Structure
Article

Geometry-Based Sampling of Conformational Transitions in Proteins

Daniel Seeliger,[1] Jürgen Haas,[2] and Bert L. de Groot[1,*]
[1]Computational Biomolecular Dynamics Group
[2]Department of Theoretical and Computational Biophysics
Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany
*Correspondence: bgroot@gwdg.de
DOI 10.1016/j.str.2007.09.017

SUMMARY

The fast and accurate prediction of protein flexibility is one of the major challenges in protein science. Enzyme activity, signal transduction, and ligand binding are dynamic processes involving essential conformational changes ranging from small side chain fluctuations to reorientations of entire domains. In the present work, we describe a reimplementation of the CONCOORD approach, termed tCONCOORD, which allows a computationally efficient sampling of conformational transitions of a protein based on geometrical considerations. Moreover, it allows for the extraction of the essential degrees of freedom, which, in general, are the biologically relevant ones. The method rests on a reliable estimate of the stability of interactions observed in a starting structure, in particular those interactions that change during a conformational transition. Applications to adenylate kinase, calmodulin, aldose reductase, T4-lysozyme, staphylococcal nuclease, and ubiquitin show that experimentally known conformational transitions are faithfully predicted.

INTRODUCTION

Regardless of whether a protein functions as an enzyme, molecular motor, transport protein, or receptor, its function is often coupled to motion. These motions range from side chain fluctuations to reorientations of domains and partial unfolding and refolding. An understanding of protein function is thus strongly coupled to insight into dynamics and flexibility. X-ray crystallography, which is still the major source of structural information of proteins, provides mainly static pictures of one conformation, even though a number of proteins have been resolved in different conformations, providing insights into protein flexibility directly from experimental data (Gerstein and Krebs, 1998). Structures resolved by NMR spectroscopy are usually published as an ensemble of conformations that fulfill the experimentally determined restraints and provide more information about protein flexibility. However, the method is still restricted to proteins of limited size.

Knowledge about protein structures in different conformational substates, either from experimental data or simulation, has been proven to enhance protein-protein docking (Bonvin, 2006; Mustard and Ritchie, 2005; Ehrlich et al., 2005) and structure-based drug design (SBDD) (Knegtel et al., 1997; Carlson, 2002; Meagher and Carlson, 2004; McGovern and Shoichet, 2003; Teague, 2003). However, proteins often undergo conformational changes upon ligand binding. Therefore, molecular docking or the derivation of pharmacophore models from a single receptor structure often leads to unsatisfying results, either by excluding known binders due to overdefinition of the binding site when using a holo structure, or by not identifying the correct binding pose when using an apo structure or protein model (McGovern and Shoichet, 2003).

Among the computational approaches used to tackle protein flexibility, molecular dynamics (MD) simulations are predominantly employed. However, despite the enormous increase in computer power and advances in algorithm techniques and parallelization, MD simulations are computationally expensive; moreover, high-energy barriers are often not overcome within accessible time. In order to alleviate the resulting sampling problem, several advanced simulation methods based on MD, including replica-exchange molecular dynamics (REMD) (Sugita and Okamoto, 1999), conformational flooding (Grubmueller, 1995; Lange et al., 2006), and targeted molecular dynamics (TMD) (Schlitter et al., 1994; van der Vaart and Karplus, 2005), have been developed and successfully applied to numerous problems within the field of protein research. However, even these methods are not routinely applicable to the efficient sampling of conformational transitions. Computationally more efficient, but less accurate, methods are based on Gaussian network models (Bahar et al., 1998; Haliloglu et al., 1997), normal mode analysis (Go et al., 1983; Brooks and Karplus, 1983; Krebs et al., 2002; Alexandrov et al., 2005), or graph theoretical approaches (Jacobs et al., 2001).

A different approach is the CONCOORD method (de Groot et al., 1997), which is based on geometrical considerations for the prediction of protein flexibility. A given input structure is analyzed and translated into a geometric description of the protein. Based on this description, the structure is rebuilt, commonly several hundreds of times,

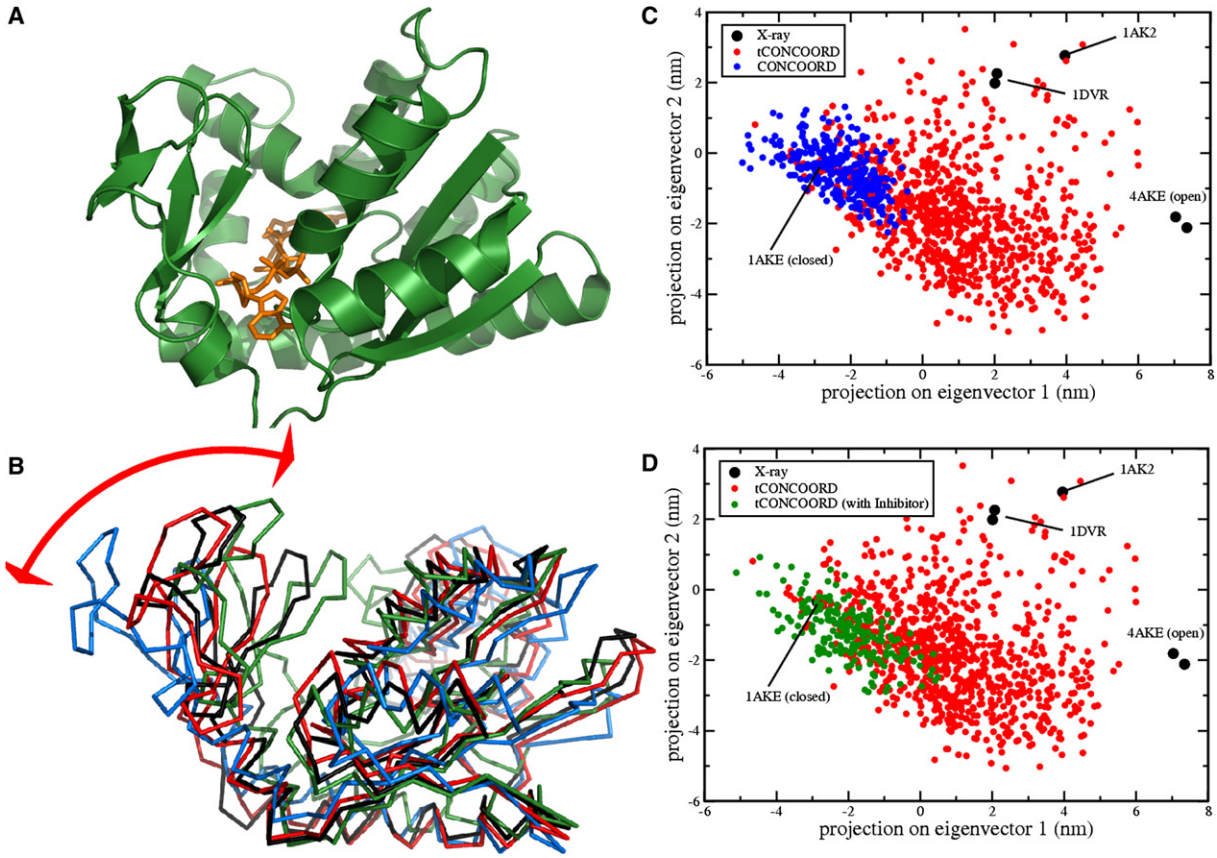

Figure 1. Adenylate Kinase

(A) Crystal structure (PDB code: 1AKE) of adenylate kinase (green) with bound inhibitor AP_5A (orange).

(B) Superimposition of several X-ray structures in different conformations, indicating the induced fit motion.

(C and D) Principal components analysis. Experimental structures (black circles) and three simulation ensembles (blue, red, and green circles) are projected onto the first two eigenvectors. The blue ensemble was generated with CONCOORD, and the red one was generated with tCONCOORD. tCONCOORD correctly predicts the induced fit motion and samples open conformations when they are started from the closed conformation with the ligand removed. If the ligand remains in the input structure, the conformational space is restricted to conformations around the closed state (green).

leading to an ensemble that can be analyzed, and essential degrees of freedom (Amadei et al., 1993), often representing the biological, relevant motions in proteins, may be extracted. Whereas the original implementation of CONCOORD was developed to predict conformational ensembles around a known structure, in this work we present a major extension termed tCONCOORD (Seeliger and de Groot, 2007a) that allows for the prediction of conformational transitions of proteins. tCONCOORD has been completely parameterized based on experimental data, from which, for example, a novel set of protein-specific atomic radii has been derived (Seeliger and de Groot, 2007b) to ensure optimal geometry. Moreover, the constraint definition has been calibrated to also allow for the prediction of large-scale conformational transitions. An integral part of tCONCOORD is a newly developed approach for estimating hydrogen-bond stability via a thorough analysis of the environment. Its incorporation into the constraint definition significantly enhances the prediction quality of conformational transitions. We show simulation results for adenylate kinase, calmodulin, aldose reductase, T4-lysozyme, ubiquitin, and staphylococcal nuclease to assess the prediction quality for different applications ranging from flexible to rigid protein structures, including large conformational transitions. Additionally, the influence of ligands on conformational flexibility is investigated.

RESULTS

Adenylate Kinase

Adenylate kinase displays a distinct induced fit motion upon binding to its substrate (ATP/AMP) or an inhibitor (see Figure 1B). Structures in different conformations have been resolved (Müller and Schulz, 1992; Müller et al., 1996; Schlauderer and Schulz, 1996; Schlauderer et al., 1996), contributing significantly to the understanding of the catalytic mechanism of this class of enzymes. We carried out two tCONCOORD simulations by using the closed conformation of adenylate kinase (Protein Data Bank [PDB] code: 1AKE, see Figure 1A) as input. In one simulation, the ligand (AP_5A) was removed.

Figures 1C and 1D show the result of a principal components analysis (PCA) applied to the experimental structures. The first eigenvector (x axis) corresponds to the induced fit motion indicated by the red arrow in Figure 1B. Every dot in the plots represents a single structure. The red dots represent the ensemble that has been generated with tCONCOORD by using the closed conformation of adenylate kinase without ligand as input. The blue dots in Figure 1C represent an ensemble that has been generated by using CONCOORD (version 1.2), with the same input. As can be seen, the CONCOORD ensemble (blue) basically samples the conformational space around the input structure, leaving out open conformations. The tCONCOORD ensemble (red) behaves differently. It almost completely samples the conformational space that is covered by the experimental structures, thereby clearly producing open conformations (high x values). The experimental structures were reached with a deviation of 2.4, 2.6, and 3.1 Å C_α-rmsd for 1DVR, 1AK2, and 4AKE, respectively. For comparison, for the CONCOORD cluster these RMSD values are much higher with 3.4, 4.4, and 5.9 Å. In SBDD, the reverse problem, predicting induced-fit structures from an open conformation, often needs to be addressed. A tCONCOORD run with an open conformation (PDB code: 4AKE) as input produces structures that approach the closed conformations with rmsds of 2.5, 2.9, and 3.3 Å for 1DVR, 1AK2, and 1AKE, respectively. Thus, conformations close to the experimentally determined ligand-bound states are present within the ensemble that was generated by using the apo structure (PDB code: 4AKE) as input.

The conformational flexibility changes significantly if the ligand remains in the input structure. Figure 1D shows a comparison of an ensemble with the ligand present in the input structure (green dots) with the previously discussed ensemble, generated without ligand (red dots). As can be seen, the presence of the ligand leads to a reduction of the conformational space that is sampled by the protein, and open conformations are not sampled anymore.

Calmodulin

The structure and dynamics of calmodulin have been studied extensively by X-ray crystallography and NMR. In its activated (Ca^{2+}-bound) conformation (Chattopadhyaya et al., 1992), calmodulin exposes hydrophobic residues to the solvent, enabling binding to a target, either a protein or an inhibitior. The binding process itself requires a large conformational change involving the unfolding of the central helix in order to allow for rotation of the C-terminal domain to form the binding site (Cook et al., 1994) (Figures 2A and 2B).

A tCONCOORD simulation of this particularly challenging test case has been carried out. The instability of a number of hydrogen bonds in the central helix of the activated form (PDB code: 1CLL) was correctly identified (see Figure 2C) and incorporated into the constraint definition.

The resulting ensemble (Figure 2E, left) can be described as two freely rotating domains connected by a linker. These results are in good agreement with NMR studies of calmodulin (Elshorst et al., 1999) (Figure 2E, right). In Figure 2F, the projections of the tCONCOORD ensemble (green cloud), the NMR ensemble (red dots), the X-ray structures of the activated form (orange dot), and the ligand-bound conformation (blue dot) onto the first three eigenvectors of a PCA are shown. The tCONCOORD ensemble represents an extended sampling of the conformational space, comprising all experimentally determined structures.

The rmsd between the activated conformation of calmodulin and the bound conformation is 14.6 Å. The closest match of a structure from the ensemble, generated with tCONCOORD, to the experimentally known ligand bound conformation is as low as 2.8 Å (Figure 2D). This is an example of a case in which a ligand-bound conformation of the protein is predicted by using only the structurally completely different unbound state as input. The possibility of such predictions is of obvious interest for applications in the field of SBDD.

Aldose Reductase

Aldose reductase (AR) is believed to play an important role in diabetes and therefore is a potential drug target (Brownlee, 2001; Steuber et al., 2006). It adopts a TIM barrel fold and uses NADPH as a cofactor to reduce various aldehydes. AR has been crystallized with different inhibitors. A remarkable fact concerning these inhibitors is that they have very different structures, sizes, and molecular weights (Steuber et al., 2006). AR is able to bind these structurally different inhibitors because of a very flexible binding site.

Figure 3 shows the structure of AR (PDB code: 2FZD) with bound cofactor (red) and the inhibitor Tolrestat (orange). The regions that are responsible for the formation of a hydrophobic subpocket are labeled with A and C. The B loop is responsible for binding the cofactor. In order to study the influence of both the ligand and the cofactor on the conformational flexibility of AR, tCONCOORD simulations were carried out for the entire complex (AR+NADP+Tolrestat), the complex with removed inhibitor (AR+NADP), and free AR. To compare the flexibility of the different systems, a PCA was applied to the combined ensembles of all three runs. Subsequently, the ensembles for each system were projected onto the eigenvectors with the largest eigenvalues.

Eigenvectors 1 and 2 mainly correspond to movements of the A loop in AR, as indicated in Figure 4 (right panel). The projection of the ensembles onto these eigenvectors (Figure 4, left panel) reveals the same flexibility along these eigenvectors for the free AR (σ_1^{free} = 5.15 nm, σ_2^{free} = 4.34 nm) and the AR with bound cofactor (σ_1^{holo} = 5.07 nm, σ_2^{holo} = 4.25 nm). In the third system, in which Tolrestat is also bound, the flexibility is reduced significantly due to interaction of the ligand with the A and C loops (σ_1^{tol} = 3.13 nm, σ_2^{tol} = 3.28 nm).

Figure 5 compares the motions along eigenvectors 3 and 4. The motions corresponding to eigenvector 3 predominantly represent a movement of loop B, which is

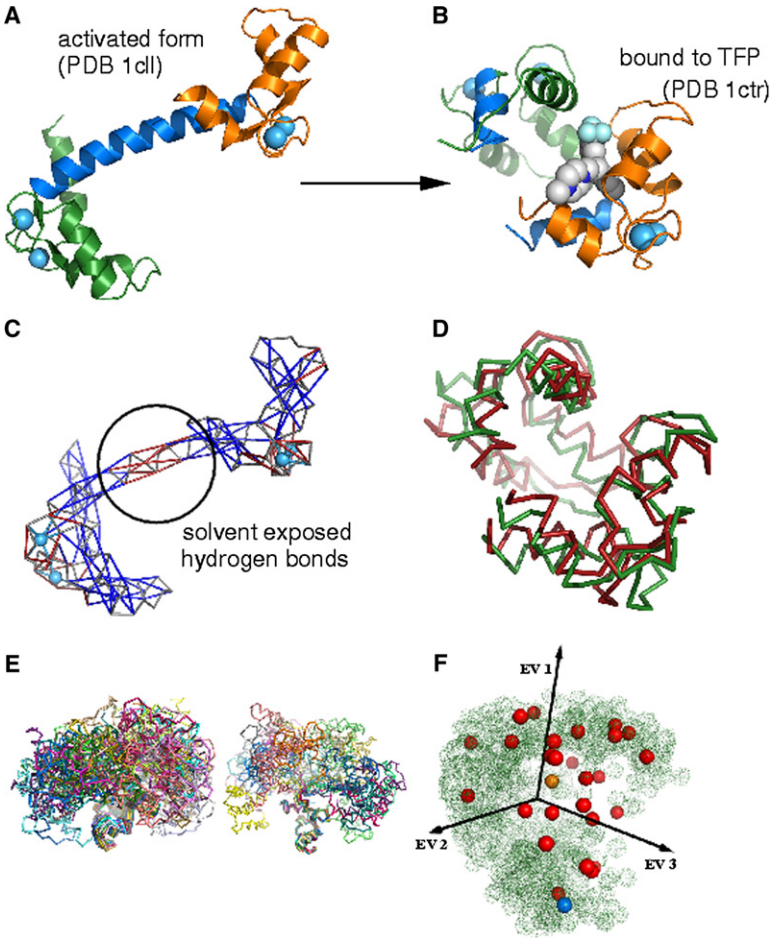

Figure 2. Calmodulin

(A) The activated form of calmodulin (PDB code: 1CLL) used as input for tCONCOORD.

(B) The structure of calmodulin bound to Trifluoroperazine (TFP). The rmsd between these two structures is 14.6 Å.

(C) The result of the hydrogen-bond analysis of tCONCOORD. Red sticks represent hydrogen bonds with high solvation probabilities and are not regarded as constraints in the tCONCOORD simulation.

(D) The superimposition of the ligand-bound conformation (green) and the closest match of a structure from the tCONCOORD ensemble (red) with an rmsd of 2.8 Å.

(E) A tCONCOORD ensemble and an NMR ensemble (PDB code: 1CFF) fitted onto the C-terminal domain.

(F) The projection onto the three eigenvectors with the largest eigenvalues of a PCA. The tCONCOORD ensemble is shown as a green cloud, and the NMR ensemble is shown as red dots. The orange dot represents the X-ray structure of the open (activated) conformation, and the blue dot represents the closed (ligand-bound) state.

involved in binding the cofactor. Here, we observe high flexibility for the free AR (σ_3^{free} = 3.41 nm), whereas the fluctuation for the holo form (σ_3^{holo} = 2.76 nm) and the entire complex (σ_3^{tol} = 2.96 nm) along this mode is comparable. Eigenvector 4 again reveals a clear difference between the holo form and the complete complex systems. As the main component of this mode is a movement of the C loop, flexibility of this region is dramatically reduced by Tolrestat (σ_4^{tol} = 1.30 nm), whereas free AR and holo AR display comparable and somewhat higher flexibility along this eigenvector (σ_4^{free} = 2.01 nm, σ_4^{holo} = 2.10 nm).

T4-Lysozyme

Bacteriophage T4-lysozyme (T4L) is one of the rare cases in which conformational flexibility can be directly estimated from X-ray structures (de Groot et al., 1998). It has been crystallized in many different conformational states, shedding light on the dynamical behavior. The main collective motion is a hinge-bending mode that is necessary for entrance and release of the substrate. This mode is described by the first eigenvector of a PCA, carried out on the experimental data.

In order to predict open conformations by using the closed conformation as input for tCONCOORD, the correct detection of unstable hydrogen bonds is mandatory. As can be seen in Figure 6, a hydrogen bond that is formed between Glu22 and Arg137 in the closed conformation (PDB code: 2LZM, left structure) is not present in the open conformation (PDB code: 149L, right structure), and the distance from the Cδ of Glu22 to Cζ of Arg137 changes from 3.8 Å to more than 18 Å. The hydrogen-bond analysis method of tCONCOORD correctly predicts the instability of this hydrogen bond, as indicated in the picture in the central upper panel of Figure 6. The blue sticks represent stable hydrogen bonds, whereas red sticks mark those that display high probabilities of water attack. These hydrogen bonds are not defined as constraints.

Figure 6 also shows the projection of the experimental data, a tCONCOORD ensemble, and three MD trajectories, which have started from different conformational states, onto the first two eigenvectors.

It can be seen that the tCONCOORD ensemble, started from a closed state (PDB code: 2LZM), also samples open conformations. A closer look at the MD trajectories reveals that the longest trajectory (cyan, 184 ns) does not sample open conformations at all, whereas the shorter simulations (red and green) cover more of the conformational space. The phase space density produced by the MD simulations indicates an energy barrier between the closed and the open conformations that is not overcome in the simulation

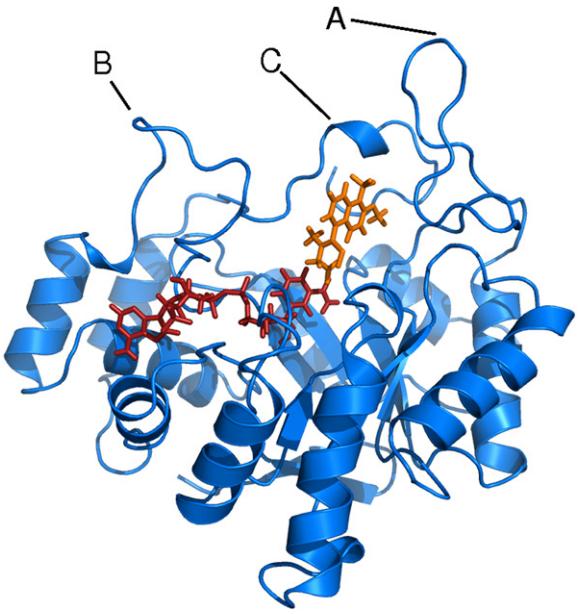

Figure 3. Aldose Reductase
The loops labeled A and C form parts of the Tolrestat-binding site. Loop B interacts with the cofactor.

represented by the cyan circles. The tCONCOORD sampling, however, is not affected by energy barriers and samples most of the space covered by the MD trajectories. Although the tCONCOORD ensemble samples both open and closed conformations, it does not completely sample the conformational space sampled by the MD simulations that started from open conformations. This is due to the fact that tCONCOORD defines constraints from a single input structure, in this particular case a closed conformation. If unstable interactions are not entirely detected in the constraint definition process, this can lead to an exclusion of regions of the conformational space.

The tCONCOORD ensemble furthermore samples regions of the conformational space that are not visited by the MD simulations and the experimental structures. This could be either due to an energy barrier that is too high to be overcome by MD simulations within the accessible timescale, or to the energy of this region of the conformational space being too high to be part of the relevant conformational space.

Rigid and Flexible Regions in Proteins

Functional studies on protein structures benefit significantly from information about the flexibility and rigidity of protein parts. The calculation of root-mean-square fluctuations (rmsf) from tCONCOORD ensembles can provide valuable hints regarding these properties. To test the reliability of flexibility predictions, we chose two test cases with completely different structure and flexibility properties, which have been experimentally determined. As the first test case, we chose ubiquitin, a small 70 residue protein of which 46 X-ray structures are available in the PDB

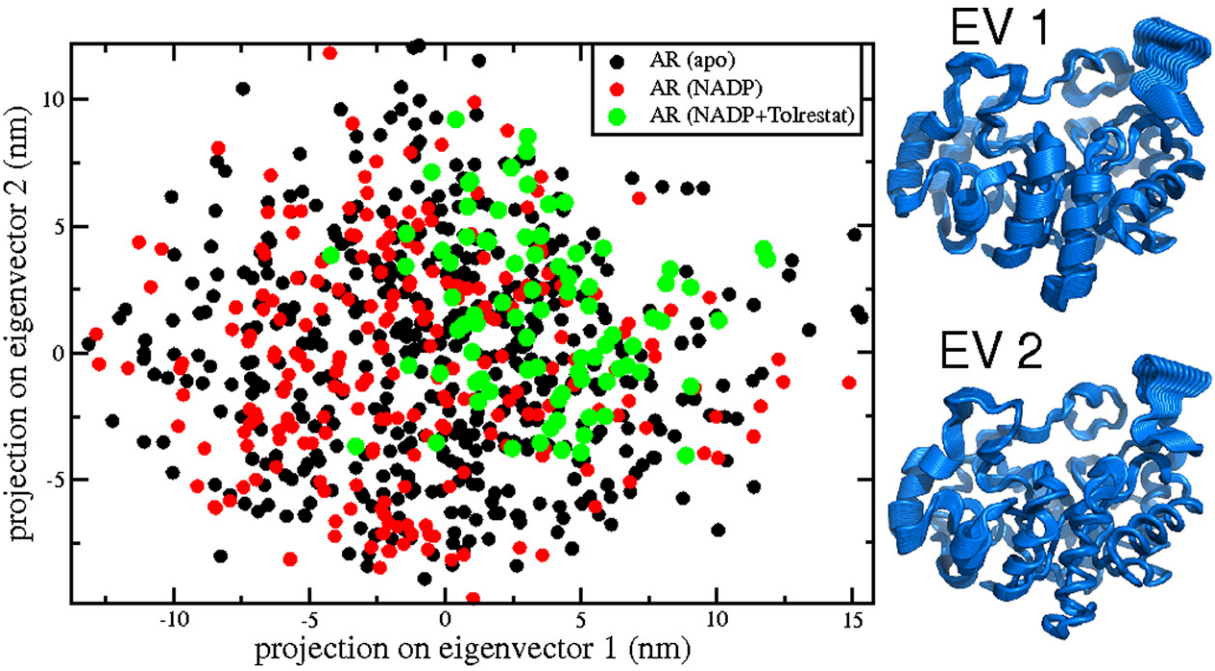

Figure 4. Projection of tCONCOORD Ensembles of Aldose Reductase onto Eigenvectors 1 and 2 of a Principal Components Analysis
The structures on the right represent the predominant motions along these vectors. On the left, the two-dimensional projection of three different ensembles is shown. The green dots represent the ensemble of the entire complex, the red dots represent the holo form, and the black dots represent the apo form. The projection shows the reduced flexibility of the binding site in the presence of Tolrestat. Binding of NADP, however, has no effect on these modes.

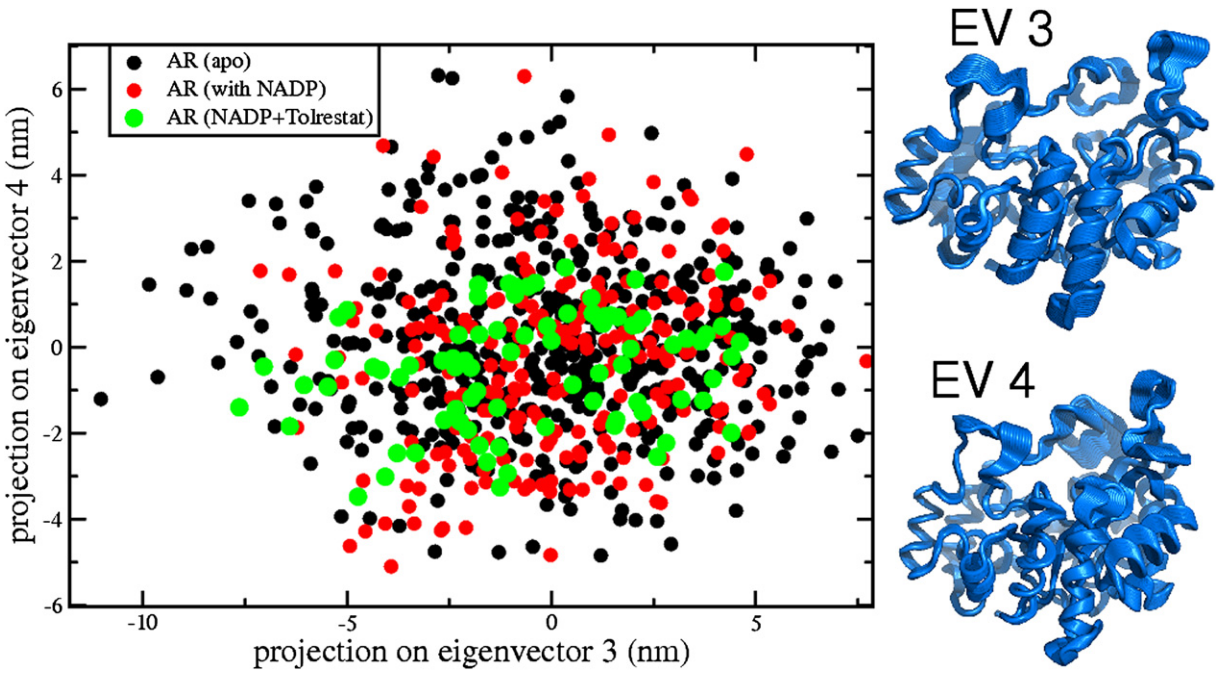

Figure 5. Projection of tCONCOORD Ensembles of Aldose Reductase onto Eigenvectors 3 and 4 of a Principal Components Analysis

The structures on the right represent the predominant motions along these vectors. On the left, the two-dimensional projection of three different ensembles is shown. The green dots represent the ensemble of the entire complex, the red dots represent the holo form, and the black dots represent the apo form. The projection shows increased flexibility along eigenvector 3 if NADP is removed, because loop B is predominantly involved in this motion. Eigenvector 4 mainly represents a movement of loop C, which leads to decreased flexibility for the ensemble with Tolrestat bound.

(see the Supplemental Data available with this article online). The rmsf determined from the X-ray structures (Figure 7, red curve) shows that the protein is relatively rigid, and that the only noteworthy flexibility is at the C terminus and a loop. The rmsf calculated from the tCONCOORD ensemble generated by using PDB code 1UBI (Love et al., 1997) as input (Figure 7, black curve) represents the same flexibility properties as the experimental data. Although the flexibility level of the tCONCOORD ensemble is constantly above the X-ray ensemble, the overall picture of a rigid protein with a flexible C terminus is reproduced (correlation coefficient of 0.95). For comparison, the rmsf of an ensemble generated with an elastic network model (Suhre and Sanejouand, 2004a, 2004b) is shown (Figure 7, green curve). This fast and efficient method is routinely employed to predict protein flexibility and reproduces the experimental fluctuations only slightly worse than tCONCOORD (correlation coefficient of 0.9). However, the structures from the tCONCOORD ensemble all have reasonable geometry (bond lengths, angles, dihedrals, and interatomic distances), which is not always the case for single structures derived from elastic network models.

As a second test case, we chose staphylococcal nuclease, of which an NMR ensemble (Wang et al., 1997) (PDB code: 1JOR) provides information on the flexibility of the protein. The rmsf calculated from the NMR ensemble (Figure 8, red curve) renders mainly one loop around residue 42 very flexible. Furthermore, the loops around residues 80 and 110 show increased flexibility. The rmsf calculated from a tCONCOORD ensemble (Figure 8, black curve), by using an X-ray structure (PDB code: 1EY4) (Chen et al., 2000) as input, qualitatively yields the same picture. The most flexible regions detected by the tCONCOORD ensemble are in good agreement with the experimental data (correlation coefficient of 0.8) and, again, are slightly better than those predicted by the elastic network model (green curve, correlation coefficient of 0.78). The tCONCOORD ensemble predicts higher flexibility for some parts of the protein than observed in the NMR ensemble. This might be due either to interactions that tCONCOORD underestimates, or to an overly tight representation of the NMR data, which is sometimes caused by imposing time- and ensemble-averaged experimental properties onto single structures during refinement (Spronk et al., 2003; Bonvin and Brünger, 1995; Cuniasse et al., 1997).

DISCUSSION

We report a novel, to our knowledge, approach to accurately predict large conformational transitions in proteins and its application to selected systems with biological relevance. The method rests on a thorough analysis of the interactions in proteins and their translation into constraints. In particular, hydrogen bonds are investigated, and their stability is estimated by analyzing their surroundings in respect to hydrophobic protection. Using

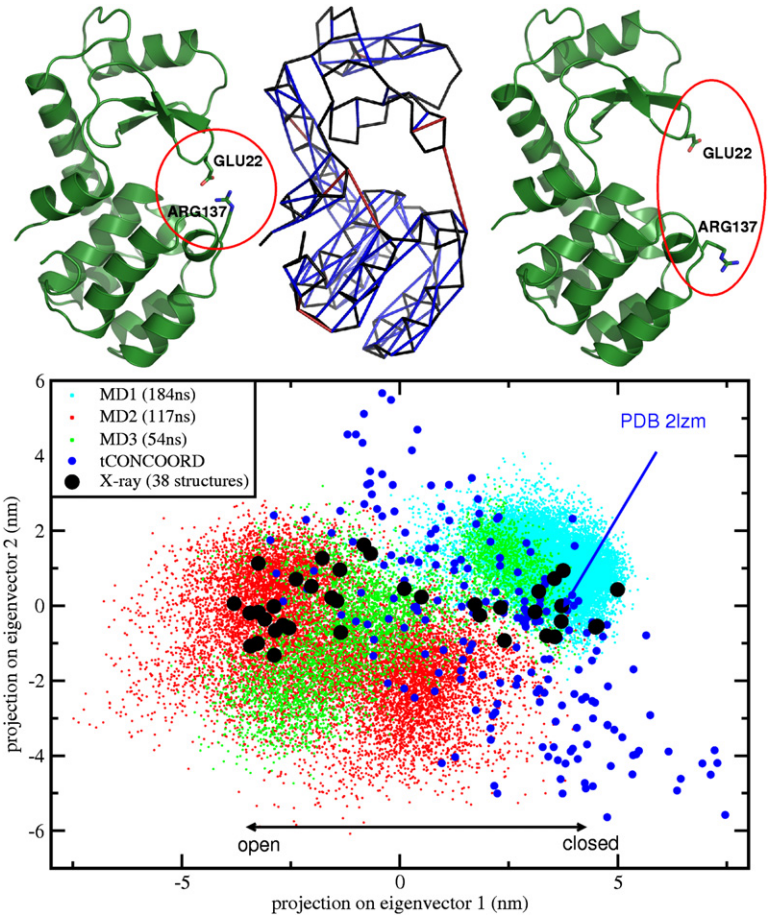

Figure 6. T4-Lysozyme

The upper left panel shows the structure of the closed conformation of T4-lysozyme (PDB code: 2LZM). This structure has been used as input for tCONCOORD. The picture in the middle shows the hydrogen-bond stability analysis carried out by tCONCOORD. Red-marked hydrogen bonds, like the bond between GLU22 and ARG137, are predicted to be unstable. The picture on the right shows the structure of an open conformation of T4-lysozyme (PDB code: 149L). Indeed, in this conformation, this hydrogen bond is not present anymore. The lower panel shows the result of a principal components analysis applied to the experimental structures. The experimental structures (black), the tCONCOORD ensemble (blue), and three MD trajectories (cyan, red, and green) are projected on the first two eigenvectors.

the predefined constraints, structures are built from random starting conditions by iteratively correcting atomic coordinates. The resulting ensemble covers the conformational space that is available within those constraints, regardless of potential energy barriers between different conformations, which usually preclude efficient sampling with other methods.

Information about conformational transitions is often a prerequisite for understanding protein function. With tCONCOORD, we provide an efficient simulation approach

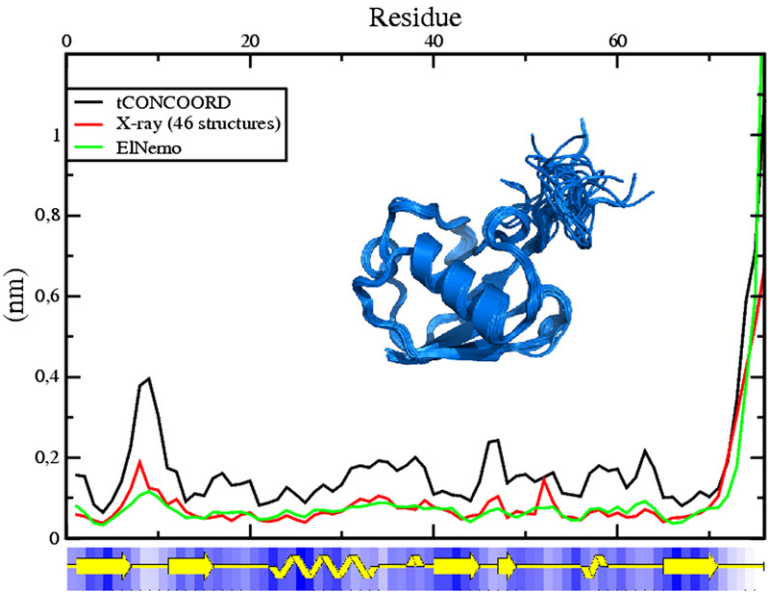

Figure 7. Root-Mean-Square Fluctuation in Ubiquitin

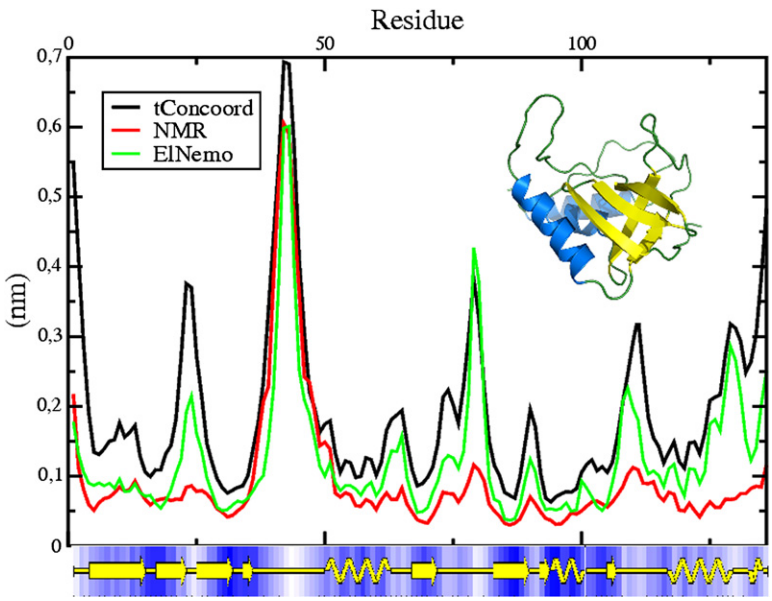

Figure 8. Root-Mean-Square Fluctuation in Staphylococcal Nuclease

to predict protein conformational transitions. The resulting ensemble can be used to study the essential degrees of freedom of a protein, to identify flexible and rigid parts in a structure, or to obtain different starting points for other simulation protocols. Furthermore, incorporation of protein flexibility by tCONCOORD ensembles, e.g., in docking protocols, is expected to enhance the efforts of SBDD.

EXPERIMENTAL PROCEDURES

Structure Analysis

Interactions in protein structures are rigorously analyzed and translated into a set of geometrical constraints that can be compared to a construction plan of the protein. This set consists of topological constraints (e.g., bonds, angles, planarities) and noncovalent constraints like hydrogen bonds, salt bridges, and hydrophobic clusters. During the analysis of experimentally known conformational transitions, it was found that they routinely involve opening of one or more hydrogen bonds. tCONCOORD therefore attempts to predict unstable hydrogen bonds by estimating the solvation probability. This approach is based on the work of Fernandez et al. (Fernandez et al., 2002a, 2002b, 2004; Fernandez and Berry, 2002), who showed that keeping a hydrogen bond "dry" is a prerequisite for its stability, and that protein folding is associated with a systematic desolvation of hydrogen bonds by surrounding hydrophobic groups. Thus, analyzing the neighborhood of a particular hydrogen bond should provide hints for the probability of a water molecule attacking it, which is directly correlated to the opening probability.

To this end, we have analyzed 35 large-scale molecular dynamics trajectories from different proteins (Table 1) and calculated for each protein atom type *i* (a total of 167 atom types, hydrogen atoms were not taken into account) the radial distribution function (RDF) for water-oxygen (O_{wat}). Integrating the weighted RDFs according to $P_i = \int_0^d R_{i-Owat}(r)\,dr$ (with d = 6 Å) yields a value that may serve as a solvation parameter and allows for the estimation of the probability of finding a water molecule within a certain distance to the particular atom. Because these values were obtained by analyzing a very limited number of trajectories, an accurate statistical error estimation is difficult. Additionally, there is a systematic error, resulting from the low number of different folds and sequences taken into account for this work. However, previous studies on hydrophobic protection showed that even more simple approaches, such as counting hydrophobic residues around a hydrogen bond, provide valuable hints for predicting unstable hydrogen bonds (Fernandez et al., 2002a, 2002b, 2004; Fernandez and Berry, 2002).

The obtained solvation parameters are used to evaluate the surroundings of a particular hydrogen bond. We consider all atoms within

Table 1. Molecular Dynamics Trajectories that Were Used for the Derivation of Solvation Parameters

PDB Code	Simulation Time, ns	PDB Code	Simulation Time, ns
1TUX	110	1RAT	110
1PGS	110	1UBI	110
1CNV	110	1UNE	110
135L	110	1VCC	110
153L	110	1WBA	110
1A3D	110	1A3H	110
1AST	110	4ICB	110
1BJ7	110	1CLM	110
1BM8	110	1CSP	198
1CPN	110	1EXR	77
1DSL	110	1EZM	110
1GBG	110	2CHE	113
1HYP	174	1MLA	110
2APR	110	4AKE	110
1CHD	110	1HKA	110
1AAJ	110	1KOE	110
1ELT	110	1OSA	110
1GBS	110		

All simulations were carried out by using the GROMACS suite and the OPLS-AA force field with TIP4P water.

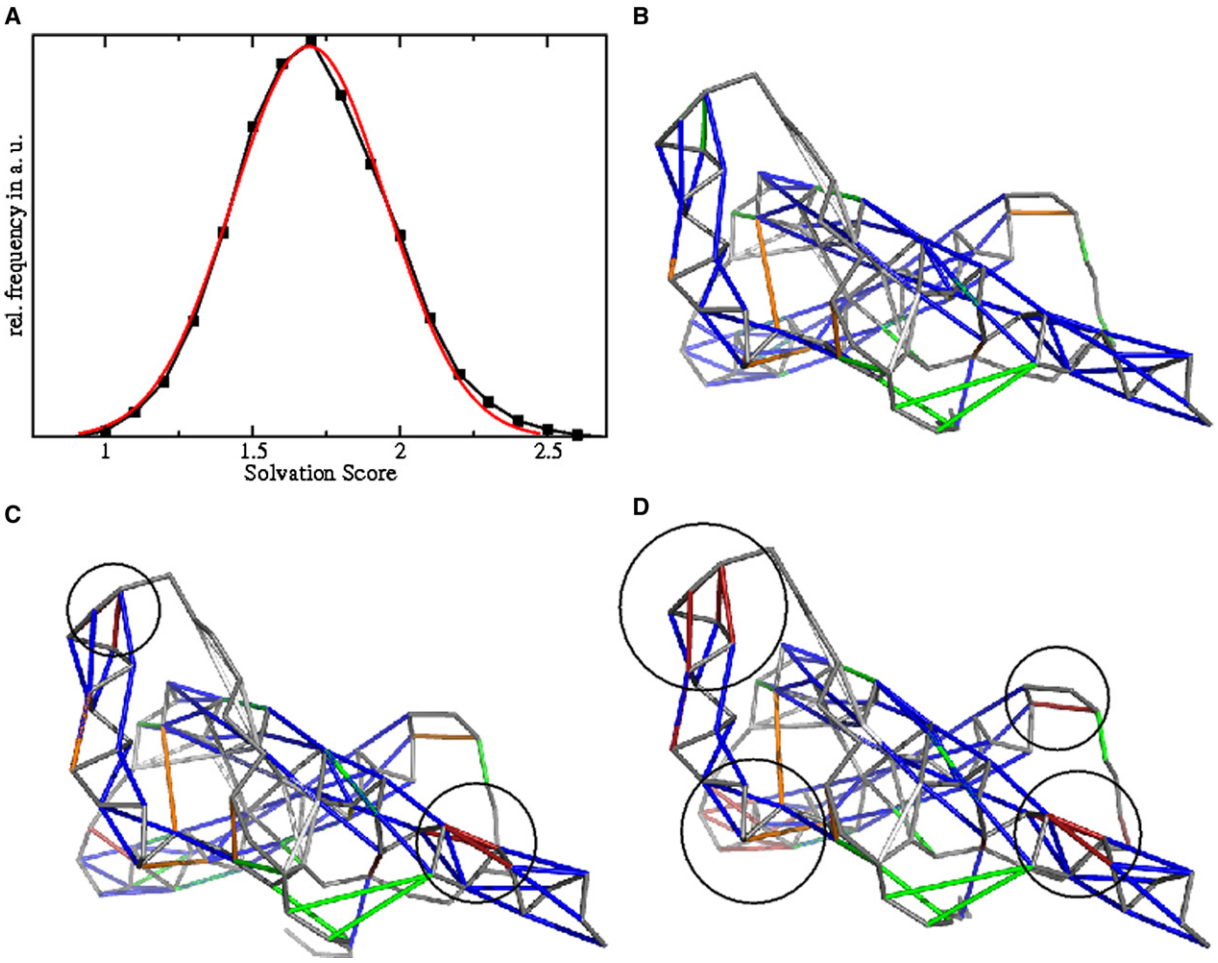

Figure 9. Hydrogen Bond Solvation

(A) Distribution of solvation scores in a subset of 290 protein structures from the Protein Data Bank.

(B) All hydrogen bonds in the human prion protein (PDB code: 1QM0). Blue sticks represent backbone-backbone hydrogen bonds, orange sticks represent backbone-side chain hydrogen bonds, and green sticks represent side chain-side chain hydrogen bonds.

(C) Detection of unstable hydrogen bonds with tCONCOORD by using a threshold of 2.2. Red sticks represent hydrogen bonds that are not turned into constraints.

(D) The same picture calculated with a threshold of 2.1. The number of unstable hydrogen bonds is larger than in (C).

two intersecting spheres, with radii of 6 Å, one centered at the hydrogen and the other one centered at the acceptor atom, for the nearest neighbors of a hydrogen bond, thereby excluding atoms that are less than three bonds away from the hydrogen or acceptor atom. Using the solvation parameters from these nearest neighbors, we calculate a solvation score, *S*, according to

$$S = \frac{1}{N}\sum_{i=0}^{N} P_i ; N : Number\ of\ neighbors, \quad (1)$$

which denotes the average of the solvation parameters of the neighboring atoms. This score is high if the neighborhood mostly consists of hydrophilic groups.

In order to incorporate this evaluation method into the constraint definition in tCONCOORD, we calculated the distribution of the introduced solvation score for all hydrogen bonds in 290 protein structures (Supplemental Data) from the Protein Data Bank (PDB) (Berman et al., 2000) with a resolution higher than 1.6 Å (Figure 9). For the constraint definition in tCONCOORD, we use thresholds between 2.1 and 2.2. A threshold of 2.2 means that hydrogen bonds with a score higher than 2.2, and thus exceeding that of 97% of the hydrogen bonds in the analyzed subset of the PDB, are considered to be unstable. Hence, they are disregarded and not translated into constraints.

The conformational space sampled by tCONCOORD is very sensitive to the identification of unstable hydrogen bonds and thus to small changes of coordinates (see Figure 9). We therefore provide default values for multiple simulation parameters but enable the user to change them. Moreover, constraints can be defined and undefined via a graphical user interface in order to study the influence of single interactions on conformational flexibility.

Structure Generation

tCONCOORD uses the CONCOORD algorithm (de Groot et al., 1997) for structure generation. Based on the predefined constraints, structures are built starting from random coordinates by iteratively correcting the coordinates to fulfill the constraints. Distances, angles, planarities, and chiralities are corrected simultaneously until all constraints

are fulfilled. Depending on the size of the protein and the number of constraints, this procedure takes from seconds to hours. For example, generating an ensemble of 100 structures for staphylococcal nuclease takes about 5 hr on a single Athlon4600+ cpu.

Because each run starts from random coordinates, each newly generated structure is completely independent from the previous one. On the one hand, this means that neither information about the path from one conformation to the other nor about potential energy barriers between two conformational states is obtained. On the other hand, the insensitivity to energy barriers means that, like CONCOORD, tCONCOORD does not suffer from sampling problems like, for example, MD simulations.

Structure Preparation and Ensemble Analysis

The quality of tCONCOORD-generated structures depends on the quality of the input structure. Therefore, structures should be checked, either by WHATIF (Vriend, 1990) or PROCHECK (Laskowski et al., 1993), prior to tCONCOORD simulations. Also, energy minimization prior to simulation can improve simulation results. The structures used in this work were either protonated by using the HB2NET module (Hooft et al., 1996) of WHATIF or the pdb2gmx program from the GROMACS 3.3.1 (Lindahl et al., 2001) suite. The GROMACS package has also been used for analyzing the generated structure ensembles.

Supplemental Data

Supplemental Data include the PDB codes of the structures that were used to derive hydrogen bond statistics and the PDB codes of the 46 X-ray structures of ubiquitin and are available at http://www.structure.org/cgi/content/full/15/11/1482/DC1/.

ACKNOWLEDGMENTS

We thank Ira Tremmel for carefully reading the manuscript.

Received: July 16, 2007
Revised: September 5, 2007
Accepted: September 17, 2007
Published: November 13, 2007

REFERENCES

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. (2005). Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. Protein Sci. *14*, 633–643.

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

Bahar, I., Atilgan, A.R., Demirel, M.C., and Erman, B. (1998). Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. Phys. Rev. Lett. *80*, 2733–2736.

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

Bonvin, A.M., and Brünger, A.T. (1995). Conformational variability of solution nuclear magnetic resonance structures. J. Mol. Biol. *250*, 80–93.

Bonvin, A.M.J.J. (2006). Flexible protein-protein docking. Curr. Opin. Struct. Biol. *16*, 194–200.

Brooks, B., and Karplus, M. (1983). Harmonic dynamics in proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. USA *80*, 6571–6575.

Brownlee, M. (2001). Biochemistry and molecular cell biology of diabetic complications. Nature *414*, 813–820.

Carlson, H.A. (2002). Protein flexibility and drug design: how to hit a moving target. Curr. Opin. Chem. Biol. *6*, 447–452.

Chattopadhyaya, R., Meador, W.E., Means, A.R., and Quiocho, F.A. (1992). Calmodulin structure refined at 1.7 Å resolution. J. Mol. Biol. *228*, 1177–1192.

Chen, J., Lu, Z., Sakon, J., and Stites, W.E. (2000). Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. J. Mol. Biol. *303*, 125–130.

Cook, W.J., Walter, L.J., and Walter, M.R. (1994). Drug binding by calmodulin: crystal structure of a calmodulin-trifluoperazine complex. Biochemistry *33*, 15259–15265.

Cuniasse, P., Raynal, I., Yiotakis, A., and Dive, V. (1997). Accounting for conformational variability in NMR structure of cyclopeptides: ensemble averaging of interproton distance and coupling constant restraints. J. Am. Chem. Soc. *119*, 5239–5248.

de Groot, B.L., van Aalten, D.M.F., Scheek, R.M., Amadei, A., Vriend, G., and Berendsen, H.J.C. (1997). Prediction of protein conformational freedom from distance constraints. Proteins Struct. Funct. Genet. *29*, 240–251.

de Groot, B.L., Hayward, S., van Aalten, D.M.F., Amadei, A., and Berendsen, H.J.C. (1998). Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. Proteins Struct. Funct. Genet. *31*, 116–127.

Ehrlich, L.P., Nilges, M., and Wade, R.C. (2005). The impact of protein flexibility on protein-protein docking. Proteins *58*, 126–133.

Elshorst, B., Hennig, M., Försterling, H., Diener, A., Maurer, M., Schulte, P., Schwalbe, H., Griesinger, C., Krebs, J., Schmid, H., et al. (1999). NMR solution structure of a complex of calmodulin with a binding peptide of the Ca^{2+} pump. Biochemistry *38*, 12320–12332.

Fernandez, A., and Berry, S. (2002). Extend of hydrogen-bond protection in folded protein: a constraint on packing architectures. Biophys. J. *83*, 2475–2481.

Fernandez, A., Colubri, A., and Berry, R.S. (2002a). Three-body correlations in protein folding: the origin of coopertivity. Physica A *307*, 235–259.

Fernandez, A., Sosnick, T.R., and Colubri, A. (2002b). Dynamics of hydrogen bond desolvation in protein folding. J. Mol. Biol. *321*, 659–675.

Fernandez, A., Rogale, K., Scott, R., and Scheraga, H.A. (2004). Inhibitor design by wrapping packing defects in HIV-1 proteins. Proc. Natl. Acad. Sci. USA *101*, 11640–11645.

Gerstein, M., and Krebs, W. (1998). A database of macromolecular motions. Nucleic Acids Res. *26*, 4280–4290.

Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. USA *80*, 3696–3700.

Grubmueller, H. (1995). Predicting slow structural transitions in macromolecular systems: conformational flooding. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics *52*, 2893–2906.

Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian dynamics of folded proteins. Phys. Rev. Lett. *79*, 3090–3093.

Hooft, R.W.W., Sander, C., and Vriend, G. (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins Struct. Funct. Genet. *26*, 363–376.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. (2001). Protein flexibility predictions using graph theory. Proteins Struct. Funct. Genet. *44*, 150–165.

Knegtel, R.M.A., Kuntz, I.D., and Oshiro, C.M. (1997). Molecular docking to ensembles of protein structures. J. Mol. Biol. *266*, 424–440.

Krebs, W., Alexandrov, V., Wilson, C., Echols, N., Yu, H., and Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins *48*, 682–695.

Lange, O.F., Schaefer, L.V., and Grubmueller, H. (2006). Flooding in GROMACS: accelerated barrier crossings in molecular dynamics. J. Comput. Chem. *27*, 1693–1702.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). Procheck: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. *26*, 283–291.

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. *7*, 306–317.

Love, S.G., Muir, T.W., Ramage, R., Shaw, K.T., Alexeev, D., Sawyer, L., Kelly, S.M., Price, N.C., Arnold, J.E., Mee, M.P., and Mayer, R.J. (1997). Synthetic, structural and biological studies of the ubiquitin system: synthesis and crystal structure of an analogue containing unnatural amino acids. Biochem. J. *323*, 727–737.

McGovern, S.L., and Shoichet, B.K. (2003). Information decay in molecular docking screens against holo, apo and modeled conformations of enzymes. J. Med. Chem. *46*, 2895–2907.

Meagher, K.L., and Carlson, H.A. (2004). Incorporating protein flexibility in structure-based drug design: using HIV-1 protease as a test case. J. Am. Chem. Soc. *126*, 13276–13281.

Müller, C.W., and Schulz, G.E. (1992). Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. J. Mol. Biol. *224*, 159–177.

Müller, C.W., Schlauderer, G.J., Reinstein, J., and Schulz, G.E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure *4*, 147–156.

Mustard, D., and Ritchie, D.W. (2005). Docking essential dynamics eigenstructures. Proteins *60*, 269–274.

Schlauderer, G.J., and Schulz, G.E. (1996). The structure of bovine mitochondrial adenylate kinase: comparison with isoenzymes in other compartments. Protein Sci. *5*, 434–441.

Schlauderer, G.J., Proba, K., and Schulz, G.E. (1996). Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP. J. Mol. Biol. *256*, 223–227.

Schlitter, M., Engels, M., and Kruger, P. (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. J. Mol. Graph. *12*, 84–90.

Seeliger, D., and de Groot, B.L. (2007a). tCONCOORD (http://www.mpibpc.mpg.de/groups/grubmueller/start/people/dseelig/tconcoord.html).

Seeliger, D., and de Groot, B.L. (2007b). Atomic contacts in protein structures: a detailed analysis of atomic radii, packing and overlaps. Proteins *68*, 565–601.

Spronk, C.A.E.M., Nabuurs, S.B., Bonvin, A.M.J.J., Krieger, E., Vuister, G.W., and Vriend, G. (2003). The precison of NMR structure ensembles revisited. J. Biomol. NMR *25*, 225–234.

Steuber, H., Zentgraf, M., Gerlach, C., Sotriffer, C.A., Heine, A., and Klebe, G. (2006). Expect the unexpected or caveat for drug designers: multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. J. Mol. Biol. *363*, 174–187.

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. *314*, 141–151.

Suhre, K., and Sanejouand, Y.H. (2004a). ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res. *32*, 610–614.

Suhre, K., and Sanejouand, Y.H. (2004b). On the potential of normal mode analysis for solving difficult molecular replacement problems. Acta Crystallogr. D Biol. Crystallogr. *60*, 796–799.

Teague, S.J. (2003). Implications of protein flexibility for drug discovery. Nat. Rev. Drug Discov. *2*, 527–541.

van der Vaart, A., and Karplus, M. (2005). Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. J. Chem. Phys. *122*, 114903.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. *8*, 52–56.

Wang, J., Truckses, D.M., Abildgaard, F., Dzakula, Z., Zolnai, Z., and Markley, J.L. (1997). Solution structures of staphylococcal nuclease from multidimensional, multinuclear NMR: nuclease-h124l and its ternary complex with Ca^{2+} and thymidine-3′,5′-bisphosphate. J. Biomol. NMR *10*, 143–164.

