in the context of Ramachandran distributions from two other sources: those derived from the static structures of proteins in the ASTRAL-40 dataset (Brenner *et al.*, 2000; Chandonia *et al.*, 2002, 2004) and for comparison to a sterically unrestrained Ile distribution: data from our simulations at 298 K of the end-capped (acetylated and amidated) pentapeptide of sequence Gly-Gly-Ile-Gly-Gly (GGIGG) (Beck *et al.*, 2008).

There are 1342 Ile residues in the 188 targets, of which 382 are in helices, 303 in beta sheets and 657 in non-α, non-β classifications in their starting structures. We chose Ile for depiction in Fig. 3 because it is fairly evenly distributed between helices and β-sheets. In contrast, Ala residues in secondary structure elements are predominantly in helical conformations in the starting structures: of 1793 residues in the target set, the distribution was 650 and 187 for helices and beta sheets, respectively. Asn residues had the most unbalanced distribution of all the residues: of the 1067 Asn residues, 183 were in helices while only 55 were in β-structures.

Residue-based contact preferences

Pairwise residue contact preferences differed between the starting structures and the simulations (Fig. 4). In our simulations, disulfide bonds are modeled using covalently bound Cys, while free Cys residues were modeled with a complete

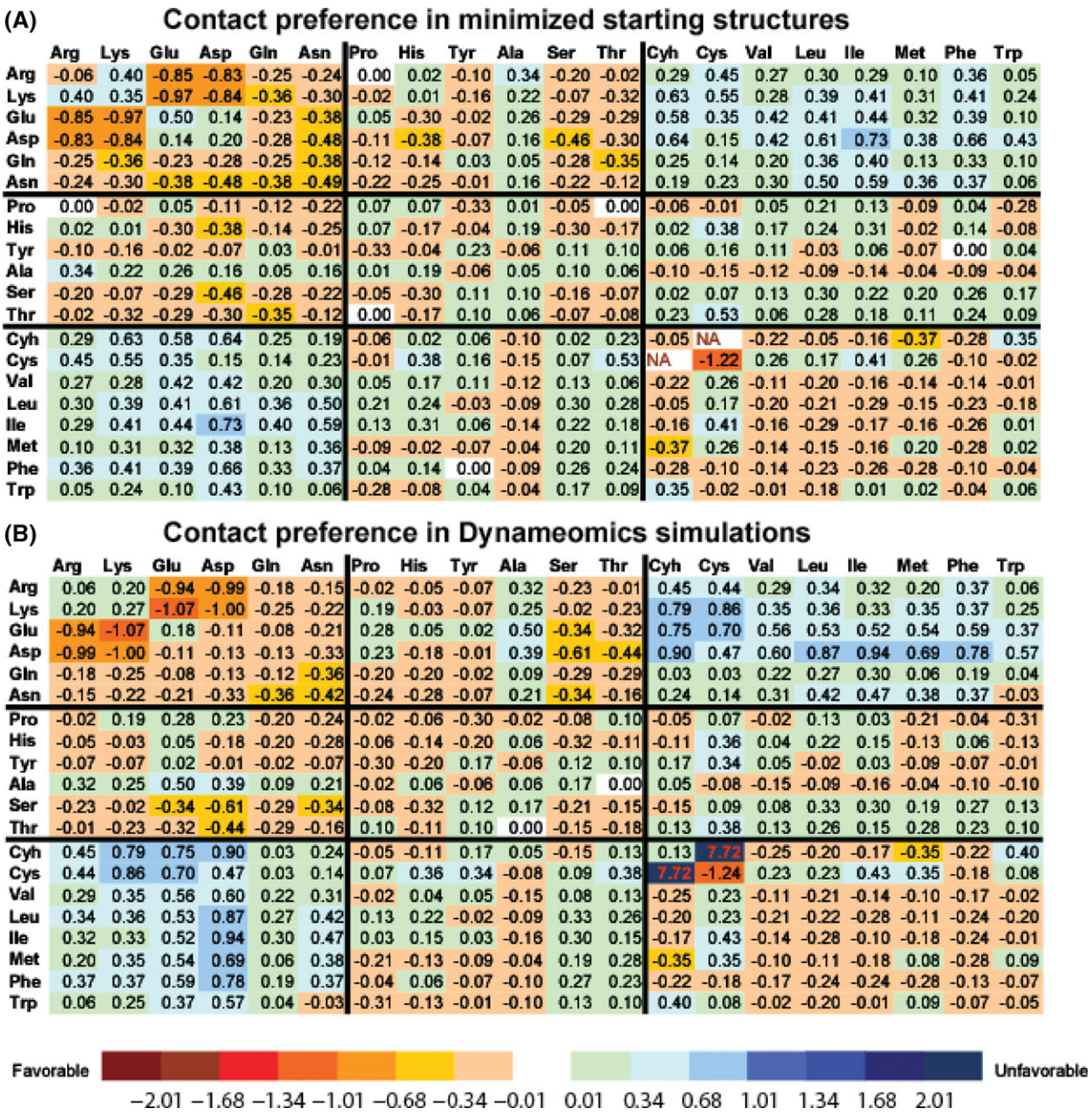

Fig. 4. Pairwise contact preferences for dynamic and static structures. Contact preferences observed between residue pairs for (**A**) minimized starting structures and (**B**) simulations. Preferences were calculated as described in the methods and are in units of *k*T. Favorable contacts have a ratio of observed contacts to expected contacts that exceeds unity. No Gly preferences can be observed given the side chain heavy atom contact criteria. No Cys-Cyh contacts were observed in the starting structures because all Cys in contact at the start of a simulation are in disulfide bonds, whereas Cyh pairs are necessarily distant. The charge–charge interactions and interactions between polar and charged residues become more favorable during the simulations.

thiol group (Cyh). For obvious reasons, the most favorable contact in both sets was between disulfide-bonded Cys, -1.22 *k*T in the starting structures and -1.24 *k*T in the simulations. The second most favorable group of interactions involved charged residues (Glu, Asp, Lys, Arg). Contacts between oppositely charged residues were favorable in both datasets, averaging -0.87 ± 0.05 *k*T in the starting structures and -1.00 ± 0.04 *k*T in the simulations. Interestingly, contacts between like-charged residues became more favorable as well, shifting -0.22 ± 0.09 *k*T from the starting structures to the simulations, due to some relief of the repulsion from the slight expansion of the protein in response to the solvent environment and kinetic energy. Overall, there was an average favorable shift of -0.16 ± 0.12 *k*T in all types of charged residue contacts. In contrast, the hydrophobic residues exhibited only a small average favorable shift of -0.04 ± 0.05 *k*T.

Relative motions of helices in the native state

By mining our Dynameomics database, we identified 390 pairs of helices in contact in the starting structures for the 188 targets. Of these, 17 helix pairs had angular distributions with large standard deviations (greater than 10°) over time. One example is the change in orientation of helices α3 and α4 of Chemotaxis protein Y (CheY, [PDB:3chy]).

CheY is a 128 residue protein with a three-layer α/β/α Rossmann fold of rank 2 (Fig. 1, Step 2). While the 36 ns native-state simulation of CheY maintained the overall fold of

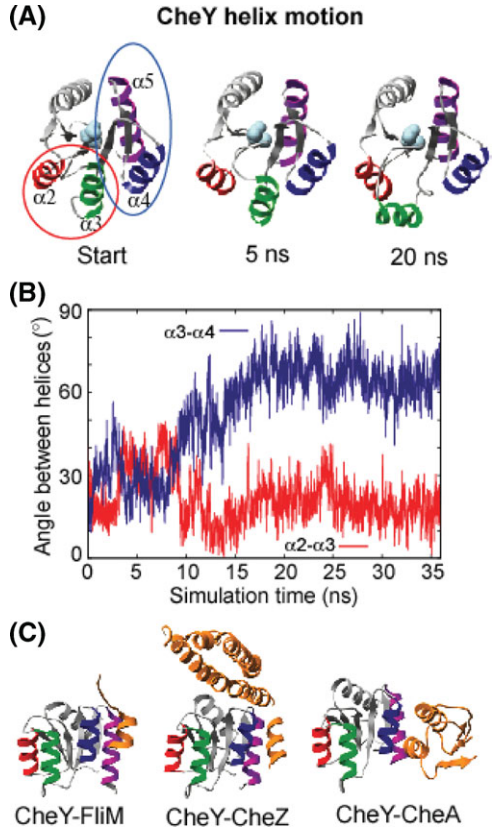

Fig. 5. Dynamics of the Chemotaxis Protein Y. (**A**) Snapshots of the starting structure, 5 and 20 ns timepoints of CheY showing the movement of helices α2 and α3 circled in red. There is change in orientation of helix α2. Helix α3 pulls away from the core of the protein. The protein is displayed in ribbon looking down the plane of the β-sheet, α2 is colored red, α3 is colored green, α4 is colored blue and α5 is colored purple. Helices α4 and α5 appear stable throughout the simulation and are circled in blue. The phosphorylation site, Asp 57, is shown colored cyan. All other protein and solvent atoms are removed for clarity. (**B**) The angle between helices α2 and α3 changes over time and both helices change orientation. The angle between helices α3 and α4 increases over time as helix α3 moves away from the core of the protein. Chemotaxis binding partners of CheY. CheY is displayed in ribbon with helices α2, α3, α4 and α5 colored red, green, blue and purple, respectively. Binding partners FliM ([PDB: 1f4v]), CheA ([PDB: 1eay]) and CheZ ([PDB: 1kmi]) are shown in orange. The stable face of α4-β5-α5 is an important binding interface for all its binding partners.

the protein, there was an increase of approximately 35° in the angle between α3 and α4 due to the movement of α3 (Fig. 5A and B). α3 changed orientation to move 5 Å away from Asp 57 (a critical phosphorylation site in CheY), extending the loop between β3 and α3 (Fig. 5A). The α2 and α3 helix pair also had heightened dynamics throughout the simulation with an increase in angle of approximately 20°. In contrast, α4 and α5 were stable throughout the simulation. This latter pair does not meet our 10 contacts criterion as described in the methods, so we measure structural stability in terms of distance between the ends of the helices; the standard deviations for these distances are ± 1.15 and ± 1.03 Å.

Discussion

Assessments of simulation quality reveal simulations to be stable

The extent of energy drift in our Dynameomics simulations is quite minimal and indicates that the simulation engine is working correctly and that the simulations are computationally correct. The drift is a result of numerical round-off error inherent to any numerical integration by limited precision computing, even using 64-bit precision. To the best of our knowledge, these are among the most stable, most rigorous, protein simulations available. A more complete discussion of the origin of energy drift can be found elsewhere (Beck and Daggett, 2004).

Given ergodic sampling, each simulation should reproduce the available experimental observables such as those from NMR. However, in our experience, most proteins require many tens to hundreds of ns to sample a sufficiently broad set of conformations to satisfy all the NOEs (Beck and Daggett, 2007). We attribute many of the unsatisfied NOEs from our simulations to this limitation. Another factor was that some target domains were abstracted from larger protein entities. It is common for these domains to change structure, sometimes radically, in the absence of their binding partners or the rest of their chain. In addition, as we illustrate with ubiquitin, the simulations are in good agreement with NOEs and chemical shifts from NMR, S^2 order parameters from NMR relaxation experiments and crystallographic B-factors.

Summary statistics depict stable proteins across structural classes and experimental sources

In the selected summary statistics, presented in Tables V and VI, there were no significant differences in overall protein properties between any of the structural subsets. This was affirmed when the complete set of summary statistics from all of the analytical properties were examined (data not shown). Our interpretation of these data is that the general protein structural classes (i.e. all α, all β, α/β, α+β) are free from any systematic bias as a result of the potential function or the simulation engine (such as overestimation of helical propensity). Further, NMR structures were indistinguishable from X-ray crystal structures for the purposes of simulation.

Several properties calculated from the trajectories, such as the count of intramolecular hydrogen bonds and main-chain to main-chain contacts, showed a decrease relative to the values calculated from the minimized starting structure (data not shown). Similarly, other properties, such as SASA, RMSD and radius of gyration exhibited an increase relative to the values from the minimized starting structures (Tables V and VI). The change is an expected consequence of increased thermal energy in the system and move to an explicit solvent environment.

An extreme example of this kind of change in compactness can be found in the excised DNA-binding domain HU [PDB:1huu] (Fig. 6A). This system is a dimer in its biological form (Fig. 6B). We initially simulated it as a monomer of 80 residues. At the start of that simulation, the hydrophobic dimer interface was exposed. The interface was buried in the first 10 ns by contortion of the domain's second helix (Fig. 6A). This burial resulted in a drop in the total SASA of approximately 13%, a collapse in the monomer's Cα radius of gyration of nearly 20% and a large Cα RMSD to the crystal structure of 7.5 ± 1.2 Å. In subsequent simulations of the dimer [PDB:1hue] (Fig. 6B), the equivalent monomer remained stable with a lower Cα RMSD to crystal (3.2 ± 0.4 Å).

Where we observed a greater than two standard deviations of change in SASA or radius of gyration, it was exclusively

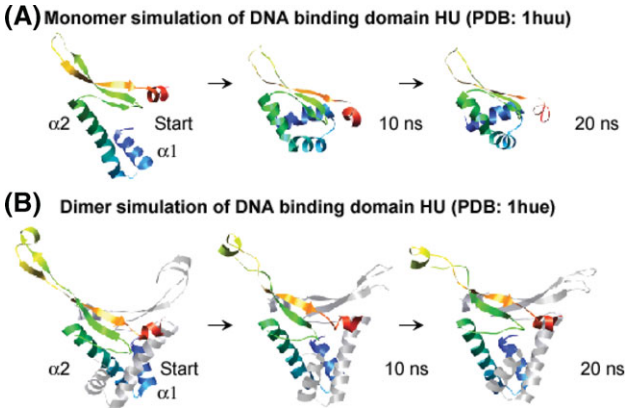

Fig. 6. Excised monomer and complete dimer simulation snapshots of DNA-binding domain HU. Comparisons of snapshots from two simulations of the target DNA-binding domain HU at simulation start, 10 ns and 20 ns. The proteins are shown in ribbon, colored from blue to red. Protein and solvent atoms included in the simulation are removed for clarity. (**A**) The monomer was excised from the dimer (**B**) and simulated by itself. During the MD simulation, the excised monomer rapidly (by 10 ns) reorganized itself, by breaking the second helix (α2) roughly in the middle allowing the first helix (α1) to pack against and shield the exposed hydrophobic dimeric interface. (**B**) In contrast, the simulation of the dimer was stable with no radical rearrangement in packing.

the case that these targets had been excised from larger proteins. For these cases, our initial approach to excising target domains may not have been appropriate for all folds, particularly where it resulted in cutting peptide bonds and removing a covalently bound sub-chain from a larger protein or protein assembly. In one case, we identified 2 folds, glutathione S-tranferase's N-terminal domain (fold rank 12) and its C-terminal domain (fold rank 27), that were amenable to simulation together as a single monomer with one continuous chain. As a proactive measure to avoid the problems related to excision of these domains, we simulated both of the target domains as a complete monomer (and subsequently the biological dimer) [PDB:1ev4]. We are now simulating all excised domains alone and in their biological units and eventually, their complete assemblies. The latter approach allows us to simulate multiple folds per target, thereby enhancing our understanding of the interplay between domains.

Comparison of Ramachandran maps from dynamic and static sources

For most residues, the location of the α-helical and β-strand peaks in the Ramachandran distribution may differ slightly depending on the origin of the data. This difference is illustrated by the Ramachandran maps for Ile residues given in Fig. 3. Relative to the ASTRAL-40 data from static structures, the Dynameomics data at 298 K are slightly more dispersed; however, there is little change in the location of the peaks. The difference in β-strand population between ASTRAL-40 and Dynameomics is primarily a result of the wider peak region in the dynamic dataset, not a significant change away from β-strand, as can be seen in the data from β-strand only residues in the Dynameomics dataset at 298 K. Similarly, from the Ramachandran maps of Ile residues starting in helical space, we see little change and a strong peak at the maximum observed in the ASTRAL-40 dataset. In contrast, the distribution for the Ile in the GGIGG peptide simulation shows marked shifts in the maxima for α-helix and β-strand, in addition to an overall more dispersed distribution when compared to those from Dynameomics and ASTRAL-40. In the well-hydrated sterically unhindered environment of GGIGG, the balance of energetic contributions for Ile is very different than when the residue resides in the context of folded proteins. Similar results are seen for the other 19 naturally occurring amino acids (Beck *et al.*, 2008).

A shift in contact preferences in the context of stable structures

We observed several general trends in contact preferences when comparing the simulations and the starting structures. First, the increased association of oppositely charged residues is indicative of the formation of salt bridges. Second, the increase in the favorability of like-charged interactions is due to formation of multi-body salt bridge networks, the presence of water-mediated hydrogen bond networks and the expansion of the structures compared with the static starting structures. Third, we found no significant change in the preferences for hydrophobic residues, suggesting that even in dynamic ensembles, the strength of the hydrophobic effect is similar to the effect in experimentally derived structures. The residue interactions derived here from a broad sample of dynamic native states provide a more complete view of the intermolecular interactions in proteins in solution and how they differ from those in static, averaged structures from experimental sources.

Native-state dynamics and implications for protein function

The movement of secondary structural elements can be important for protein function. For example, the breathing motion of the helices in myoglobin allows oxygen to diffuse through the protein to bind heme (Brunori, 2000). Movement of helices is also important in ion channel opening and closing (Spencer and Rees, 2002). Then, the fundamental questions are whether such motions are apparent in other proteins and whether they are important to function. The Dynameomics database [accompanying paper, Simms *et al.*, 2008)] provides us with a framework to investigate this question through the measurement of the angles of pairs of helices. In mining the database, we found that unphosphorylated CheY showed heightened dynamics of α2 and α3 causing them to move away from the rest of the protein, opening a cleft between helices α3 and α4 (Fig. 5). The movement of α3 leads to increased exposure of Asp 57, which oscillates with the helical motion. In contrast, α4 and α5 are stable over the course of the simulation.

To better understand the possible biological consequences of the heightened dynamics of α2 and α3 and the stability of α4 and α5 on its function in *E. coli*, we consider the chemotaxis cascade. CheY interacts with three proteins: chemotaxis protein A (CheA), chemotaxis protein Z (CheZ) and flagella motor switch protein M (FliM) (Fig. 5C) (McEvoy *et al.*, 1998; Lee *et al.*, 2001; Zhao *et al.*, 2002). In the absence of attractants, CheA is autophosphorylated and transfers the phosphoryl group to CheY at Asp 57. Phosphorylated CheY dissociates from CheA and binds FliM, a component of the flagella motor complex, causing the flagella to rotate clockwise and the cell to tumble randomly. Increasing amounts of attractants inhibit CheA's autophosphorylation, decreasing the cellular concentration of phosphorylated CheY. The

CheY signal is terminated by the phosphatase CheZ. Unphosphorylated CheY does not bind the flagella complex, allowing the flagella to rotate counterclockwise and the cell to swim in a smooth, directed fashion towards increasing concentrations of attractants (Wadhams and Armitage, 2004).

Structures of CheY in complex with CheA, CheZ and FliM have revealed a common binding site on the α4-β5-α5 face of CheY (McEvoy *et al.*, 1998; Lee *et al.*, 2001; Zhao *et al.*, 2002). The movements of α2 and α3 do not disrupt this binding site and the distance between α4 and α5 remains stable over the course of the simulation. We propose that the movements of α2 and α3 may serve as an entropy sink to help stabilize the protein and compensate for the relatively immobile α4 and α5, which serve as a scaffold for protein–protein interactions. Furthermore, the motion of α3 leads to heightened, but transient, exposure of the phosphorylation site: Asp 57.

Interestingly, previous MD simulations of another member of this fold, catechol *O*-methlytransferase (COMT), have shown movement of helices α6, α7 and α8 of COMT (Rutherford *et al.*, 2006), which correspond to α2, α3 and α4, respectively, of CheY (Fig. 7A), which suggests that motion in this region may be a conserved feature of this fold. The soluble form of COMT has 221 residues with a central β-sheet surrounded by 8 α-helices (Fig. 7A). At position 108, the human COMT sequence has a common single nucleotide polymorphism substituting Val (108V) for Met (108M). This polymorphism is associated with a decrease in enzymatic activity (Spielman and Weinshilboum, 1981; Boudikova *et al.*, 1990; Chen *et al.*, 2004) and an increased risk for diseases such as breast cancer (Goodman *et al.*, 2002) and obsessive-compulsive disorder (Karayiorgou *et al.*, 1997). The differences in activities between 108M and 108V COMT have been attributed to the lower thermostability of 108M COMT (Lotta *et al.*, 1995). However, the addition of *S*-adenosylmethionine (SAM) rescues its activity. The polymorphic site is located in a loop between α5 and β3, approximately 16 Å from the active site.

MD simulations of human COMT models revealed that there is heightened motion of α6 and α7 of COMT (corresponding to α2 and α3 in CheY) and that this motion becomes dramatic in the case of the 108M polymorph (Rutherford *et al.*, 2006), leading to a chasm between the helices and disruption of the active site (Fig. 7) similar to what occurs with α3 and α4 of CheY (Fig. 5B). The simulations also showed that residue 108 had a greater variability of interactions within the polymorphic site and was more solvent exposed in 108M than in 108V COMT. The 108M COMT simulations sampled a broader distribution of structures with a larger average Cα-RMSD to the starting structure than 108V COMT simulations at physiological temperature (Rutherford *et al.*, 2006). Changes in the polymorphic site were propagated across the protein fold to affect the co-substrate SAM binding site. Movement of helix α6 in 108M COMT had a larger effect on the solvent accessibility of the SAM binding site than in 108V (Fig. 7B). Helix α6 moved away from the core of the protein in 108M simulations, increasing the angle between helices α6 and α7 by approximately 30° (Fig. 7C). The angle between helices α7 and α8 also changed by approximately 20° (Fig. 7C). Interestingly, the green helix, as depicted in Fig. 7A, went to the right in COMT and left in CheY. We hypothesize that a larger

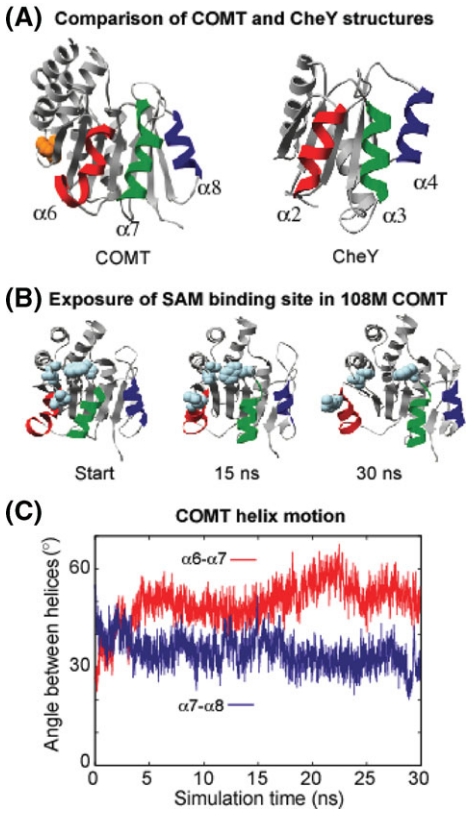

Fig. 7. Exposure of the SAM binding site in 108M COMT. (**A**) Comparison of the simulation starting structures for COMT and CheY. The proteins are shown in side-view orientation in ribbon. Helices α6, α7 and α8 of COMT (left) are colored red, green and blue, respectively. The corresponding helices of CheY (right) are colored red (α2), green (α3) and blue (α4). Residue 108 of COMT where the V/M polymorphism occurs is shown in orange with space-filling atoms. The phosphorylation site of CheY, Asp 57, is shown in space-filling magenta. (**B**) Snapshots from the start, 15 ns and 30 ns of a representative 310 K simulation of 108M COMT. Structures show movement of helices α6 to α8 and exposure of the SAM binding site. The protein is displayed in ribbon with the helices α6, α7 and α8 colored red, green and blue, respectively. Residues Glu 90, Gln 120 and Trp 143 of the SAM binding site are displayed and colored cyan. Other protein and solvent atoms included in the simulation are removed for clarity. (**C**) The angle between helices α6 and α7 (red) and helices α7 and α8 (blue) change over time. Helices α6 and α7 both change orientation with helix α6 moving away from the core of the protein.

proportion of 108M COMT exists in inactive conformations at physiological temperature (Rutherford *et al.*, 2006).

MD simulations were also performed in the presence of SAM for both 108V and 108M COMT, which decreased the Cα RMSF about the mean structure for α3 to α6. The presence of SAM in the simulations minimized the helix motions and stabilized the SAM binding site, thereby increasing the population of the catalytically productive conformer. Therefore, MD reveals conserved dynamical behavior within a fold family, reinforcing the proposal that a single representative of a fold can describe the dominant, general features of the dynamics of other members of the family, although this may not always be the case. In CheY, the dynamic behavior appears to increase the entropy and offset the cost of maintaining a scaffold for binding, as well as to facilitate or regulate phosphorylation. However, this is a fine balance and the same motion in COMT, in concert with a mutation, leads to dramatic effects on stability and activity. Knowledge of

such deleterious dynamics may be useful for rescuing proteins, as illustrated through the protective effect of SAM on COMT.

Conclusions

The Dynameomics project has generated the largest collection of protein MD simulations to date. We are continuing to expand the dataset by simulation of new fold space representatives in their native state and along their unfolding pathways. The top 30 targets of the Dynameomics database are available via our web site: www.dynameomics.org. The database can be mined for various features of native-state dynamics, such as we have demonstrated in this work by our identification of helix motions. The results from the data mining show how heightened dynamics relevant to function and stability in one protein (CheY) are implicated in disease-related mutations for a fold-space neighbor: COMT. We expect such patterns of knowledge-based discovery to continue as more data mining is conducted on this database.

Acknowledgements

Most simulations in this study were performed using Department of Energy grants of processor time at the National Energy Research Supercomputing Center (NERSC), initially through an Innovative and Novel Computational Impact on Theory and Experiment (INCITE) award and later through the Office of Biological and Environmental Research (OBER); we are grateful for the support of Dr Marvin Stodolsky. Protein structures in Figs 1, 2, 5, 6 and 7 were rendered with UCSF Chimera (Pettersen *et al*., 2004). The PDBSUM website (Laskowski *et al*., 2005) was used at various stages of this work.

Funding

In the early phase of this project, seed financial support was provided by the Biomedical and Health Informatics Program at the University of Washington and the eScience program of Microsoft Research. We are also grateful for the financial support provided by Technical Computing @ Microsoft; in particular, we thank Dr Tony Hey and Dan Fay. R.D.S. and A.L.J. were supported by the NIH Molecular Biophysics Training Grant (National Research Service Award 5 T32 GM 08 268-17) at the beginning of this study.

References

Allen,M.P. and Tildesley,D.J. (1987) *Computer Simulation of Liquids.* Oxford University Press, Oxford.
Beck,D.A.C. and Daggett,V. (2004) *Methods*, **34**, 112–120.
Beck,D.A.C. and Daggett,V. (2007) *Biophys. J.*, **93**, 3382–3391.
Beck,D.A.C., Alonso,D.O.V. and Daggett,V. (2000–2008) *in lucem molecular mechanics*. University of Washington, Seattle, WA.
Beck,D.A.C., Alonso,D.O. and Daggett,V. (2003) *Biophys. Chem.*, **100**, 221–237.
Beck,D.A.C., Armen,R. and Daggett,V. (2005) *Biochemistry*, **44**, 609–616.
Beck,D.A.C., Alonso,D.O.V., Inoyama,D. and Daggett,V. (2008), submitted for publication.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Boudikova,B., Szumlanski,C., Maidak,B. and Weinshilboum,R. (1990) *Clin. Pharmacol. Ther.*, **48**, 381–389.
Brenner,S.E., Koehl,P. and Levitt,M. (2000) *Nucleic Acids Res.*, **28**, 254–256.
Brunori,M. (2000) *Biophys. Chem.*, **86**, 221–230.
Carlson,H.A. (2002) *Curr. Opin. Chem. Biol.*, **6**, 447–452.
Carugo,O. and Pongor,S. (2001) *Protein Sci.*, **10**, 1470–1473.
Chandonia,J.M., Walker,N.S., Conte,L.L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) *Nucleic Acids Res.*, **30**, 260–263.
Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) *Nucleic Acids Res.*, **32**, D189–D192.
Chen,J., Lips,B.K., Halim,N., Ma,Q.D., Matsumoto,M., Melhem,S., Kolachana,B.S., Hyde,T.M., Herman,M.M. and Apud,J. (2004) *Am. J. Hum. Genet.*, **75**, 807–821.
Clarke,J., Cota,E., Fowler,S.B. and Hamill,S.J. (1999) *Structure*, **7**, 1145–1153.
Daggett,V., Kollman,P.A. and Kuntz,I.D. (1991) *Biopolymers*, **31**, 1115–1134.
Daggett,V. (2006) *Acc. Chem. Res.*, **39**, 594–602.
Day,R., Beck,D.A.C., Armen,R.S. and Daggett,V. (2003) *Protein Sci.*, **12**, 2150–2160.
Dietmann,S., Park,J., Notredame,C., Heger,A., Lappe,M. and Holm,L. (2001) *Nucleic Acids Res.*, **29**, 55–57.
Doreleijers,J.F., Mading,S., Maziuk,D., Sojourner,K., Yin,L., Zhu,J., Markley,J.L. and Ulrich,E.L. (2003) *J. Biomol. NMR*, **26**, 139–146.
Doreleijers,J.F., Nederveen,A.J., Vranken,W., Lin,J.D., Bonvin,A.M.J.J., Kaptein,R., Markley,J.L. and Ulrich,E.L. (2005) *J. Biomol. NMR*, **32**, 1–12.
Gianni,S., Guydosh,N.R., Khan,F., Caldas,T.D., Mayor,U., White,G.W., DeMarco,M.L., Daggett,V. and Fersht,A.R. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 13286–13291.
Godzik,A., Kolinski,A. and Skolnick,J. (1995) *Protein Sci.*, **4**, 2107–2117.
Goodman,J.E., Jensen,L.T., He,P. and Yager,J.D. (2002) *Pharmacogenetics*, **12**, 517–528.
Gunasekaran,K., Eyles,S.J., Hagler,A.T. and Gierasch,L.M. (2001) *Curr. Opin. Struct. Biol.*, **11**, 83–93.
Hom,K., Ma,Q.F., Wolfe,G., Zhang,H., Storch,E.M., Daggett,V., Basus,V.J. and Waskell,L. (2000) *Biochemistry*, **39**, 14025–14039.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Karayiorgou,M., Altemus,M., Galke,B.L., Goldman,D., Murphy,D.L., Ott,J. and Gogos,J.A. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 4572–4575.
Kearsley,S.K. (1989) *Acta Crystallogr. A*, **45**, 208–210.
Kehl,C., Simms,A.M., Toofanny,R.D. and Daggett,V. (2008) *Prot. Eng. Des. Sel.*, **21**, 379–386.
Kell,G.S. (1967) *J. Chem. Eng. Data*, **12**, 66.
Lattman,E.E. (2005) *Proteins*, **54**, i.
Laskowski,R.A., Chistyakov,V.V. and Thornton,J.M. (2005) *Nucleic Acids Res.*, **33**, D266–D268.
Lee,B. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
Lee,S.Y., Cho,H.S., Pelton,J.G., Yan,D., Henderson,R.K., King,D.S., Huang,L., Kustu,S., Berry,E.A. and Wemmer,D.E. (2001) *Nat. Struct. Biol.*, **8**, 52–56.
Levitt,M., Hirshberg,M., Sharon,R. and Daggett,V. (1995) *Comput. Phys. Commun.*, **91**, 215–231.
Levitt,M., Hirshberg,M., Sharon,R., Laidig,K.E. and Daggett,V. (1997) *J. Phys. Chem. B*, **101**, 5051–5061.
Lin,J.H., Perryman,A.L., Schames,J.R. and McCammon,J.A. (2003) *Biopolymers*, **68**, 47–62.
Lotta,T., Vidgren,J., Tilgmann,C., Ulmanen,I., Melen,K., Julkunen,I. and Taskinen,J. (1995) *Biochemistry*, **34**, 4202–4210.
McEvoy,M.M., Hausrath,A.C., Randolph,G.B., Remington,S.J. and Dahlquist,F.W. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7333–7338.
Meagher,K.L. and Carlson,H.A. (2004) *J. Am. Chem. Soc.*, **126**, 13276–13281.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
Osapay,K. and Case,D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.
Petsko,G.A. (1996) *Nat. Struct. Biol.*, **3**, 565–566.
Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) *J. Comput. Chem.*, **25**, 1605–1612.
Rueda,M., Ferrer-Costa,C., Meyer,T., Perez,A., Camps,J., Hospital,A., Gelpi,J.L. and Orozco,M. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 796–801.
Rutherford,K., Bennion,B.J., Parson,W.W. and Daggett,V. (2006) *Biochemistry*, **45**, 2178–2188.
Schneider,D.M., Dellwo,M.J. and Wand,A.J. (1992) *Biochemistry*, **31**, 3645–3652.
Scouras,A. and Daggett,V. (2008) *J. Mater. Sci.* DOI: http://dx.doi.org/10.1021/jp710546e.
Simms,A.M., Toofanny,R.D., Kehl,C., Benson,N.C. and Daggett,V. (2008) *Prot. Eng. Des. Sel.*, **21**, 369–377.
Spencer,R.H. and Rees,D.C. (2002) *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 207–233.
Spielman,R.S. and Weinshilboum,R.M. (1981) *Am. J. Med. Genet.*, **10**, 279–290.

Storch,E.M. and Daggett,V. (1995) *Biochemistry*, **34**, 9682–9693.
Storch,E.M., Grinstead,J.S., Campbell,A.P., Daggett,V. and Atkins,W.M. (1999a) *Biochemistry*, **38**, 5065–5075.
Storch,M.M., Daggett,V. and Atkins,W.M. (1999b) *Biochemistry*, **38**, 5054–5064.
Vijay-Kumar,S., Bugg,C.E. and Cook,W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.
Wadhams,G.H. and Armitage,J.P. (2004) *Nat. Rev. Mol. Cell Biol.*, **5**, 1024–1037.
Wolfram Research I. (2005) *Mathematica.*, 5.2 edition. Wolfram Research, Inc., Champaign, IL.
Wong,K.B. and Daggett,V. (1998) *Biochemistry*, **37**, 11182–11192.
Yee,D.P. and Dill,K.A. (1993) *Protein Sci.*, **2**, 884–899.
Zhao,R., Collins,E.J., Bourret,R.B. and Silversmith,R.E. (2002) *Nat. Struct. Biol.*, **9**, 570–575.

Received February 23, 2008; revised February 23, 2008; accepted February 25, 2008

Board Member: Jane Clarke

Edited by Alan Fersht