Opinion TRENDS in Biotechnology Vol.23 No.9 September 2005

Simple but predictive protein models

Feng Ding and Nikolay V. Dokholyan

Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

The traditional approach to computational biophysics studies of molecular systems is brute force molecular dynamics simulations under the conditions of interest. The disadvantages of this approach are that the time and length scales that are accessible to computer simulations often do not reach biologically relevant scales. An alternative approach, which we call intuitive modeling, is hypothesis-driven and based on tailoring simplified protein models to the systems of interest. Using intuitive modeling, the length and time scales that can be achieved using simplified protein models exceed those of traditional molecular-dynamic simulations. Here, we describe several recent studies that signify the predictive power of simplified protein models within the intuitive-modeling approach.

Introduction

Since the determination of the structure of myoglobin by Kendrew in 1959, our view of the typical protein has evolved from that of a static structure to one of a dynamic ensemble of protein conformations [1]. Given that only some of these conformations are active functionally, the dynamics with which these multiple conformations interchange determines how often functionally important structures appear during the life of a protein. Despite recent revolutionary advances in experimental methods, we are limited in our ability to sample and decipher the structural and dynamic aspects of single molecules that are crucial for their biological function. Thus, there is a need for new, unorthodox techniques to uncover the fundamentals of molecular structure and interactions. These unorthodox techniques might come from computational approaches that balance the physical realism of molecular models and their computational effectiveness.

The traditional approach to computational biophysics studies of molecular systems is brute force molecular dynamics simulations of these systems under conditions of interest. With such an approach, the protein models – geometries and the interaction parameters – that underlie computer simulations are developed separately from studies of the actual biophysical systems. The thrust of this traditional approach is to develop an 'ideal' protein model that can be applied to any biophysical system. The advantages of this approach are its rigor and cross-system transferability; the disadvantages are that the time- and length scales that are accessible to computer simulations do not, in many cases, reach biologically-relevant scales (Figure 1), and that the interaction parameters that are derived are inaccurate. Molecular dynamics simulations have been pivotal in studying protein dynamics [2–5]. However, the intrinsic limitation of traditional molecular dynamics is its inability to span time scales from femtoseconds (10^{-15}) to seconds. Sampling this time scale, especially the milliseconds to seconds range, is often required to decipher the dynamics of proteins and protein complexes. Computational approaches that use simplified protein models offer a unique opportunity to study large-scale protein folding dynamics and aggregation, and offer experimentally testable predictions.

An alternative approach, which we call intuitive modeling, is hypothesis-driven and based on tailoring simplified protein models to the systems of interest. The disadvantages are that these models do not always transfer across systems, that the quality of biophysical studies depends on human intuition and that these models rely on experimental verification. The advantage of intuitive modeling simplified protein models is that the length and time scales exceed those of traditional molecular dynamics simulations. We describe the use of simplified protein models in conjunction with the rapid simulations methodology, discrete molecular dynamics (DMD). Despite the use of DMD in simulating polymer fluids [6,7], single homopolymers [7,8], proteins [9–25], protein aggregates [13,17,23,24,26], and gases and liquids [27,28], we believe it is significantly underutilized in the molecular-modeling field.

Importantly, intuitive modeling refers to a particular bottom-up, molecular-modeling approach to capture the driving forces that govern dynamics of complex systems, in which the brute-force approach is often inapplicable. This approach was adopted in physics in the mid-20th century to study complex systems. Biological systems are complex and the identification of driving forces for a particular system often results from physical and/or biological intuition. In this respect, DMD is a fast dynamic engine that we use for modeling. Although DMD is designed specifically to tailor interaction potentials, in principle, the intuitive modeling approach can also be used in other molecular dynamics (MD) engines.

DMD

Several protein models that have striking predictive power have been developed in the last six years [15,17,18,22,25]. In most cases, predictions made using these models were not validated by experiments at the time of publication. However, recent experimental data reveals that these predictions are true. These models are based on DMD [29–31], which incorporates a radically different philosophy of molecular dynamics simulations than has been

Corresponding author: Dokholyan, N.V. (dokh@med.unc.edu).
Available online 20 July 2005

 0167-7799/$ - see front matter doi:10.1016/j.tibtech.2005.07.001

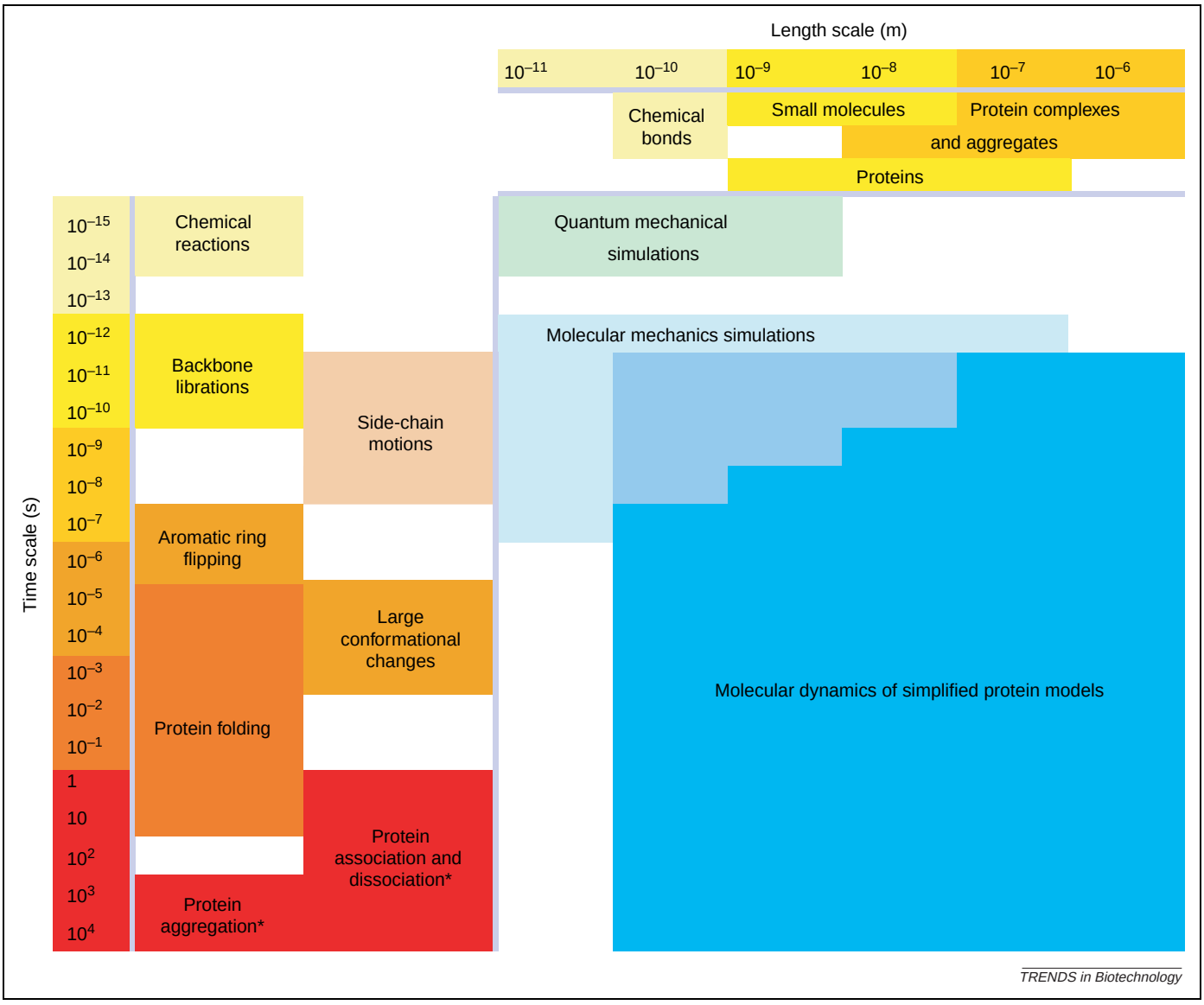

Figure 1. Time and length scales. The left of the figure includes examples of processes that occur at different time scales. Protein association, dissociation and aggregation (*) are concentration-dependent and might span longer times than presented here. Examples of molecular sizes are at the top. Three simulation approaches, quantum mechanical, molecular mechanics and molecular dynamics simulations of simplified protein models, outline the time and length scales that are accessible to these approaches. The area of the time–length scale that corresponds to molecular dynamics simulations signifies a range of simplified protein models that are used in simulations. Thus, accessing all the scales outlined might require the use of several, mutually-consistent simplified protein models.

used previously in the field of molecular modeling. Most molecular dynamics simulations are based on iterative solutions of the Newtonian equations of motions. The physical realism and the speed of these simulations depend heavily on the force-field parameters of atomic interactions. Despite striking progress in determining these parameters and many successes in biophysical studies using these force-fields, the speed and the accuracy remain a paramount challenge. By contrast, DMD simulations do not rely on finely-tuned physical force-field parameters. Instead, they mimic natural interactions by assigning the simplest interaction potential between atoms, a 'square-well' potential. A square-well potential of interaction is a step-wise potential, so its derivatives are either zero (acceleration is zero) or infinity (collision occurs with instantaneous velocity change). Thus, DMD simulations are based on iterative solutions of ballistic equations of motions; such simplification makes these simulations sufficiently rapid to access the dynamics of large molecular complexes on a time-scale of seconds. Although the actual speed up of DMD compared with traditional MD varies from system to system, it can be 10^8–10^{10}-times faster.

The difference in philosophy of DMD versus traditional molecular dynamics simulations is apparent in the way biological problems are approached. In traditional molecular dynamics simulations, we aim to predict either structures or dynamics of the states of interest *ab initio*, whereas in DMD we make a hypothesis of the kinds of forces that drive the system, mimic these forces by tuning square-well potentials, and produce molecular states of interest with DMD. The latter states, and their structure and dynamics, offer experimentally-testable hypotheses. Experimental validations of such hypotheses herald the

viability of the DMD philosophy to decipher complex biological systems. In the following section we describe five such studies.

Transition-state ensemble

The protein transition-state ensemble is a set of conformations adopted by proteins upon the folding–unfolding transition that are characterized by the largest free energy on the protein folding–unfolding path. The thermodynamic manifestation of the transition states, which constitutes the top of the free-energy barrier, is the minimal probability of observing them on the protein folding–unfolding path. However, these states determine the folding kinetics of two-state proteins because the expectation time for crossing the free-energy barrier is related to the height of the barrier. A method to probe the role of various amino acids in the formation of the transition-state ensemble has been proposed by Fersht [32,33], who suggests measuring the impact of amino acid substitutions on the free energy gap between transition and unfolded states. The value that characterizes the impact of an amino acid substitution is called an ϕ-value. This normally ranges between 0 and 1, indicating no to high impact. Although ϕ-values provide invaluable information about transition-state ensembles, protein-engineering experiments are difficult, error-prone and have intrinsic limitations. One such limitation is that many glycine residues cannot be probed adequately because they lack side-chains. As a result, because several experimental ϕ-values are not reasonable (e.g. $\phi \ll 0$ and $\phi \gg 1$), other experiments are needed to further elucidate the role of specific residues in the protein-folding kinetics.

We have determined ϕ-values for an experimentally-characterized Src homology 3 (SH3) domain [15]. Correlating all simulations and experimental ϕ-values, including those that are negative experimentally, we find a statistically significant correlation. We also predicted that the two most kinetically important residues in folding the SH3 domain are L24 and G54. Experimentally, L24 is important kinetically. Although G54 was determined experimentally to have a negative f-value, we predict that it is the most important residue. Recent experimental work by Di Nardo *et al.* [34] recognizes that G54 is important for SH3 folding kinetics, and the authors suggest that there is evolutionary pressure to preserve G54 because of this role [34].

Using an alternative approach, Vendruscolo *et al.* [35] related the fraction of native contacts formed by specific residues to their ϕ-values measured in protein-engineering experiments [32,33]. Using only three, high ϕ-values for protein acylphosphatase as experimental constraints, the authors have reconstructed the transition-state ensemble. The computed ϕ-values for other residues from the reconstructed ensemble agree with the experimental values.

SH3-domain swapping

Domain swapping between similar proteins is an interesting phenomena. In domain swapping, two complementary parts of similar proteins are exchanged to form an oligomer that is comprised of native-like proteins. Eisenberg and co-workers [36] have proposed that domain swapping is a generic mechanism for protein oligomerization. In our attempts to reproduce aggregation of the SH3 domain, which has been observed experimentally, we performed DMD simulations of the Src SH3 model [17] in which we assigned interactions between amino acids that are in proximity in the native state to bias the SH3 domain to the native states at low simulation temperatures. This is termed the Gō model [37]. In addition, we assigned similar interactions between similar proteins, thus, if the interaction potential between residue A and residue B is ε_{AB}, then the interaction potential between residue A of one protein and residue B' of another is $\varepsilon_{AB'} = \varepsilon_{AB}$. DMD simulations of this simple protein model results in aggregates that are consistent with all known experimental data. However, to our surprise, we found that SH3 featured two forms of dimers on aggregation; an 'open' dimer that has higher energy and is capable of further assembly, and a 'closed' dimer that has lower energy and does not assemble further because it features domain-swapped SH3. The RT loops (structural elements of SH3) were exchanged between two proteins, resulting in a stable dimer with two, complementary, intertwined domains.

In crystallographic studies of the Eps8 SH3 domain, Kishan *et al.* [38] uncovered a domain swapped Eps8 SH3 that is similar visually to our domain swapped Src SH3 dimer (Figure 2) [17], and a root-mean-square deviation between the predicted and observed structures of <6 Å determined by Pymol (http://www.pymol.org). Similarity between predicted and experimental values for domain-swapped SH3 homologs indicates that the principal assumptions in our model reflect actual interactions between similar proteins. Furthermore, Yang *et al.* [39] have recently reproduced the domain-swapped dimer structures of the SH3 domain using molecular-dynamics simulations of a simplified protein model and Gō interaction potentials that are similar to those used by Ding *et al.* [17].

Salt bridges modulate the folding of superoxide dismutase

Misfolding and aggregation of copper, zinc superoxide dismutase (SOD1) are implicated in the familial form of the neurodegenerative disease amyotrophic lateral sclerosis. To understand how misfolding occurs, we studied the folding of SOD1 using DMD simulations of two models of interactions [18]. In the first model, we assigned interactions between amino acids based on the native structure of SOD1 (the 'unscaled' Gō model). In the second model, we scaled the square-well interactions between amino acids according to the interaction energies derived from a pair-wise decomposition of the thermodynamic stability of SOD1 (the 'scaled' Gō model). The decomposition was performed using traditional molecular-dynamics simulations. The difference between the scaled and unscaled models revealed 16 long-range electrostatic interactions on the surface of SOD1, scaling of which makes SOD1 fold cooperatively without stable intermediates. These salt-bridge contacts form between oppositely charged residues on the surface of SOD1. Based on these results, we predicted that changes in pH that lead to titration of salt-bridge residues should destabilize SOD1 and increase the

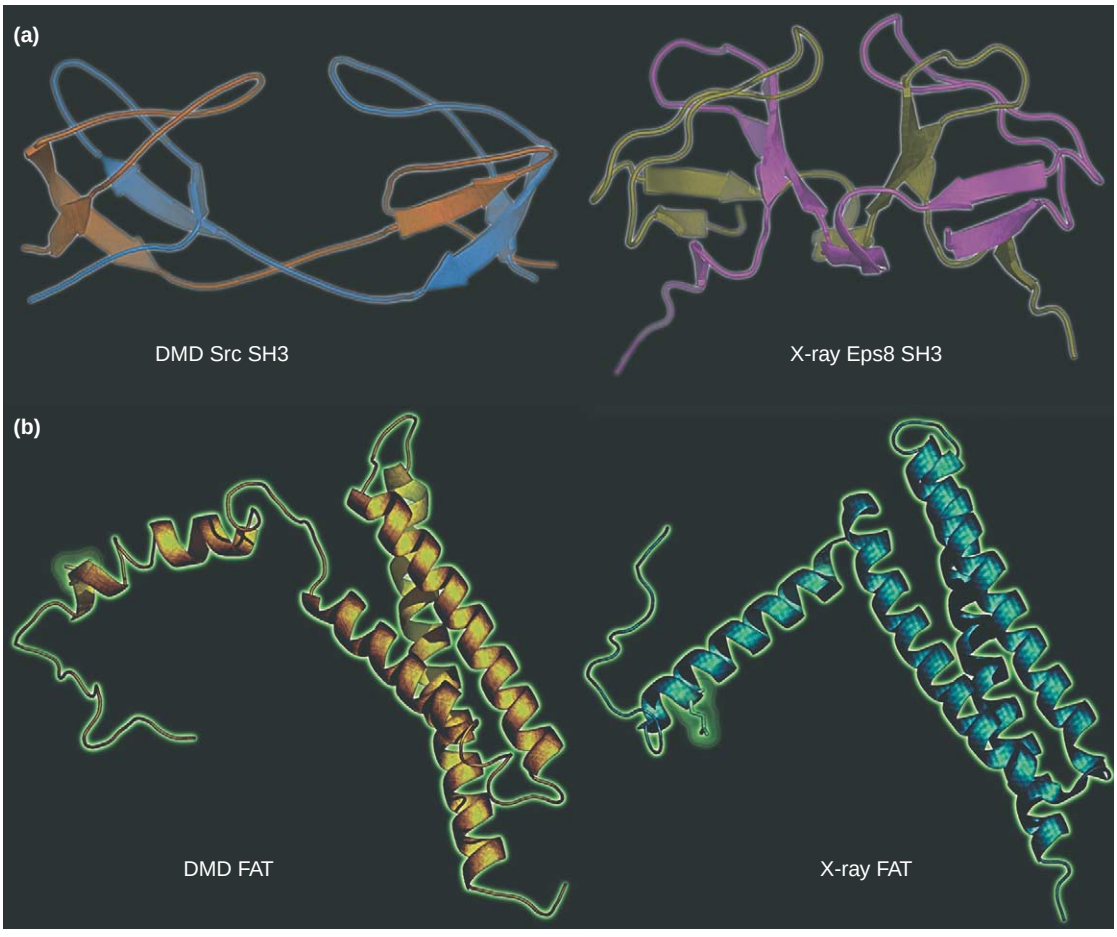

Figure 2. DMD simulations and experiments. **(a)** Comparison of the DMD simulations of Src SH3 domain swapping (left) with the x-ray crystallographic structure of the Eps8 SH3 domain (right). The structure of Eps8 SH3 is homologous but not identical to Src SH3. **(b)** The intermediate structure of the FAT derived from DMD simulation (left) compared with the x-ray crystallographic structure of the domain-swapped dimer (right).

population of folding intermediates. Recent experimental studies [40] and our observations of the pH-dependent destabilization of SOD1 [41] confirm this prediction.

Catching evasive protein states

Recently, we have developed a method to reconstruct functionally important protein states that are 'invisible' to nuclear magnetic resonance (NMR). This method is based on a set of computational approximations led by experimental data obtained in a kinetically-accessible regime. We used NMR-derived amide–hydrogen-exchange-protection factors, which measure amino acid (backbone amide proton) solvent protection, to bias rapid DMD simulations. This unveiled an ensemble of protein conformations that is consistent with observed protection factors. We have applied our method to the targeting domain (FAT) of focal adhesion kinase (FAK). FAK is a tyrosine kinase that acts downstream of growth factor and integrin receptors. FAK is activated by localization to focal adhesions, which, in turn, leads to phosphorylation and binding to cell-signaling proteins. The phosphorylation-competent domain states of FAT are weakly populated and, therefore, are not amenable to structural determination by NMR. Using the developed method, we have identified these weakly populated intermediate states of the FAT domain and found that Y926, which belongs to helix 1 of the four-helix bundle that comprises the FAT domain, is phosphorylated in these states [22]. Experimental observation of domain-swapped FAT dimers, in which helix 1 is exchanged with helix 1 of another FAT molecule, provides experimental evidence of intermediate domain states of FAT (Figure 2). Based on the computationally reconstructed FAK intermediates, we have suggested mutations that should alter the population of these conformations and, thus, FAK function. Recently, Dr Michael Schaller demonstrated *in vitro* that these mutants do alter FAK function (M. Schaller pers. commun.).

A similar approach has also been proposed by Vendruscolo *et al.* [42], who has related the number of native contacts per residue to its equilibrium hydrogen-exchange-protection factor and reconstructed the ensemble of protein conformations that are consistent with the experimental condition. The reconstructed conformation ensemble of α-lactalbumin features rare fluctuations around the native state.

DMD simulations of the folding of Trp-cage

To validate the reliability and accuracy of DMD, we have used DMD simulations to fold a small, 20-residue protein Trp-cage from a fully extended conformation using a simplified interaction model between amino acids. Importantly, the amino acid interaction model is not based on

knowledge of the native structure of Trp-cage, but has just three key, stabilizing interactions: the aromatic proline; the salt-bridge; and the hydrogen-bond interactions. We have demonstrated the ability of the Trp-cage model to consistently reach conformations within a 2 Å backbone root-mean-square distance from the corresponding NMR structures [25], which is similar to recent molecular dynamics simulations of this protein [43–48]. The minimum root-mean-square distance of Trp-cage conformations in the simulation is <1.0 Å.

Concluding remarks

The essence of this philosophy of approaching biological problems is its dependence on scientific intuition, which, following translation to molecular models, allows the formulation of hypotheses that are testable experimentally. The validation of these hypotheses, and further iterative refinement of our biophysical intuition and molecular models, is crucial for the intuitive modeling approach. The advantage of this approach is the ability to ascertain time and length scales that are not amenable to traditional molecular dynamics approaches.

Acknowledgements

We thank B. Kuhlman for suggestions. This work is supported in part by Muscular Dystrophy Association grant MDA3720, Research Grant No. 5-FY03–155 from the March of Dimes Birth Defect Foundation and the UNC/IBM Junior Investigator Award.

References

1 Lindorff-Larsen, K. *et al*. (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433, 128–132
2 Pearlman, D.A. *et al*. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91, 1–41
3 Brooks, B.R. *et al*. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4, 187–217
4 Mann, G. *et al*. (2002) The Sigma MD program and a generic interface applicable to multi-functional programs with complex, hierarchical command structure. In *Computational Methods for Macromolecules: Challenges and Applications - Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modelling* (Schlick, T. and Gan, H.H., eds), pp. 129–145, Springer-Verlag
5 van Gunsteren, W.F. and Berendsen, H.J.C. (1990) Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angew. Chem. Int. Ed.* 29, 992–1023
6 Smith, S.W. *et al*. (1997) Molecular dynamics for polymeric fluids using discontinuous potentials. *J. Comput. Phys.* 134, 16–30
7 Zhou, Y.Q. *et al*. (1997) Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of systems with square-well interactions. *J. Chem. Phys.* 107, 10691–10708
8 Dokholyan, N.V. *et al*. (2002) Glassy behavior of a homopolymer from molecular dynamics simulations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65, 030801
9 Dokholyan, N.V. *et al*. (1998) Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.* 3, 577–587
10 Zhou, Y. and Karplus, M. (1999) Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. *J. Mol. Biol.* 293, 917–951
11 Zhou, Y. and Karplus, M. (1999) Interpreting the folding kinetics of helical proteins. *Nature* 401, 400–403
12 Dokholyan, N.V. *et al*. (2000) Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* 296, 1183–1188
13 Smith, A.V. and Hall, C.K. (2001) Assembly of a tetrameric α-helical bundle: computer simulations on an intermediate-resolution protein model. *Proteins* 44, 376–391
14 Dokholyan, N.V. *et al*. (2002) Topological determinants of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 99, 8637–8641
15 Ding, F. *et al*. (2002) Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys. J.* 83, 3525–3532
16 Borreguero, J.M. *et al*. (2002) Thermodynamics and folding kinetics analysis of the SH3 domain from discrete molecular dynamics. *J. Mol. Biol.* 318, 863–876
17 Ding, F. *et al*. (2002) Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J. Mol. Biol.* 324, 851–857
18 Khare, S.D. *et al*. (2003) Folding of Cu, Zn superoxide dismutase and familial amyotrophic lateral sclerosis. *J. Mol. Biol.* 334, 515–525
19 Ding, F. *et al*. (2003) Mechanism for the alpha-helix to beta-hairpin transition. *Proteins* 53, 220–228
20 Dokholyan, N.V. *et al*. (2003) Identifying importance of amino acids for protein folding from crystal structures. *Methods Enzymol.* 374, 616–638
21 Borreguero, J.M. *et al*. (2004) Multiple folding pathways of the SH3 domain. *Biophys. J.* 87, 521–533
22 Dixon, R.D.S. *et al*. (2004) New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate. *Structure* 12, 2161–2171
23 Peng, S. *et al*. (2004) Discrete molecular dynamics simulations of peptide aggregation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69, 041908
24 Urbanc, B. *et al*. (2004) Molecular dynamics simulation of amyloid beta dimer formation. *Biophys. J.* 87, 2310–2321
25 Ding, F. *et al*. (2005) Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* 88, 147–155
26 Nguyen, H.D. and Hall, C.K. (2004) Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16180–16185
27 Franzese, G. *et al*. (2001) Generic mechanism for generating a liquid-liquid phase transition. *Nature* 409, 692–695
28 Skibinsky, A. *et al*. (1999) Quasicrystals in a monodisperse system. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 60, 2664–2669
29 Alder, B.J. and Wainwright, T.E. (1959) Studies in molecular dynamics. I. General method. *J. Chem. Phys.* 31, 459–466
30 Allen, M.P. and Tildesley, D.J. (1987) *Computer Simulations of Liquids*, Clarendon Press
31 Grosberg, A.Y. and Khokhlov, A.R. (1997) *Giant Molecules*, Academic Press
32 Matouschek, A. *et al*. (1990) Transient folding intermediates characterized by protein engineering. *Nature* 346, 440–445
33 Matouschek, A. *et al*. (1989) Mapping the transition-state and pathway of protein folding by protein engineering. *Nature* 340, 122–126
34 Di Nardo, A.A. *et al*. (2004) Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7954–7959
35 Vendruscolo, M. *et al*. (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature* 409, 641–645
36 Bennett, M.J. *et al*. (1994) Domain swapping - entangling alliances between proteins. *Proc. Natl. Acad. Sci. U. S. A.* 91, 3127–3131
37 Go, N. and Abe, H. (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers* 20, 991–1011
38 Kishan, K.V.R. *et al*. (2001) Effect of pH and salt bridges on structural assembly: Molecular structures of the monomer and intertwined dimer of the Eps8 SH3 domain. *Protein Sci.* 10, 1046–1055
39 Yang, S. *et al*. (2004) From the cover: domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. U. S. A.* 101, 13786–13791
40 DiDonato, M. *et al*. (2003) ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. *J. Mol. Biol.* 332, 601–615
41 Khare, S.D. *et al*. (2004) The rate and equilibrium constants for a

multistep reaction sequence for the aggregation of superoxide dismutase in amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15094–15099

42 Vendruscolo, M. *et al*. (2003) Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *J. Am. Chem. Soc.* 125, 15686–15687

43 Simmerling, C. *et al*. (2002) All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* 124, 11258–11259

44 Snow, C.D. *et al*. (2002) The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* 124, 14548–14549

45 Zagrovic, B. and Pande, V. (2003) Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comput. Chem.* 24, 1432–1436

46 Zhou, R. (2003) Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13280–13285

47 Pitera, J.W. and Swope, W. (2003) Understanding folding and design: replica-exchange simulations of "Trp-cage" fly miniproteins. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7587–7592

48 Chowdhury, S. *et al*. (2003) *Ab initio* folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* 327, 711–717