Title
The effects of fluctuations on the native state structure, stability, and flexibility of proteins

Author
Tang, Karen En-Hwa Silverstein

Publication Date
1996

Peer reviewed|Thesis/dissertation

The Effects of Fluctuations on the Native State Structure, Stability, and Flexibility of Proteins

by

Karen En-Hwa Silverstein Tang

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

To my parents
and my husband, Kevin,
for their love, support, and encouragement

Acknowledgments

Foremost, I'd like to thank my thesis advisor, Ken Dill, for his unflagging faith in me; for his patience in letting me do science my own way, thereby allowing me to develop my own style (including making many mistakes) and to spend time "just" thinking without feeling the pressure to "produce"; for his enthusiasm which kept up my motivation level; for his training me to always keep the "big picture" in mind; and, of course, for his excellent scientific and career advice and direction. There are few people like Ken—an excellent scientist, mentor, and manager.

I also thank my parents for believing in me, for their advice, and for keeping my spirits up when I was greatly discouraged about graduate school. My husband, Kevin, brought a great light into my life and has always given me something to be happy about, even when research wasn't going well. He helped me keep a well-balanced life, and in doing so, made me a lot happier about doing science. I could always count on him to give me his most honest opinion on anything.

The entire Dill lab has been wonderful for career and scientific help and also for good companionship. In particular, Klaus Fiebig taught me how to choose good projects and how to write good computer code; he also made his programs freely available to the lab (and I sure got a lot of good use out of them). I had many long and candid conversations with Sarina Bromberg and Paul Thomas about being a scientist. And of course my work would have been much more difficult without our excellent computer system managers, Danny Heap, Dave Yee, and Rick Rodgers.

I thank my thesis and orals committees, Tack Kuntz, Fred Cohen, Peter Kollman, and Robert Fletterick, as well as David Agard (who managed to get off both committees by being on sabbatical), for their excellent scientific comments and criticisms. I especially thank Tack for being the chair of my orals committee, for putting up with my endless questions on NMR, and for his advice on my post-doctoral career. I am grateful to Celia Schiffer and Melissa Starovasnik for taking me under their wings and helping me to develop as a scientist. I also thank the members of the biophysics graduate group for teaching me about the experimental world. And Julie Ransom was tremendously helpful by making the practicalities of graduate life ever

so much easier.

"Karen, the scientist" first took root with the help of the unofficial "Physics Graduate Student Support Group". I am indebted to Storrs Hoen and the rest of the group for helping me overcome my insecurities and making me realize that I wasn't the only one doubting myself. I also want to thank Dan Rokhsar of the physics department at Berkeley for listening and helping out when I was "between a rock and a hard place".

In an ironic way, my difficult experiences in the Berkeley physics department and with the various scientists I worked with during that period taught me a most invaluable lesson: to have faith in my own judgment and to take charge of the course of my own life.

And of course I am most thankful for my many friends (whom I won't name here) who were always willing to listen when life wasn't going well and for keeping me a happy person.

Financial support for this work was provided by an NIH training grant, a UCSF Graduate Dean's Health Science Fellowship, and a UCSF Regent's Fellowship, for which I am very grateful.

The Effects of Fluctuations on the Native State Structure, Stability, and Flexibility of Proteins

Karen En-Hwa Silverstein Tang

December 1996

Abstract

Fluctuations of proteins are important for many biological processes such as enzyme catalysis and ligand binding. However, protein fluctuations are not completely understood. In particular, little is known about *large* fluctuations—in which proteins make large excursions from their native conformations. If proteins have rugged potential energy landscapes, they could occasionally sample very non-native structures, called "conformational distance relatives" (CDRs) [Miller, D. W. and Dill, K. A., *Prot. Sci.* **9**:1860 (1995)]. I ask what effects CDRs might have on structural measures of proteins, on structure-determination experiments, and on the flexibility and stability of proteins. Here, using a simple two-dimensional lattice model and employing exact methods, I study the consequences of having a wide range of fluctuations.

First, I find that most measures of structure, like crystallographic diffraction patterns and many nuclear magnetic resonance (NMR) spectra, are insensitive to CDR fluctuations because of the averaging over the many fluctuation conformations. As a consequence, the presence of CDRs does not greatly affect the outcome of standard structure-determination experiments; they are robust for determining

the dominant native conformation. At the same time, determining a well-defined structure implies neither few nor small fluctuations. If the total CDR population is considerable, as may occur for proteins near denaturation or for peptides, structural constraints are mutually inconsistent and determined structures may be biased. Second, CDRs may be the underlying cause of the experimentally observed biphasic temperature dependence of crystallographic Debye-Waller factors [Rasmussen, B. F., et al., *Nature* **357**:423 (1992)]. The observed "transition" may not be glassy. Lastly, I find an inverse correlation between protein stability and flexibility. The more stable the protein, the fewer are its CDR fluctuations and hence the lower the average flexibility.

Contents

List of Figures

Chapter 4

Chapter 1

Introduction

It is commonly believed that proteins designed by nature have a well-defined, unique, structure which is called the "native" conformation. One of the strongest pieces of evidence to support this view is the fact that there are hundreds of high resolution crystal structures of proteins. Some proteins have been crystallized with different space groups yet still show essentially the same conformation [2,6]. In addition, the native conformations predicted by nuclear magnetic resonance (NMR) experiments generally have well-defined structures with only small fluctuations in surface loops. Comparisons of native structures determined by both NMR and crystallography have indicated a close similarity, particularly in the core regions [7,40]. From this evidence, one can safely say that biologically-relevant proteins have one dominant conformation.

But proteins are not frozen in one conformation. They must fluctuate to allow ligands to bind and catalysis to occur. What do we know about protein fluctuations? That is, what other conformations does the molecule occasionally sample? There is ample evidence of small fluctuations, motions that don't change the chain conformation very much. Crystallographic Debye-Waller (B) factors which measure mean-square deviations in atomic positions indicate the existence of vibrational motions [38]. Models of protein motions based on hydrogen-deuterium exchange (HX) experiments suggest that proteins undergo "breathing" motions [27,37,48] or "local unfolding" [15,21]. Molecular dynamics computer simulations indicate the presence of many small motions: side-chain movements, ring flips, motions of surface loops, etc. [28]. What about large-scale fluctuations? Do proteins make excursions to conformations that are structurally quite different from the native? At the present time, there seems to be little experimental evidence for such motions. However, if we review current experiments designed to measure fluctuations, we see that they are actually not very good at detecting large movements:

- Binding of ligands provides some of the best evidence *for* large-scale motions. Crystal structures of proteins with and without ligands bound indicate great changes in structure [17]. However, one cannot know for sure whether the protein would make such large excursions in the absence of ligand.

- HX experiments provide another piece of evidence *for* the existence of large fluctuations. They can signal large conformational changes [4]. However, they cannot actually predict the corresponding fluctuation conformations, so we don't really know what the protein is "doing". The HX rate of a proton measures the ability of solvent to reach that proton. But, knowing which protons do/don't have access to solvent still doesn't give information on the actual protein conformations that allowed solvent access. For example, if both proton A and proton B have fast exchange rates, one doesn't know if one fluctuation allowed solvent access to both A and B, or if two different fluctuations occurred, one exposing A and the other exposing B to solvent. And when the number of fluctuation conformations is large, predicting fluctuation conformations becomes intractable. Currently, researchers use models to explain HX data [4, 15, 21, 27, 29, 37, 48]. Unfortunately, many models assume that fluctuations are small. To not malign HX experiments completely, they are useful for showing which regions of protein are flexible and solvent-exposed and they can suggest the existence of large fluctuations [4, 47], even if they can't (uniquely) predict actual conformations.

- Crystallographic B factors measure atomic motions as an expansion around $\Delta x = x - x_o$, the deviation of each atom's position relative to its equilibrium position. B factors are mean-square deviations, the first-order terms of the expansion. Anisotropic B factors are higher order terms. These measurements are *average* deviations in *small* Δx and are not designed for measuring large-scale motions.

- Fluorescence spectra detect changes in the local environment of fluorescing species. They signal when a conformational change has occurred, but, again, they are not useful for predicting actual fluctuation conformations and may not even be able to tell whether the fluctuations are large or small. Fluorescence quenching experiments are similar to HX in that they detect where solvent molecules can gain access.

- NMR experiments are able to detect *highly-populated* non-native conformations.

In fact, there are several nice experiments showing that under certain conditions, a few proteins have more than one dominant conformation [39, 50]. However, conformations whose populations are less than about 5% do not give a strong enough signal to be measured [35]. Weakly-populated fluctuation conformations would not be directly detected by NMR.

NMR dynamics experiments indicate the regions of protein which are moving and the time scales and the amplitudes of the motions [44], but they do not detect actual fluctuation conformations.

- Limited proteolysis is yet another experiment which yields information on the regions of a protein which have access to solvent. Again, the fluctuation conformations cannot be predicted. These experiments have a further difficulty for studying fluctuations. Proteases cleave peptide bonds which are solvent accessible [46]. In principle, to detect non-native fluctuations, one might look for cleavages in places that are not exposed to solvent in the native conformation. Unfortunately, most cleavages would occur in native loops, and after a cleavage has occurred, the subsequent loss of stability might allow a cleaved protein to fluctuate (or even denature) in ways that it would not have had it been intact.

What do theoretical simulations tell us about fluctuations? Because of computational limitations, molecular dynamics simulations are limited to short time scales. They are useful for examining small motions, but not for investigating large motions.

To summarize, current experimental and theoretical studies on protein motions yield information on small fluctuations but they are not good at determining the nature of large fluctuations. Hence, the fact that the current experimental evidence for large fluctuations is limited does not necessarily mean that such fluctuations do not exist. The best evidence against the presence of large fluctuations *seems* to be the existence of the hundreds of well-defined crystal and NMR structures of proteins.

Interestingly enough, the HP lattice model [10, 13, 24, 25], which has been used for many protein studies by the Dill lab [13, 29], and other statistical models [8, 9, 11, 20] have the property that the conformations which are low in energy can be

quite different from the native in structure. (For the HP model, this is true in both two- and three-dimensions.) These models have large fluctuations. Another way to put it is these models have bumpy potential energy landscapes; there are many local minima which are conformationally distant from the global (native) minimum.

The combination of (1) the limited amount of direct experimental information on large fluctuations, (2) the many high quality protein structures which suggest that there are no such fluctuations, and (3) the seemingly contradictory fact that the HP lattice model (and other statistical models) have many such fluctuations is what prompted the work in the second chapter of this thesis. One obvious "solution" to the conundrum is that the HP lattice model has an energy surface which is not like that of real proteins[1]. But this is too easy an escape. So, my thesis advisor, Ken Dill, posed the questions, "If proteins had large fluctuations, what hallmark would these fluctuations have in standard crystallographic or NMR structure-determination experiments? What evidence would one look for to detect their existence?" I went off and simulated structure-determination experiments and found that there is no such hallmark. As a matter of fact, quantities which are ensemble averages (which includes most experimentally measured quantities) are insensitive to the presence of large fluctuations. Since standard structure-determination experiments use such experimentally measured quantities as constraints for determining the "best-fit" structure, the output structures strongly resemble the native conformation, regardless of whether there are large fluctuations present. Consequently, the ability to determine a good structure implies neither the presence nor absence of large fluctuations. This work is presented in chapter 2.

In addition, I also looked into the effects of large fluctuations on the temperature dependence of crystallographic B factors. Studies of ribonuclease A at nine different temperatures showed that molecule-averaged B factors have a biphasic temperature dependence. At temperatures below/above about $220K$, the B factor has a mild/strong dependence on temperature [41]. Other measures of atomic mean-square

[1] Actually, I'm inclined to believe that this is to some extent true. The HP lattice model does have a terribly bumpy energy landscape, probably more so than real protein potential surfaces. But the model is useful for examining what happens for those proteins which have even somewhat bumpy landscapes. The model presents a "worst case" scenario.

displacements from Mössbauer scattering [5, 23, 33, 34], and from inelastic neutron scattering [14] show this behavior; so do excited-state quenching rates [31] and viscoelastic properties [30]. The underlying physical explanation for this biphasic behavior is not understood [3, 5, 14, 16, 18, 19, 23, 26, 30, 33, 34, 36, 41, 42]. I calculated a quantity similar to a molecule-averaged B factor in an attempt to explain the physics underlying this phenomenon. This work is presented in the beginning of chapter 3.

The second half of chapter 3 is a fairly natural extension of the work performed for the second chapter. In the fall of 1995, Greg Petsko gave a seminar to the Dill group, part of which included some research on proteins of thermophilic organisms. In particular, he observed that these proteins are more stable and also less flexible at room temperature than their mesophilic counterparts [1,12,22,32,43,45,49]. The reasons for this inverse correlations are not understood. At the time I was examining quite closely the effects of fluctuations on measures of structure, so it was immediately obvious to me one possible explanation: flexibility and stability might be linked by the entropy of the fluctuations. Chapter 3 discusses the evidence and underlying reasons for such a connection. Chapter 4 presents the theory behind the relationship.

References

[1] M. Abe, Y. Nosoh, M. Nakanishi, and M. Tsuboi. Hydrogen-deuterium exchange studies on guanidinated pig heart lactate dehydrogenase. *Biochim. Biophys. Acta*, 746:176–181, 1983.

[2] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, pages 107–32. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[3] A. Ansari, C. M. Jones, E. R. Henry, J. Hofrichter, and W. A. Eaton. The role of solvent viscosity in the dynamics of protein conformational changes. *Science*, 256:1796–1798, 1992.

[4] Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander. Protein folding intermediates: Native-state hydrogen exchange. *Science*, 269:192–197, 1995.

[5] E. R. Bauminger, S. G. Cohen, I. Nowik, S. Ofer, and J. Yariv. Dynamics of heme iron in crystals of metmyoglobin and deoxymyoglobin. *Proc. Natl. Acad. Sci. USA*, 80:736–740, 1983.

[6] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[7] M. Billeter. Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. *Quart. Rev. Biophys.*, 3:325–77, 1992.

[8] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins*, 21(3):167–195, 1995.

[9] C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA*, 90:6369–6372, 1993.

[10] H. S. Chan and K. A. Dill. "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.*, 95(5):3775–3787, 1991.

[11] D. G. Covell and R. L. Jernigan. Conformations of folded proteins in restricted spaces. *Biochemistry*, 29:3287–3294, 1990.

[12] M. Delepierre, C. M. Dobson, S. Selvarajah, R. E. Wedin, and F. M. Poulsen. Correlation of hydrogen exchange behaviour and thermal stability in lysozyme. *J. Mol. Biol.*, 168:687–692, 1983.

[13] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding. A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[14] W. Doster, S. Cusack, and W. Petry. Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature*, 337:754–756, 1989.

[15] S. W. Englander. Measurement of structural and free energy changes in hemoglobin by hydrogen exchange methods. *Ann. NY Acad. Sci.*, 244:10–27, 1975.

[16] H. Frauenfelder, F. Parak, and R. D. Young. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451–479, 1988. And references therein.

[17] M. Gerstein, A. M. Lesk, and C. Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, 33(22):6739, 1994.

[18] J. L. Green, J. Fan, and C. A. Angell. The protein–glass analogy: some insights from homopeptide comparisons. *J. Phys. Chem.*, 98(51):13780–13790, 1994.

[19] S. J. Hagen, J. Hofrichter, and W. A. Eaton. Protein reaction kinetics in a room-temperature glass. *Science*, 269:959–962, 1995.

[20] D. A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668–682, 1994.

[21] A. Hvidt and S. O. Nielsen. Hydrogen exchange in proteins. *Adv. Prot. Chem.*, 21:287–386, 1966.

[22] R. Jaenicke. Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.*, 202:715–728, 1991.

[23] H. Keller and P. G. Debrunner. Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys. Rev. Lett.*, 45(1):68–71, 1980.

[24] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[25] K. F. Lau and K. A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA*, 87:638–642, 1990.

[26] R. J. Loncharich and B. R. Brooks. Temperature dependence of dynamics of hydrated myoglobin: Comparison of force field calculations with neutron scattering data. *J. Mol. Biol.*, 215:439–455, 1990.

[27] R. Lumry and A. Rosenberg. The mobile defect hypothesis of protein function. *Colloq. Int. CNLRS*, 246:55–63, 1975.

[28] J. A. McCammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, 1987.

[29] D. W. Miller and K. A. Dill. A statistical mechanical model of hydrogen exchange in globular proteins. *Prot. Sci.*, 4(9):1860–1873, 1995.

[30] V. N. Morozov and S. G. Gevorkian. Low-temperature glass transition in proteins. *Biopolymers*, 24:1765–1799, 85.

[31] J. M. Nocek et al. Low-temperature, cooperative conformational transition within [Zn-cytochrome *c* peroxidase, cytochrome *c*] complexes: Variation with cytochrome. *J. Am. Chem. Soc.*, 113:6822–6831, 1991.

[32] S. Ohta, M. Nakanishi, M. Tsuboi, K. Arai, and Y. Kaziro. Structural fluctuation of the polypeptide-chain elongation factor Tu. *Eur. J. Biochem.*, 78:599–608, 1977.

[33] F. Parak, E. N. Frolov, A. A. Kononenko, R. L. Mössbauer, V. I. Goldanskii, and A. B. Rubin. Evidence for a correlation between the photoinduced electron transfer and dynamic properties of the chromatophore membranes from *rhodospirillum rubrum*. *FEBS Letts.*, 117(1):368–372, 1980.

[34] F. Parak, E. N. Frolov, R. L. Mössbauer, and V. I. Goldanskii. Dynamics of metmyoglobin crystals investigated by nuclear gamma resonance absorption. *J. Mol. Biol.*, 145:825–833, 1981.

[35] G. Párraga and R. E. Klevit. Multidimensional nuclear magnetic resonance spectroscopy of DNA-binding proteins. *Meth. Enzym.*, 208:63, 1991.

[36] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko. Crystalline ribonuclease A loses function below the dynamical transition at 220K. *Nature*, 357:423–424, 1992.

[37] F. M. Richards. Packing defects, cavities, volume fluctuations, and access to the interior of proteins. *Carlsberg Res. Commun.*, 44:47–63, 1979.

[38] D. Ringe and G. A. Petsko. Mapping protein dynamics by X-ray diffraction. *Prog. Biophys. Molec. Biol.*, 45:197–235, 1985.

[39] N. J. Skelton, K. C. Garcia, D. V. Goeddel, C. Quan, and J. P. Burnier. Determination of the solution structure of the peptide hormone guanylin: observation of a novel form of topological stereoisomerism. *Biochemistry*, 33:13581–92, 1994.

[40] L. J. Smith et al. Comparison of four independently determined structures of human recombinant interleukin-4. *Nature Struct. Biol.*, 1(5):301, 1994.

[41] R. F. Tilton, Jr., J. C. Dewan, and G. A. Petsko. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease A at nine different temperatures from 98 to 320K. *Biochemistry*, 31(9):2469–2481, 1992.

[42] M. G. Usha and R. J. Wittebort. Orientational ordering and dynamics of the hydrate and exchangeable hydrogen atoms in crystalline crambin. *J. Mol. Biol.*, 208:669–678, 1989.

[43] P. G. Varley and R. H. Pain. Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and *thermus thermophilus*. *J. Mol. Biol.*, 220:531–538, 1991.

[44] G. Wagner. The importance of being floppy. *Nat. Struct. Biol.*, 2(4):255–257, 1995.

[45] G. Wagner and K. Wüthrich. Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitor. *J. Mol. Biol.*, 130:31–37, 1979.

[46] J. E. Wilson. The use of monoclonal antibodies and limited proteolysis in elucidation of structure-function relationships in proteins. *Meth. Biochem. Analysis*, 35:207, 1991.

[47] C. K. Woodward, L. M. Ellis, and A. Rosenberg. Solvent accessibility in folded proteins: studies of hydrogen exchange in trypsin. *J. Biol. Chem.*, 250:432–9, 1975.

[48] C. K. Woodward and A. Rosenberg. Studies of hydrogen exchange in proteins. VI. Urea effects on RNase hydrogen exchange kinetics leading to a general model for hydrogen exchange from folded proteins. *J. Biol. Chem.*, 246:4114–4121, 1971.

[49] A. Wrba, A. Schweiger, V. Schultes, R. Jaenicke, and P. Závodskzky. Extermely thermostable D-Glyceraldehyde-3-phosphate dehydrogenase from the eubacterium *thermotoga maritima*. *Biochemistry*, 29(33):7584–7592, 1990.

[50] O. Zhang and J. D. Forman-Kay. Structural characterization of folded and unfolded states of an SH3 domain in equilibrium in aqueous buffer. *Biochemistry*, 34:6784–6794, 1995.

Chapter 2

Fluctuations of Native Proteins: Are There Large Motions?

Karen E. S. Tang and Ken A. Dill

2.1 Abstract

The fluctuations of native proteins are often regarded as "small wiggles". But if proteins have rugged energy landscapes, thermal motions could occasionally sample quite non-native structures, called "conformational distant relatives" (CDRs). If real proteins have CDR fluctuations, could standard experiments detect them? Here, using the HP lattice model, which produces rugged landscapes, we study native-state fluctuations, small to large, by exact methods. We simulate structural experiments, like nuclear magnetic resonance (NMR) nuclear Overhauser enhancement spectroscopy (NOESY). We find that current experiments are unlikely to detect occasional large excursions from the native structure because of heavy averaging over the many fluctuation conformations. In short, proteins may occasionally undergo large conformational deviations from the native structure, but new methods will be needed to find them. We also consider "structure inversion", the process of determining one native structure from ensemble-averaged NMR or X-ray constraints. These algorithms are robust for determining the dominant native structure, regardless of the presence of CDRs. However, predictions about conformational fluctuations, as implied by the Debye-Waller factors in crystallography or by the multiple structures of NMR, may not well represent the true fluctuations and may incorrectly suggest the absence of CDRs. Under conditions with many fluctuations, structure inversion can lead to inconsistent constraints, those which cannot be satisfied by any single physically-viable structure.

2.2 The thermal motions of proteins

Proteins in their native states move, wiggle, and fluctuate. Fluctuations are important for biological processes such as enzyme catalysis and ligand binding, induced-fit mechanisms [36,37] or binding to buried active sites such as when oxygen binds hemoglobin [51]. Fluctuations can be detected by experimental measurements [28] like hydrogen-deuterium exchange (HX) [21,30,62], Debye-Waller factors in X-ray crystallography [25,54] or NMR dynamics experiments [50,61]. Fluctuations are the

"conformational noise" that can complicate structure determination in NMR or X-ray crystallographic experiments, by affecting the numbers and precisions of experimental constraints.

Much common wisdom holds that the fluctuations of native proteins are "small wiggles". For example, two models of HX interpret motions as small local unfolding [20, 30] or solvent penetration events [44, 53, 63]. Normal-mode computer simulations are based on assuming that protein motions can be treated by spring-like forces around a protein's native conformation [26]. Debye-Waller factors in X-ray crystallography assume that atomic motions are Gaussian distributed [25]. The restraining potentials used in structural refinement methods are designed so that the highest population is centered on the native structure, and populations decrease monotonically with increasing deviation of a conformation from the native structure. That is, very native-like conformations are the most populated, and very non-native conformations are rarely populated.

Although this common wisdom that fluctuations are small wiggles might be true, there is little direct evidence to prove it. Most experiments only measure averages over the fluctuations, not the individual fluctuation conformations themselves. HX provides some evidence to suggest the presence of large fluctuations [1], but a model is required to interpret the conformational change. Because of the lack of computational power, molecular dynamics simulations are limited to short time scales and hence to small motions. Even if there were large motions, current simulations would not see them.

Occasional large motions are expected if proteins have "rugged energy landscapes", as explored in some recent statistical mechanical models [8,9,11,15,17,29,45, 55]. By definition, fluctuations (under native conditions) must involve small changes in energy, according to the Boltzmann distribution law. But, being near-native in *energy* does not imply that a conformation must be near-native in *structure* [8, 45]. Fluctuations having energies only slightly higher than the native could, in principle, have very different structures. An extreme example would be if a protein whose native conformation is α-helical undergoes occasional brief excursions to a β-sheet conformation to form a hydrophobic core that is nearly as good.

There is some experimental evidence for rugged energy landscapes, but it is mostly indirect. Some proteins, including prions [14,48] and amyloidogenic proteins [33] have two dramatically different native conformations. Others, including α-lytic protease [3] and serine protease plasminogen activator inhibitor-1 [46], can get stuck in deep kinetic traps. The SH3 domains of *Drosophila* drk and GRB2 exist in equilibrium between folded and unfolded states [27,66]. Cytochrome c, in low concentrations of denaturant, undergoes HX via partially unfolded forms in which entire helices or omega loops are unstable [1]. Influenza haemagglutinin makes a large change of its native conformation with only a small change in solution conditions [10]. Also, the denatured states of some proteins have persistent non-native structure [56].

Is there more direct evidence that proteins have rugged energy landscapes? To determine what types of experiments might probe fluctuation conformations, we need a model that can treat fluctuations, large and small. Since no model having atomic detail can do this yet, simplifications are necessary. Here we use a lattice model to determine the effects of a wide range of fluctuation conformations on experimental observables. We focus on three questions of how fluctuations affect measures of structure and structure-determination experiments. First, we ask whether measures of structure like nuclear Overhauser effect (NOE) spectra or Patterson maps could detect large fluctuations.

Second, we consider the structure-determination or "structure inversion" problem. This process involves two steps: one might be called the forward process and the other the inverse process. The forward process is the experiment itself: the native structure and its ensemble of fluctuations give rise to a set of Boltzmann-averaged measurements of structure. The inverse process is the subsequent processing of that data: using the measurements as constraints, along with a model, to determine the native structure which produces these constraints. If there were no fluctuations or averaging, this inversion process would be simple and unambiguous: one native structure gives constraints which, upon inversion, return the one native structure. But reality involves fluctuations, which have the potential to distort the inversion process: many conformations give averaged constraints which, upon inversion, return one "average" native structure [32,52]. We use the simplified model to

study the distortions that are introduced by fluctuations in the inversion process. Third, structure-determination experiments often produce some measure of the fluctuations, e.g., the Debye-Waller factors in crystallography and the family of related structures in NMR. We ask whether the predicted fluctuations accurately represent the underlying ensemble of conformations.

2.3 The model

2.3.1 The HP lattice model of proteins

We model the fluctuations of proteins using the HP lattice model [13,17,42, 43]. Proteins are represented as specific sequences of H (hydrophobic) and P (other) residues on a two-dimensional (2D) square lattice or a three-dimensional (3D) cubic lattice. Each amino acid can only occupy one lattice site, and no two amino acids may reside on the same site. The energetic interactions, designed to capture the essence of the hydrophobic interaction, consist of a single term: there is a favorable interaction, $\varepsilon < 0$, whenever two non-bonded H residues are on adjacent lattice sites, i.e., "in contact". Hence, the free energy of any conformation is $h\varepsilon$, where h is the number of HH contacts. The lowest energy conformation, with h_{nat} HH contacts, is the native conformation. In this work, we study only sequences with a single native conformation.

The energies of a protein are like the rungs on a ladder. The lowest rung represents the one native conformation. The next higher rung on the energy ladder corresponds to all of the "first-excited" conformations, those having $h_{nat} - 1$ HH contacts. The first-excited conformations are the dominant fluctuations under native conditions. On the next rung up the ladder are the second-excited conformations, etc. These fluctuations and those at higher rungs of the energy ladder become increasingly important as the system approaches denaturing conditions.

The limitations of the model are obvious: the chains are short, two-dimensional, and have low resolution, with limited bond angles and bond distances; there are only two amino acid types; the interactions are simplified. Nevertheless, we

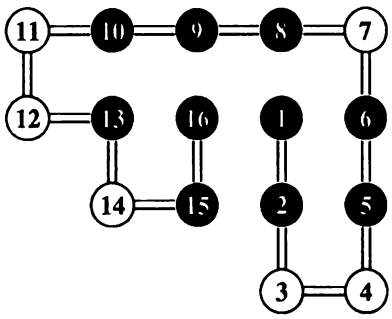


Figure 1: The native conformation of sequence *R*. Dark beads are hydrophobic (H); light beads are other (P). HH interactions are favorable.

believe the model captures the basic components of protein folding—the hydrophobic interactions, conformational freedom of the chain, and the steric restrictions imposed by excluded volume. This model has been shown to have many protein-like properties [17, 45], for example, collapse to compact states having unique folds and secondary structure. Here we use the lattice model to study structure-determination procedures. For this purpose, we do not need the molecular details of an all-atom model or an accurate energy function. Rather, our aim here is simply a study of principle, of what kinds of distortions or false predictions (if any) might result from the neglect of the full ensemble of fluctuations. For this purpose, it is more important to have a way of treating both large and small fluctuations. The advantage of this model is that, since the chains are short and in 2D, all conformations can be enumerated exactly, so we can avoid making any further assumptions about the nature of the fluctuations. Within the model, all fluctuations are included exactly with their proper Boltzmann weights and without restriction to small-amplitude motions or short time scales.

We study residue sequences of lengths 16 and 18 in 2D, and one 27-mer sequence in 3D. Figure 1 shows sequence *R*, HHPPHHPHHHPPHPHH, in its single native conformation. The behavior of this sequence is typical of the others we studied, so it will serve as our main example here.

2.3.2 Defining the fluctuation conformations

The fluctuation conformations consist of all conformations with non-zero population, minus the native conformation. The equilibrium population of each conformation, c, is given by its Boltzmann probability:

$$probability\ of\ c = p(c) = \frac{e^{-E_c/kT}}{Q} \tag{1}$$

where E_c is the energy of conformation c (equal to $h\varepsilon$ in the HP model), T is the absolute temperature, k is Boltzmann's constant. Q is the partition function:

$$Q = \sum_{c=1}^{N} e^{-E_c/kT} \tag{2}$$

where the sum is over all N possible conformations. At low temperatures, only the low energy conformations are populated. Therefore under native conditions, the fluctuation conformations consist of the lowest energy non-native conformations.

Figure 2 shows how the populations of different excited states change with temperature. At $T = 0$ (subfigure A), every chain is in its native conformation (ground state); there are no fluctuations. At low T (subfigures B and C), the native conformation is still dominant but other low-energy conformations are also populated to small degrees. Ultimately, at high T, higher levels up the energy ladder (more open conformations) become populated and the protein denatures. However, we focus here only on "native" conditions, i.e., temperatures below the denaturation midpoint. Under native conditions, the main fluctuation conformations are the first-excited conformations.

2.4 What conformations are visited by thermal motions?

Figure 3 shows a sample of the 37 first-excited conformations of sequence R. These fluctuation conformations are compact and have hydrophobic cores and secondary structure. (For the lattice model, we use the definitions of secondary

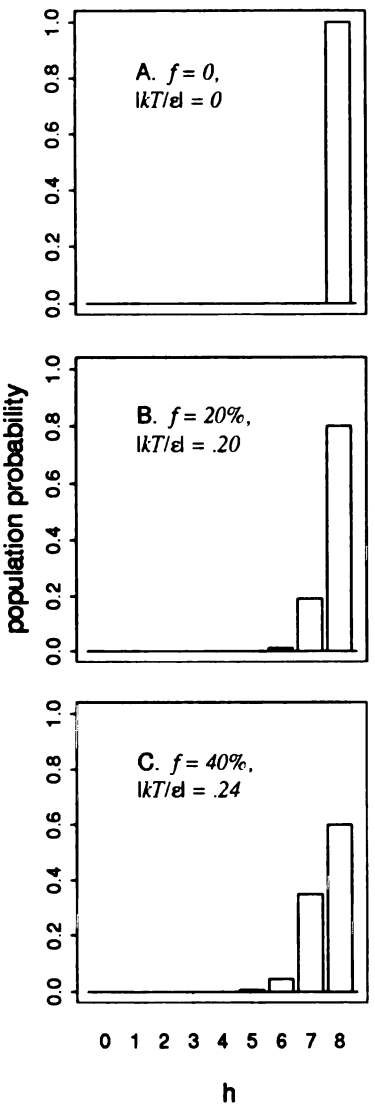


Figure 2: Populations of different energy states, at different temperatures, for sequence R. Energy $= h\varepsilon$, where h is the number of HH contacts and $\varepsilon < 0$. All conformations with the same energy are grouped together. A: At zero temperature, the ensemble consists entirely of the native conformation ($h = 8$). At higher temperatures (B: $|kT/\varepsilon| = .20$; C: $|kT/\varepsilon| = .24$), the first-excited conformations ($h = 7$) are also populated and comprise the majority of fluctuation conformations.

structure in [12].) They are all low in energy; every one of the 37 conformations has only one less HH contact than the native conformation. Some of the first-excited conformations, like those near the top of figure 3, are somewhat similar to the native conformation. The upper left conformation is essentially native, except with residues 10, 11, and 12 shifted clockwise. Others, like those near the bottom of the figure, have completely different folds. Being near-native in energy does not imply being near-native in conformation.

The fluctuations range from being similar to the native structure to being very dissimilar. To be more quantitative, we use the structural dissimilarity measure of Yee and Dill [64] to compare the fluctuations to the native conformation (figure 4). The Yee measure computes a score between any pair of conformations by comparing their distance maps[1]. A score of zero means that two conformations are identical. To give a reference for comparison, the distribution of dissimilarities between all pairs of native conformations of the set of $n = 16$ sequences is shown in figure 5. For two arbitrarily chosen native structures, it is most probable they will differ by a score of about 0.7. As noted before [45], comparing figure 4 to 5 shows that: (i) fluctuations have a broad range of structures, but (ii) fluctuation conformations are more similar to their corresponding native conformation than to other compact structures from different sequences of residues.

Fluctuations that are *near-native in energy* but *distant from native in conformation* have been called "conformational distant relatives" (CDRs) [45] to distinguish them from fluctuations involving only small conformational changes like vibrations, side-chain movements, loop wiggles, small shifts in orientation of secondary structure, etc. The latter are intrinsically below the resolution of the lattice model. However, the conformations that are represented within the lattice model can then be classified as being either relatively small deviations from the native structure, or relatively large, i.e., CDRs. CDRs are a common feature of the HP lattice model and of other models with rugged energy landscapes [8,9,11,15,17,29,45,55]. We cannot draw a precise definition of what "large" fluctuations might be for real proteins, but

[1] A distance map is the set of inter-residue distances, $\{d_{ij}\}$, between all residue pairs (i, j).

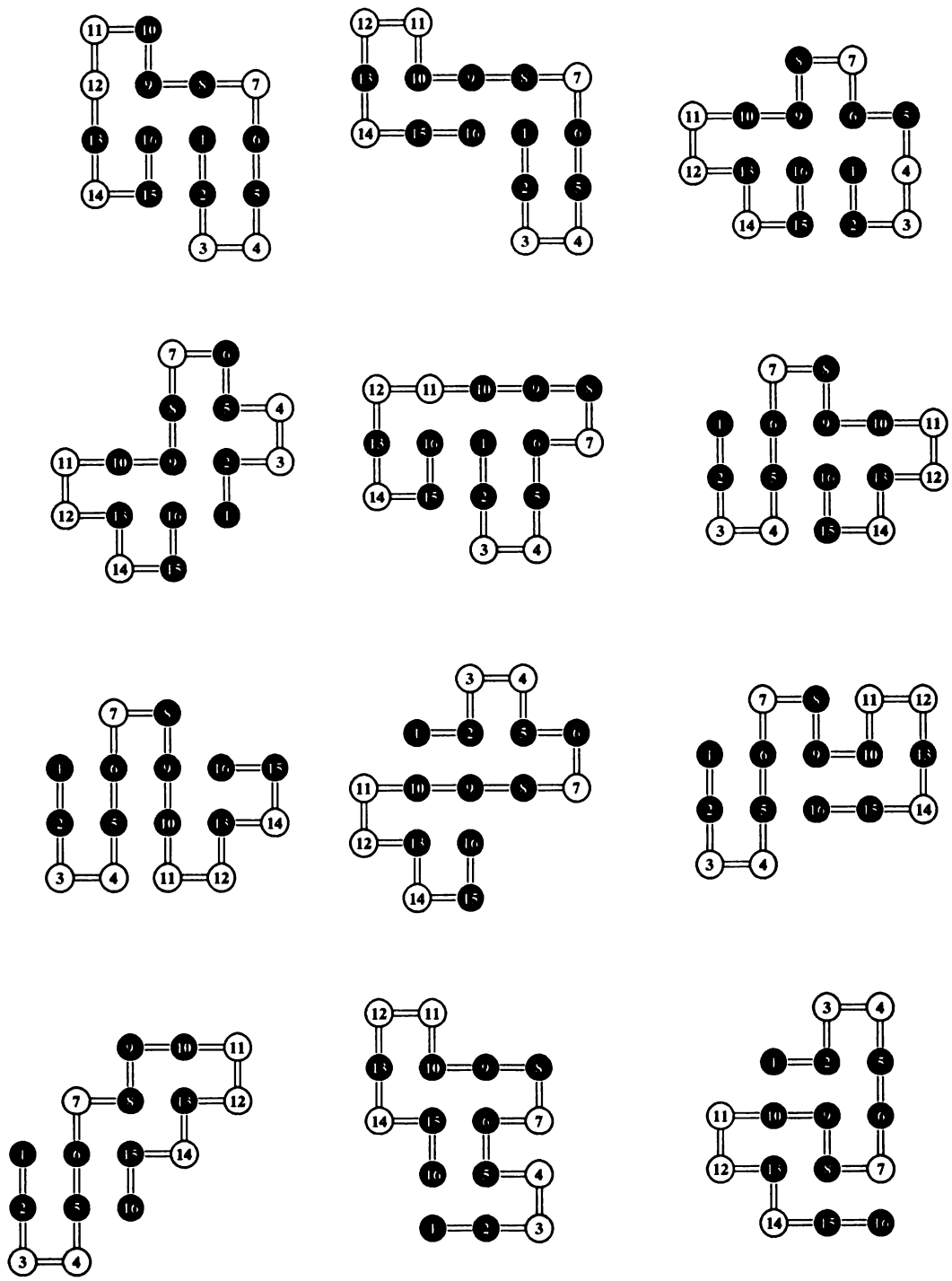


Figure 3: A sample of the 37 first-excited conformations of sequence R. The conformations are ordered from the highest structural similarity to the native conformation at the upper left to the lowest structural similarity at the lower right.

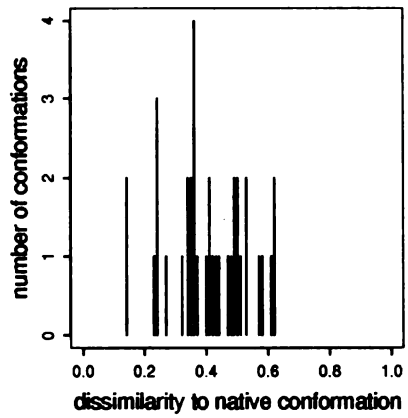


Figure 4: The conformational dissimilarity, using the measure of [64], between every first-excited conformation and the native conformation of sequence R. A distance of 0 indicates that two conformations are identical.

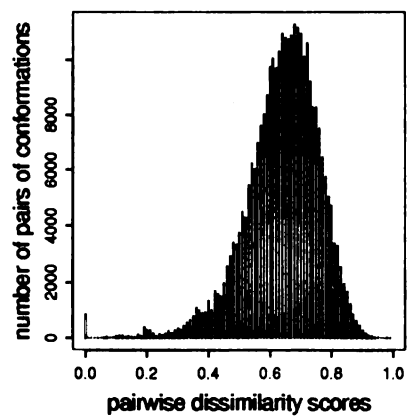


Figure 5: The conformational dissimilarities between all pairs of native conformations of sequences of length $n = 16$. This figure is replotted from figure 8B of [45].

they would undoubtedly have non-native structural elements. Whether real proteins have rugged energy landscapes with CDR fluctuations remains to be determined experimentally. The conclusions that arise from the HP model are not unique to this model; they should apply whenever protein energy landscapes are not smooth.

2.5 Do structural experiments detect CDR fluctuations?

If an α-helical protein occasionally fluctuates into a β-sheet conformation, would the change be detected by structural experiments? To answer this question, we calculate various measures of structure and observe how they change as the population of CDR fluctuations increases.

Most experiments measure ensemble averages. To describe the effects of ensemble averaging, let f represent the total fraction of the population that are in fluctuation conformations,

$$f = 1 - p(native) \ . \tag{3}$$

According to equation 1, f increases with temperature (see also figure 2). We use temperature here simply as the most convenient way to control the range of conditions from native to denaturing, i.e., from few to many fluctuations. Of course, in practice, conditions are more commonly controlled by denaturant concentrations. Angle brackets, $\langle \cdots \rangle$, indicate Boltzmann-averaged (i.e., ensemble-averaged) quantities:

$$\langle X \rangle = \sum_{c=1}^{N} X_c \, p(c) \tag{4}$$

where X_c is the value of some property, X, for conformation c. Curly brackets $\{\ldots\}$ denote "the set of all", as in the set of all inter-residue distances $\{d_{ij}\}$. All distances are in terms of "lattice units", the distance between adjacent lattice sites.

Although we perform all calculations under equilibrium conditions, this restriction is not necessary. For proteins whose folding is under "kinetic control" [2], experiments still make measurements averaged over the populated conformations. Equations 3 and 4 are unchanged except that the $p(c)$ are set to the fractional population of each conformation, c.

2.5.1 Average distance maps are insensitive to CDRs.

The first property, X, we consider is the distance map, the set of distances $\{d_{ij}\}$ between all pairs of residues (i,j). If experiments give us ensemble-averaged distance maps, could we tell if non-native conformations are occasionally populated? The distance map of the native state at $T = 0$ is sufficient to reconstruct the native conformation. But at higher temperatures, the measured distance map will be a Boltzmann-averaged composite of the distance maps of all of the conformations of the ensemble. A Patterson map (see, e.g., [5,49]), obtained from X-ray scattering on protein crystals, measures all *average* pairwise distances, $\{\langle d_{ij}\rangle\}$, which we denote as ⟨distance map⟩. Figure 6 shows the changes in ⟨distance map⟩s as the temperature, T, is increased and the population, f, of fluctuations becomes significant. The ⟨distance map⟩s are similar at all temperatures where the proteins are folded. Therefore, a ⟨distance map⟩ (or a Patterson map), gives little evidence for fluctuations of a native protein into very different conformations.

To get an idea why there is little influence from the CDR fluctuations, figure 7 shows a distance map, averaged *only* over the first-excited conformations; the native conformation is not included in the average. This distance map has a resemblance to the native distance map (figure 6A), despite the fact that the first-excited conformations can be quite different from the native conformation. Apparently, the non-native-like structures cancel upon averaging, leaving behind a weakly native-like signal.

2.5.2 NOE spectra are also insensitive to CDRs.

As a second measure of structure, we simulate NOE spectra. The intensity of the NOE signal between any two protons decays as r^{-6} where r is the inter-proton distance. When multiple conformations interconvert slowly, NOE intensities are proportional to the average $\langle r^{-6}\rangle$ [34, 59]. To simulate an NOE spectrum, we calculate $\langle d_{ij}^{-6}\rangle$ for all residue pairs (i,j). In figure 8 we show the simulated spectra of sequence R at different fluctuation populations. A peak between any pair of residues indicates a strong NOE signal, signifying a short average distance between them.

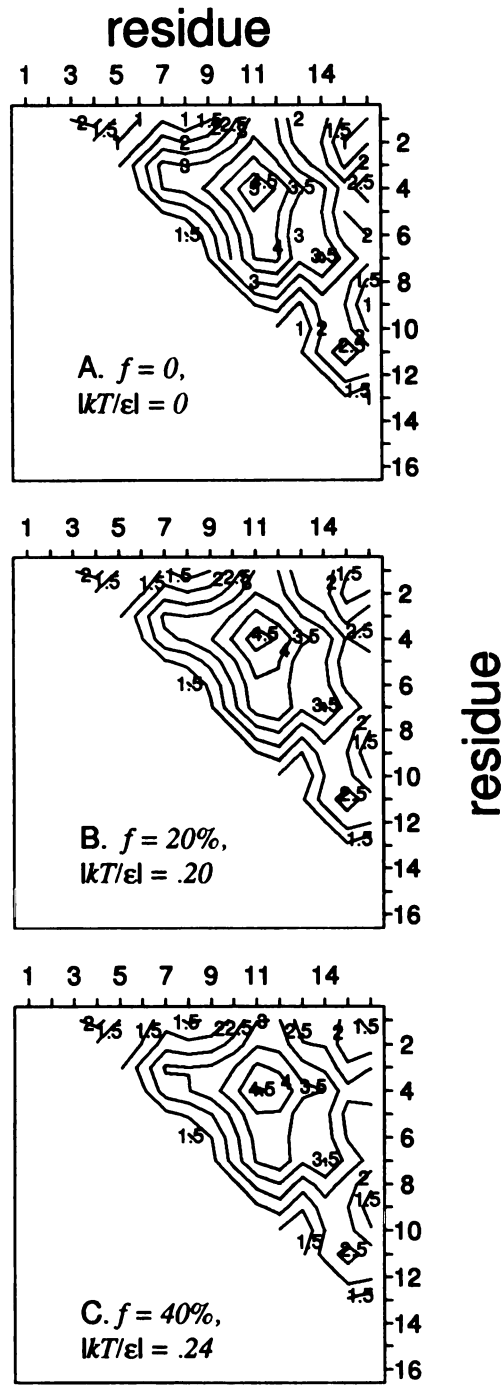


Figure 6: Contour plots of ⟨distance map⟩s for sequence R. A: The native distance map ($f = T = 0$). B and C: ⟨Distance map⟩s at increasing temperature and f: $|kT/\varepsilon| = .20$, $f = 20\%$ and $|kT/\varepsilon| = .24$, $f = 40\%$, respectively. The vertical and horizontal axes are the residue numbers.

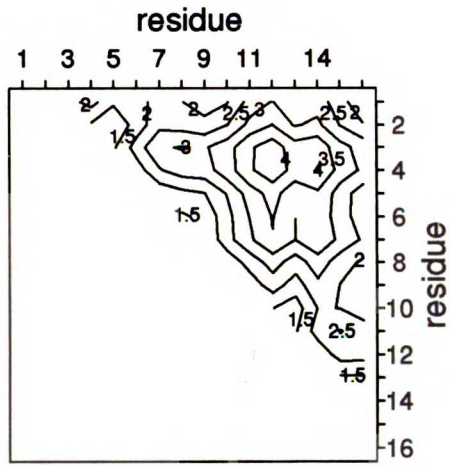


Figure 7: Contour plot of the distance map, averaged only over the first-excited conformations, of sequence R. The vertical and horizontal axes are the residue numbers.

Because of noise in real NOE spectra, only signals from proton pairs closer than about 6Å apart are measurable. To simulate the noise, we assume that any calculated NOE less than $(\sqrt{2})^{-6}$ is undetectable. We choose $\sqrt{2}$ as the cutoff distance because it is the shortest non-contact distance on a square lattice. The noise is represented in the figure by the background plateau.

Figure 8 shows that the calculated NOE spectra differ little from the native spectrum, even when there are significant populations of CDR fluctuations. For sequence R, non-native peaks first appear when $f = 26\%$, i.e., when the ensemble of molecules is only three-quarters native. Therefore, just as with ⟨distance map⟩ information, our simulated NOE spectra are found to be insensitive to CDR fluctuations.

A caveat is that since we cannot simulate chemical shifts, we assume that the chemical shift of each residue is independent of conformation. In a more realistic spectrum, there might be additional non-native peaks.

We conclude that ensemble-averaged measures of structure are insensitive to CDR conformations. The reason for this is that each sequence has a substantial number of CDR fluctuations, each of which has some parts that are native-like and some that are different from native. When an average is taken over all the CDRs,

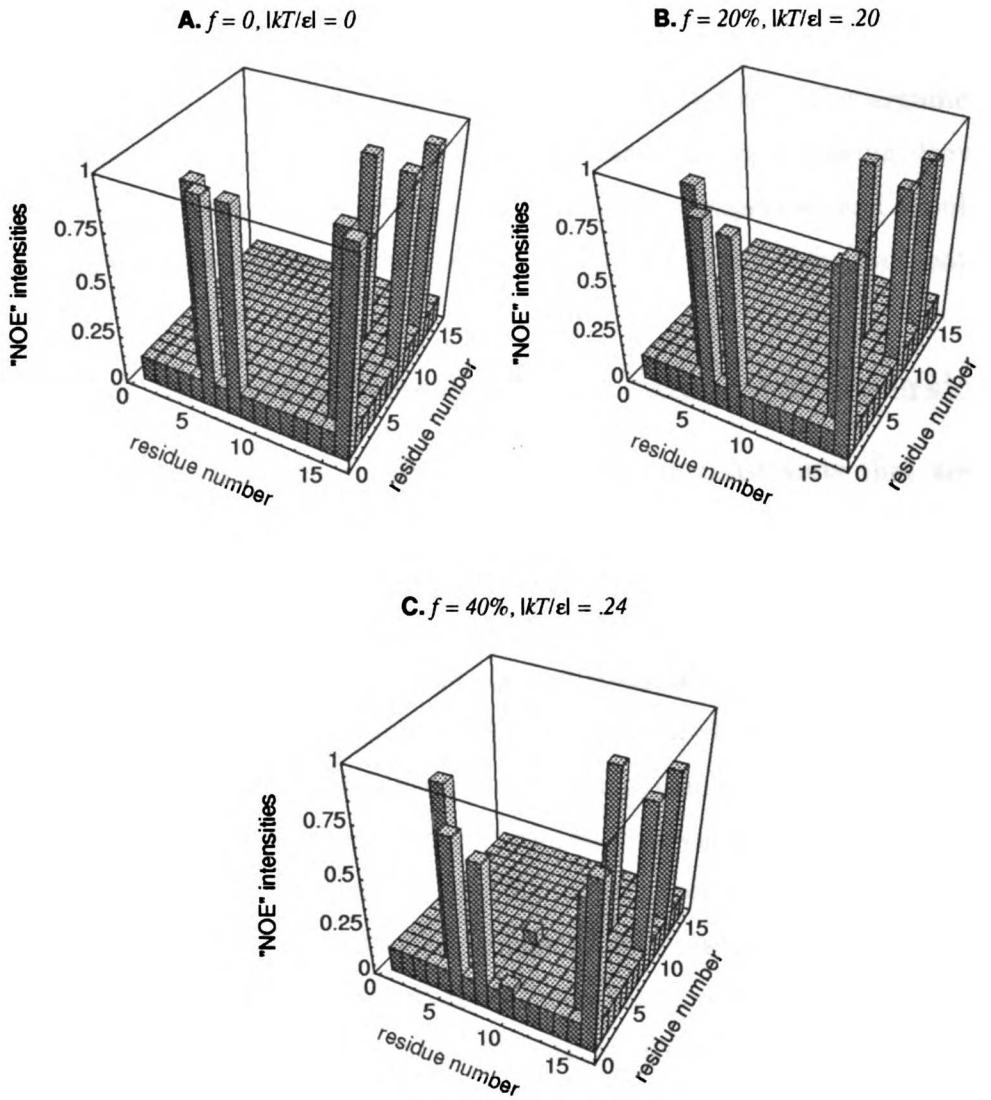


Figure 8: Simulated NOE spectra of sequence R. The axes in the horizontal plane represent the residue numbers. The vertical axis shows the simulated NOE intensities, $\langle d_{ij}^{-6} \rangle$, between all nonbonded residue pairs (i, j). A peak indicates that two residues are close together in space, on average. Any NOE intensity which is below the background noise level of $(\sqrt{2})^{-6}$ is graphed *at* the noise level; hence the flat plateau at the bottom. A: The native NOE spectrum ($f = T = 0$). B and C: "NOE" spectra at increasing temperature and f: $|kT/\varepsilon| = .20$, $f = 20\%$ and $|kT/\varepsilon| = .24$, $f = 40\%$, respectively.

the differences mostly cancel, leaving an average signal that is weakly native-like. Of course, under native conditions, the signal from the native conformation overwhelmingly dominates and the overall average signal is native-like. The argument that signals from non-native fluctuations will mostly cancel upon averaging, leaving behind the dominant native signal, is very general, is not unexpected, and should apply to any ensemble-averaged structural measure, including X-ray diffraction patterns.

2.5.3 How do fluctuations influence "structure inversion"?

"Structure inversion" is the process of taking constraints that are Boltzmann-averaged and using them to predict a "single" best structure. (Some recent efforts improve upon structure inversion by fitting constraints to a group of conformers, rather than to one [6, 7, 35, 39, 47, 57, 58].) In reality, these constraints do not come from a single best structure; other conformations also contribute. What distortions are caused by trying to predict a single best structure from averaged constraints? When CDRs are present, the origins of the constraints can be quite complex. Figure 3 shows instances in which very non-native conformations can give native-like distances (bottom right conformation, residues (1, 13)), or in which native-like conformations can have non-native distances (top left conformation, residues (10, 13)). As a consequence, fitting these ensemble-averaged constraints to *one* conformation may very well result in a distorted and physically nonviable structure [32,52,57,58]. This structure is some "average" and is not necessarily the true native conformation (nor even a member of the underlying ensemble which determined the constraints) [32,52]. We explore here the extent to which "single" structures derived from ensemble-averaged constraints resemble the true native conformation.

To test structure-inversion procedures, we begin with a given HP sequence and its proper Boltzmann-weighted ensemble of conformations. We then calculate the average inter-residue distances, $\{\langle d_{ij} \rangle\}$ (i.e., an $\langle$distance map$\rangle$). This reflects what an experiment might measure. One way to turn this information into a structure is to use the $\{\langle d_{ij} \rangle\}$ as distance-constraint inputs to a distance geometry (DG) embedding algorithm [16] and then to project into 2D (since the model's conformations are all

two dimensional). The output, which we call a "distance-averaged" structure, is a single structure that satisfies the constraints. Our process is idealized: (i) There are no artifacts due to insufficient constraints or due to noise. The constraints on *all* pairwise distances are calculated and are precise; there is no need for the creation of a "bounds matrix". Consequently, we determine only one structure. (ii) Unlike experimental NMR structure determination, we are able to include long distance information. (iii) We do not include an optimization/refinement step because there is no direct correspondence of typical methods for the 2D lattice model. Our aim here is not to test a particular refinement method but rather to explore the consequences of fitting a single structure to constraints derived from an ensemble average.

Figure 9 shows the distance-averaged structures. At low temperatures, where the single native conformation dominates strongly and the population of fluctuations is not too large, the structure-inversion procedure always returns the correct native structure. Inversion is robust, even in the presence of CDRs.

But when the population of fluctuations is great, as would occur under conditions approaching denaturation, the structure-inversion process becomes increasingly poor, giving bond angles and lengths and steric clashes that violate the "lattice chemistry" that was obeyed by all the conformations in the underlying ensemble. The structures do not accurately reflect any conformation in the ensemble that generated the constraints. When the deviations from ideal chemistry are small enough, the distortions might be corrected in a subsequent optimization step, but for larger deviations, refinement might not be able to find the native structure and will likely lead to poor agreement between the final structure and the input constraints (see also [40]).

The poor quality of these structures is due to internally inconsistent constraints, as indicated by the fact that the DG embedding algorithm outputs more than two non-zero eigenvalues (two for two dimensions). The inconsistency fundamentally arises because there is no single structure that corresponds to the entire ensemble [7, 32, 34, 40, 52, 57, 58]. Inconsistent constraints may be more of a problem under conditions with many fluctuations, e.g., for "conformationally-ambivalent" proteins like prions or amyloidogenic proteins, for proteins having disordered or struc-

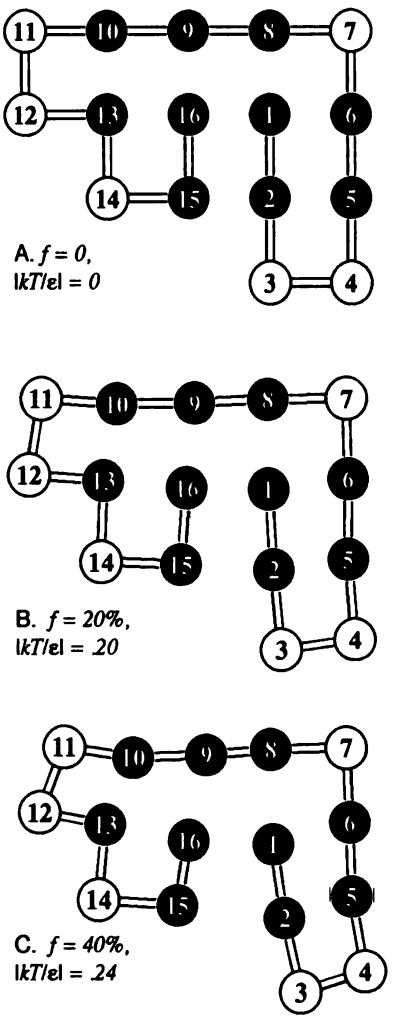


Figure 9: Distance-averaged structures of sequence R. These structures were calculated using average inter-residue distances, $\{\langle d_{ij} \rangle\}$, as constraints to a standard DG embedding algorithm [16]. A: The native conformation ($f = T = 0$). B and C: Structures at increasing temperature and f: $|kT/\varepsilon| = .20$, $f = 20\%$ and $|kT/\varepsilon| = .24$, $f = 40\%$, respectively.

turally unresolved regions, or for peptides.

2.5.4 Are fluctuations Gaussian/harmonic variations?

In addition to determining the dominant native conformation, structure-determination procedures also produce some representation of the fluctuations of the ensemble—Debye-Waller factors in crystallography or some family of related structures in NMR. These fluctuation conformations are assumed to resemble the native conformation. The Debye-Waller factor, B, is based on the assumption that the position of each atom obeys a Gaussian distribution around its equilibrium (see, e.g., [25]), i.e., that the atom shouldn't stray very far. In addition, during the optimization/refinement steps of both NMR and crystallography, output structures are restrained to obey all experimentally measured constraints. This is done by imposing a restraining potential or penalty function such that the lowest energy is centered at the experimental value and increases monotonically with increasing deviation from that value. The function is often a simple harmonic potential [60]. Hence, *each* of the structures generated during structure-determination obeys *all* of the constraints. This is true of all of the "related structures" produced by NMR structure-determination. But what if proteins have rugged energy landscapes and CDR fluctuations? In this section, we ask whether this assumption will result in correct predictions of the fluctuation conformations.

To test the assumption, we see if each of the distances, d_{ij}, of the first excited conformations is close to the constraining value, $\langle d_{ij} \rangle$. If the assumption is good, $\Delta d_{ij} = d_{ij} - \langle d_{ij} \rangle$, should be small, i.e., within 1 or 2 standard deviations, $\sigma_{ij} = \sqrt{\langle \Delta d_{ij}^2 \rangle}$. σ_{ij} is a measure of the variability in the i-to-j distance over the whole ensemble. Figure 10 shows the distributions of Δd_{ij} for two different first-excited conformations and for all the first-excited conformations combined. The deviation of each i-to-j distance (in units of σ_{ij}) from its ensemble-averaged mean, $\Delta d_{ij}/\sigma_{ij}$, is shown for all nonbonded (i, j) pairs. A Gaussian distribution is drawn for comparison (dotted line).

Figure 10A shows a conformation having inter-residues distances that are

well fit by a Gaussian approximation. Most of its distances are close to the constraining values. Figure 10B shows a case in which the Gaussian approximation is poorer. If one were to use a restraining potential to force every distance to lie within 1 or 2σ of the native structure, this CDR conformation would be discarded as having an energy that is too high. The distributions of Δd_{ij} for all 37 first-excited conformations were averaged together (figure 10C). Considerably more inter-residue distances are 2 or 3σ away from their mean values than would be expected if the distribution were Gaussian, even though *all* the CDR fluctuation conformations contributed to the σ_{ij}.

Our results show that there can be fluctuations having distances that are not close to the native distances. Distances averaged over the entire ensemble are close to native distances, but some of the distances of individual fluctuations are not. Using a restraining potential that forces *each* fluctuation to lie close to the measured average can lead to the false prediction that there are no CDR fluctuations. Similarly, Debye-Waller factors do not imply that residues are limited to distances $\sim\sqrt{B}$ from their equilibrium positions.

What fluctuations are predicted by assuming that every conformation's properties must be close to the corresponding experimental values? We create 20 structures, each of whose distances lie within 1 or $2\sigma_{ij}$ of the mean, $\langle d_{ij} \rangle$. These structures are shown in figure 11 aligned to the distance-averaged structure (the same as figure 9C) which is the dark line. Figure 11A shows that when the inter-residue distances are within 1σ of their mean values, the predicted fluctuations form an envelope around the distance-averaged structure. The fluctuations all have a native-like fold. These predicted fluctuations are not an accurate representation of the real fluctuations. For example, residue 1 would seem to be completely buried but is actually exposed in 70% of the first-excited conformations. Residue 9, which is completely exposed to the solvent according to the predicted fluctuations, is actually fully buried in the core in 54% of the conformations that represent the true underlying fluctuations. When the constraints on the distances are loosened to be within 2σ of the mean (measured) values, some of the conformations actually have a non-native fold (figure 11B).

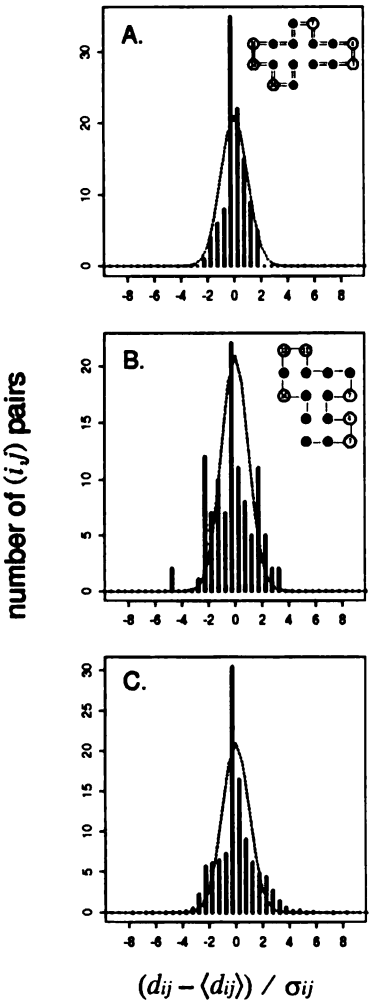


Figure 10: A: The distribution of $\Delta d_{ij}/\sigma_{ij}$ for the first-excited conformation shown in the upper right. The average values $\langle d_{ij} \rangle$ and σ_{ij} are calculated at $|kT/\varepsilon| = .24$ (f=40%). For comparison, a Gaussian distribution with standard deviation of 1 is drawn (dotted line). B: The same as figure A, except for a different fluctuation conformation. Some i-j distances are several standard deviations from their mean values. The use of a harmonic restraining potential in such cases (which results in a Gaussian distribution of distances) might lead to this conformation's being incorrectly discarded as having a too high apparent energy. C: The $\Delta d_{ij}/\sigma_{ij}$ distribution, averaged over all 37 first-excited conformations (over $37 \times 105 = 3885$ residue pairs). The first-excited conformations, on average, have a non-Gaussian distribution of distances with i-j pairs several standard deviations away from their mean values.

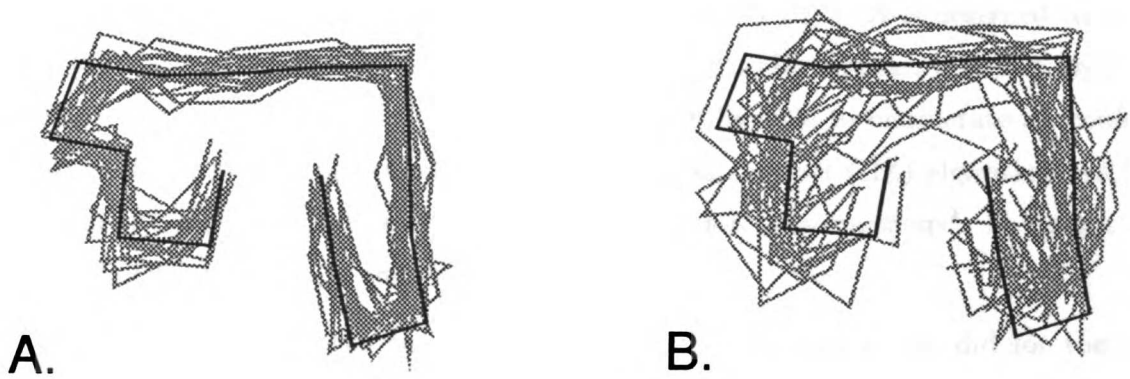


Figure 11: Structural variations predicted by assuming that all fluctuation conformations must closely obey measured constraints. Calculations are done with data averaged at $|kT/\varepsilon| = .24$, $f = 40\%$. To calculate structures whose i-j distances lie approximately within $\langle d_{ij} \rangle \pm \sigma_{ij}$ and $\pm 2\sigma_{ij}$, we imitate NMR structure determination using the following distance bounds: the lower and upper bounds are set to $\max(\langle d_{ij} \rangle - w\sigma_{ij}, 1)$ and $\min(\langle d_{ij} \rangle + w\sigma_{ij}, |j-i|)$, respectively, and then smoothed [19]. 1 and $|j-i|$ are the minimum and maximum distances between i and j on a lattice, analogous to the sum of van der Waal's distances and to a stretched out chain for real proteins. w is 1 or 2 for tighter or looser bounds. A random metrization procedure [38] is used to consistently choose inter-residue distances from within the bounds for input to the DG embedding algorithm and to create a diverse sampling of structures. Again, no optimization step is performed. This procedure is repeated 20 times to create 20 structures. The multiple structures are aligned to the distance-averaged structure (that of figure 9C) shown in black. Subfigures A and B show structures determined with distances chosen from within the 1σ and 2σ bounds, respectively.

2.5.5 The results are similar in 3D.

To test whether our conclusions are limited to 2D, we also study a 27-mer sequence in 3D: PHPHHHPHHHPPHHPPPHPHHPHPPPH. It is designed to have few (20) native conformations [65] which are structurally similar, differing only by surface loop flips. Since it is not computationally feasible to enumerate all conformations for a 27-mer in 3D, we use the hydrophobic zipper (HZ) algorithm [18,24] to generate a representative sample of the fluctuation conformations[2]. As in 2D, we find CDR fluctuations in this 3D simulation.

We calculate the same average physical properties as we did for the 2D sequences. Figures 12 and 13 show the ⟨distance map⟩s and the simulated NOE spectra at different f. Figure 14 shows the distance-averaged structures. Again, the distribution of $\Delta d_{ij}/\sigma_{ij}$ is non-Gaussian. Some of the residue pair distances are several σ_{ij} from their means. (Data not shown.)

There are no qualitative differences between the results in 3D and those in 2D. There are CDRs, but they are masked by average measures of structure and by standard structure-determination procedures. Structure-inversion determines the correct native structure under native conditions, but may produce distorted structures as conditions approach the denaturation point.

---

[2] HZ is designed to generate low energy conformations rapidly for any sequence. HZ conformations are often clearly non-native and thus make a useful model of CDR fluctuations. We make two approximations: (i) We assume that the HZ conformations are a representative sample of all low energy conformations. (ii) HZ generates a set of HH contact maps, not a set of conformations. (An HH contact map is the set of all pairs of H residues which are in contact.) Instead of generating all the n_m conformations consistent with each contact map, m, we generate one sample conformation, s_m. Then when calculating any average quantity, $\langle X \rangle$, s_m is weighted by n_m:

$$\langle X \rangle = \frac{\sum_m X(s_m)\, n_m\, e^{-E_m/kT}}{\sum_m n_m\, e^{-E_m/kT}} \tag{5}$$

where the sum is over all HZ contact maps, m, and $X(s_m)$ is the value of X for s_m, and E_m is the energy of m. The justification for the second approximation is that conformations which have the same contact map are more similar to each other than to conformations with different contact maps [41].

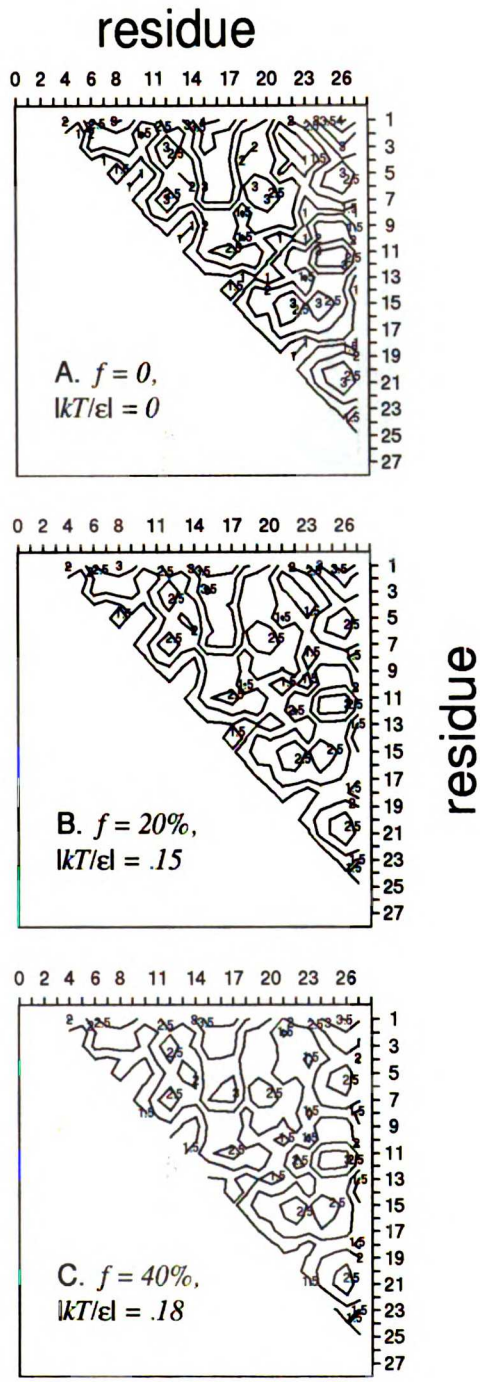


Figure 12: Contour plots of the ⟨distance map⟩s for the 27-mer sequence in 3D. A: The distance map of the native conformation ($T = f = 0$). B and C: ⟨Distance map⟩s at increasing T and f: $|kT/\varepsilon| = .15$, $f = 20\%$ and $|kT/\varepsilon| = .18$, $f = 40\%$, respectively.

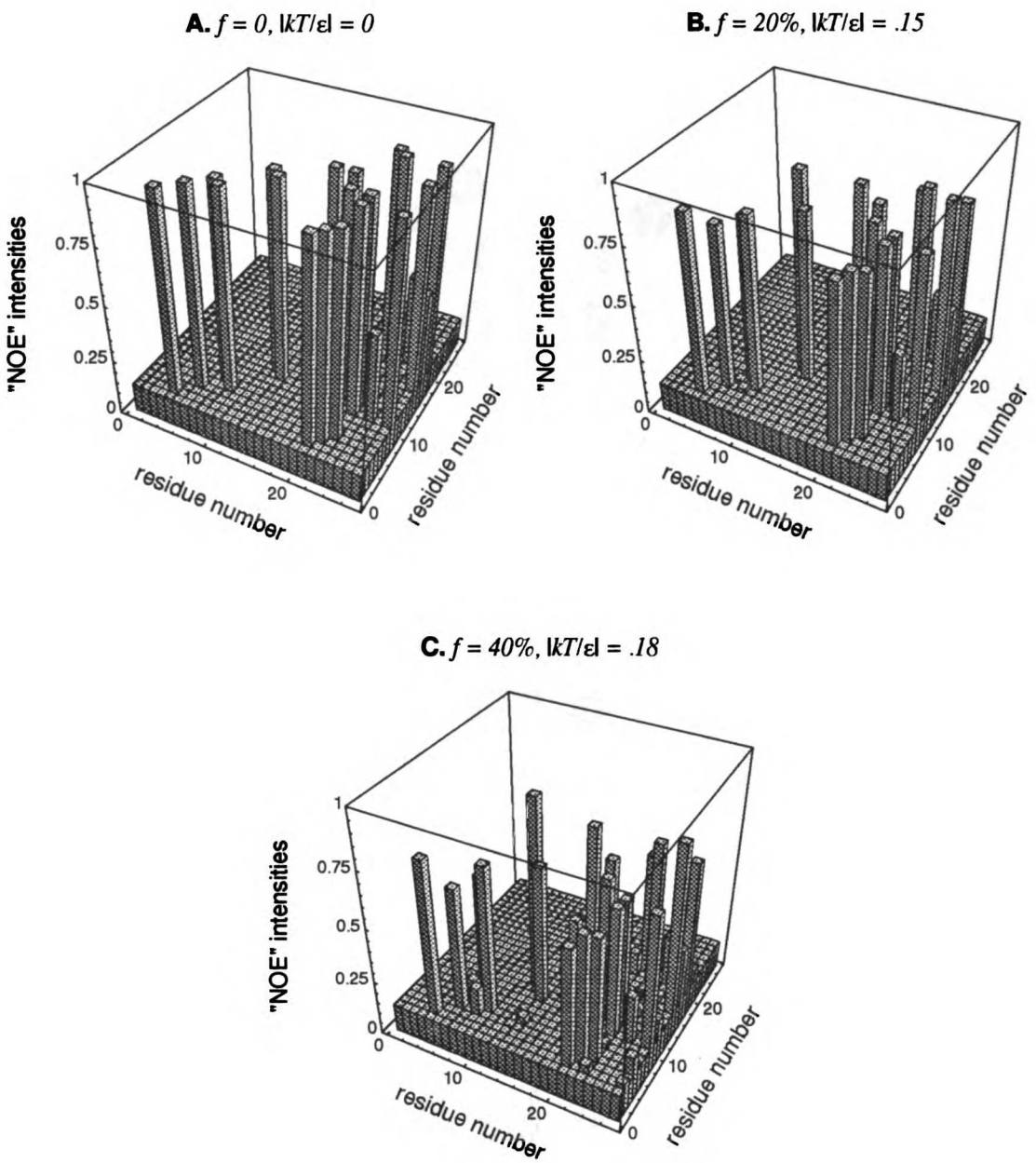
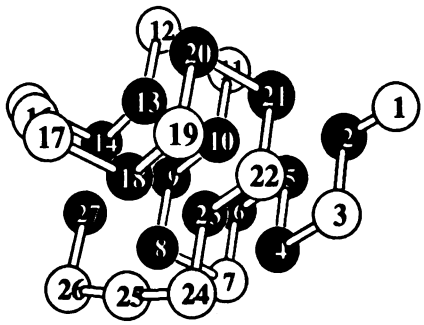
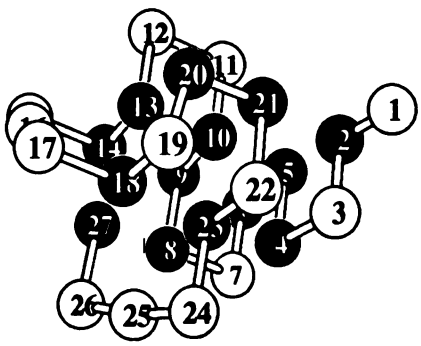


Figure 13: "NOE" spectra of the 27-mer sequence in 3D. The figure is interpreted as for figure 8. A: The native "NOE" spectrum ($T = f = 0$). B and C: "NOE" spectra at increasing T and f: $|kT/\varepsilon| = .15$, $f = 20\%$ and $|kT/\varepsilon| = .18$, $f = 40\%$, respectively.

A. $f = 0, |kT/\varepsilon| = 0$



B. $f = 20\%, |kT/\varepsilon| = .15$



C. $f = 40\%, |kT/\varepsilon| = .18$

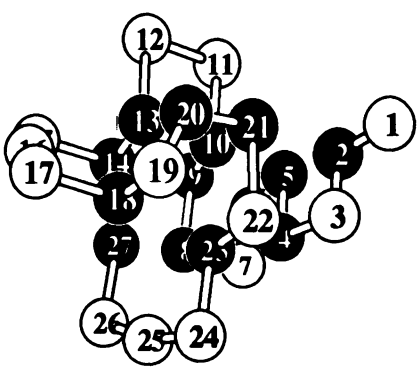


Figure 14: Distance-averaged structures of the 27-mer sequence in 3D. A: The native conformation ($T = f = 0$). B and C: Structures at increasing T and f: $|kT/\varepsilon| = .15$, $f = 20\%$ and $|kT/\varepsilon| = .18$, $f = 40\%$, respectively.

2.5.6 How might experimentalists find CDR fluctuations?

To detect CDRs requires a method that either does not measure ensemble averages or that is not overwhelmed by a signal from the native conformation under native conditions. A possibility is a method developed by Ermácora et al. [22, 23] for mapping the structures of non-native equilibrium conformations. EDTA-Fe is covalently attached to a cysteine residue. Whenever the EDTA-Fe complex hits the peptide backbone, the backbone undergoes a single self-cleavage reaction. Mapping these cleavage sites might show non-native points of contact with the EDTA-Fe complex. Perhaps this technique could be used with mass spectrometry to identify the infrequently occurring cleavage sites, corresponding to weakly populated conformations. CDR conformations are indicated when cleavage occurs at a site which is distant from or inaccessible to the EDTA-Fe complex in the native conformation. Fluorescence energy transfer experiments might also be able to detect non-native inter-residue distances [4, 31].

The molecules that are more likely to have CDRs are peptides, "conformationally-ambivalent" proteins like prions and amyloidogenic proteins, and perhaps proteins with slow or multi-state folding kinetics.

2.6 Summary

The fluctuations of proteins in their native states are often regarded as small wiggles. But there is an important distinction between a *small difference in energy* relative to the native conformation versus a *small difference in conformation.* According to the Boltzmann distribution law, under native conditions, fluctuations must involve small differences in energy, but they need not necessarily involve small changes in conformation. Recent statistical mechanical models [8, 9, 11, 15, 17, 29, 55] predict the possibility of some ruggedness in protein folding energy landscapes. It follows directly (see also [45]), independently of the details of the particular model used, that protein fluctuations may occasionally involve "conformational distant relatives" (CDRs), conformations that are near native in energy but quite different than native

in structure.

We have used a simple model to study how a wide range of fluctuations, including CDRs, affect experimental measures of structure and structure-determination experiments. This study is not meant to be a model of real proteins nor a detailed study of structure-determination. Our intention is to test a principle, to understand how fluctuations influence structure (as determined by experiments).

We find that standard measures of structure, like NMR NOE spectra, are unlikely to be able to detect CDRs and the ruggedness of an energy landscape. These experiments average over the entire ensemble of conformations and the signal from the non-native character of the CDRs cancels out. As a consequence, structure-determination procedures, which use these experimentally measured constraints as inputs, produce structures which are very close to the native conformation. Of course, this result has a positive aspect, namely that structures determined under native conditions are robust and insensitive to CDR fluctuations. On the other hand, being able to determine a good structure implies neither few nor small fluctuations. In addition, the fluctuations predicted by structure-determination—the implied fluctuations of Debye-Waller factors and the multiple structures produced by NMR—may not accurately reflect the true ensemble of fluctuations, if CDRs are present.

As a protein approaches its denaturation point, the population of fluctuations will be great. Then, the problem of structure-inversion, determining a single structure from ensemble-averaged constraints, can be challenging because the constraints may not be internally consistent. The lack of self-consistency may simply indicate that the basic strategy of fitting the constraints to a "single" structure is flawed [7, 32, 34, 40, 52, 57, 58]. Constraints come from an ensemble of multiple conformations. This may be particularly problematic for proteins with rugged energy landscapes and CDR fluctuations.

2.7 Acknowledgements

We thank Klaus Fiebig for the use of many computer programs and for help with NMR, David Miller for generating the data for figure 5, and Kai Yue for

designing the 3D sequence and finding its native conformations. We are also grateful to Tom James, Tack Kuntz, Uli Schmitz, and Melissa Starovasnik for teaching us about NMR, and to David Agard, Eric Anderson, and Sarina Bromberg for helpful discussions. K.E.S.T. was supported by an NIH training grant, a UCSF Graduate Dean's Health Science Fellowship, and a UCSF Regent's Fellowship.

References

[1] Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander. Protein folding intermediates: Native-state hydrogen exchange. *Science*, 269:192–197, 1995.

[2] D. Baker and D. A. Agard. Kinetics versus thermodynamics in protein folding. *Biochemistry*, 33(24):7505–7509, 1994.

[3] D. Baker, J. L. Sohl, and D. A. Agard. A protein-folding reaction under kinetic control. *Nature*, 356:263–265, 1992.

[4] J. M. Beechem and E. Haas. Simultaneous determination of intramolecular distance distributions and conformational dynamics by global analysis of energy transfer measurements. *Biophys. J.*, 55:1225–1236, 1989.

[5] T. L. Blundell and L. N. Johnson. *Protein Crystallography.* Academic Press, New York, 1976.

[6] A. M. J. J. Bonvin and A. T. Brünger. Conformational variability of solution nuclear magnetic resonance structures. *J. Mol. Biol.*, 250(1):80–93, 1995.

[7] R. Brüschweiler, M. Blackledge, and R. R. Ernst. Multi-conformational peptide dynamics derived from NMR data: A new search algorithm and its application to antamanide. *J. Biomol. NMR*, 1:3–11, 1991.

[8] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins*, 21(3):167–195, 1995.

[9] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[10] P. A. Bullough, F. M. Hughson, J. J. Skehel, and D. C. Wiley. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature*, 371:37–43, 1994.

[11] C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA*, 90:6369–6372, 1993.

[12] H. S. Chan and K. A. Dill. Compact polymers. *Macromolecules*, 22:4559–4573, 1989.

[13] H. S. Chan and K. A. Dill. "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.*, 95(5):3775–3787, 1991.

[14] F. E. Cohen, K.-M. Pan, Z. Huang, M. Baldwin, R. J. Fletterick, and S. B. Prusiner. Structural clues to prion replication. *Science*, 264:530–531, 1994.

[15] D. G. Covell and R. L. Jernigan. Conformations of folded proteins in restricted spaces. *Biochemistry*, 29:3287–3294, 1990.

[16] G. M. Crippen and T. F. Havel. Stable calculation of coordinates from distance information. *Acta Crys.*, A34:282–284, 1978.

[17] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding. A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[18] K. A. Dill, K. M. Fiebig, and H. S. Chan. Cooperativity in protein folding kinetics. *Proc. Natl. Acad. Sci. USA*, 90:1942–1946, Mar. 1993.

[19] A. W. M. Dress and T. F. Havel. Shortest-path problems and molecular conformation. *Discrete Appl. Math.*, 19:129–144, 1988.

[20] S. W. Englander. Measurement of structural and free energy changes in hemoglobin by hydrogen exchange methods. *Ann. NY Acad. Sci.*, 244:10–27, 1975.

[21] S. W. Englander and N. R. Kallenbach. Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Quart. Rev. Biophys.*, 16(4):521–655, 1983.

[22] M. R. Ermácora, J. M. Delfino, B. Cuenoud, A. Schepartz, and R. O. Fox. Conformation-dependent cleavage of staphylococcal nuclease with a disulfide-linked iron chelate. *Proc. Natl. Acad. Sci. USA*, 89:6383–6387, 1992.

[23] M. R. Ermácora, D. W. Ledman, H. W. Hellinga, G. W. Hsu, and R. O. Fox. Mapping staphylococcal nuclease conformation using an EDTA–Fe derivative attached to genetically engineered cysteine residues. *Biochemistry*, 33(46):13625–13641, 1994.

[24] K. M. Fiebig and K. A. Dill. Protein core assembly processes. *J. Chem. Phys.*, 98(4):3475–3487, 1993.

[25] H. Frauenfelder. The Debye-Waller factor: from villain to hero in protein crystallography. *Int. J. Quant. Chem.*, 35:711–715, 1989.

[26] N. Gō, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80:3696–3700, 1983.

[27] N. Goudreau et al. NMR structure of the N-terminal SH3 domain of GRB2 and its complex with a proline-rich peptide from Sos. *Nat. Struct. Biol.*, 1(12):898–907, 1994.

[28] F. R. N. Gurd and M. Rothgeb. Motions in proteins. *Adv. Prot. Chem.*, 33:73–165, 1979.

[29] D. A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668–682, 1994.

[30] A. Hvidt and S. O. Nielsen. Hydrogen exchange in proteins. *Adv. Prot. Chem.*, 21:287–386, 1966.

[31] V. Ittah and E. Haas. Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. *Biochemistry*, 34(13):4493–4506, 1995.

[32] O. Jardetzky. On the nature of molecular conformations inferred from high-resolution NMR. *Biochim. Biophys. Acta*, 621:227–232, 1980.

[33] J. W. Kelly. Alternative conformations of amyloidogenic proteins govern their behavior. *Curr. Opi. Struct. Biol.*, 6(1):11–17, 1996.

[34] H. Kessler, C. Griesinger, J. Lautz, A. Müller, W. F. van Gunsteren, and H. J. C. Berendsen. Conformational dynamics detected by nuclear magnetic resonance NOE values and J coupling constants. *J. Am. Chem. Soc.*, 110:3393–3396, 1988.

[35] Y. Kim and J. H. Prestegard. A dynamic model for the structure of acyl carrier protein in solution. *Biochemistry*, 28(22):8792–8797, 1989.

[36] D. E. Koshland, Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44:98–104, 1959.

[37] D. E. Koshland, Jr. Enzyme flexibility and enzyme action. *J. Cellular Comp. Physiol.*, 54:245–258, 1959.

[38] J. Kuszewski, M. Nilges, and A. T. Brünger. Sampling and efficiency of metric matrix distance geometry: A novel partial metrization algorithm. *J. Biomol. NMR*, 2:33–56, 1992.

[39] C. Landis and V. S. Allured. Elucidation of solution structures by conformer population analysis of NOE data. *J. Am. Chem. Soc.*, 113:9493–9499, 1991.

[40] E. E. Lattman. Why are protein crystallographic R-values so high? *Proteins*, 25(1):i–ii, 1996.

[41] E. E. Lattman, K. M. Fiebig, and K. A. Dill. Modeling compact denatured states of proteins. *Biochemistry*, 33:6158–6166, 1994.

[42] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[43] K. F. Lau and K. A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA*, 87:638–642, 1990.

[44] R. Lumry and A. Rosenberg. The mobile defect hypothesis of protein function. *Colloq. Int. CNLRS*, 246:55–63, 1975.

[45] D. W. Miller and K. A. Dill. A statistical mechanical model of hydrogen exchange in globular proteins. *Prot. Sci.*, 4(9):1860–1873, 1995.

[46] J. Mottonen et al. Structural basis of latency in plasminogen activator inhibitor-1. *Nature*, 355(6357):270–273, 1992.

[47] G. V. Nikiforovich, O. Prakash, C. A. Gehrig, and V. J. Hruby. Conformations of the dermenkephalin backbone in DMSO solution by a new approach to the solution conformations of flexible peptides. *J. Am. Chem. Soc.*, 115(9):3399–3406, 1993.

[48] K.-M. Pan et al. Conversion of α-helices into β-sheets features in the formation of the scrapie prion proteins. *Proc. Natl. Acad. Sci. USA*, 90:10962–10966, 1993.

[49] A. L. Patterson. A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.*, 46:372, 1934.

[50] J. W. Peng and G. Wagner. Investigation of protein motions via relaxation measurements. *Meth. Enzym.*, 239:563–596, 1994.

[51] M. F. Perutz and F. S. Mathews. An X-ray study of azide methaemoglobin. *J. Mol. Biol.*, 21:199–202, 1966.

[52] G. A. Petsko. Not just your average structures. *Nat. Struct. Biol.*, 3(7):565–566, 1996.

[53] F. M. Richards. Packing defects, cavities, volume fluctuations, and access to the interior of proteins. *Carlsberg Res. Commun.*, 44:47–63, 1979.

[54] D. Ringe and G. A. Petsko. Mapping protein dynamics by X-ray diffraction. *Prog. Biophys. Molec. Biol.*, 45:197–235, 1985.

[55] E. I. Shakhnovich and A. M. Gutin. Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.*, 93(8):5967–5971, 1990.

[56] D. R. Shortle. Structural analysis of non-native states of proteins by NMR methods. *Curr. Opi. Struct. Biol.*, 6:24–30, 1996. And references therein.

[57] A. E. Torda, R. M. Scheek, and W. van Gunsteren. Time-dependent distance restraints in molecular dynamics simulations. *Chem. Phys. Lett.*, 157(4):289–294, 1989.

[58] A. E. Torda, R. M. Scheek, and W. F. van Gunsteren. Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J. Mol. Biol.*, 214:223–235, 1990.

[59] J. Tropp. Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. *J. Chem. Phys.*, 72(11):6035–6043, 1980.

[60] W. F. van Gunsteren, R. M. Brunne, P. Gros, R. C. van Schaik, C. A. Schiffer, and A. E. Torda. Accounting for molecular mobility in structure determination based on nuclear magnetic resonance spectroscopic and X-ray diffraction data. *Meth. Enzym.*, 239:619–654, 1994. And references therein.

[61] G. Wagner. The importance of being floppy. *Nat. Struct. Biol.*, 2(4):255–257, 1995.

[62] C. Woodward, I. Simon, and E. Tüchsen. Hydrogen exchange and the dynamic structure of proteins. *Mol. Cell. Biochem.*, 48:135–160, 1982.

[63] C. K. Woodward and A. Rosenberg. Studies of hydrogen exchange in proteins. VI. Urea effects on RNase hydrogen exchange kinetics leading to a general model for hydrogen exchange from folded proteins. *J. Biol. Chem.*, 246:4114–4121, 1971.

[64] D. P. Yee and K. A. Dill. Families and the structural relatedness among globular proteins. *Prot. Sci.*, 2:884–899, 1993.

[65] K. Yue and K. A. Dill. Sequence-structure relationships in proteins and copolymers. *Phys. Rev. E*, 48(3):2267–2278, 1993.

[66] O. Zhang and J. D. Forman-Kay. Structural characterization of folded and unfolded states of an SH3 domain in equilibrium in aqueous buffer. *Biochemistry*, 34:6784–6794, 1995.

Chapter 3

The Relationship Between Protein Flexibility and Stability

Karen E. S. Tang and Ken A. Dill

3.1 Abstract

We study the fluctuations of native proteins by exact enumeration using the HP lattice model. First, we investigate how fluctuations grow with temperature. We observe a low-temperature point below which large fluctuations of the protein are frozen out. The behavior resembles that observed by Petsko and colleagues [56] who showed that the thermal motions of ribonuclease A increase sharply above about $220K$. This "rigor mortis point" may not be a glass transition. Second, we find an inverse correlation between the stability of a protein and its "flexibility" as determined by Debye-Waller-like factors and solvent accessibilities of core residues to hydrogen exchange. Proteins having high stability have fewer large fluctuations and hence lower flexibility. The model allows us to conjecture why proteins from thermophilic organisms, which are exceptionally stable, may be catalytically inactive at normal temperatures and why there is yet no dominant theme underlying thermostability.

3.2 The temperature dependence of thermal motions in globular proteins

Proteins in their native state are not in a single conformation. They consist of a Boltzmann ensemble dominated by the native conformation, with smaller populations of fluctuation conformations. The fluctuations are the conformations that are occasionally visited as the protein responds to Brownian motion. Fluctuations are important for the catalytic functions of globular proteins, for induced fit mechanisms of ligand binding [29,30], and for allosteric regulation. The thermal fluctuations play a role in the Debye-Waller factors in x-ray crystallography and in hydrogen-deuterium exchange (HX) rates, properties which are sometimes thought to be measures of "flexibility".

Flexibility can be defined either as a static equilibrium property or as a dynamic property. Static flexibility refers to the size of the conformational ensemble of fluctuations; it refers to the number and structural diversity of fluctuation conforma-

tions at equilibrium. These conformations are populated according to the Boltzmann distribution law, irrespective of the barriers that must be overcome to interconvert between them. Dynamic flexibility is determined by how quickly the protein can hop from one such conformation to another, and is a measure of the energy barriers between the native and fluctuation conformations. Because the time scales of HX and of x-ray experiments are long relative to the fast motions, and sometimes even long relative to the folding times, Debye-Waller factors and HX properties probably reflect mainly the static flexibilities. For example, the intrinsic time scale of HX for random-coil poly-alanine varies from .1 msec to 10 min (from pH 1 to 9) [16]. We focus here on the static flexibilities.

A most remarkable observation was made a few years ago by Greg Petsko and his colleagues. Through a crystallographic tour de force, Tilton et al. [56] performed x-ray diffraction experiments on ribonuclease A at nine different temperatures. They observed a sort of "transition" around $220K$. Crystallographic Debye-Waller factors are small and have little dependence on temperature below about $220K$, but grow more rapidly with temperature above $220K$ [56]. Similar results appear in earlier studies of different properties. Atomic mean-square displacements, as measured by Mössbauer scattering [5, 28, 43, 44] and inelastic neutron scattering [15] have this behavior; so do quenching rates [39] and viscoelastic properties [37]. In addition, carbon monoxide rebinding to myoglobin shows non-exponential relaxation and non-Arrhenius rates below about $200K$ [4, 18, 25]. At low temperatures, certain motions are frozen out, although the nature of these motions is unclear [2, 5, 15, 18, 21, 23, 28, 33, 37, 43, 44, 47, 56, 58]. It has been suggested that the protein/solvent system may be in a glass-like state [14, 20, 25, 37]. However, we find that equilibrium (not glassy) fluctuations can cause similar behavior. Hence we refer to the "transition" as the "rigor mortis point". Our interest here is in understanding what types of conformational fluctuations might give rise to it.

In this paper, we also address another issue. Studies suggest that various protein properties correlate with protein flexibility: (i) The greater the stability of a protein, the less flexible it is, as measured by HX [1, 12, 27, 40, 61, 62] and by fluorescence quenching [59]. There is some evidence along the same lines from proteolysis

experiments [11,34,54] but the data are not conclusive (e.g., [6,38]). (ii) Proteins with few "flexible residues" tend to be more thermostable [35,60]. (iii) It has been pointed out that active sites are not optimized for stability [52] and are more flexible than other parts of the chain [57]. We use the HP lattice model to explore the possible physical bases for relationships between the flexibilities and the stabilities of proteins.

3.3 The model

3.3.1 The HP lattice model of proteins

We model the fluctuations of proteins using the two-dimensional (2D) HP lattice model [10, 13, 31, 32]. Proteins are represented as specific sequences of H (hydrophobic) and P (other) residues on a 2D square lattice. Each amino acid can only occupy one lattice site, and no two amino acids may reside on the same site. The energetic interactions, designed to capture the essence of the hydrophobic interaction, consist of a single term: there is a favorable interaction, $\varepsilon < 0$, whenever two non-bonded H residues are on adjacent lattice sites, i.e., "in contact". Hence, the free energy of any conformation is $h\varepsilon$, where h is the number of HH contacts. The lowest energy conformation, with h_{nat} HH contacts, is the "native" conformation. We call those conformations with $h_{nat} - 1$ contacts the "first-excited" conformations since they have the next-to-lowest energy. Conformations with $h_{nat} - 2$ contacts belong to the "second-excited" states, etc. In this work, we study only sequences with a single native conformation.

The limitations of the model are obvious: the chains are short, two-dimensional, and low resolution, with limited bond angles and bond distances; there are only two amino acid types; and the interactions are simplified. But we believe the model captures the major components of protein folding—the hydrophobic interactions, the conformational freedom of the chain, and the steric restrictions imposed by excluded volume. This model has many protein-like properties [13,36]. In particular, the native conformations are compact with a hydrophobic core and secondary structure [9,31]. The advantage of this model is that, since the chains are short and in 2D, one can

enumerate all conformations exactly. Therefore we are able to consider fluctuations of all sizes, without restriction to small-amplitude motions or to short time scales.

3.3.2 The Fluctuation Conformations

We define the "fluctuation conformations" to be all conformations that have non-zero populations, excluding the native conformation. What determines the fluctuation conformations? Under equilibrium conditions, every conformation, c, is populated according to its Boltzmann probability:

$$probability\ of\ c = p(c) = \frac{e^{-E_c/kT}}{Q} \tag{1}$$

where E_c is the energy of conformation c (equal to $h\varepsilon$ in our model), T is the absolute temperature, and k is Boltzmann's constant. Q is the partition function:

$$Q = \sum_{c=1}^{N} e^{-E_c/kT} \quad , \tag{2}$$

the sum being over all N possible conformations. At low temperatures, only the low energy conformations are populated; high energy conformations are not. Therefore the fluctuation conformations are low-lying on the energy landscape, but they are non-native.

By exhaustively enumerating all conformations and using equations 1 and 2, we calculate exactly the population of each conformation, for any temperature. Within the framework of the model, no further approximations or assumptions are made about which fluctuations are important or their relative magnitudes.

Figure 1 shows how the populations of low-energy conformations change with temperature, for sequence R[1], HHPPHHPHHHPPHPHH, shown in its native conformation in figure 2. At $T = 0$, every chain is in the native conformation (ground state); there are no fluctuations. At low T, the native conformation is still dominant but other low-energy conformations are also populated. Ultimately, at high T, the proteins denature, but we focus here always on "native" conditions, i.e., temperatures below the denaturation midpoint.

[1] R is the same sequence as studied in [55].

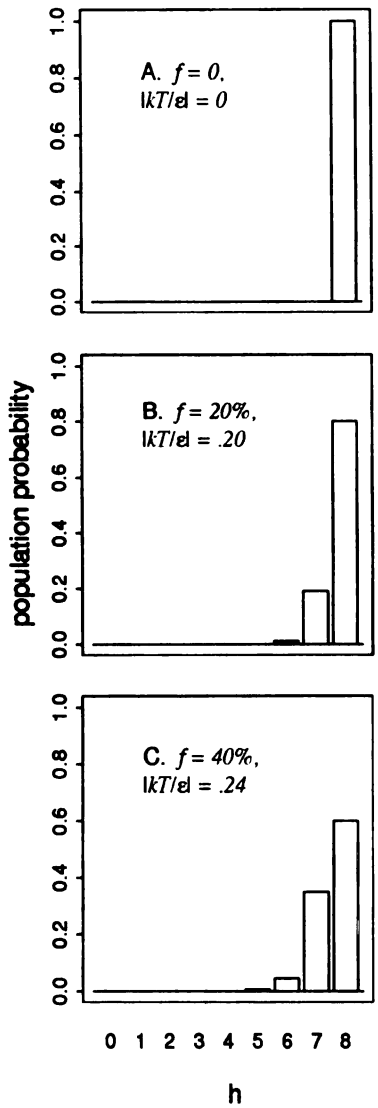


Figure 1: Populations of different energy states, at different temperatures, for sequence R. Energy $= h\varepsilon$, where h is the number of HH contacts and $\varepsilon < 0$. All conformations with the same energy are grouped together. A: At low temperatures, the ensemble consists almost entirely of the native conformation ($h = 8$). At higher temperatures (B: $|kT/\varepsilon| = .20$; C: $|kT/\varepsilon| = .24$), the first-excited conformations ($h = 7$) are also populated and comprise the majority of fluctuation conformations. f is the fraction of the total population which are in fluctuation conformations (i.e., non-native).

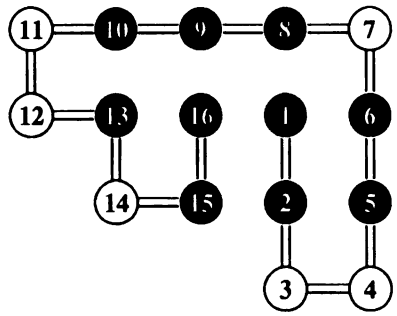


Figure 2: The native conformation of sequence R. Dark beads are hydrophobic (H); light beads are other (P). HH interactions are favorable.

Previous work with this model has shown that the fluctuation conformations under native conditions are mainly the first-excited conformations [36,55]. A sample of R's first-excited conformations is shown in figure 3. These conformations are compact and have hydrophobic cores and secondary structure. Their folds range from being somewhat similar to the native (with a few residues shifted relative to the native conformation) to strongly non-native-like [36,55]. We regard all of the first- (and higher-)excited conformations to be "non-native-like" or "large" fluctuations. For real proteins, a "large" fluctuation would be any conformation which would not be confused for the native conformation. These fluctuations exclude vibrations and small loop/side-chain wiggles.

We have recently explored how experiments might detect non-native-like fluctuations [55]. The HP model tends to have rugged energy landscapes and therefore best models those proteins with similar landscapes, perhaps those that have slow or multistate folding kinetics, or are metastable.

3.3.3 Static equilibrium flexibility measures

Experiments that involve intrinsically long time scales (such as x-ray crystallographic or HX experiments) measure equilibrium properties, that is, averages over the Boltzmann ensemble. For any property, X, the ensemble average is

$$\langle X \rangle = \sum_{c=1}^{N} X_c \, p(c) \tag{3}$$

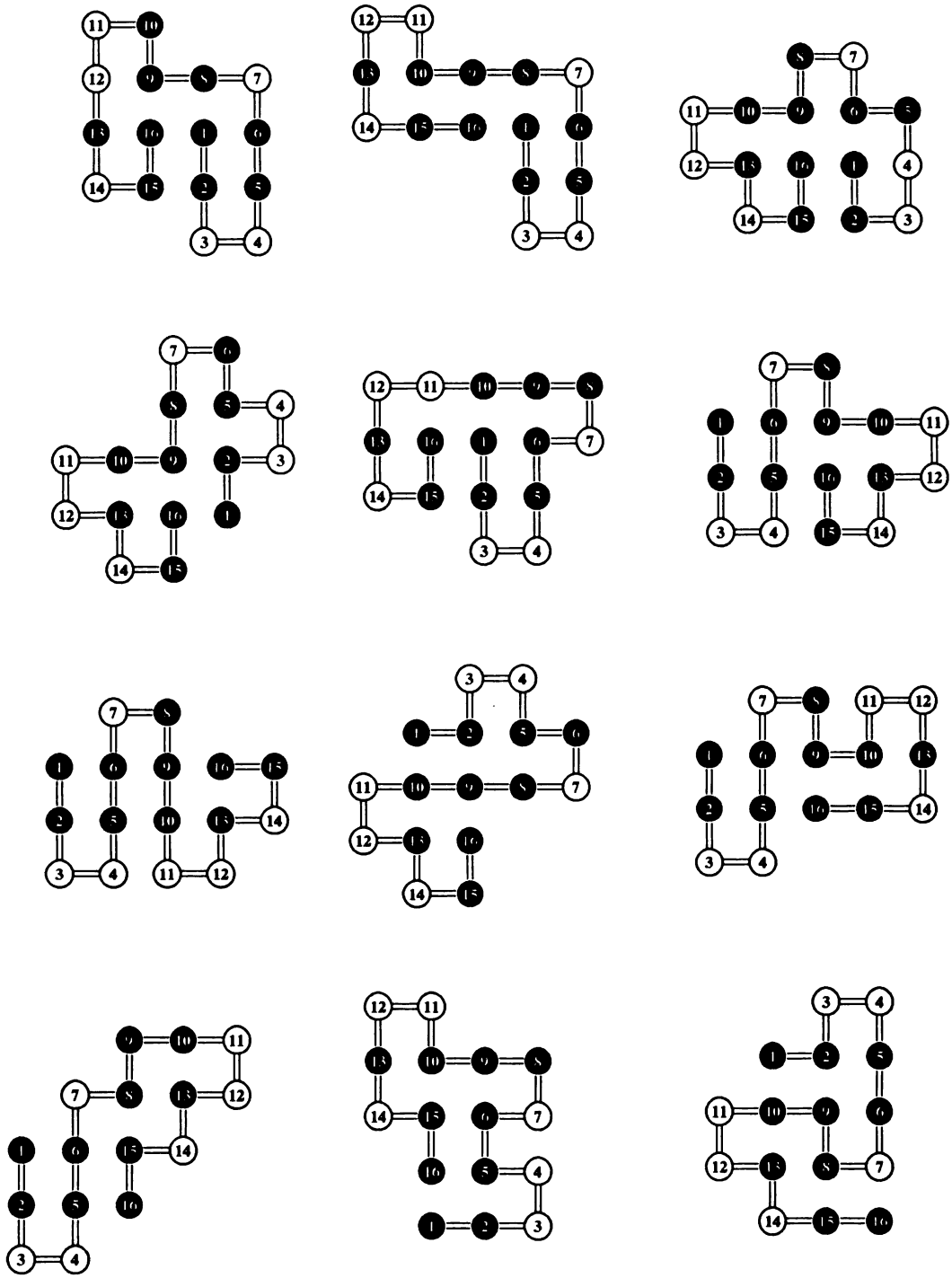


Figure 3: A sample of the 37 first-excited conformations of sequence R. The conformations are ordered from the highest structural similarity to the native conformation at the upper left to the lowest structural similarity at the lower right.

where $p(c)$ is the Boltzmann probability for conformation c, defined in equation 1 and X_c is the value of X for conformation c. The angle brackets, $\langle \cdots \rangle$, indicate the ensemble average.

We simulate two measures of flexibility, crystallographic Debye-Waller-like factors and the HX rate. We compute a quantity, b_i, similar to a Debye-Waller factor for the lattice model:

$$b_i = \frac{1}{n_{nb}} \sum_{j=1}^{n} \langle \Delta d_{ij}^2 \rangle \tag{4}$$

where $\Delta d_{ij} = d_{ij} - \langle d_{ij} \rangle$, is the deviation of the i-j distance away from its mean value and $\langle \Delta d_{ij}^2 \rangle$ is the mean-square deviation. The sum is over all n_{nb} residues which are not bonded to i, i.e., $j \neq i$, $i \pm 1$; n is the length of the chain. b_i is related to the Debye-Waller factor of residue i, $B_i = 8\pi^2 \langle \Delta x_i^2 \rangle$: if residue i is in many different conformations or if the conformations that i is in are very different, b_i is large. b_i has the same dimensions as an atomic mean-square displacement. We introduce a doubly-averaged quantity, $\bar{b}$; the bar indicates an average over all the residues of the chain, in addition to the ensemble average. $\bar{b}$ gives a measure of the overall mean-square displacement of the chain. It is similar to a molecule-averaged Debye-Waller factor. $\bar{b}$ is a measure of the conformational diversity of the ensemble. If the ensemble is dominated by one conformation, $\bar{b}$ is small; if there are many conformations, $\bar{b}$ is large.

The HP lattice model has been used before to study HX protection [36]. The exchange competence of a residue is given by its accessibility to solvent: the accessibility, $A_{c,i}$, of residue i is defined to be 0 if i is completely surrounded by four other residues in conformation c; $A_{c,i} = 1$ if i is adjacent to a lattice site that is occupied by a solvent molecule (i.e., not occupied by a residue). The HX rate of any residue i is proportional to its average accessibility, $\langle A_i \rangle$ [36].

3.4 How fluctuations depend on temperature: the "rigor mortis" point.

To explore the temperature dependence of the fluctuations, we calculated $\bar{b}$ as a function of temperature for many different HP sequences of lengths $n = 16$ to 20. All the sequences we studied show similar behavior. We present data mainly on sequence *R*.

Figure 4 shows the behavior of $\bar{b}$ as a function of temperature for three different $n = 16$ sequences. The model predicts two temperature regimes separated by a temperature T_{rm}. The behavior resembles that of the molecule-averaged Debye-Waller factors in the experiments of Tilton et al. [56]. The basis for this behavior is simple to explore in the model. Figure 1 shows that at low temperatures, $kT \ll |\varepsilon|$, only the native conformation is populated; $\bar{b} = 0$. Vibrational modes are not included in our model and thus the slope for the model is zero at low temperatures. The experimental data, on the other hand, show a small non-zero slope that is due to vibrational motions [15, 19, 44, 49]. As the temperature is increased, the population of first-excited conformations becomes non-negligible and then grows with increasing T; $\bar{b}$ also grows with T. A breakpoint appears at the temperature, T_{rm}, where the first-excited conformations just begin to be populated:

$$\frac{population\ of\ first-excited\ conformations}{population\ of\ native\ conformation} = g(1)\, e^{-|\varepsilon|/kT_{rm}} \sim 1\% \qquad (5)$$

or

$$\frac{kT_{rm}}{|\varepsilon|} \sim \frac{1}{-\ln(.01/g(1))} \qquad (6)$$

where $g(1)$ is the number of first-excited conformations. In our model, the rigor mortis point occurs at the temperature at which the first-excited (i.e., low energy, non-native-like) fluctuations begin to be populated.

Is the rigor mortis point a "glass transition"? By definition, a glass below its transition temperature is a metastable system and is not in equilibrium. But in our model the rigor mortis point is due to *equilibrium* fluctuations and hence is not a glass transition. Nocek et al. [39] have also explained their data by an equilibrium model.

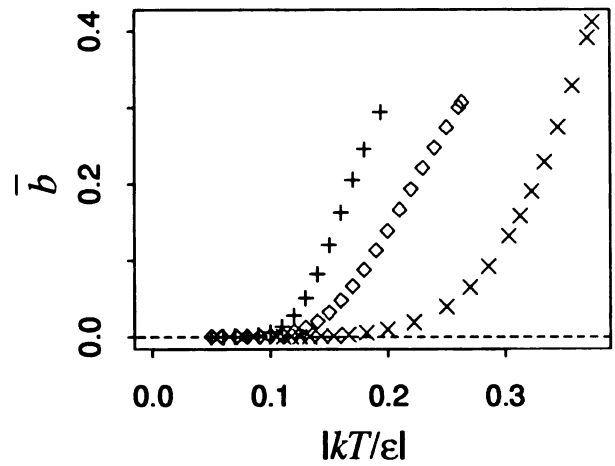


Figure 4: $\bar{b}$ as a function of temperature for three different $n = 16$ sequences. $\bar{b}$ is similar to a molecule-averaged Debye- Waller factor. Data are plotted up to each sequence's denaturation midpoint. (+) Sequence Q, $|kT_m/\varepsilon| = .19$; ($\diamond$) sequence R, $|kT_m/\varepsilon| = .26$; ($\times$) sequence S, $|kT_m/\varepsilon| = .37$. Sequence R was also the primary example in [55]; sequence S was studied in [36].

Angell and coworkers [21] have noted that "[t]his phenomenon itself cannot be what is normally understood by relaxation phenomenologists as the glass transition (though it may well be the triggering mechanism...) because it can be observed in extremely short (picosecond) time scale studies, e.g., [molecular dynamics simulations]." The time scales associated with glass transitions in liquid or plastic crystals are in the range of 100 sec [21]. Mössbauer experiments and inelastic neutron scattering experiments explore much shorter time scales, those faster than 10^{-7} seconds [42] and between 10^{-13} to 10^{-10}sec [15], respectively. Parak and Knapp [45] point out that thermal equilibrium is reached during the time needed for x-ray investigations; hence the Debye-Waller factor is an equilibrium property. On the other hand, experiments on CO binding to myoglobin show non-exponential relaxation and non-Arrhenius reaction rates below $\sim 200K$ [25]. The actual temperature of this "transition" depends crucially on the glass temperature of the solvent; hence the transition is thought to be "slaved" to the solvent [3, 25]. The non-equilibrium behavior has been attributed to the freezing of the solvent or to the solvent's high viscosity or glassy behavior

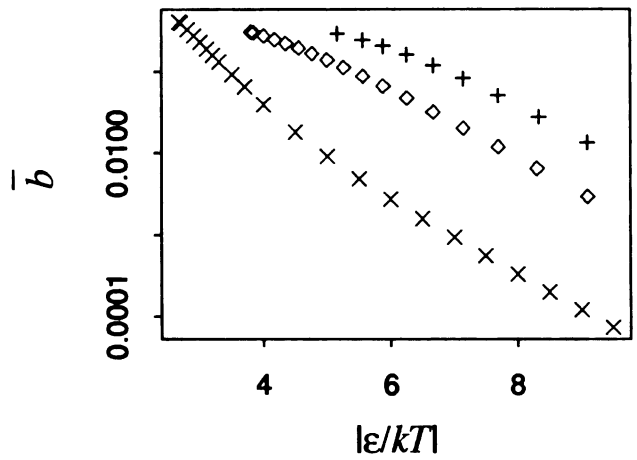


Figure 5: $\ln \bar{b}$ as a function of $1/T$ for three different $n = 16$ sequences. The data are the same as for figure 4: (+) sequence Q; ($\diamond$) sequence R; ($\times$) sequence S.

[2, 14, 23, 58]. However, it is possible that these experiments and the x-ray data are measuring different motions.

Our results show that the fluctuations that are populated above the rigor mortis point may be simply a new class of fluctuations. This has been suggested before. These motions have been attributed to conformational substates [4, 18, 45], anharmonic vibrational modes [33, 47], large-scale collective motions [5], the coupling of fast local motions to slower collective motions [15], cooperative loosening and rapid dynamics of a ligand binding interface [39], or a change in the structure and motion of the solvent around the protein [14, 37, 58]. kT_{rm} is the intrinsic energy of the fluctuations (or of the barriers between them).

The experimental data have also been explained by activated process models [4, 5, 15, 28, 45]. We note that the lattice model results are consistent with two-state activated processes in the equilibrium limit. Figure 5 shows that $\ln \bar{b}$ is approximately linear with $1/T$. Activated processes can explain both equilibrium and non-equilibrium conditions.

We have made an additional comparison of the model with experiments. Tilton et al. [56] observed that increasing the temperature results not only in an increase in the overall Debye-Waller factor but also in a broadening of the distribution

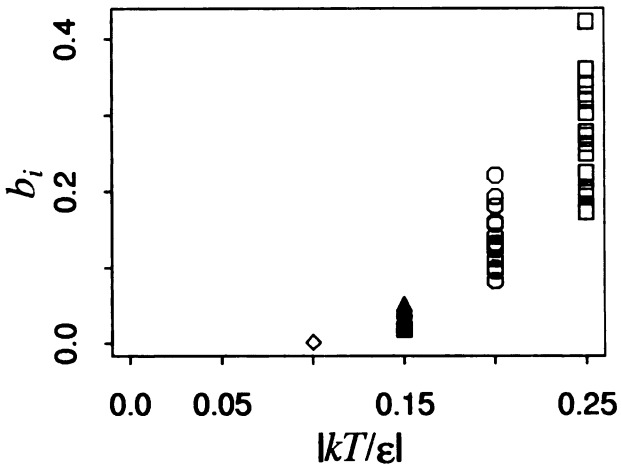


Figure 6: Sequence R: The b_i of every residue i are plotted at different temperatures. The b_i are Debye-Waller-like measures of distance variations for residue i. The distribution of b_i widens as temperature is increased.

of Debye-Waller factors throughout the molecule. We observe the same behavior (figure 6). This indicates that the model gives at least a plausible basis for the chain motions underlying the rigor mortis point.

3.5 Protein flexibility decreases with stability

We also used the lattice model to explore the relationship between flexibility and stability for different HP sequences. To monitor protein stability, we use the denaturation temperature, T_m, the temperature at which the free energy of folding is zero, $\Delta G = 0$, i.e., where 50% of the molecules are in the native conformation. More stable sequences have higher T_m values.

We use three measures of flexibility: the average solvent accessibilities (which correlate with HX rates [36]) of core residues, a Debye-Waller-like measure of distance variations, and the rigor mortis point temperature. We studied ~30 different HP sequences of chain lengths 16 to 20.

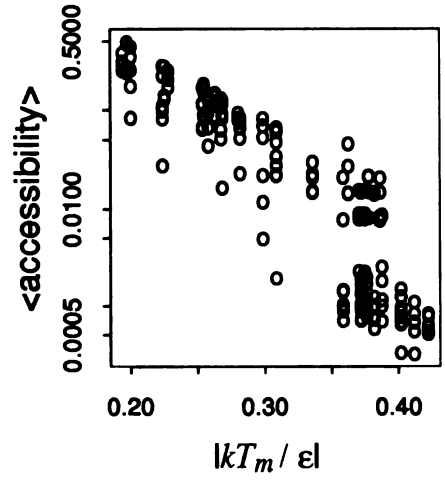


Figure 7: The average accessibility (which is proportional to the HX rate [36]) of core residues versus the denaturation temperature, T_m for $\sim$ 30 different sequences of lengths 16 to 20. A core residue is defined to be one which is completely buried (surrounded by other residues) in the native conformation. Each point represents one core residue of one sequence. As average accessibilities decrease (less flexibility), T_m's increase (greater stability). The average accessibilities are calculated at a fixed temperature of $|kT/\varepsilon| = .20$.

3.5.1 Hydrogen exchange rates

For each sequence, the average solvent accessibility, $\langle A_i \rangle$, is calculated for every residue i that is buried in the native conformation. Figure 7 shows these average accessibilities for all of the sequences versus T_m. Each point represents one residue of one sequence. We find that buried core residues tend to have less solvent accessibility, corresponding to a slower HX rate [36], if they are in more stable proteins.

3.5.2 Debye-Waller factors

We also studied how $\bar{b}$, our analog of the molecule-averaged Debye-Waller factor, correlates with stability. Figure 8 shows that, on average, more stable proteins tend to have smaller and/or fewer thermal motions, by this measure.

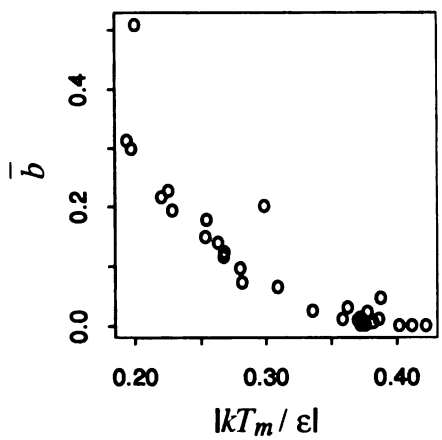


Figure 8: $\bar{b}$ versus T_m. $\bar{b}$ is a quantity similar to a molecule-averaged Debye-Waller factor. Each point represents one sequence. As $\bar{b}$ decreases (less flexibility), the T_m's increase (greater stability). The $\bar{b}$ are calculated at a fixed temperature of $|kT/\varepsilon| = .20$.

3.5.3 The "rigor mortis" point temperature

The rigor mortis point temperature, T_{rm}, is another measure of the flexibility of the molecule. The lower T_{rm}, the more flexible is the protein. The T_{rm} is computed as the point of intersection of two straight lines on the plots of $\bar{b}$ versus $|kT/\varepsilon|$ (like figure 4): one drawn to the limit of low temperatures, and the other at the limit of high temperatures (the highest temperature we used is at the denaturation midpoint). Figure 9 shows the rigor mortis temperature versus the denaturation temperature. This measure of flexibility also correlates inversely with stability. More stable proteins are found to have higher rigor mortis temperatures.

3.5.4 Why are stable proteins less flexible?

By our three measures of flexibility, the results above indicate that HP proteins with greater stability have less flexibility. Why? The non-native-like fluctuations are the link between both properties. If a sequence has many first-excited conforma-

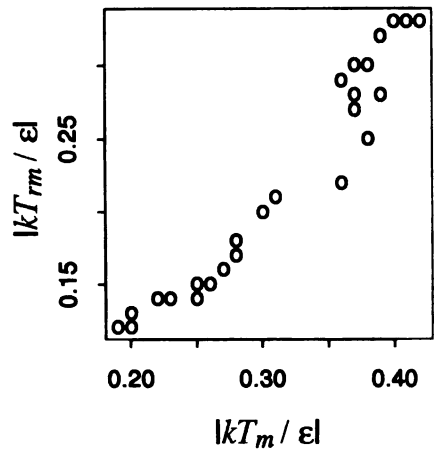


Figure 9: The rigor mortis temperature, T_{rm}, versus the denaturation midpoint temperature, T_m. Lower rigor mortis temperatures correspond to more flexible sequences. As the rigor mortis temperature increases (less flexiblility), T_m's increase (greater stability).

tions, or the energy of that level is low, the protein will be both unstable and have many populated fluctuations. But if the energy gap to the first-excited state is large, or the number of first-excited conformations is small, the protein will be stable and there will be few populated non-native-like fluctuations. Hence, stability and static equilibrium flexibility inversely correlate.

3.5.5 Speculations on why thermophilic proteins have low enzyme activities at room temperature and why there is no dominant theme underlying thermostability.

It follows that whatever stabilizing factors lower the native-state free energy relative to the non-native-like conformations will also diminish the population of these conformations (hence decreasing the flexibility of the protein). It has been argued that proteins need big stability gaps in order to have stable native folds and rapid folding kinetics [50]. But proteins may also need flexibility for enzyme function, induced fits of ligands, and for allostery. Evolution may have had to balance these

needs, resulting in biological proteins with marginal stabilities.

Greg Petsko has suggested an interesting possibility [46], for which we find support in the model. There is evidence that proteins from thermophilic organisms ("thermophilic proteins"), which are unusually stable at high temperatures (see, e.g., [27]), are catalytically active mainly at high temperatures and less active at lower temperatures at which their mesophilic counterparts function well [8,26,51,59]. Since greater stability correlates with less flexibility, cooling to room temperature may freeze out the non-native-like fluctuations that are necessary for catalytic action. For example, from figure 4, when $|kT/\varepsilon|$ is between .15 and .20, sequence Q (+ symbols) might be active whereas sequence S ($\times$ symbols) might not be. It is interesting that ribonuclease A binds substrate and inhibitor $10°C$ above its rigor mortis temperature but not $10°C$ below [47], indicating that some binding-friendly modes of the protein are not populated below the rigor mortis temperature.

What forces might stabilize thermophilic proteins? Several studies of native thermophiles compared to their counterpart mesophiles have not yet revealed a dominant type of interaction. The differences in stability are attributed to the cumulative effects of many subtle interactions [17,22,27,48]. But another possibility is that the message may not be in the native structures, but rather may also involve the denatured conformations that are the dominant fluctuations. The secret to extreme stability in thermophilic proteins could reside in the first- (or higher-)excited states. Pertinent to this point is the observation of a "reverse hydrophobic effect" [7,24,41,53], whereby increasing the hydrophobicity of some surface residues in proteins destabilizes them, presumably because they affect the denatured states more strongly than the native states.

3.6 Conclusions

We have studied the fluctuations of native states of model proteins. For all HP sequences we have examined, the model has a "rigor mortis point", as has been observed experimentally for ribonuclease A [56]. At temperatures below this point large fluctuations are frozen out. Heating above the rigor mortis point temperature

recruits new classes of conformational fluctuations that contribute to the thermal motions. We have also studied the correlation of protein stability with three measures of static equilibrium flexibility—average solvent accessibilities of core residues (related to their HX rates), a Debye-Waller-like measure of distance variations, and the rigor mortis point temperature. We find that, on average, protein stability correlates inversely with flexibility. The greater the stability of a sequence, the fewer are the accessible large fluctuations, resulting in a lower equilibrium flexibility.

3.7 Acknowledgements

We thank Greg Petsko for motivating this work. Klaus Fiebig provided some of the computer code for generating the data, for which we are very grateful. K.E.S.T. was supported by a UCSF Graduate Dean's Health Science Fellowship and a UCSF Regent's Fellowship.

References

[1] M. Abe, Y. Nosoh, M. Nakanishi, and M. Tsuboi. Hydrogen-deuterium exchange studies on guanidinated pig heart lactate dehydrogenase. *Biochim. Biophys. Acta*, 746:176–181, 1983.

[2] A. Ansari, C. M. Jones, E. R. Henry, J. Hofrichter, and W. A. Eaton. The role of solvent viscosity in the dynamics of protein conformational changes. *Science*, 256:1796–1798, 1992.

[3] A. Ansari et al. Rebinding and relaxation in the myoglobin pocket. *Biophys. Chem.*, 26:337–355, 1987.

[4] R. Austin, K. Beeson, L. Eisenstein, H. Frauenfelder, and I. Gunsalus. Dynamics of ligand binding in myoglobin. *Biochemistry*, 14:5355–5373, 1975.

[5] E. R. Bauminger, S. G. Cohen, I. Nowik, S. Ofer, and J. Yariv. Dynamics of heme iron in crystals of metmyoglobin and deoxymyoglobin. *Proc. Natl. Acad. Sci. USA*, 80:736–740, 1983.

[6] M. C. Bonaccorsi di Patti et al. The multidomain structure of ceruloplasmin from calorimetric and limited proteolysis studies. *J. Biol. Chem.*, 265(34):21016–21022, 1990.

[7] B. E. Bowler, K. May, T. Zaragoza, P. York, A. Dong, and W. S. Caughey. Destabilizing effects of replacing a surface lysine of cytochrome *c* with aromatic amino acids: implications for the denatured state. *Biochemistry*, 32(1):183–190, 1993.

[8] T. D. Brock. Life at high temperatures. *Science*, 230:132–138, 1985.

[9] H. S. Chan and K. A. Dill. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA*, 87:6388–6392, 1990.

[10] H. S. Chan and K. A. Dill. "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.*, 95(5):3775–3787, 1991.

[11] R. M. Daniel, D. A. Cowan, H. W. Morgan, and M. P. Curran. A correlation between protein thermostability and resistance to proteolysis. *Biochem. J.*, 207:641–644, 1982. And references therein.

[12] M. Delepierre, C. M. Dobson, S. Selvarajah, R. E. Wedin, and F. M. Poulsen. Correlation of hydrogen exchange behaviour and thermal stability in lysozyme. *J. Mol. Biol.*, 168:687–692, 1983.

[13] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding. A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[14] W. Doster, A. Bachleitner, R. Dunau, M. Hiebl, and E. Lüscher. Thermal properties of water in myoglobin crystals and solutions at subzero temperatures. *Biophys. J.*, 50:213–219, 1986.

[15] W. Doster, S. Cusack, and W. Petry. Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature*, 337:754–756, 1989.

[16] S. W. Englander and L. Mayne. Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annu. Rev. Biophys. Biomol. Struct.*, 21:243–65, 1992.

[17] A. Fontana. Structure and stability of thermophilic enzymes: Studies on thermolysin. *Biophys. Chem.*, 29:181–193, 1988.

[18] H. Frauenfelder, F. Parak, and R. D. Young. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451–479, 1988. And references therein.

[19] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, 280:558–563, 1979.

[20] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[21] J. L. Green, J. Fan, and C. A. Angell. The protein–glass analogy: some insights from homopeptide comparisons. *J. Phys. Chem.*, 98(51):13780–13790, 1994.

[22] M. G. Grütter, R. B. Hawkes, and B. W. Matthews. Molecular basis of thermostability in the lysozyme from bacteriophage T4. *Nature*, 277:667–669, 1979.

[23] S. J. Hagen, J. Hofrichter, and W. A. Eaton. Protein reaction kinetics in a room-temperature glass. *Science*, 269:959–962, 1995.

[24] L. Herrmann, B. E. Bowler, A. Dong, and W. S. Caughey. The effects of hydrophilic to hydrophobic surface mutations on the denatured state of iso-1-cytochrome *c*: investigation of aliphatic residues. *Biochemistry*, 34(9):3040–3047, 1995.

[25] I. E. T. Iben et al. Glassy behavior of a protein. *Phys. Rev. Lett.*, 62(16):1916–1919, 1989. And references therein.

[26] R. Jaenicke. Enzymes under extreme conditions. *Annu. Rev. Biophys. Bioeng.*, 10:1–67, 1981. And references therein.

[27] R. Jaenicke. Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.*, 202:715–728, 1991.

[28] H. Keller and P. G. Debrunner. Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys. Rev. Lett.*, 45(1):68–71, 1980.

[29] D. E. Koshland, Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44:98–104, 1959.

[30] D. E. Koshland, Jr. Enzyme flexibility and enzyme action. *J. Cellular Comp. Physiol.*, 54:245–258, 1959.

[31] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[32] K. F. Lau and K. A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA*, 87:638–642, 1990.

[33] R. J. Loncharich and B. R. Brooks. Temperature dependence of dynamics of hydrated myoglobin: Comparison of force field calculations with neutron scattering data. *J. Mol. Biol.*, 215:439–455, 1990.

[34] M. Matsumura, S. Yasumura, and S. Aiba. Cumulative effect of intragenic amino-acid replacements on the thermostability of a protein. *Nature*, 323:356–358, 1986.

[35] L. Menéndez-Arias and P. Argos. Engineering protein thermal stability. *J. Mol. Biol.*, 206:397–406, 1989.

[36] D. W. Miller and K. A. Dill. A statistical mechanical model of hydrogen exchange in globular proteins. *Prot. Sci.*, 4(9):1860–1873, 1995.

[37] V. N. Morozov and S. G. Gevorkian. Low-temperature glass transition in proteins. *Biopolymers*, 24:1765–1799, 85.

[38] G. Musci et al. Unusual stability properties of a reptilian ceruloplasmin. *Arch. Biochem. Biophys.*, 279(1):8–13, 1990.

[39] J. M. Nocek et al. Low-temperature, cooperative conformational transition within [Zn-cytochrome *c* peroxidase, cytochrome *c*] complexes: Variation with cytochrome. *J. Am. Chem. Soc.*, 113:6822–6831, 1991.

[40] S. Ohta, M. Nakanishi, M. Tsuboi, K. Arai, and Y. Kaziro. Structural fluctuation of the polypeptide-chain elongation factor Tu. *Eur. J. Biochem.*, 78:599–608, 1977.

[41] A. A. Pakula and R. T. Sauer. Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature*, 344:363–364, 1990.

[42] F. Parak and H. Formanek. Investigation of the contributions of vibrations and crystal lattice faults to the temperature factors in myoglobin through comparison of Mössbauer absorption measurements with x-ray structure data. *Acta Crystallogr. sect. A*, 27:573–578, 1971.

[43] F. Parak, E. N. Frolov, A. A. Kononenko, R. L. Mössbauer, V. I. Goldanskii, and A. B. Rubin. Evidence for a correlation between the photoinduced electron transfer and dynamic properties of the chromatophore membranes from *rhodospirillum rubrum*. *FEBS Letts.*, 117(1):368–372, 1980.

[44] F. Parak, E. N. Frolov, R. L. Mössbauer, and V. I. Goldanskii. Dynamics of metmyoglobin crystals investigated by nuclear gamma resonance absorption. *J. Mol. Biol.*, 145:825–833, 1981.

[45] F. Parak and E. W. Knapp. A consistent picture of protein dynamics. *Proc. Natl. Acad. Sci. USA*, 81:7088–7092, 1984.

[46] G. A. Petsko. Personal communication.

[47] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko. Crystalline ribonuclease A loses function below the dynamical transition at 220K. *Nature*, 357:423–424, 1992.

[48] D. C. Rees and M. W. W. Adams. Hyperthermophiles: taking the heat and loving it. *Structure*, 3:251–254, 1995.

[49] D. Ringe and G. A. Petsko. Mapping protein dynamics by X-ray diffraction. *Prog. Biophys. Molec. Biol.*, 45:197–235, 1985.

[50] A. Šali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369(6477):248–251, 1994.

[51] J. Schumann, A. Wrba, R. Jaenicke, and K. O. Stetter. Topographical and enzymatic characterization of amylases from the extremely thermophilic eubacterium *thermotoga maritima*. *FEBS Lett.*, 282(1):122–126, 1991.

[52] B. K. Shoichet, W. A. Baase, R. Kuroki, and B. W. Matthews. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA*, 92:452–456, 1995.

[53] D. Shortle, H. S. Chan, and K. A. Dill. Modeling the effects of mutations on the denatured states of proteins. *Prot. Sci.*, 1(2):201–215, 1992.

[54] Y. Suzuki and T. Imai. Abnormally high tolerance against proteolysis of an exo-oligo-1,6-glucosidase from a thermophile *bacillus thermoglucosidius* KP 1006, compared with its mesophilic counterpart from *bacillus cereus* ATCC 7064. *Biochim. Biophys. Acta*, 705:124–126, 1982.

[55] K. E. S. Tang and K. A. Dill. Fluctuations of native proteins: are there large motions? Preprint, 1996.

[56] R. F. Tilton, Jr., J. C. Dewan, and G. A. Petsko. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein

ribonuclease A at nine different temperatures from 98 to $320K$. *Biochemistry*, 31(9):2469–2481, 1992.

[57] C.-L. Tsou. Conformational flexibility of enzyme active sites. *Science*, 262:380–381, 1993.

[58] M. G. Usha and R. J. Wittebort. Orientational ordering and dynamics of the hydrate and exchangeable hydrogen atoms in crystalline crambin. *J. Mol. Biol.*, 208:669–678, 1989.

[59] P. G. Varley and R. H. Pain. Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and *thermus thermophilus*. *J. Mol. Biol.*, 220:531–538, 1991.

[60] M. Vihinen. Relationship of protein flexibility to thermostability. *Prot. Eng.*, 1(6):477–480, 1987.

[61] G. Wagner and K. Wüthrich. Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitor. *J. Mol. Biol.*, 130:31–37, 1979.

[62] A. Wrba, A. Schweiger, V. Schultes, R. Jaenicke, and P. Závodskzky. Extermely thermostable D-Glyceraldehyde-3-phosphate dehydrogenase from the eubacterium *thermotoga maritima*. *Biochemistry*, 29(33):7584–7592, 1990.

Chapter 4

Theory and Future Directions

In chapter 3, I showed that the static equilibrium flexibility and the stability of proteins are inversely correlated. I present here a simple theory which predicts how measures of fluctuations should behave as a function of the denaturation temperature, T_m.

Let F be some flexibility measure which has a value of 0 for the native conformation, e.g., the solvent accessibility, A_i, of a native-buried residue. What we want is $\langle F \rangle$, the ensemble-averaged value of F (which is what's measured in an experiment), as a function of T_m:

$$\langle F(T) \rangle = \frac{\sum_c F_c \, e^{-E_c/kT}}{Q} \tag{1}$$

where the sum is over all conformations, and F_c and E_c are the value of F and the energy of conformation c, respectively. Q is the partition function.

But, $F = 0$ for the native conformation so the sum is only over non-native or fluctuation conformations.

$$\langle F(T) \rangle = \frac{\sum_{c=fluct} F_c \, e^{-E_c/kT}}{Q} \ . \tag{2}$$

Now I make my first approximation. I assume that the *dominant* fluctuation conformations are the first-excited conformations. For every $N = 16$ sequence I studied, the second-excited conformations contribute only when the temperature is near the denaturation point, or when the number of first-excited conformations is small. The third-excited states were always insignificant under native conditions. With this approximation, then,

$$\langle F(T) \rangle \cong \frac{\sum_{c=1st\ exc} F_c \, e^{-E_c/kT}}{Q} \ . \tag{3}$$

Since all the first-excited conformations have the same energy, E_1,

$$\langle F(T) \rangle \cong \frac{e^{-E_1/kT} \sum_{c=1st\ exc} F_c}{Q} = \langle F \rangle_1 \, g(1) \, p_1(T) \ , \tag{4}$$

where $\langle F \rangle_1$ is the average of F over all the first-excited conformations, $g(1)$ is the number of first-excited conformations, and $p_1(T) = e^{-E_1/kT}/Q$ is the probability of populating any single first-excited conformation at temperature T.

The sequence dependence of $\langle F(T)\rangle$ comes from $\langle F\rangle_1\ g(1)$. I now make my second assumption, that $\langle F\rangle_1$ is, to *first order*, sequence *independent*. Then, the sequence dependence of $\langle F(T)\rangle$ lies only in $g(1)$. Right away, we see that the average fluctuations are then proportional to $g(1)$ which inversely correlates with stability (the larger $g(1)$, the more low-energy non-native conformations, and the smaller $|\Delta G|$). At first, this approximation might seem a bit unintuitive, but for the moment, let me give a simple reason for its validity. The first-excited conformations tend to be quite different from each other and from the native conformation (see chapter 2). When one averages over all of them, details of individual conformations get washed out and one is left with an overall variability of the fluctuations which doesn't change much from sequence to sequence in the lattice model. I will further discuss the validity of both approximations later.

But, let me continue with the derivation, assuming that both approximations are good, to show how $\langle F(T)\rangle$ should depend on T_m. With this second assumption,

$$\langle F(T)\rangle \cong g(1) \times (\textit{sequence independent factors}) \ . \tag{5}$$

What is T_m's sequence dependence? T_m is defined to be the temperature at which $\Delta G = 0$,

$$\Delta G(T_m) = 0 = kT_m \ln \frac{[\textit{non - native conformations}]}{[\textit{native conformation}]} \ . \tag{6}$$

Again, assuming the dominant non-native conformations are the first-excited conformations,

$$0 \cong \ln \frac{[\textit{first - excited conformations}]}{[\textit{native conformation}]} \tag{7}$$

$$\cong \ln(g(1)\ e^{-|\varepsilon|/kT_m}) \tag{8}$$

or

$$\ln g(1) = \frac{|\varepsilon|}{kT_m} \ . \tag{9}$$

Combining with equation 4,

$$\ln \langle F(T)\rangle \cong \frac{|\varepsilon|}{kT_m} + \ln \langle F\rangle_1 + \ln p_1(T) \tag{10}$$

$$= \frac{|\varepsilon|}{kT_m} + (\textit{sequence independent factors}) \ . \tag{11}$$

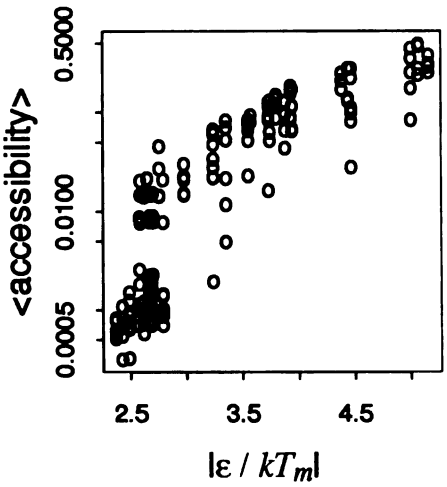


Figure 1: The average accessibility of core residues as a function of $1/T_m$ for ~30 different sequences of lengths 16 to 20. Each point represents one native-buried residue of one sequence; its T_m is that of the corresponding sequence. The data are the same as for figure 7 of the previous chapter.

The logarithm of the average fluctuation measure should be proportional to $1/T_m$.

We see in figure 1 that this is indeed approximately true when the fluctuation measure, F, is the average accessibility of residues which are buried in the native conformation. When the fluctuation measure is $\bar{b}$ (figure 2), we see that the relationship is linear at large $|\varepsilon|/kT_m$, corresponding to large $g(1)$ (where the first approximation is good).

Let me now quantitate the validity of the two approximations. First, do the first-excited conformations dominate the fluctuations? If so, then we expect from equation 9 that the plot of $|\varepsilon|/kT_m$ versus $\ln g(1)$ should be linear, with a slope of 1. Figure 3 shows that the $|\varepsilon|/kT_m$ does asymptotically approach the theoretically predicted behavior (solid line). In particular, the approximation is fairly good when $g(1) \gtrsim 50$, corresponding to $|\varepsilon|/kT_m \gtrsim 4$. To check the second approximation, figure 4 shows $\langle F \rangle_1$ as a function of $g(1)$, where F is the molecule-averaged "B" factor, $\bar{b}$. Each point represents a different $N = 16$ sequence. $\bar{b}$ hovers around .4 for sequences with $g(1) \gtrsim 20$. To first order, assuming that $\langle F \rangle_1$ is sequence independent is a good approximation within the lattice model. In general, the quality of this

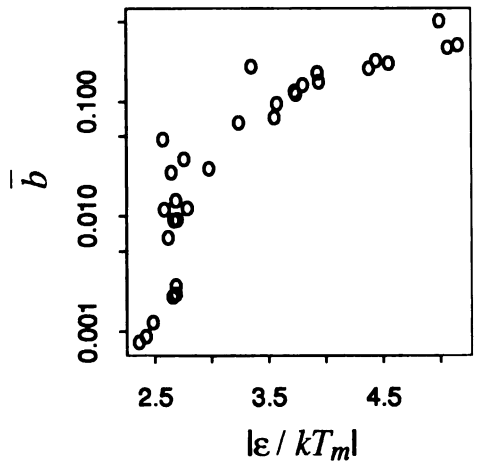


Figure 2: $\bar{b}$ as a function of $1/T_m$. Each point represents one sequence. The data are the same as figure 8 of the previous chapter.

second approximation depends on (1) the sequences one is comparing and (2) the fluctuation measure, F. For the lattice model, most first-excited conformations tend to be quite different from each other and from the native conformation, especially when $g(1)$ is large (the regime where the first approximation is valid) [1]. $\langle F \rangle_1$ is essentially an average over many compact but different conformations. As long as F is a fairly broad measure of conformational dissimilarity (see below), the details of these fluctuation conformations get averaged out and $\langle F \rangle_1$ is then a measure of the overall variability of the fluctuations. The sequence dependence is not very strong, at least for the lattice model. Thus, this is a good first-order approximation and we consequently saw clean inverse correlations between flexibility and stability (see chapter 3 and figures 1 and 2 above), even when comparing very different sequences of varying lengths. For real proteins, the approximation is probably less good and most likely breaks down when comparing molecules with different native conformations. For example, if molecule A were designed to have a rock solid core with a flexible active site, whereas molecule B has moderate flexibility overall, A and B might have the same average overall flexibility (as measured by, say, a molecule-averaged Debye-Waller factor), but different stabilities. There might also be chain length effects. In general, it would be best to compare sequences with the same native structure (as

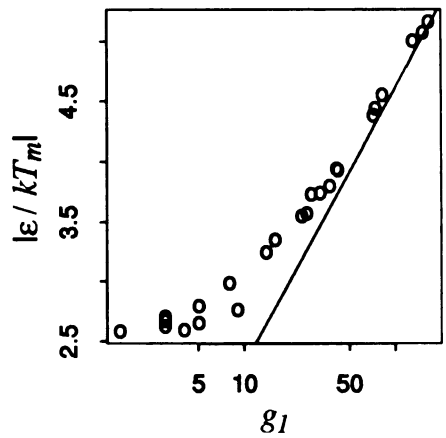


Figure 3: $1/T_m$ plotted as a function of $\ln g(1)$. Each point represents a different sequence. The data are for ~30 sequences of lengths 16 to 20.

is currently done in studies of thermophiles and their counterpart mesophiles). In addition, the choice of flexibility measure F shouldn't be too detailed. For example, the molecule-averaged Debye-Waller factor is probably more sequence-independent than is the Debye-Waller factor of serine residues. The latter may sample onliy a specific environment of the protein.

I can improve the first approximation by including second-excited conformations. Equation 8 then becomes:

$$0 \cong \ln(g(1)e^{-|\varepsilon|/kT_m} + g(2)e^{-2|\varepsilon|/kT_m}) \tag{12}$$

where $g(2)$ is the number of second-excited conformation. Simplifying,

$$|\varepsilon|/kT_m \cong \ln(g(1) + g(2)e^{-|\varepsilon|/kT_m}) \ . \tag{13}$$

In the plot of $|\varepsilon|/kT_m$ versus $\ln g(1)$ (figure 3), we see now that the deviation of the points from the theoretical line is the contribution from the second-excited conformations. Not surprisingly, as $g(1)$ increases, the second-excited conformations play a smaller role. In summary, my two approximations are good when $g(1)$ is large, $\gtrsim 50$ or $|\varepsilon|/kT_m \gtrsim 4$.

What does the theory predict for the behavior of the "rigor mortis" temperature, T_{rm}, with T_m? In our model, the rigor mortis point occurs when the total

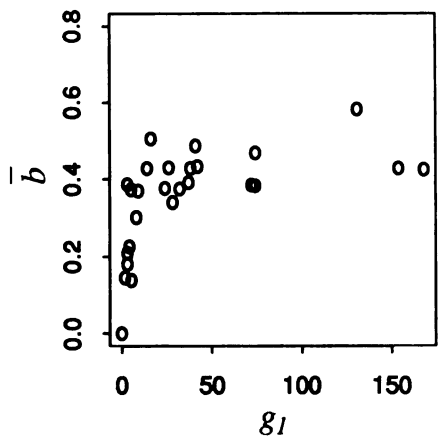


Figure 4: $\bar{b}$ as a function of $g(1)$. Each point represents a different sequence.

population of the first-excited state is approximately constant at P_{rm}. (P_{rm} is somewhere between 0 to 20% for most sequences. Data not shown.) Then,

$$P_{rm} = \frac{g(1)e^{-|\varepsilon|/kT_{rm}}}{Q} \quad . \tag{14}$$

Again, assuming that the first-excited conformations are the dominant contributors,

$$P_{rm} \cong \frac{g(1)\ e^{-|\varepsilon|/kT_{rm}}}{1 + g(1)\ e^{-|\varepsilon|/kT_{rm}}} \quad . \tag{15}$$

Using equation 9 to replace $g(1)$,

$$P_{rm} \cong \frac{e^{|\varepsilon|/kT_m}\ e^{-|\varepsilon|/kT_{rm}}}{1 + e^{|\varepsilon|/kT_m}\ e^{-|\varepsilon|/kT_{rm}}} \quad , \tag{16}$$

which simplifies to

$$\frac{|\varepsilon|}{kT_{rm}} = \frac{|\varepsilon|}{kT_m} - \ln \frac{P_{rm}}{1 - P_{rm}} \quad . \tag{17}$$

Hence, we expect that $|\varepsilon|/kT_{rm}$ to be linear with respect to $|\varepsilon|/kT_m$ with a slope of 1. Figure 5 shows that this is indeed true under conditions where the first approximation is valid, i.e., when $|\varepsilon|/kT_m \gtrsim 4$. The dotted lines are the theoretical predictions with P_{rm} set to different values. Fitting the $|\varepsilon|/kT_m \gtrsim 4$ data to a straight line, we obtain $-\ln[P_{rm}/(1 - P_{rm})] \cong 3$ or $P_{rm} \cong 5\%$, which is a reasonable value.

Can we use this theory to make predictions on the density of states, $g(E)$, for real proteins? As noted, the first-excited conformations are the dominant fluctuations

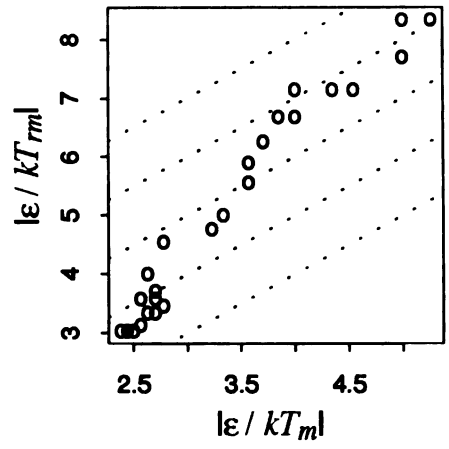


Figure 5: $1/T_{rm}$ as a function of $1/T_m$. The dotted lines are the theoretical curves (slope of 1) with different values of P_{rm}.

when $g(1)$ is large. When this condition is not met, there are deviations from the predicted theory. It would be interesting to find out whether one could use the non-ideality to back predict the number of second-excited conformations, $g(2)$, and/or their total population relative to the first-excited conformations, $g(2)e^{|\varepsilon|/kT}/g(1)$. If this were possible, one might then extend the theory to a model with a continous density of states and perhaps make predictions for real proteins. The long term goal of this study would be to use experimental information on average fluctuations and stability, perhaps with data from several temperatures, to predict (the low energy end of) $g(E)$ for real proteins. But I leave that for future work.

References

[1] D. W. Miller and K. A. Dill. A statistical mechanical model of hydrogen exchange in globular proteins. *Prot. Sci.*, 4(9):1860, 1995.