PROTEINS: Structure, Function, and Genetics 35:447–452 (1999)

Correlation Between Knowledge-Based and Detailed Atomic Potentials: Application to the Unfolding of the GCN4 Leucine Zipper

Debasisa Mohanty,[1] Brian N. Dominy,[1] Andrzej Kolinski,[1,2] Charles L. Brooks III,[1]* and Jeffrey Skolnick[1]*
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **The relationship between the unfolding pseudo free energies of reduced and detailed atomic models of the GCN4 leucine zipper is examined. Starting from the native crystal structure, a large number of conformations ranging from folded to unfolded were generated by all-atom molecular dynamics unfolding simulations in an aqueous environment at elevated temperatures. For the detailed atomic model, the pseudo free energies are obtained by combining the CHARMM all-atom potential with a solvation component from the generalized Born, surface accessibility, GB/SA, model. Reduced model energies were evaluated using a knowledge-based potential. Both energies are highly correlated. In addition, both show a good correlation with the root mean square deviation, RMSD, of the backbone from native. These results suggest that knowledge-based potentials are capable of describing at least some of the properties of the folded as well as the unfolded states of proteins, even though they are derived from a database of native protein structures. Since only conformations generated from an unfolding simulation are used, we cannot assess whether these potentials can discriminate the native conformation from the manifold of alternative, low-energy misfolded states. Nevertheless, these results also have significant implications for the development of a methodology for multiscale modeling of proteins that combines reduced and detailed atomic models. Proteins 1999;35:447–452.** © **1999 Wiley-Liss, Inc.**

Key words: reduced protein model; CHARMM/GCN4 leucine zipper; protein unfolding; knowledge-based potentials

INTRODUCTION

The ab initio prediction of the three-dimensional structure of a protein, starting from its amino acid sequence, has long been a major challenge[1–4] for computational biophysics. The majority of such ab initio approaches are based on the hypothesis that the native structure of a protein corresponds to the global free energy minimum of the protein–solvent system.[5,6] Hence, these computational methods construct a potential energy function[4,7–10] that describes the interactions between various constituents of the protein–solvent system. However, depending on the application, different methods employ different levels of detail to describe the energy or pseudo free energy function. The most straightforward approaches use a detailed atomic representation for the potential energy function.[7,8] Here, both the protein and the solvent are described in atomic detail, and the laws of physical chemistry govern interactions between the constituent atoms. The time evolution of the system is simulated using molecular dynamics (MD) techniques.[11,12] In principle, it is possible to simulate the folding process of a protein using detailed atomic potentials and molecular dynamics techniques provided that the simulation can cover a sufficiently long time scale. However, with the computing power available today, with very few exceptions, only simulations of nanosecond time scales are possible for real protein–water systems.[13] On the other hand, proteins fold on the time scale of milliseconds to seconds[14]; therefore, it has not been possible to fold even relatively small proteins using detailed atomic potentials. Nevertheless, MD with detailed atomic potentials has been successful in simulating fast events involving local or small distance structural rearrangements. For example, ab initio folding of short peptides has been possible in a few cases,[15–18] with the most recent example being the folding of the villin headpiece from the denatured state.[19] By starting from two parallel α helices aligned with the correct registration,[20] it has also been possible to obtain the detailed structure of coiled coils. Furthermore, using high-temperature MD and detailed atomic potentials, unfolding simulations have provided interesting insights into possible unfolding pathways.[21–24] Using detailed atomic potentials and umbrella sampling, information about the folding landscape has also been obtained by carrying out free-energy calculations[25–27] along an assumed reaction coordinate. However, due to time scale limitations, detailed atomic models cannot be routinely used for ab initio folding or for the simulation of folding thermodynamics. Thus, complementary approaches are required.

Grant sponsor: National Institutes of Health; Grant number: P41 RR12255.

*Correspondence to: Dr. Jeffrey Skolnick and Dr. Charles L. Brooks III, Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037. E-mail: skolnick@scripps.edu *or* E-mail: brooks@scripps.edu

Received 12 October 1998; Accepted 11 February 1999

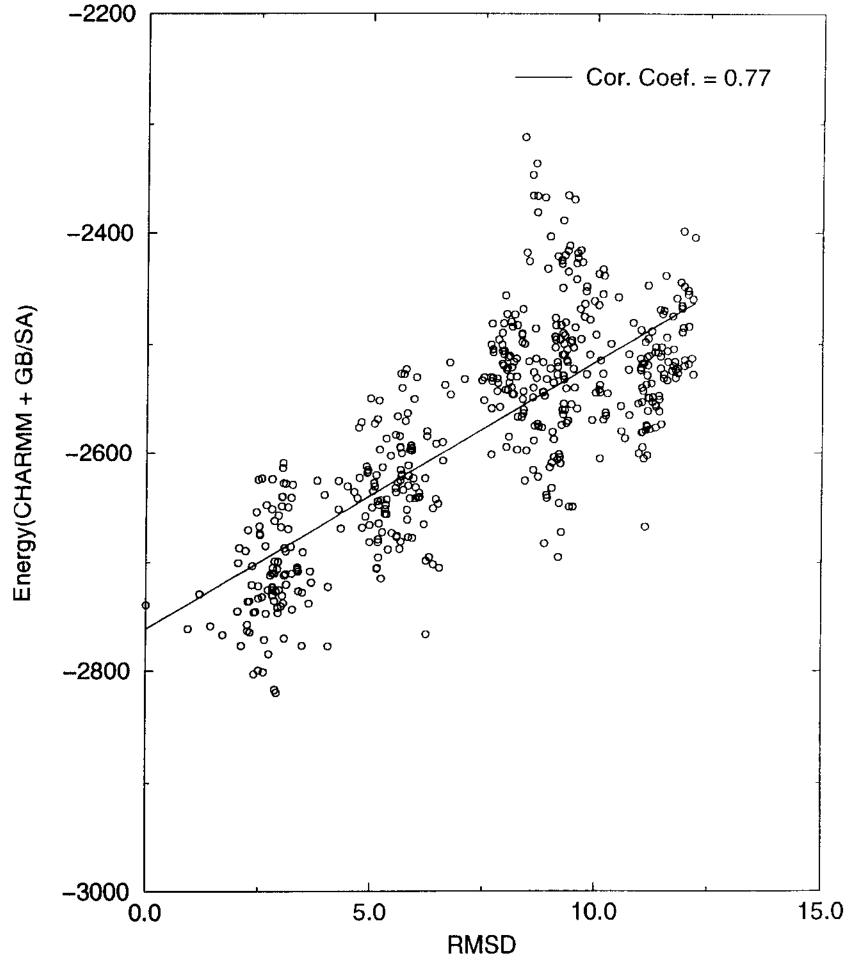

Fig. 1. Plot of the all-atom energy versus the C_α RMSD from native for a set of GCN4-lz structures.

To access longer time scales, reduced protein models[4,9,10,28] have been developed. In these reduced models,[9,28] most atomic details of the polypeptide chain are completely ignored. Rather, each residue is represented by two interaction sites and corresponds, for example, to the C_α and/or the center of mass of the amino acid side chain. Similarly, explicit solvent is completely ignored. The interactions between the interaction sites in reduced models are described by a knowledge-based potential[4,9,10] derived from a statistical analysis of a database of known native structures of proteins and that implicitly describes the effects of solvent. In these reduced models, the interaction sites are either confined to a set of lattice points[9,28] or may lie in continuous space.[29–31] A lattice representation has the advantage that many quantities can be precomputed, and hence the computational speed of a simulation can be considerably increased. Using a lattice model and knowledge-based potentials, the ab initio folding of several proteins has been achieved.[28,32,33] Furthermore, insights into the folding thermodynamics of a number of real systems have been obtained.[34–36] A recent statistical mechanical analysis of the thermodynamics of the conformational transition of the GCN4 leucine zipper, GCN4-lz, indicates that these knowledge-based potentials can describe a number of conformational properties of both the native and denatured states and can explain the physical basis of the experimentally observed two-state folding transition.[36]

Given any reduced model, there is always an inherent limitation to its accuracy. In a small number of cases, this accuracy has been improved when structures obtained from lattice simulations are further refined[32] using a detailed atomic model. Hence, it might be possible to develop hybrid models that combine the advantages of reduced and detailed atomic models. For example, one can use a reduced model in the early stages of folding where one has to distinguish between low-energy compact states and the very large number of unfolded states. Then, a detailed atomic model could be used for distinguishing the native state from other compact alternatives. However, a minimal requirement for success is that the two potentials similarly rank a set of conformers according to their energies. Therefore, in the present work, we present a comparison of an ensemble of structures for GCN4-lz. This molecule has been chosen because it has been extensively

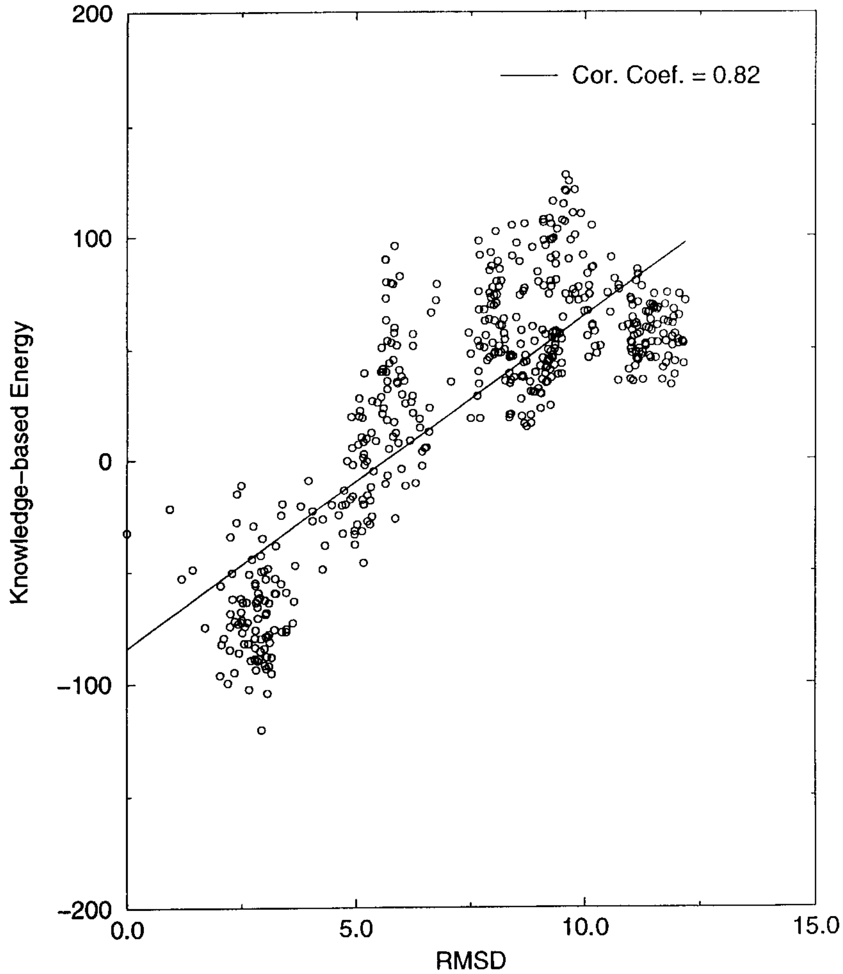

Fig. 2. Plot of the knowledge-based energy versus the C_α RMSD from native for a set of GCN4-lz structures.

studied using both reduced[32,36,37] and detailed atomic models.[20,38]

This comparison between knowledge-based potentials and detailed atomic potentials is also important from another point of view. Recently, questions have been raised in the literature[39,40] about whether or not knowledge-based potentials have any physical basis. Hence, one might also ask whether or not these knowledge-based potentials, derived from native structures of proteins, are applicable to nonnative states and whether or not questions related to the thermodynamics of folding can be addressed. Since the detailed atomic potentials in CHARMM are based on fundamental laws of physics and chemistry, they should be better able to describe both native and denatured states; this is consistent with a wide body of literature.[13,41,42] A good correlation between the two potentials would indicate that knowledge-based potentials could also describe the properties of native and nonnative states.

METHOD

To investigate the correlation between the two classes of potentials, we need an ensemble of native and nonnative structures of GCN4-lz. These could be generated from a reduced or detailed atomic model simulation. The disadvantage of the former is that one must then reconstruct a detailed atomic model; this could introduce errors that obscure the correlation between the two classes of energetic terms. On the other hand, the energetic terms in the reduced model could equally well be applied to an atomic model, and this is the strategy we shall pursue here. Hence, the ensemble of structures was generated by starting from the native X-ray structure[43] and by carrying out MD simulations at elevated temperatures. Both the protein and water were treated at atomic detail, and the CHARMM force field[7] was used to describe the interactions. From the MD trajectory, a set of structures for GCN4-lz was obtained that varied from the completely folded state to the completely unfolded state. The degree of unfolding was measured by calculating the root-mean-square deviation (RMSD) of the C_α atoms from the native GCN4-lz. The knowledge-based energies for this set of structures were calculated from the coordinates of the C_α's and the centers of mass of the side chains. The interaction scheme and force field parameters for the calculation of the knowledge-based energy were the same as in our recent

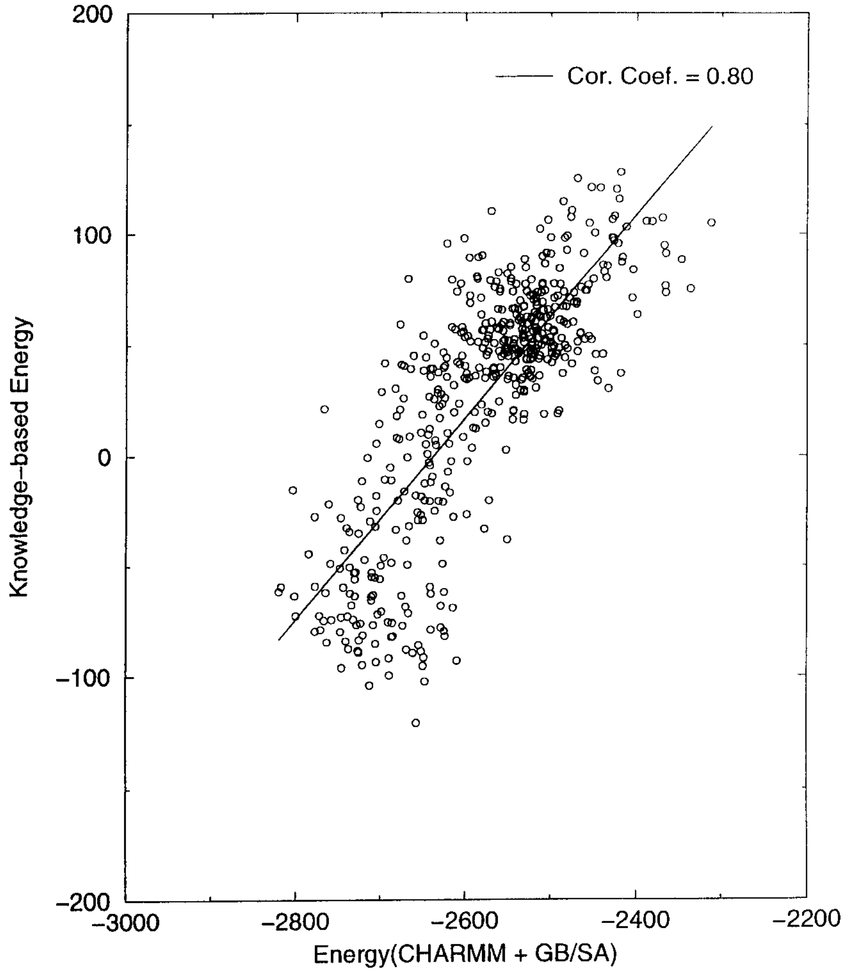

Fig. 3. Plot of knowledge-based energy versus all-atom energy for a set of GCN4-lz structures.

study of the GCN4-lz folding thermodynamics and correctly account for the amount of secondary structure in the denatured state.[36] In the detailed atomic model, the energy of a given conformational state of GCN4-lz is given by the sum of the intrinsic potential energy and the solvation energy arising from protein–water and water–water interactions. Since the protein–water—and particularly water–water—interactions exhibit large fluctuations during MD simulations at constant temperature, these components of the all-atom energy were estimated by adding a generalized Born/accessible surface area, GB/SA, term[44] to the CHARMM all-atom protein energy.[7] This GB model has been specifically parameterized to reproduce electrostatic forces and energies for proteins and nucleic acids with the CHARMM force field.[45] After obtaining the knowledge-based energies and all-atom energies, their respective correlations with the backbone C_α RMSD from native as well as the correlation between the reduced and atomic energies were calculated.

RESULTS

Figure 1 shows the correlation between the CHARMM plus GB/SA energy with the RMSD from native GCN4-lz. As can be seen, there is a quite reasonable correlation with a correlation coefficient of 0.77. This indicates that the solvation energy term, computed using GB/SA, is able to mimic the effect of explicit waters and that the detailed atomic potential is able to discriminate these nativelike structures from other nonnative states that are generated in the GCN4 unfolding simulation. Since other alternative low-energy states are not evaluated here, we cannot tell whether or not either the detailed atomic or reduced model potentials can successfully discriminate against them. Similarly, Figure 2 shows a plot of the knowledge-based energy versus RMSD for the same set of GCN4-lz structures, with the correlation coefficient being 0.82. This indicates that the knowledge-based potential can also distinguish the native state from other nonnative states. In both cases, it should be observed that below about 3 Å rms, neither potential has a particularly strong correlation with the rms of the particular conformation from native. In particular, the native conformation is not the lowest energy state. This most likely reflects the resolution of both classes of potentials as well as their inaccuracies.

The above results do not directly answer the question of whether or not detailed atomic potentials and knowledge-based potentials are correlated for the set of generated conformations. In Figure 3, we address this point by plotting the detailed atomic potential (CHARMM + GB/SA) versus the knowledge-based potential. These two quantities are highly correlated, with a correlation coefficient of 0.80. Moreover, the correlation extends from folded to unfolded conformations. This provides direct evidence that knowledge-based potentials can be used to describe many features of nonnative states of proteins, even though they were derived from a database of native structures. The good correlation suggests that it should indeed be possible to use hybrid models for protein structure prediction.

In a recent work, O'Donoghue and Nilges[46] concluded that their knowledge-based potential describing residue burial and pair interactions could not discriminate between native and nonnative states of coiled coils. Rather it must be combined with an all-atom protein backbone potential to achieve such discrimination. Although our knowledge-based potential contains burial and pair interaction contributions, it also includes a backbone term that describes local interactions.[9] Thus, their results and ours are entirely consistent.

CONCLUSION

A comparison of a knowledge-based potential and detailed atomic potential indicates that there is a significant correlation between them that extends over the whole range of conformations, from folded to unfolded states. Since it is generally believed that detailed atomic potentials can describe the properties of folded and unfolded states of proteins, the excellent correlation suggests that knowledge-based potentials can also describe the properties of the full range of conformational states. Hence, statistical thermodynamic calculations can be carried out using knowledge-based potentials to gain insight into folding thermodynamics. The results presented here also demonstrate that it should be feasible to combine knowledge-based potentials with detailed atomic potentials in order to develop hybrid models for protein structure prediction. However, these observations are based on our analysis of the GCN4-lz unfolding trajectory. A similar comparison between reduced and detailed atomic models must be carried out for other proteins to demonstrate the generality of these conclusions. These studies are now under way.

Finally, we comment on a recent study from Hermans and coworkers[47] that uses potentials similar in sprit to those employed here to distinguish folded from misfolded protein structures using the Holm and Sander EMBL database of misfolded proteins.[48] In the study of Hermans and coworkers, 12 of the 25 structures from this database were examined using a protocol that relied on molecular dynamics using explicit solvent to estimate configurational entropy and an implicit solvent model to provide an energetic assessment of free energy differences between the misfolded and folded structures. In their study, they correctly identified all of the structures they examined. In recent work from our laboratory, using the energy function described here (CHARMM + GB/SA), we also explored the use of implicit solvent models to identify the correct fold from misfolded proteins for the entire Holm and Sander database (B.N. Dominy and C.L. Brooks, unpublished data). Our findings parallel those of Hermans in that we correctly identify 24 of the 25 structures, but fail for the specific case of a protein containing an iron sulfur cluster (not included in protein structures for fold assessment). These findings are similar to those from Hermans and coworkers and further validate the energy functions used here. Furthermore, given that our studies used only minimization to "prepare" the protein structures before fold assessment suggests, as noted from the Hermans findings, that configurational entropy differences between this set of misfolded proteins is not a determining factor in identifying the correctly folded protein. However, we note that the presence/absence of cofactors and or metal ligands in calculations of energetics can complicate the identification of correct folds. Thus, studies such as those of Hermans and coworkers, as well as our recent unpublished work, provide further evidence that implicit solvent models can be of great utility in fold identification. Consequently, a more complete understanding of the relationship between these potentials and the reduced representation, generally lattice-based, potentials as examined in this article provides a useful path for integration of modeling tools for structure prediction.

ACKNOWLEDGMENTS

D.M. thanks Dr. M. Vieth and Dr. L. Zhang for valuable discussions.

REFERENCES

1. Creighton TE. Protein folding. Biochem J 1990;270:131–146.
2. Levitt M. Protein folding. Curr Opin Struct Biol 1991;1:224–229.
3. Dill KA. Folding proteins: finding a needle in a hay-stack. Curr Opin Struct Biol 1993;3:99–103.
4. Friesner RA, Gunn JR. Computational studies of protein folding. Annu Rev Biophys Biomol Struct 1996;25:315–342.
5. Anfinsen CB. Principles that govern folding of protein chains. Science 1973;181:223–230.
6. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. Adv Protein Chem 1988;39:191–235.
7. Brooks BR, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy minimizations and dynamics calculations. J Comp Chem 1983;4: 187–217.
8. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1995;117:5179–5197.
9. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
10. Sippl M. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
11. Karplus M, Petsko GA. Molecular dynamics simulations in biology. Nature 1990;347:631–639.
12. van Gunsteren WF, Weiner PK. Computer simulations of biomolecular systems: theoretical and experimental applications. Leiden: ESCOM Science Publishers, 1989.
13. Brooks CL III. Methodological advances in molecular dynamics simulations of biological systems. Curr Opin Struct Biol 1995;5: 211–215.

14. Creighton TE. Protein folding. New York: W.H. Freeman, 1992.
15. Mohanty D, Elber R, Thirumalai D, Beglov D, Roux B. Kinetics of peptide folding: computer simulations on SYPFDV and peptide variants in water. J Mol Biol 1997;272:423–442.
16. Tobias DJ, Mertz JE, Brooks CL III. Nanosecond time scale folding dynamics of a pentapeptide in water. Biochemistry 1991;30: 6054–6058.
17. Demchuk E, Bashford D, Case DA. Dynamics of type VI reverse turn in a linear peptide in aqueous solution. Fold Design 1997;2: 35–46.
18. Daura X, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280:925–932.
19. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95: 9897–9902.
20. Nilges M, Brunger AT. Successful prediction of the coiled coil geometry of the GCN4 leucine zipper domain by simulated annealing: comparison to the X-ray structure. Proteins 1993;15: 133–146.
21. Daggett V, Levitt M. Molecular dynamics simulations of helix denaturation. J. Mol. Biol. 1992;223:1121–1138.
22. Hirst JD, Brooks CL III. Molecular dynamics simulations of isolated helices of myoglobin. Biochemistry 1995;34:7614–7621.
23. Tirado-Rives J, Jorgensen WL. Molecular dynamics simulations of the unfolding of apomyoglobin in water. Biochemistry 1993;32: 4175–4184.
24. Caflisch A, Karplus M. Acid and thermal denaturation of barnase investigated by molecular dynamics simulations. J Mol Biol 1995;252:672–708.
25. Boczko EM, Brooks CL III. First principles calculation of the folding free energy of a three-helix bundle protein. Science 1995;269:393–396.
26. Sheinerman FB, Brooks CL III. Molecular picture of folding of a small α/β protein. Proc Natl Acad Sci USA 1998;95:1562–1567.
27. Sheinerman FB, Brooks CL III. Calculations on folding of segment B1 of Streptococcal protein G. J Mol Biol 1998;278:439–455.
28. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. II. Applications to protein A, ROP and crambin model and interaction scheme. Proteins 1994;18:353–366.
29. Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
30. Rey A, Skolnick J. Computer modeling and folding of four-helix bundles. Proteins 1993;16:8–28.
31. Rey A, Skolnick J. Computer simulation of the folding of coiled coils. J Chem Phys 1994;100:2267–2276.
32. Vieth M, Kolinski A, Brooks CL III, Skolnick J. Prediction of the folding pathways and structure of the GCN4 leucine zipper. J Mol Biol 1994;237:361–367.
33. Skolnick J, Kolinski A, Ortiz A. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
34. Kolinski A, Galazka W, Skolnick J. On the origin of the cooperativity of protein folding: implications from model simulations. Proteins 1996;26:271–287.
35. Kolinski A, Galazka W, Skolnick J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: applications to small helical, β, and α/β proteins. J Chem Phys 1998;108: 2608–2617.
36. Mohanty D, Kolinski A, Skolnick J. De novo simulations of the folding thermodynamics of the GCN4 leucine zipper. Biophys J 1999; in press.
37. Vieth M, Kolinski A, Brooks CL III, Skolnick J. Prediction of quaternary structure of coiled coils. Applications to mutants of the GCN4 leucine zipper. J Mol Biol 1995;251:448–467.
38. Zhang L, Hermans J. Molecular dynamics study of structure and stability of a model coiled coil. Proteins 1993;16:384–392.
39. Thomas P, Dill KA. Statistical potentials extracted from protein structure: how accurate are they? J Mol Biol 1996;257:457–469.
40. Ben-Naim A. Statistical potentials extracted from protein structures: are these meaningful potentials? J Chem Phys 1997;9:3698–3706.
41. Brooks CL III. Molecular simulations of peptide and protein unfolding: in quest of a molten globule. Curr Opin Struct Biol 1993;3:92–98.
42. Brooks CL III, Karplus M, Pettitt BM. Proteins: a theoretical prospective of dynamics, structure and thermodynamics. Adv Chem Phys 1988;71:1–249.
43. O'Shea EK, Klemm JD, Kim PS, Alber T. X-ray structure of GCN4 leucine zipper, a two stranded, parallel coiled coil. Science 1991; 254:539–544.
44. Qui D, Shenkin P, Hollinger F, Still W. The GB/SA model for solvation: a fast analytical method for the calculation of approximate Born radii. J Phys Chem 1997;101:3005–3014.
45. Dominy BN, Brooks Charles L III. Development of a generalized Born model parameterization for proteins and nucleic acids. J Phys Chem 1998; in press.
46. O'Donoghue SI, Nilges M. Tertiary structure prediction using mean-force potentials and internal energy functions: successful prediction for coiled-coil geometries. Fold Design 1997;2:S47–S52.
47. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:399–413.
48. Holm L, Sander C. Evaluation of protein models by atomic solvation preferences. J Mol Biol 1992;225:93–105.