Jianyong Wang
Gordon M. Crippen
College of Pharmacy,
University of Michigan,
Ann Arbor, MI 48109-1065

Statistical Mechanics of Protein Folding with Separable Energy Functions

Received 27 January 2004;
accepted 26 February 2004
Published online 26 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.bip.20077

Abstract: *We have initiated an entirely new approach to statistical mechanical models of strongly interacting systems where the configurational parameters and the potential energy function are both constructed so that the canonical partition function can be evaluated analytically. For a simplified model of proteins consisting of a single, fairly short polypeptide chain without cross-links, we can adjust the energy parameters to favor the experimentally determined native state of seven proteins having diverse types of folds. Then 497 test proteins are predicted to have stable native folds, even though they are also structurally diverse, and 480 of them have no significant sequence similarity to any of the training proteins.* © 2004 Wiley Periodicals, Inc. Biopolymers 74: 214–220, 2004

Keywords: *thermal denaturation; globular proteins; canonical partition function; conformational sampling*

INTRODUCTION

Basic classical statistical mechanics over the canonical ensemble is an appealing way to relate theories or models at the molecular level to macroscopic observables, such as equilibrium thermodynamic properties. For N interacting point particles having positions and momenta described in Cartesian coordinates, the complete Hamiltonian H is a function of $6N$ degrees of freedom, and the thermodynamics of the system can be derived from the partition function Z, which is the integral over all degrees of freedom of $\exp(-H/kT)$. Integrating over the $3N$ momenta components is easy because each degree of freedom is used in exactly one additive term in H, so that part breaks up into a product of simple integrals. The trouble arises from trying to integrate over the $3N$ position components, because the potential energy part of the Hamiltonian is not necessarily what we will refer to as a *separable energy* function, namely a sum of terms, each of which depends on a separate, disjoint, small subset of the position parameters. Of course a great deal of cleverness has been devoted to getting around this problem by studying systems of weakly interacting particles where adequate approximations to the partition function can be devised.

Unfortunately, useful models of protein folding all involve a large number of particles all linked together in the potential function by terms involving nearly all possible pairs of particles plus yet more complicated terms. Direct integration of the partition function is infeasible because one must integrate over all degrees of configurational freedom simultaneously, whether those are Cartesian coordinates of (united) atoms or torsion angles for rotatable bonds. Another approach is to drastically reduce the number of degrees of

Correspondence to: Gordon M. Crippen: email: gcrippen@umich.edu

Contract grant sponsor: University of Michigan Bioinformatics Program and Howard Hughes Medical Institutes

freedom, at least for a somewhat local examination of the conformation space, by combining them into a few collective variables, such as the low-frequency modes in normal mode analysis.[1–5] Alternatively, many variations on Monte Carlo can sample the configuration space widely and insightfully,[6–10] resulting in a Boltzmann distribution of states in the limit of infinite computer time. Molecular dynamics relies on the ergodic hypothesis to reach the same distribution of states.[11–14] In either case, agreement with experiment requires an adequate approximation to the true energy of the system as a function of the configuration, and for large systems it remains difficult to determine whether equilibrium has been reached: ". . .lack of sampling [is] an important concern for peptide simulations."[15]

Here we instead take the radical approach of choosing a set of variables that allow a global treatment of all conformations of a polypeptide chain, while simultaneously inventing an empirical potential function that is separable in these variables so that it is feasible to evaluate the partition function. One might call this: Statistical Mechanics Enabled Using Separable Energies = SMEUSE = "(noun) a hole in a hedge, wall, etc."[16] If the potential can be adjusted to give agreement with experimental thermodynamic results, then SMEUSE lets us slip through the stone wall of concern over adequacy of sampling.

METHODS

Haar Transform of Coordinates

The first question is how to parameterize protein conformations. We need invariants under translation and rotation because the internal energy of proteins in dilute solution is not affected by these operations. The conformational parameters should also be simple to relate to amino acid sequence features. There are many possibilities one can try beyond the customary atomic Cartesian coordinates or torsion angles about rotatable bonds. For example, a Fourier transform of atomic Cartesian coordinates as a function of sequence position requires that the ends of the chain are very close in order to be a periodic signal, which is by no means the case for real proteins. A discrete cosine transform avoids the periodicity requirement and gives terms associated with periodicities in structure that could be associated with corresponding periodicities in sequence, but the period of a particular term is relative to the full chain length, whereas there are always 3.6 residues per turn in an α-helix, regardless of the total size of the protein. However, a wavelet transform of atomic Cartesian coordinates as a function of sequence position gives us a hierarchical description of the positions of fixed length subsegments of the chain without requiring any sort of periodicity, and these terms can be related to the sequence of the corresponding subsegment. Applications of wavelets to proteins have been mostly analyses of structural features.[17–20] Of the many different kinds of wavelets,[21] the simple Haar transform[22] seems best suited to the present application.

Let $[x_i, y_i, z_i]$ be the Cartesian coordinates of the C^α atom of the ith residue in a polypeptide chain, for $i = 1, \ldots, N$. When N is a power of 2, the standard Haar wavelet is

$$\psi_{w,j}(i) = \begin{cases} (2w)^{-1/2} & \text{for } i = j, \ldots, j + w - 1 \\ -(2w)^{-1/2} & \text{for } i = j + w, \ldots, j + 2w - 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_{N,1}(i) = N^{-1/2} \qquad \text{for} \qquad \text{i} = 1, \ldots, N \tag{1}$$

where the half-width $w = 1, 2, 4, \ldots, N/2$ increases by factors of 2, and the start of the wavelets having that half-width $j = 1, 2w + 1, 4w + 1, \ldots, N - 2w + 1$ increase by $2w$. There are altogether N wavelets, and they constitute a complete orthonormal basis for vectors of data $\boldsymbol{x} = [x_1, \ldots, x_N]^T$. In other words, associated with each wavelet is a Haar transform coefficient

$$\hat{x}_{w,j} = \sum_{i=1}^{N} x_i \psi_{w,j}(i) \tag{2}$$

and the exact original signal can be recovered from them.

If N is not a power of 2, then Haar wavelets may be readily generalized[23] to arbitrary N by using Eq. (1) when $j + 2w - 1 \leq N$, but otherwise using

$$\psi_{w,j}(i) = \begin{cases} \left(\dfrac{N + 1 - j - w}{(N + 1 - j)w}\right)^{1/2} & \text{for } i = j, \ldots, j + w - 1 \\ -\left(\dfrac{w}{(N + 1 - j)(N + 1 - j - w)}\right)^{1/2} & \text{for } i = j + w, \ldots, N \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $w = 1, 2, 4, \ldots, 2^k$ and $k > 0$ is the largest integer such that $2^k < N$. With this arrangement, there are always exactly N transform coefficients, and their values are proportional to the difference between the mean x_i on the positive side and the mean x_i on the negative side.

Using center of mass coordinates guarantees that $\hat{x}_{N,1} = \hat{y}_{N,1} = \hat{z}_{N,1} = 0$, but otherwise the transform coefficients

are rotated if the original points are rotated about the origin. Consequently, we will use the transform distance

$$\hat{d}_{w,j} = (\hat{x}_{wj}^2 + \hat{y}_{w,j}^2 + \hat{z}_{w,j}^2)^{1/2} \quad (4)$$

which is independent of rotation for all the other choices of w, j. Conceptually, $\hat{d}_{w,j}$ is proportional to the distance between the centroid of the w residues starting at sequence position j and the centroid of the following w residues. Since we are using only N conformational parameters to describe $3N - 6$ conformational degrees of freedom, one can always calculate the $\hat{d}_{w,j}$ for a given conformation, but there may be multiple conformations corresponding to a given set of $\hat{d}_{w,j}$ parameters.

Segment Composition Vector

Let $\boldsymbol{s}_{w,j}$ be a vector of 21 elements consisting of the constant 1 concatenated with the scaled residue type composition vector $\boldsymbol{c}_{w,j}$ of (the nonzero positions of) $\psi_{w,j}$, namely,

$$\boldsymbol{s}_{w,j} = [1, \boldsymbol{c}_{w,j}] \quad (5)$$

where $c_{w,j,i}$ is the number of residues of type i (Ala, say) in the $2w$ positions of $\psi_{w,j}$ divided by $2w$. Thus, $\sum_{i=1}^{20} c_{w,j,i} = 1$. This is the simplest imaginable correspondence between structure expressed in terms of the $\hat{d}_{w,j}$ and the sequence composition $\boldsymbol{c}_{w,j}$ of that segment of chain. More sophisticated approaches would also reflect the specific sequence within the segment, but that remains for future studies.

Separable Energy and Partition Function

Let the potential energy depend on conformation and sequence as

$$E = \sum_{w,j} (\hat{d}_{w,j} - \boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w)^2 \quad (6)$$

where for each w the first component of the adjustable parameters in $\boldsymbol{a}_w$ is a different variable, while the following 20 components are the same for each w. For short chains, we need only wavelets having $w = 2$, 4, 8, 16, 32, and 64, so there are $6 + 20 = 26$ adjustable parameters in all. There is no special physical justification for Eq. (6), but rather it is the simplest conceivable form that lends itself to the required integration. It implies that the spatial extent of a segment of polypeptide chain depends primarily on how many residues are involved, and secondarily on the amino acid composition of that segment, and there is a single ideal value of the extent given the composition. The scaling of the $\boldsymbol{c}_{w,j}$ is essential in order to use the same 20 composition parameters for all wavelet widths.

The partition function of the corresponding polypeptide chain can be written as

$$Z = \int_{\hat{d}_{w,j}} \cdots \int \exp(-\beta E) = \prod_{w,j} \int_0^\infty \exp(-\beta(\hat{d}_{w,j} - \boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w)^2)\, d\hat{d}_{w,j} \quad (7)$$

where $\beta = (k_B T)^{-1}$ for some arbitrary temperature T. The range of integration for each $\hat{d}_{w,j}$ is taken to be zero to infinity for simplicity and with negligible error, although there are tighter bounds for real polypeptides. The separability of E permits us to convert the multivariate integral into a product of single variable integrals that are easy to evaluate once the adjustable parameters are determined.

Optimization of Parameters

Since our model considers only a single polypeptide chain, we needed to adjust the parameters to favor the native conformation of proteins stabilized strictly by intrachain interactions The Protein Data Bank (PDB)[24] contains the experimentally determined three-dimensional structures of well over 24,000 proteins, but many of these database entries illustrate slight conformational changes between proteins having very similar sequences, or even the same protein interacting with different small ligands. PDB Select[25] is a subset of these where the proteins have sequences that differ by at least a small amount, according to a formula that permits a smaller fraction of identical residues for longer chains. Out of the 5416 entries in the PDB Select 90% list of April 2002, we found 96 x-ray crystal structures apparently involving only one, short polypeptide chain without substantial ligands, such as heme groups. Further scrutiny resulted in only 32 entries consisting of only a single polypeptide chain of length no greater than 128 residues that seem to fold as monomers under reasonably standard conditions to a compact structure having a radius of gyration no more than 30% greater than the minimum for the given chain length.[26] Furthermore, no pair of these 32 chains has greater than 90% sequence identity after optimal sequence alignment, and the root mean square deviation (RMSD) between matching aligned residues after the usual optimal rigid body superposition[27] is greater than 3 Å. When adjusting the parameters of our energy function, we discovered that only 7 of the 32 contribute to the training. The final training set of seven proteins shown in Table I involves considerable diversity of fold types.[28]

If we can calculate the canonical partition function Z by integrating over all microscopic states corresponding to some macroscopic state, then the Helmholtz free energy of that macroscopic state at the temperature corresponding to $\beta = 1/(k_B T)$ is easily calculated by $A = -\beta^{-1} \ln Z$. Experimental data for protein folding is in terms of Gibbs free energy, but for such aqueous solutions, the difference is small. Let the partition function corresponding to the native state be

Table I Training Set Proteins

PDB Entry	No. Residues	Type[a]
1COA.I	64	$\alpha + \beta$
1ENH	54	α
1JWO.A	97	$\alpha + \beta$
1OPS	64	β
1PGB	56	$\alpha + \beta$
1PTF	87	$\alpha + \beta$
1TMY	118	α/β

[a] SCOP classification.[28]

$$Z^0 = \prod_{w,j} \int_{\hat{d}^0_{w,j}/2}^{2\hat{d}^0_{w,j}} \exp(-\beta(\hat{d}_{w,j} - \boldsymbol{s}_{w,j} \cdot \boldsymbol{s}_w)^2) d\hat{d}_{w,j} \qquad (8)$$

where the $\hat{d}^0_{w,j}$ are the transform distances [Eq. (4)] of the native conformation. The integration is over the native region, which is generously taken to run from $\hat{d}^0_{w,j}/2$ to $2\hat{d}^0_{w,j}$. This certainly includes the given crystal structure plus some degree of flexibility, as seen experimentally in solution,[29,30] but is much smaller than the total range of conformations in the denatured state. Future work may refine the native range of integration to better agree with experiment. The free energy of folding is simply $\Delta A = -\beta^{-1} \ln(Z^0/Z_{\text{denatured}})$, and if the given temperature corresponds to the midpoint of the thermal folding transition, $\Delta A = 0$ and the two partition functions are equal. Here we simply calculate the total partition function Z integrated over all states [Eq. (7)], and since $Z = Z^0 + Z_{\text{denatured}}$, at the midpoint $Z^0/Z = 0.5$. A polypeptide chain is considered to be stable at its native conformation at the given temperature if $Z^0/Z > 0.5$. Then we can optimize the parameters $\boldsymbol{a}_w$ by minimizing the objective function

$$F_{\text{obj}} = \sum_p \begin{cases} (Z^0_p/Z_p - 0.5)^2 & \text{if } Z^0_p/Z_p < 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

summing over all 7 polypeptide chains p in the training set.

The wavelets of the seven selected chains may have seven possible different widths. However, $\hat{d}_{1,j}$ corresponds to the transform distances between two sequentially adjacent residues, which are relatively constant along the chains. So we only consider six widths, $w = 2, 4, 8, 16, 32, 64$. Furthermore, we use the same 20 parameters corresponding to the scaled residue composition inside the wavelet independent of w, resulting in $6 + 20 = 26$ adjustable parameters $\boldsymbol{a}_w$. Noting that Z^0 will be a relatively large fraction of Z if $\boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w$ is located inside the native region, we first minimized

$$F'_{\text{obj}} = \sum_{w,j,p} \left(\frac{\boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w}{\hat{d}^0_{w,j}} \right)^2 + \left(\frac{\hat{d}^0_{w,j}}{\boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w} \right)^2 \qquad (10)$$

so that $\hat{d}^0_{w,j} \approx \boldsymbol{s}_{w,j} \cdot \boldsymbol{a}_w$ for most wavelets of most proteins. Then the $\boldsymbol{a}_w$ were further refined by minimizing F_{obj} in Eq. (9), eventually reaching $F_{\text{obj}} = 0$, when all the seven training proteins were stable. All these calculations were carried out in MOE using the SVL computer language.[31]

RESULTS AND DISCUSSION

Optimized Parameters

Success at fitting and prediction depends on many features of the model, such as the functional form of the energy, although we have not experimented yet with alternative forms. The range of conformations assumed for the somewhat flexible native state in Eq. (8) has an effect on results, but shows no obvious trends (data not shown). Since our conformational parameters, the $\hat{d}_{w,j}$, do not fully specify the conformation, refining the ranges of integration is not yet warranted.

Obviously choosing different arbitrary sets of proteins for training will produce different results. However, of the 32 small and distinct proteins found in

Table II Energy Parameters

w or Residue Type	a_i
2	−6.11
4	0.068
8	11.57
16	15.78
32	20.32
64	41.39
A	9.94
C	14.62
D	11.65
E	10.49
F	12.78
G	13.39
H	11.47
I	11.50
K	10.42
L	11.48
M	10.58
N	10.62
P	12.98
Q	9.48
R	10.62
S	11.71
T	13.39
V	12.63
W	9.16
Y	11.71

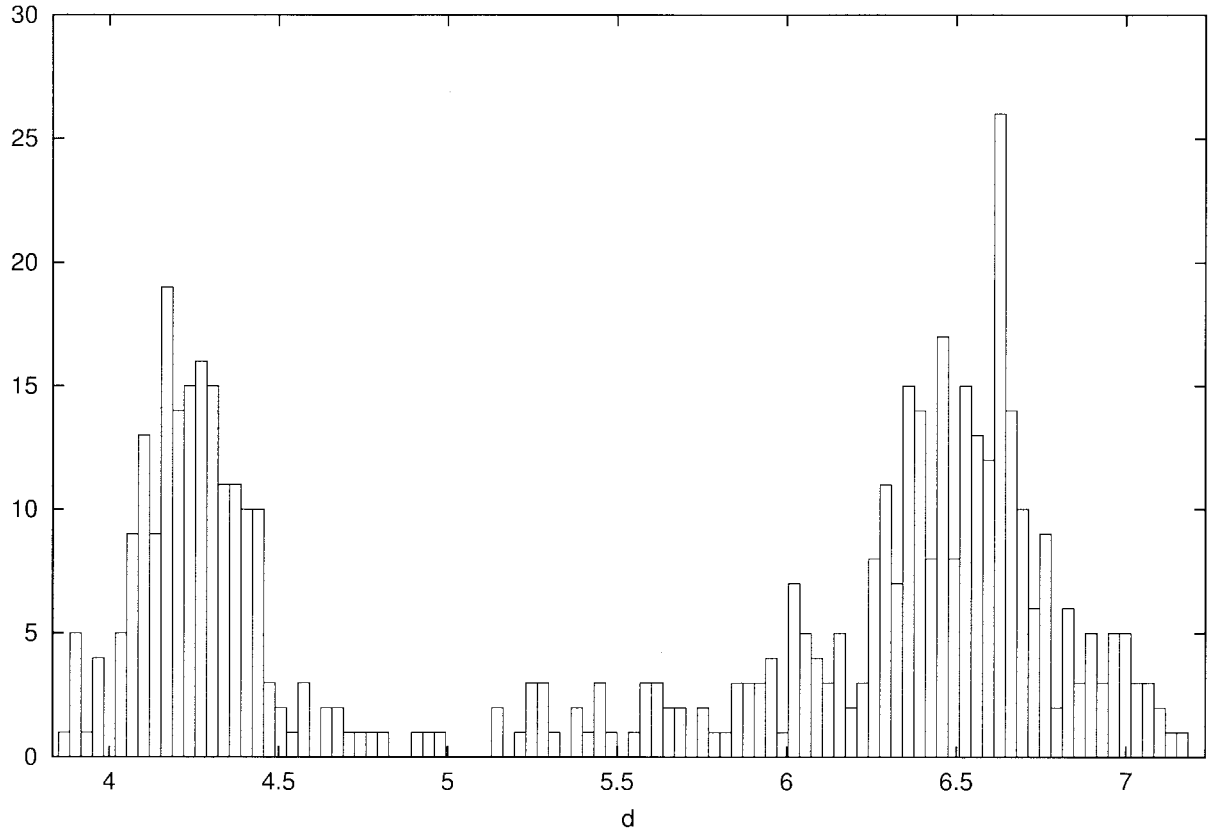

FIGURE 1 Histogram of $\hat{d}_{w,j}$ values for hydrophobic dominant wavelets of width = 2.

PDB Select that apparently fold as a single chain without large ligands or cross-links, only the seven in Table I have an effect on the training. Removing the other 32 − 7 = 25 from the training set produces the same parameters. Even these seven are not overfitted, since their final Z^0/Z values range from 0.74 to 0.98.

The final values of the $\boldsymbol{a}_w$ in Table II include a parameter associated with each of the 20 amino acid types. We found that the mean and standard deviation for clearly hydrophobic types (V, L, I, M, F, and W) are greater than those for clearly hydrophilic types (K, R, E, N, D, and Q). This is interesting because a larger positive parameter favors a larger value of $\hat{d}_{w,j}$, which corresponds to more extended local conformation. By surveying the wavelets with $w = 2$ of all 497 predicted proteins, we plotted the histogram of $\hat{d}_{w,j}$ for hydrophobic dominant (>75% clearly hydrophobic residues) and hydrophilic dominant (>75% clearly hydrophilic residues) segments. The histogram of hydrophobic dominant segments (Figure 1) shows a bimodal distribution, either very small (typically part of an α-helix) or very large (typically part of a β-sheet), whereas for hydrophilic dominant segments (Figure 2), the histogram has a high peak at small values and relatively smooth distribution at larger values. As a result, the mean and standard deviation of clearly hydrophilic parameters are smaller than those of clearly hydrophobic parameters. The bimodal distribution of $\hat{d}_{w,j}$ for hydrophobic dominant segments suggests that the hydrophobic residues tend to either form a locally compact conformation, such as α-helix, or a locally extended β-strand that is globally compact due to associating with other strands in a β-sheet.

Prediction of Short Polypeptide Chains

The parameters were adjusted so that $Z^0/Z > 0.5$ for all seven training proteins at the arbitrary temperature corresponding to $\beta = 1$. Thus, these proteins prefer their respective native states over the denatured state at this or any lower temperature. A test protein is considered to favor its native state if there is *any* temperature for which $Z^0/Z > 0.5$, not just at $\beta = 1$. After all, our very simple model is only trying to show some degree of stability for correctly folded proteins, rather than trying to match experimentally determined temperatures for the midpoint of the thermal denaturation curve. From PDB, we selected 1822 chains whose lengths vary from 30 to 128 residues. Out of these we found there are 497 proteins

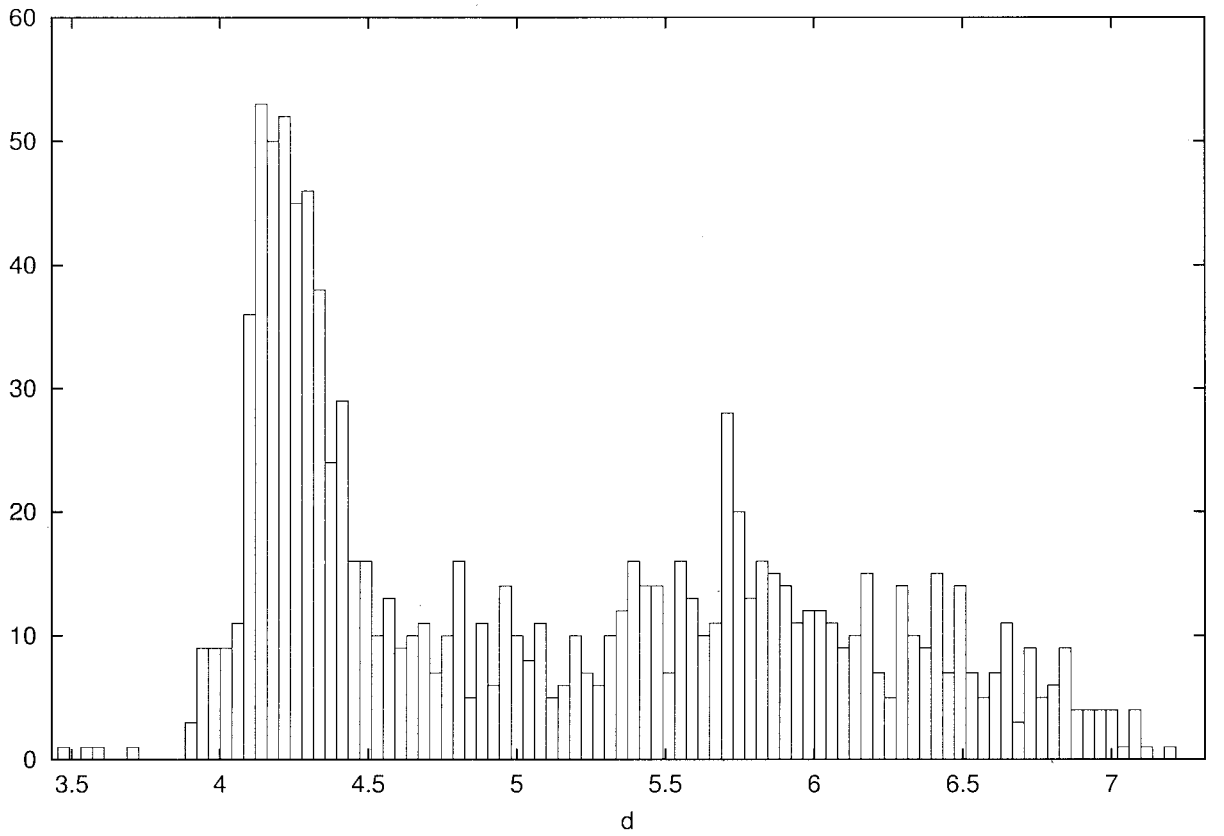

FIGURE 2 Histogram of $\hat{d}_{w,j}$ values for hydrophilic dominant wavelets of width = 2.

predicted to be stably folded, grouped according to their general fold type in Table III. Only 17 of these have noticeable sequence identity (>30%) to some of the training proteins, but the other 480 are sequentially unrelated to the training set. More α-helical proteins were predicted to be stable, but all types are represented, and there are no obvious trends in chain length or predicted Z^0/Z values. The values of Z^0/Z listed in the table were all calculated at the same temperature used in training, and most are above 0.5. For some the ratio is lower, but for these there is a lower temperature where the native is favored over the denatured state.

Table III Predicted Proteins

Type[a]	No. Proteins	Range of Chain length	Range of Z^o/Z	Range of % Sequence Identity[b]
α	166	31–122	0.381[c]–0.986	12.5–48.3
β	73	36–128	0.2[c]–0.986	10.3–27
α/β	30	61–128	0.457[c]–0.981	15.8–93.8
$\alpha + \beta$	123	37–128	0.337[c]–0.98	13.5–89.3
α and β	1	32	0.988	14.3
Small protein	90	40–112	0.329[c]–0.989	11.9–23.4
Coiled coil	1	39	0.74	18.5
Low resolution	1	79	0.792	16.9
Peptide	9	40–86	0.65–0.987	13–19.6
Designed protein	3	67–126	0.605–0.935	20.4–22.2

[a] SCOP classification.[28]
[b] The training set protein in Table I having the greatest percent sequence identity after optimal sequence alignment.
[c] $Z^o/Z > 0.5$ at a lower temperature than that used in this table.

There are many polypeptide chains in PDB for which our energy function predicts that the given conformation is not stable at any temperature. Note that our calculation takes into account only the single polypeptide chain whereas most proteins in PDB are involved in substantial associations with other polypeptide or polynucleotide chains, as well as with other substantial ligands or prosthetic groups. Covalent disulfide bridges are treated as separate Cys residues in our calculation, thus underestimating the destabilization of the denatured state.

CONCLUSIONS

The results with SMEUSE at this point amount to a proof of principle. Our initial focus has been on the most basic stability of protein folds with respect to thermal denaturation, leaving for future studies the more difficult questions of quantitative agreement with experiment on protein folding thermodynamic state functions, kinetics, and structural fluctuations. At this point it is possible to construct a separable energy function, Eq. (6), that depends on the conformation and amino acid sequence of a single polypeptide chain represented at the very low resolution of one point per residue. The parameters can be adjusted so that 7 small proteins have thermally stable native states, and we can easily find another 497 proteins that are correctly predicted to be stable in their native conformations. Both the training and test sets span the full range of general fold types for relatively short chains. This is not a matter of somehow exploiting sequence homology, because 480 of the predicted proteins have only negligible levels of sequence identity to any of the training proteins. On the other hand, many of the structures of small proteins in PDB are correctly predicted to have unstable native conformations in the sense that these proteins are significantly stabilized by factors outside the scope of the current model, such as disulfide bridges and associations between multiple polypeptide chains. Whatever the shortcomings may be of the current energy function and conformational parameterization, there is no concern about the adequacy of conformational sampling. These results come from the analytical integration of the partition function over all conformations encompassed by the model.

This work was supported in part by a grant from the University of Michigan Bioinformatics Program, and the Howard Hughes Medical Institute.

REFERENCES

1. Tirion, M. M.; ben-Avraham, D. J Mol Biol 1993, 230, 186–195.
2. ben-Avraham D.; Tirion, M. M. Biophys J 1995, 68, 1231–1245.
3. Li, G.; Cui, Q. Biophys J 2002, 83, 2457–2474.
4. Kamiya, K.; Sugawara, Y.; Umeyama, H. J Comput Chem 2003, 24, 826–841.
5. Doruker, P.; Jernigan, R. L.; Bahar, I. J Comput Chem 2002, 23, 119–127.
6. Skolnick, J.; Kolinski, A. J Mol Biol 1990, 212, 787–817.
7. Hansmann, U. H. E.; Okamoto, Y. Ann Rev Comp Phys VI 1999, 129–157.
8. Zhang, H. Proteins 1999, 34, 464–471.
9. Mitsutake, A.; Sugita, Y.; Okamoto, Y. Biopolymers 2001, 60, 96–123.
10. Zhang, Y.; Kihara, D.; Skolnick, J. Proteins 2002, 48, 192–201.
11. Guo, Z.; Brooks, C. L.; Boczko, E. M. Proc Natl Acad Sci USA 1997, 19, 10161–10166.
12. Vorobjev, Y. N.; Hermans, J. Protein Sci 2001, 10, 2498–2506.
13. Zagrovic, B.; Sorin, E. J.; Pande, V. J Mol Biol 2001, 313, 151–169.
14. Borreguero, J. M.; Dokholyan, N. V.; Buldyrev, S. V.; Shakhnovich, E. I.; Stanley, H. E. J Mol Biol 2002, 318, 863–876.
15. Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; García, A. E. Curr Opin Struct Biol 2003, 13, 168–174.
16. Oxford English Dictionary, 2nd ed., 1989.
17. Carson, A. M. J Comp-Aided Molec Design 1996, 10, 273–283.
18. Mandell, A. J.; Selz, K. A.; Shlesinger, M. F. Phys A 1997, 244, 254–262.
19. Hirakawa, H.; Muta, S.; Kuhara, S. Bioinformatics 1999, 15, 141–148.
20. Murray, K. B.; Gorse, D.; Thornton, J. M. J Mol Biol 2002, 316, 341–363.
21. Daubechies, I. Ten Lectures on Wavelets; SIAM: Philadelphia, 1992.
22. Haar, A. Math Ann 1910, 69, 331–371.
23. Crippen, G. M. Polymer 2003, 44, 4373–4379.
24. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Research 2000, 28, 235–242.
25. Hobohm, U.; Sander, C. Protein Sci 1994, 3, 522–524.
26. Maiorov, V. N.; Crippen, G. M. J Mol Biol 1992, 227, 876–888.
27. Kabsch, W. Acta Cryst 1978, A34, 827–828.
28. Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. Nucleic Acid Res 2002, 30, 264–267.
29. Englander, S. W. Ann Rev Biophys Biomol Struct 2000, 29, 213–238.
30. Maity, H.; Lim, W. K.; Rumbley, J. N.; Englander, S. W. Protein Sci 2003, 12, 153–60.
31. Molecular Operating Environment (MOE), Chemical Computing Group, Inc., http:// www.chemcomp.com.

Reviewing Editor: Dr. David A. Case