Evaluating protein structure-prediction schemes using energy landscape theory

by M. P. Eastwood
C. Hardin
Z. Luthey-Schulten
P. G. Wolynes

Protein structure prediction is beginning to be, at least partially, successful. Evaluating predictions, however, has many elements of subjectivity, making it difficult to determine the nature and extent of improvements that are most needed. We describe how the funnel-like nature of energy functions used for protein structure prediction determines their quality and can be quantified using landscape theory and multiple histogram sampling methods. Prediction algorithms exhibit a "caldera"-like landscape rather than a perfectly funneled one. Estimates are made of the expected number of effectively distinct structures produced by a prediction algorithm.

1. Introduction

Protein folding has fascinated theorists and experimenters for decades. This fascination has been driven not only by the phenomenon, being the simplest act of biological self-organization, but also by practical considerations relating to the desire to compute protein structure from sequence data. There are signs that the appropriate framework for understanding biomolecular self-organization is emerging. In considering folding theory versus structure-prediction practice, cynics, however, are likely to paraphrase the old saw "Thermodynamics owes more to the steam engine than the steam engine owes to thermodynamics."

In this paper, we initially review how folding theory based on energy landscapes has already contributed to progress on structure-prediction algorithms. We then illustrate how the computational tools used in energy landscape approaches to the study of folding kinetics can help distinguish and evaluate different energy functions used in structure prediction and how this analysis helps to quantify where improvements are most needed.

One major contribution of the last decade's study of folding to the practice of prediction is to give predictors several additional reasons for optimism, a necessary emotional stance for people to work in this field! A decade ago most people quoted folding times as being seconds to minutes. Indeed, several theories from the 1970s used these numbers as a procrustean bed for setting parameters! This would make prediction by even slavishly accurate molecular simulation, guaranteed to work by Anfinsen's experiment [1, 2], out of the question economically. Deductions from theory and heroic laboratory experiments have shown that a good deal of the self-organization in folding occurs in microseconds. Thus, if sufficiently accurate all-atom potentials are available, the IBM "Blue Gene Project"[1] should have a good chance

[1] "IBM Announces $100 Million Research Initiative to Build World's Fastest Supercomputer," press release on December 6, 1999; search for "Blue Gene" on *http://www.ibm.com* for more information; see also the paper entitled "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer," by F. Allen et al., concurrently being published in the *IBM Systems Journal*, Volume 40, No. 2, 2001.

0018-8646/01/$5.00 © 2001 IBM

to succeed in that the plan is to simulate accurately the laboratory situation.

The rapid folding without discrete traps observed experimentally suggests that the free-energy landscape (averaged over the solvent) of a protein is funnel-shaped [3, 4]. According to landscape theory [5, 6], structurally distinct traps emerge through inappropriate contacts arising from the inconsistency or "frustration" of different energy interaction terms. Evidence for the funnel point of view comes not just from detailed kinetics experiments [7–9], but also from the widespread anecdotal evidence that protein structures are robust to most single-site mutations. On a perfectly funneled landscape, barriers to folding are largely entropic and could be overcome by carrying out simulations at lower temperatures. Unfortunately, one cannot do this in complete all-atom models because there is frustration between solvent hydrogen bonds and van der Waals contacts in the protein. This leads to cold denaturation—the lowest energy (as opposed to free energy) state of protein plus solvent is actually an unfolded one. Thus, a "Cool Blue Gene" simulation with all atoms should not work any faster than a hot one. An important issue is whether solvent-averaged potentials as crafted by theorists, still with all atoms of the proteins represented, can then achieve funnel-like surfaces. The situation would be even better if reduced descriptions with less atomic detail could be used. Here there is also good news when theory and experiment are combined.

There are many indications that simplified energy functions with funnel-like landscapes can describe protein folding kinetics [10–15]. The catch is that the beautiful kinetics results require making the funnel approximation as an extra assumption or filter in order to bring the numerical predictions into conformity with experiment. To predict protein structures from sequence data, one needs simplified energy functions that do not impose this constraint and therefore allow any possible contact. For prediction we cannot *a priori* limit the model by assuming native contacts to be overwhelmingly important. The widespread failure of structure-prediction schemes until recently, in fact, was *prima facie* evidence for these schemes being based on rough, i.e., non-funneled, energy landscapes—landscapes in which topologically distinct structures can be found to lie within a few k_BT of each other near the folding temperature. Such states would represent traps in laboratory folding kinetics. They also lead to a lack of confidence about the result of a prediction, since different simulation runs should give each as a separate result. This situation is generically to be expected for random sequences using any given energy function. The trick is to find an energy function consistent with the one used in evolution. A structure-prediction scheme with a rugged non-funneled landscape for naturally evolved protein sequences is effectively hardly better than some random energy function. If proteins evolved to have a funnel-like landscape, there is also reason to believe that errors in potentials can be tolerated. Bryngelson showed that if the landscape is completely random, the potential must be known to an accuracy that scales as $1/\sqrt{N}$, making it nearly impossible to predict structures of long proteins [16]. Using similar methods, Pande et al. showed that for a funneled, minimally frustrated landscape, the percent accuracy needed did not scale with length. Using estimates of natural protein landscape ruggedness derived from experiment, they found that an accuracy of only 20–30% would be needed [17].

While a decade ago structure-prediction schemes based on simplified protein representations were nearly uniformly unsuccessful, some success can now be found. Some of the success can be viewed as arising from the development of schemes for emphasizing the funnel-like aspects of landscapes. Some of the success was explicitly motivated by the energy landscape paradigm, but sometimes the connection is only implicit. Homology modeling relies on using evolutionary information to reduce the importance of non-native-like associations [18], again assuming that evolution leads to minimal frustration. Energetic approaches for sequence-structure matching as a first step of modeling ("threading") are improved by varying parameters to distinguish "decoys" from native structures [19–24], again minimizing conflicts or frustration. This improvement can be carried out systematically using landscape-based algorithms. Averaging over multiple sequences smoothes out asperities on the landscape, again leading to a more globally funnel-like landscape [25–28]. Associative memory Hamiltonians for structure prediction using molecular dynamics as well as more purely physically based Hamiltonians used in lattice calculations have been improved by making explicit use of the quantitative form of the minimal frustration principle to optimize their parameters [29].

That some success has been achieved, while heartening to those of us who are optimists, should not blind us to the possibilities of further improvements. This requires the development of quantitative evaluation methods for assessing prediction algorithms. To do this efficiently, we must go further than merely judging final results of our own calculations. Setting up social schemes for avoiding inevitably subjective comparisons is helpful but not sufficient by itself. Both of these latter issues are well addressed by the CASP experiments [30, 31]. One approach to determining how to improve structure prediction has been to exploit the same tools used for studying folding kinetics to describe the landscapes provided by potential energy functions used in structure prediction. While we have so far done this for

understanding our own algorithms, "in-house" as it were, we believe the style of analysis can be used by others to help improve their own algorithms. We also hope to show in this paper how these schemes can be used to give an idea of the probability of success in multiple attempts to predict a structure using a stochastic algorithm.

The organization of the paper is as follows. In Section 2 we review the specific energy functions we use as well as the optimization schemes for relevant parameters. We then describe, in Section 3, sampling methods used to examine landscape topography. In Section 4 we discuss the resulting landscape for an ideal prediction scheme based on knowing the exact contact map of a protein, and for *ab initio* predictions, using associative memory energy functions which assume no homology information.

2. Model

While the method of analysis presented in this paper is generally applicable to any energy function, it is illustrated for an associative memory Hamiltonian in particular. In this section we describe the associative memory energy function. First we discuss the form of the energy function, and then we consider how the parameters in the energy function are optimized using energy landscape theory. Specific *ab initio* structure predictions from the energy function outlined here are discussed by Hardin et al. [32].

Associative memory energy function

The associative memory energy function discussed here has been designed with the objective of making *ab initio* predictions of protein structure. It was originally developed in the context of structure prediction by Friedrichs and Wolynes [33, 34], and is a reduced description, with only C^α, C^β, and O atoms explicitly represented. Search for low-energy states is carried out by molecular dynamics simulations. The full energy function is formally

$$E = E_{\text{back}} + E_{\text{amc}}\,, \tag{1}$$

where E_{back} describes the protein backbone, and E_{amc} comprises associative memory and contact terms, as discussed further below. The backbone part of the energy function is very similar to that used in previous studies [34, 35], and we relegate discussion of it to Appendix A.

Before considering the form of E_{amc}, we wish to make a comment regarding units. The unit of energy is denoted as ε, and is defined in terms of the native state energy excluding backbone contributions, E^N_{amc}, via

$$\varepsilon = \frac{|E^N_{\text{amc}}|}{4N}, \tag{2}$$

where N is the number of residues of the protein being considered. Temperatures are quoted in terms of the reduced temperature $\tilde{T} = k_B T/\varepsilon$. Distances, r, are in units of angstroms, and $\tilde{r}$ is used to denote the dimensionless $r/(1\ \text{Å})$.

The interactions described by E_{amc} depend on the sequence separation $|i - j|$ between the residues i and j involved. Specifically, they are divided into three proximity classes $x(|i - j|)$: x = short ($|i - j| < 5$), x = medium ($5 \le |i - j| \le 12$), and x = long ($|i - j| > 12$). Thus,

$$E_{\text{amc}} = E_{\text{short}} + E_{\text{med}} + E_{\text{long}}\,. \tag{3}$$

Both the short- and medium-range interactions are treated by an associative memory energy function,

$$E_{\text{AM}} = E_{\text{short}} + E_{\text{med}} = -\frac{\varepsilon}{a}\sum_{\mu=1}^{N_{\text{mem}}}\sum_{j-12\le i\le j-3}\left\{\gamma[P_i, P_j, P^\mu_{i'}, P^\mu_{j'}, x(|i-j|)] \times \exp\left[-\frac{(r_{ij} - r^\mu_{i'j'})^2}{2\sigma^2_{ij}}\right]\right\}. \tag{4}$$

The sum over i and j runs over all unique pairs of atoms (C^α–C^α, C^α–C^β, C^β–C^α, C^β–C^β) with sequence separation between 3 and 12, and r_{ij} is the distance between atoms i and j. The index μ runs over all N_{mem} memory proteins to which the protein has previously been aligned using a sequence-structure threading algorithm [36] (i.e., each i–j pair in the protein has an i'–j' pair associated with it in every memory protein; if, due to gaps in the alignment, there is no i'–j' pair associated with i–j for a particular memory, this memory protein simply makes no contribution to the interaction between residues i and j). The interaction between C^α and C^β atoms is thus a sum of Gaussian wells centered at the separations $r^\mu_{i'j'}$ of the corresponding memory atoms. The widths of the Gaussians are given by $\sigma_{ij} = |i - j|^{0.15}$ Å. The weights given to each well are controlled by $\gamma[P_i, P_j, P^\mu_{i'}, P^\mu_{j'}, x(|i - j|)]$, which depends on the identities $P_{i'}$ and $P_{j'}$ of the residues to which i and j are aligned, as well as the identities P_i and P_j of i and j themselves. A reduced four-letter (as opposed to 20-letter) code is used for the identities: P_i = hydrophilic ($i \in$ ala, gly, pro, ser, thr); P_i = hydrophobic ($i \in$ cys, ile, leu, met, phe, trp, tyr, val); P_i = acidic ($i \in$ asn, asp, gln, glu); and P_i = basic ($i \in$ arg, his, lys). Since γ also depends on proximity class, there are thus $4^4 \times 2 = 512$ different γ parameters in the associative memory term. These are optimized using energy landscape ideas as discussed below. We finally note that a is a dimensionless constant chosen so that Equation (2) is satisfied.

The energy function in the long-range proximity class is given by a three-well contact potential,

$$E_{\text{long}} = -\frac{\varepsilon}{a}\sum_{i<j-12}\sum_{k=1}^{3}\gamma(P_i, P_j, k)c_k(N)U[r_{\min}(k), r_{\max}(k), r_{ij}], \quad (5)$$

where i and j run only over all pairs of C^{β} atoms separated by more than 12 residues. The sum over k is over the three wells which are approximately square wells between $r_{\min}(k)$ and $r_{\max}(k)$. Specifically,

$$U[r_{\min}(k), r_{\max}(k), r_{ij}] = \frac{1}{4}\{[1 + \tanh(7[r_{ij} - r_{\min}(k)]/\text{Å})] \times [1 + \tanh(7[r_{\max}(k) - r_{ij}]/\text{Å})]\}. \quad (6)$$

The parameters $[r_{\min}(k), r_{\max}(k)]$ are (4.5 Å, 8.0 Å), (8.0 Å, 10.0 Å), and (10.0 Å, 15.0 Å) for $k = 1$, 2, and 3, respectively. In order to approximately account for the variation of the probability distribution of pair distances with number of residues in the protein (N), a factor $c_k(N)$ has been included in E_{long}. It is given by $c_1 = 1.0$, $c_2 = 1.0/(0.0065N + 0.87)$, and $c_3 = 1.0/(0.042N + 0.13)$. The individual wells are also weighted by γ parameters which depend on the identities of the amino acids involved, using the four-letter code defined above. Since we also enforce $\gamma(P_i, P_j, k) = \gamma(P_j, P_i, k)$ for all i and j, and there are three wells, the number of γ parameters is $10 \times 3 = 30$ in addition to the 512 from the associative memory part of the Hamiltonian.

Optimization

The 542 linear parameters, $\boldsymbol{\gamma}$, were optimized by training them on a set of ten proteins, using an optimization criterion explicitly based on energy landscape ideas. While the Hamiltonian above is rather general, it was trained for the specific task of making *ab initio* predictions for the structures of small to medium-sized alpha-helical proteins for which no structural homologs are known. Therefore, the training proteins were all alpha-helical proteins (ranging in size from 63 to 189 residues), as were the 36 memory proteins. Since our objective was to make predictions where structural homologs are not available, the memory proteins were also chosen to have a low structural (and sequence) similarity to the training proteins. Thus, it is a challenging problem to find the set of parameters $\boldsymbol{\gamma}$ that will give good "predictions" for the training proteins; however, we may have some confidence that having done so, the energy function will give good results for true unknowns outside the training set. The PDB codes of the training proteins and memories are given in Appendix B.

Full details of the optimization procedure have been presented most recently in Reference [32], and similar schemes have been discussed elsewhere [22, 29, 36, 37]; here we simply outline the main points. The basic idea of the approach is to find a $\boldsymbol{\gamma}$ that maximizes the funneling in the energy landscape of the training proteins. Specifically, we maximize the stability gap averaged over the training proteins, $\overline{\delta E_s(\boldsymbol{\gamma})}$. The stability gap of a single protein $\delta E_s(\boldsymbol{\gamma})$ is defined as the average energy gap (excluding backbone terms) between its native state and a specified set of collapsed non-native structures (these "globules" are generated in the first instance by simply taking coordinates from fragments of larger proteins). The maximization of $\overline{\delta E_s(\boldsymbol{\gamma})}$ must be subject to at least one constraint. In the simplest version of the scheme [22], this is achieved by fixing the variance of the globule energies averaged over the training set, $\overline{\Delta E^2(\boldsymbol{\gamma})}$. While this constraint effectively just sets the energy scale in the optimization procedure, the physical motivation is transparent: For a single protein, $\delta E_s/\sqrt{\Delta E^2}$ is the ratio of the native bias to the roughness of the landscape, i.e., a measure of the degree of funneling. It has also been shown that for a random energy model, maximization of this ratio is essentially equivalent to maximization of the ratio of folding to glass transition temperatures (T_f/T_g) [22].

In the current optimization procedure, the single constraint on the roughness is replaced by three separate constraints on the contribution to the roughness coming from the three proximity classes. The motivation behind this is to prevent the occurrence of a glass transition at high temperature within an individual class. For example, if the roughness in the short-range class were very large, the local in-sequence structure of the protein could become frozen (partially incorrectly), rendering the protein rigid before longer-range interactions have a chance to form. There are also three additional constraints on the contribution to the globule energy coming from the three classes. These give control over the collapse transition temperature of the resulting energy function, and also have the effect of giving a roughly even distribution of energy between proximity classes (in both native and globule states) in accordance with expectation [38].

The optimization procedure is performed iteratively; i.e., after solving for $\boldsymbol{\gamma}$, a new set of globules is generated by molecular dynamics using $E(\boldsymbol{\gamma})$ as the energy function (this is the most computationally intensive part of the procedure). Reoptimization follows, and the process is repeated until convergence is reached.

After $\boldsymbol{\gamma}$ is obtained, structure predictions are made by searching for low-energy states by molecular dynamics. An annealing schedule (temperature as a function of simulation time) must be set. With a simple linear reduction in T from $\tilde{T} = 2.0$ to $\tilde{T} = 0.0$ in 90 000 time steps, very good predictions (by current standards) are obtained for proteins outside the training set. This is detailed in the paper by Hardin et al. [32].

3. Sampling methods

Three principal cases have been investigated by the sampling method described in this section. The prediction Hamiltonian presented in the previous section has been analyzed for two different alpha-helical proteins. The first, phase 434 repressor (PDB code 1r69), is a member of the training set, and the second, HDEA (PDB code 1bg8), is neither a member of the training set nor does it have significant structural homology with any of the memory proteins. We emphasize again that the training proteins also do not have significant structural homology with any of the memory proteins. Thus, we do not expect there to be a very significant difference between the quality of the Hamiltonian for small to medium-sized alpha-helical proteins outside and inside the training set. There is of course the possibility of "over-learning" parameters in the optimization procedure if the training set is too small. While we do not attempt to address this issue in this paper, since clearly many more than two examples would be required, the techniques described below can straightforwardly be used in a systematic investigation. The main point in this paper is to quantitatively contrast the *ab initio* prediction energy landscapes of proteins 1r69 and 1bg8 with that arising for an ideal prediction scheme based on knowledge of the native structure. For this ideal scheme we use a Gō model [39] with the native structure of protein 1r69 designed to be its global energy minimum, as described in Appendix C.

One of the key thermodynamic quantities desired from simulation is the free energy as a function of temperature and one or more order parameters, **Q**. Since a protein is a finite-sized system, this could in principle be obtained from one constant-temperature simulation run (using, for example, the so-called single-histogram technique [40]). However, despite the fact that this approach can be successful in the context of protein folding (see, e.g., [41]), it is found to be inefficient for the off-lattice model studied here because the form of the free energy as a function of nativeness leads to only a very small region of phase space being explored in a reasonable simulation time. The method preferred here then is the multiple-histogram technique [42, 43], which has found widespread application in the protein folding field. As specified more precisely below, many simulations are carried out with different biasing potentials added to the Hamiltonian, each acting to constrain the protein to a chosen region of phase space. This allows efficient sampling of the funnel, while the multiple-histogram technique provides the prescription for extracting free energies (of the "bare" Hamiltonian) from this data.

The sampling procedure adopted is as follows: Initially, a temperature at which to sample must be chosen. The choice here is $\tilde{T} = 1$, which is guided by a number of considerations. First, this is close to the folding temperature of the Gō model, which is found to be $\tilde{T} = 1.06$. Second, it lies in the region $\tilde{T} = 0.8$–1.0, where structures with the highest degree of nativeness are typically found on annealing runs. Finally, at $\tilde{T} = 1$ the protein kinetics are sufficiently facile that the length of time between essentially independent samples is relatively small. At lower temperatures, slowing in the kinetics leads to a rapidly increasing computational burden for the prediction Hamiltonian. At $\tilde{T} = 1$, $n_s = 20$ constant-temperature simulations are then performed using the energy functions

$$E_i = E + V_i(\mathbf{Q}). \tag{7}$$

(E is replaced with $E_{G\bar{o}}$ for the Gō case, of course.) The functions $V_i(\mathbf{Q})$ ($i = 1, 2, \cdots, n_s$) are well-shaped potentials centered on different values of $\mathbf{Q}$ to give a good sampling of phase space along all of the reaction coordinates of interest, with care taken that "adjacent" simulations have overlap in the regions sampled. The preferred choice was $V_i(\mathbf{Q}) = V_i(Q) = 10^5\varepsilon(Q - Q_i)^4$, with $Q_i = 0$, $0.05, 0.1, \cdots, 0.95$. The order parameter, Q, measures similarity to the native state and involves a sum over *all* (except nearest-neighbor) pairs of C^α atoms,

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2} \right], \tag{8}$$

where r_{ij}^N is the C^α–C^α distance between residues i and j in the native state. Q runs between 0 (completely unfolded) and 1 (native).

During each simulation, a total of $N_i^{\text{obs}} = 400$ samples of $\mathbf{Q}$ and E are performed at regularly spaced intervals of 3000 time steps. The time interval such that successive observations may be considered independent depends on the Hamiltonian under consideration. At $\tilde{T} = 1$, it is approximately 30 000 time steps in the case of 1r69 and 1bg8, and about 6000 time steps for the Gō model; the dependence of this time on the biasing functions is weak. Thus, the total number of essentially independent samples is $20 \times 400/10 = 800$ for 1bg8a and 1r69, and $20 \times 400/2 = 4000$ for the Gō model. A histogram $N_i(E, \mathbf{Q})$ is created for each simulation, i, whence the multiple-histogram approach gives the following for the density of states, $n(E, \mathbf{Q})$:

$$n(E, \mathbf{Q}) = \sum_i w_i(E, \mathbf{Q}) \frac{N_i(E, \mathbf{Q})}{N_i^{\text{obs}}} e^{\beta_i E + \beta_i V_i(\mathbf{Q})} Z_i(\beta_i). \tag{9}$$

Here $\beta_i = 1/k_B T_i$ is the inverse temperature of simulation i, and the weights $w_i(E, \mathbf{Q})$ that minimize the error in $n(E, \mathbf{Q})$ may be expressed as

$$w_i(E, \mathbf{Q}) = \frac{A_i^{-2}}{\sum_j A_j^{-2}},$$

$$A_i^2(E, \mathbf{Q}) = n(E, \mathbf{Q})(N_i^{\text{obs}})^{-1} e^{\beta_i E + \beta_i V_i(\mathbf{Q})} Z_i(\beta_i). \tag{10}$$

The partition function, Z_i, is as usual given by

$$Z_i = \sum_{E,\mathbf{Q}} n(E, \mathbf{Q}) e^{-\beta_i E - \beta_i V_i(\mathbf{Q})}. \tag{11}$$

Equation (9) for $n(E, \mathbf{Q})$ and Equation (11) for Z_i self-consistently determine $n(E, \mathbf{Q})$ to within a multiplicative constant, and hence the free energy,

$$F(\mathbf{Q}, T) = -k_B T \log \left[\sum_{E,\mathbf{Q}} n(E, \mathbf{Q}) e^{-E/k_B T} \right], \tag{12}$$

to within an additive constant. Knowledge of the density of states allows straightforward calculation of the canonical energy and entropy, as well as expectation values of various observables, as a function of $\mathbf{Q}$ and T. Our experience is that, with proteins of a size similar to those analyzed here, the amount of sampling performed here gives good quantitative results for an extrapolation in temperature of up to roughly 10%. The temperature range may of course be extended by performing additional simulations at different temperatures.

In the case of protein 1r69, we perform an additional 20 simulations, as described above, but with the single difference that the functions $\{V_i(Q)\}$ are replaced with $\{V_i(Q_t)\}$, where Q_t measures the similarity to a local minimum ("trap") of the potential found from an annealing run. Q_t is defined as Q [Equation (8)], but with the coordinates of the native state replaced by those of the trapped state.

In the results section, thermodynamic quantities are principally displayed as a function of Q, but other order parameters are investigated as well; the use of one particular order parameter in the biasing functions does not preclude calculation of free energies as a function of another, as the multiple histogram equations above make clear. Other order parameters we consider in the following section include RMSD (the root mean square deviation of the C^α carbons from their native positions) and a "contact Q," Q_c. This is defined in a way similar to Q, but with the difference that only C^α pairs that are closer than a cutoff distance, r_c, in the native structure are included in the sum,

$$Q_c = \frac{\sum_{i<j-1} \theta(r_c - r_{ij}^N) \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2} \right]}{\sum_{i<j-1} \theta(r_c - r_{ij}^N)}. \tag{13}$$

We chose $r_c = 8$ Å. Q_c is thus more directly comparable to the typical order parameter used in lattice model studies (the fraction of native contacts) than is Q, which includes all C^α pairs in its sum. It can therefore be useful to examine the landscape using Q_c as an order parameter so that a better correspondence with lattice studies can be achieved. Finally, we also define order parameters that describe the amount of order by proximity class. These are also defined analogously to Q, but with the sum restricted to run over only the $|i - j|$ appropriate to that class. For example,

$$Q_{\text{short}} = \frac{\sum_{j-5<i<j-1} \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2} \right]}{\sum_{j-5<i<j-1} 1}. \tag{14}$$

4. Results

To give a sense of the relationship between the order parameter principally used in the thermodynamic profiles below (Q) and the commonly used RMSD order parameter, we show in **Figure 1(a)** the mean value of RMSD for protein 1r69 at $\tilde{T} = 1$ obtained with the prediction Hamiltonian, as a function of Q. The points on the figure represent instantaneous values obtained during the course of the simulation, and are included to indicate the size of fluctuations around the mean. Similarly, **Figure 1(b)** shows the dependence of the mean value of the contact order parameter Q_c as a function of Q. Several features apparent in the figure are worth noting. First, Q_c and Q are strongly correlated over their entire range. Note that $\langle Q_c(Q) \rangle > Q$ for all Q so that, for example, an ensemble of structures at $Q = 0.4$ (40% of residue pairs at the correct distance) corresponds roughly to one with $Q_c = 0.6$ (60% of contacts formed). On the other hand, Q and RMSD are strongly correlated for only $Q > 0.5$ (roughly RMSD < 5 Å). For lower values of similarity to the native state, there is a much wider spread of RMSD found at a single value of Q. For example, at $Q = 0.4$ the RMSD ranges from 5 Å to 12 Å. This reflects the fact that outside the immediate vicinity of the native state, motion of (say) a protruding arm of the protein can cause little change in the number of contacts while giving rise to large fluctuations in RMSD; this is one reason that Q (or Q_c) is often preferred to RMSD as an order parameter until only extremely good results are considered.

A rough conversion between RMSD and Q can be attempted for near-native structures, and Figure 1(a) shows that $Q = 0.75$ corresponds roughly to a 2-Å RMSD, while $Q = 0.54$ corresponds to an RMSD of 4 ± 0.5 Å. While these values and the results of Figure 1 have been obtained for the particular case of protein 1r69

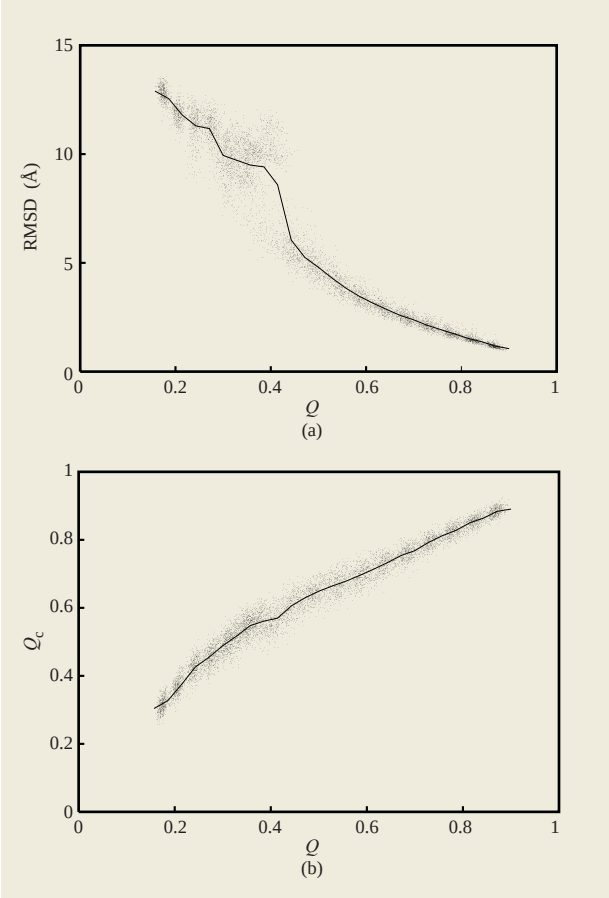

Figure 1

Mean value of order parameters (a) RMSD and (b) Q_c as a function of Q for protein 1r69 at $\tilde{T} = 1.0$ (solid lines). The points represent instantaneous values of the order parameters taken from the simulation at widely spaced time intervals.

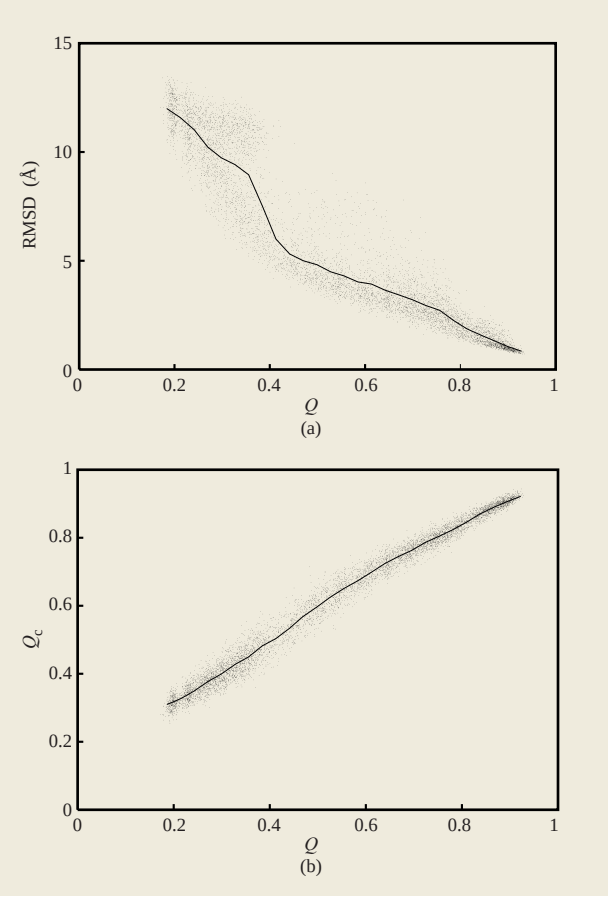

Figure 2

Mean value of order parameters (a) RMSD and (b) Q_c as a function of Q for the Gō model at $\tilde{T} = 1.0$ (solid lines). The points represent instantaneous values of the order parameters taken from the simulation at widely spaced time intervals.

at $\tilde{T} = 1.0$ with the prediction Hamiltonian, the situation is found to be insensitive to temperature variation, and very similar quantitatively for protein 1bg8. The case of the Gō model, which is shown in **Figure 2**, is also broadly similar, but there are some quantitative differences. Most notably, a significant number of structures for which $Q > 0.5$ have a large RMSD. This is probably related to the fact that, although in all cases studied here the proteins remain mostly collapsed over the entire range of Q, the fluctuations in the radius of gyration are larger in the Gō model than the prediction Hamiltonian, as illustrated in **Figure 3**. This stems from the fact that in the Gō situation no non-native interactions exist to stabilize the more fully collapsed states, and again points up the fact that when judging the quality of a structure it is desirable to have several measures available, and not rely only on the RMSD parameter.

Before examining the free-energy profiles, it is helpful to simply look at a few examples of the huge number of structures encountered during the molecular dynamics sampling. In **Figure 4** we show simulation structures for protein 1r69 with the prediction Hamiltonian at $Q = 0.25$, (top), $Q = 0.4$, $Q = 0.55$, and $Q = 1.0$ (bottom). At $Q = 0.25$ the structures are essentially random, with hardly any well-defined secondary structure; they are similar neither to each other nor to the native state. At $Q = 0.4$, helical secondary-structure elements are clearly visible; the structure to the right has some noticeable topological similarity with the native state. By $Q = 0.55$, however, the resemblance to the native state is clear. Note that while all other non-native structures were encountered at $\tilde{T} = 1$ (and in the presence of a biasing potential), the $Q = 0.4$ structure to the left is in fact a local

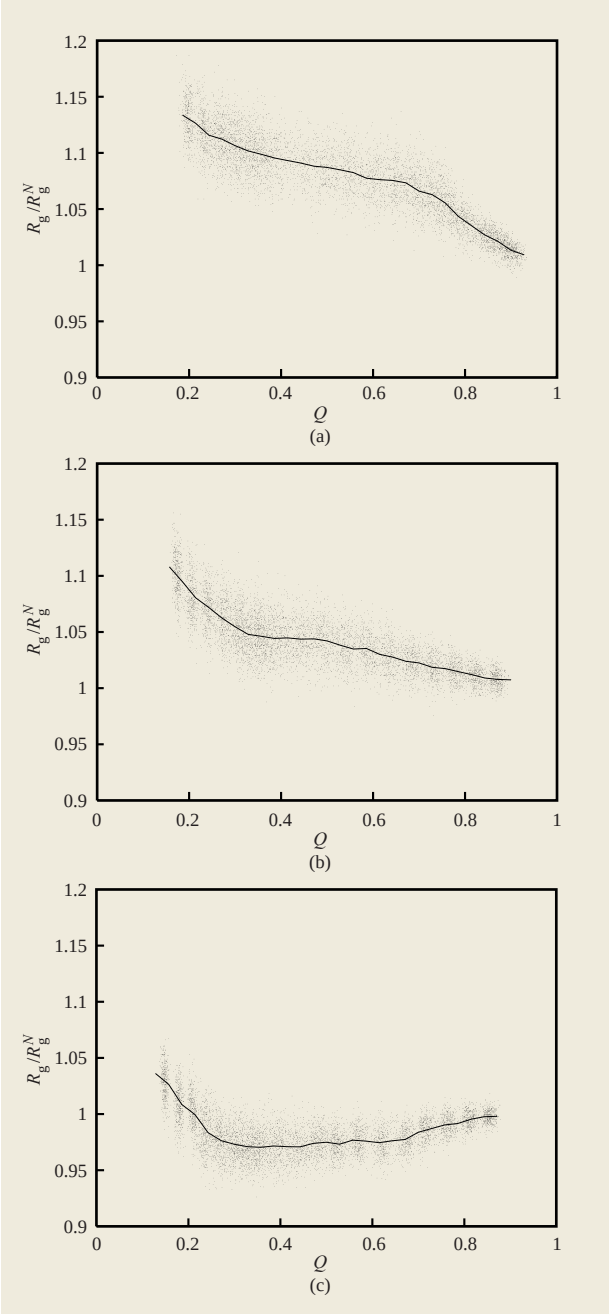

Figure 3

Mean value of radius of gyration (as a fraction of the native value) as a function of Q at $\tilde{T} = 1.0$ for (a) the Gō model, (b) lr69, and (c) 1bg8. The points represent instantaneous values taken from the simulation at widely spaced time intervals.

minimum of the (unbiased) energy function found from an annealing run. This probably accounts for the noticeable difference in helical content from the other two $Q = 0.4$ structures shown.

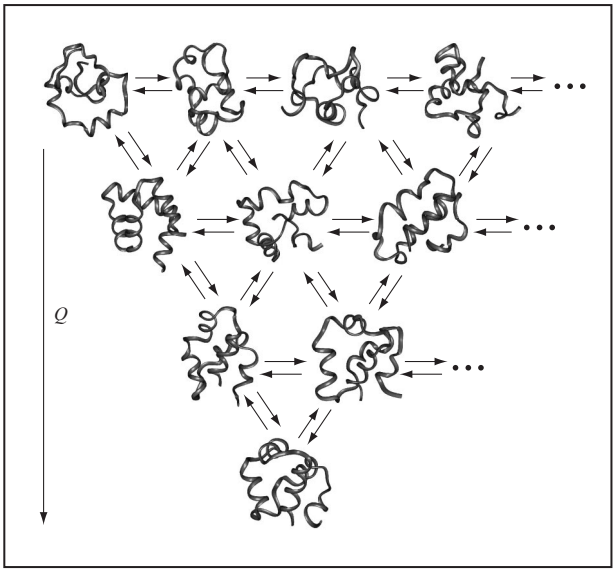

Figure 4

Simulation structures for protein 1r69 with the following degrees of nativeness: $Q = 0.25$ (top row), $Q = 0.4$ (second row), $Q = 0.55$ (third row), and $Q = 1.0$ (bottom). All structures have been rotated to minimize their RMSD with the native ($Q = 1.0$) structure shown. The $Q = 0.4$ structure to the left is the trap analyzed in the text. All other structures were obtained during the sampling procedure described in the previous section.

Figure 5 shows the free energy as a function of both Q and Q_c at two different temperatures ($\tilde{T} = 1.06$ and 0.9) for the three different Hamiltonians investigated here (Gō model, protein 1r69 with the prediction Hamiltonian, and protein 1bg8 with the prediction Hamiltonian). The behavior of the Gō model is seen to be in striking contrast to that of the prediction Hamiltonian. In the Gō case at $\tilde{T} = 1.06$ [Figure 5(a)], the free energy has a double-well structure, with a small barrier of 3.2ε ($3.0k_BT$) separating a disordered molten globule-like minimum at around $Q = 0.33$ ($Q_c = 0.41$) from a native-like minimum with $Q = 0.76$ ($Q_c = 0.79$). As the temperature is lowered, the globule rises in free energy relative to the native minimum (and in fact by $\tilde{T} = 0.9$ the globule minimum has essentially been "washed away"), which in turn moves to still higher Q ($Q = 0.87$ or $Q_c = 0.9$ at $\tilde{T} = 0.9$) and approaches $Q = 1$ as the temperature is lowered further. The free-energy minima in this case are controlled by the exchange of energy for entropy. With regard to the prediction Hamiltonian, however, only a single minimum is seen over the same temperature range for proteins 1r69 and 1bg8. Measured by Q, this minimum lies roughly at the same position as the Gō globule minimum, but measured by Q_c it is more native-like. Lowering $\tilde{T}$ from 1.06 to 0.9 does lead to a shift in the position of the

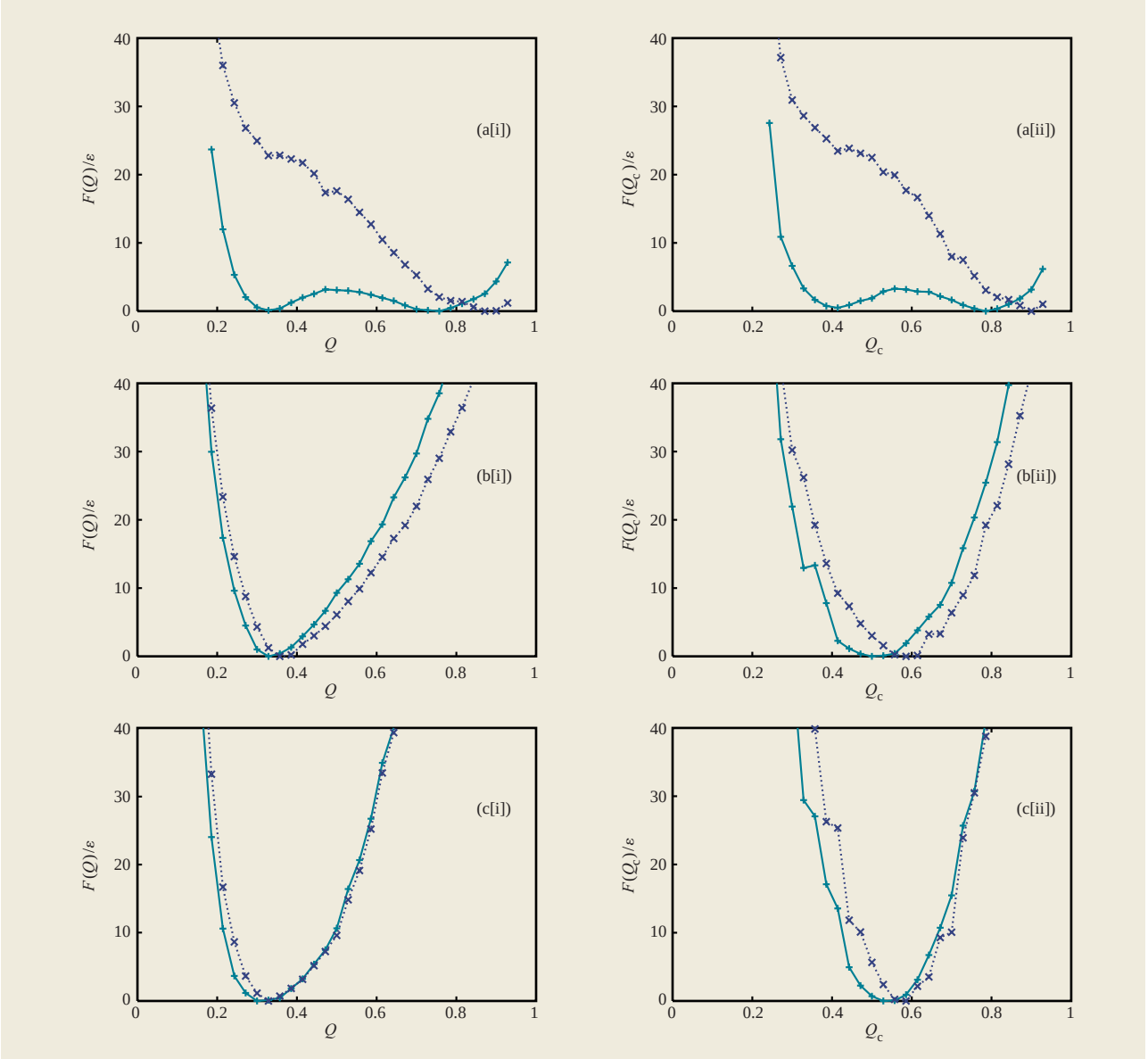

Figure 5

Free energy [expressed in units ε, as defined in Equation (2)] as a function of (i) Q and (ii) Q_c at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. Note that the curves have been shifted by a Q-independent constant so that their minima lie at $F = 0$, to aid comparison.

minima (for both proteins 1r69 and 1bg8) toward the native state (measured by both Q and Q_c); however, this shift is modest in comparison to the Gō case, and a further decrease in temperature is found not to move the minimum significantly nearer the native state. This is an indication that conflicts between different forms of energy, in contrast to a simple tradeoff between energy and entropy, are important; i.e., the prediction Hamiltonian is more frustrated than the Gō Hamiltonian.

Knowledge of the form of $F(Q, T)$ for proteins in the training set has immediate practical benefits, notwithstanding the insight it may give into improving energy functions (which is discussed further below). In particular, it aids in the design of an appropriate annealing schedule, although for a full picture all proteins in the training set, as well as kinetic issues, should be considered. On the basis of analysis of protein 1r69, it appears that search via MD efforts should be concentrated

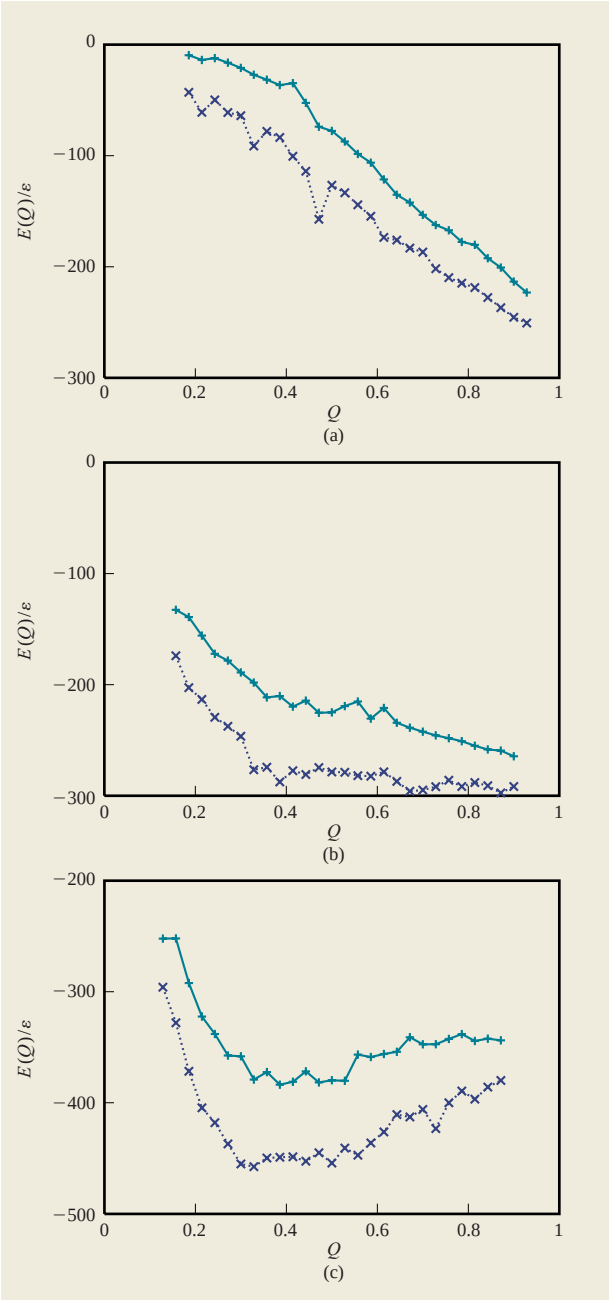

Figure 6

Energy as a function of Q at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. The caldera-like nature of the prediction energy function is seen to contrast with the strongly funneled character of the Gō model.

around $\tilde{T} = 0.9$. Not only does $F(Q, T)$ indicate the expected quality of a prediction at a given temperature, but it also trivially allows an estimate to be made of the expected sampling time required before a structure of a desired quality is found. For example, at $\tilde{T} = 1.06$, $F(Q = 0.55) \approx 13\varepsilon \approx 12k_BT$, and so the number of independent samples typically required to realize such a structure will be roughly $e^{12} \approx 10^5$ at this temperature. At $\tilde{T} = 0.9$, $F(Q = 0.55) \approx 9\varepsilon \approx 10k_BT$, and the number of independent samples required to see a $Q = 0.55$ structure is reduced to $e^{10} \approx 10^4$. For protein 1bg8 at $\tilde{T} = 1.06$, $F(Q = 0.55) \approx 18k_BT$, requiring roughly 10^8 independent samples, a number which in fact increases slightly at the lower temperature of $\tilde{T} = 0.9$, at which $F(Q = 0.55) \approx 20k_BT$.

In order to better understand the behavior of the free energy, it is helpful to consider the energetic and entropic components separately. These are illustrated in **Figures 6** and **7** respectively as a function of Q (examining the entropy and energy as functions of other order parameters is of course possible and yields essentially the same picture). Consider first the energy. In the case of the Gō model, the energy decreases monotonically, almost linearly as Q is raised from 0 to 1. Decreasing the temperature does not affect this behavior, for which the energy function provides a thermodynamic driving force for increasing Q whatever the value of Q; i.e., the energy function is funneled to the native state. In the case of protein 1r69, it is found [Figure 6(b)] that at $\tilde{T} = 1.06$, $E(Q)$ is monotonically decreasing; i.e., the energy function is funneled to the native state, although for $Q > 0.4$ the slope is less pronounced. At $\tilde{T} = 0.9$, however, for $Q > 0.4$ the energy is essentially flat and is no longer funneled to the native state. The energy landscape in this case resembles a caldera more than a funnel. In the case of protein 1bg8 the situation is similar, but more pronounced: The landscape is funneled to $Q \approx 0.4$, but a further increase in Q is accompanied by an increase in $E(Q)$; lowering the temperature only acts to make this bias away from the native more severe. In this light, the behavior of the free energy is more understandable. For example, since as stated above, $E(Q)$ is funneled down to $Q = 0.4$ at $\tilde{T} = 0.9$, lowering the temperature below $\tilde{T} = 0.9$ will not lead to a minimum in $F(Q)$ that is more native-like than at $Q = 0.4$. Since the minimum in $F(Q)$ already lies close to $Q = 0.4$ [Figure 5(b)], further reduction in temperature is unlikely to significantly improve $F(Q)$ and could even lead to deterioration.

The reason that the energy landscape for the prediction Hamiltonian is caldera-like rather than funnel-like is of course due to the existence of non-native structures that lie lower in energy than the putative native structure. In order to gauge the number of structures that are energetically competitive with the native structure, it is necessary to turn to the entropy curves. It is sobering first to consider the Gō model. As noted above, this model has a folding transition at $\tilde{T} = 1.06$, and the transition-state ensemble [defined to be the maximum in $F(Q)$] lies just below $Q = 0.5$. Thus, defining all states with $Q \geq 0.5$ to

lie in the native basin and all other states to lie in the globule minimum, we calculate the entropy difference between the native and globule basins. At T_f this is found to be $122k_B$ [as may be approximately verified by looking at the entropy difference in Figure 7(a) between $Q = 0.33$ and $Q = 0.76$]. In other words, the globules are entropically equivalent to $e^{122} \approx 10^{53}$ native basins. With regard to the size of a native basin, we note that $\langle \mathrm{RMSD}(Q = 0.76, \tilde{T} = 1.06) \rangle = 2.7$ Å. Since $10^{53} \approx 63^{29}$, this corresponds to roughly 30 states of 2.7-Å resolution per residue. Even though these numbers of states would be reduced were the Gō model confined to be more tightly collapsed, they clearly emphasize the entropic mountain faced in structure prediction.

In the case of protein 1r69 with the prediction Hamiltonian, the slope of $S(Q)$ is smaller than for the Gō Hamiltonian. As temperature is reduced, this slope is reduced still further. This again reflects the presence of non-native interactions; these naturally have more effect at small to intermediate values of Q (at $Q = 1$, of course, they are absent by definition), where they stabilize a small fraction of states which gain in thermodynamic weight as the temperature is lowered. The situation for protein 1bg8 is seen to be similar to that for protein 1r69, but it is more extreme: At $\tilde{T} = 0.9$, $S(Q)$ becomes nearly flat above $Q = 0.3$, strongly suggesting that the system is "running out of states" at this temperature; i.e., it is approaching a glass transition [44–47]. To be more quantitative, we pose the question "What is the (canonical) entropy difference ΔS^* between the two volumes of phase space separated by the surface $Q = Q^*$?" Given that we do not know the actual basin size in the prediction model, and that with this Hamiltonian the desired native state may not even lie in a localized basin, the choice of Q^* is somewhat arbitrary. We make the choice $Q^* = 0.55$, motivated by the fact that this corresponds to an RMSD of slightly less than 4 Å on average, which would currently be considered a very good *ab initio* prediction. A quick estimate of ΔS^* may be obtained from locating the minimum in $F(Q)$ (Figure 5), then calculating the difference $\Delta S^* = S(Q_{\min}) - S(Q^*)$ from Figure 7. These differences are roughly $\Delta S^*(\tilde{T} = 1.06) \approx 30k_B$ and $\Delta S^*(\tilde{T} = 0.9) \approx 20k_B$ for protein 1r69, and $\Delta S^*(\tilde{T} = 1.06) \approx 20k_B$ and $\Delta S^*(\tilde{T} = 0.9) \approx 10k_B$ for protein 1bg8. In other words, the number of states thermally occupied in protein 1r69 decreases from 10^{13} 4-Å basins at $\tilde{T} = 1.06$ to 10^8 4-Å basins at $\tilde{T} = 0.9$. [Note that the difference between the estimated number of thermally occupied states at $\tilde{T} = 0.9$ (10^8 $Q = 0.55$ basins) and the above-estimated expected number of states to search through at this temperature before finding one with $Q \geq 0.55$ (10^4) reflects the fact that $E(Q)$ is not precisely flat at this temperature, but retains a small bias toward the native.] The corresponding drop for protein 1bg8 is roughly 10^8–10^4 4-Å basins. Unfortunately, accurate calculation of these numbers beyond this temperature range is hampered by statistical noise in the calculation of the energy. However, within the range $\tilde{T} = 1.1$–0.9, the decrease in entropy is found to be

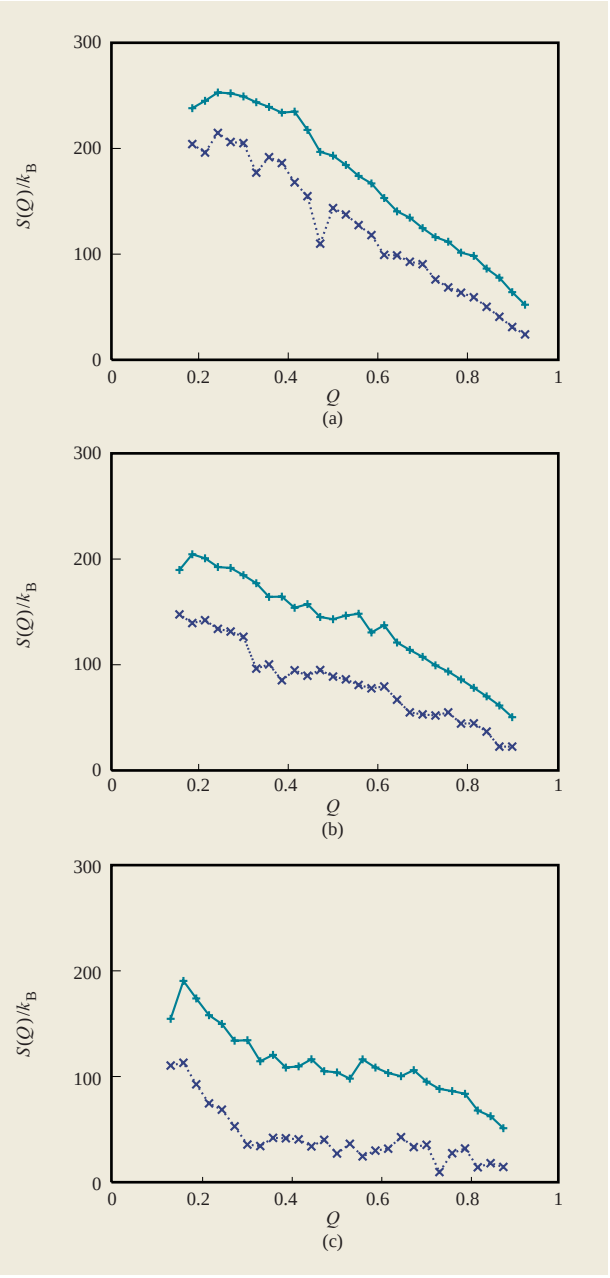

Figure 7

Entropy (divided by k_B) as a function of Q at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. Note that both curves in each part of this figure have been shifted by the same Q-independent constant so that the entropy appears to approach zero at $Q = 1$.

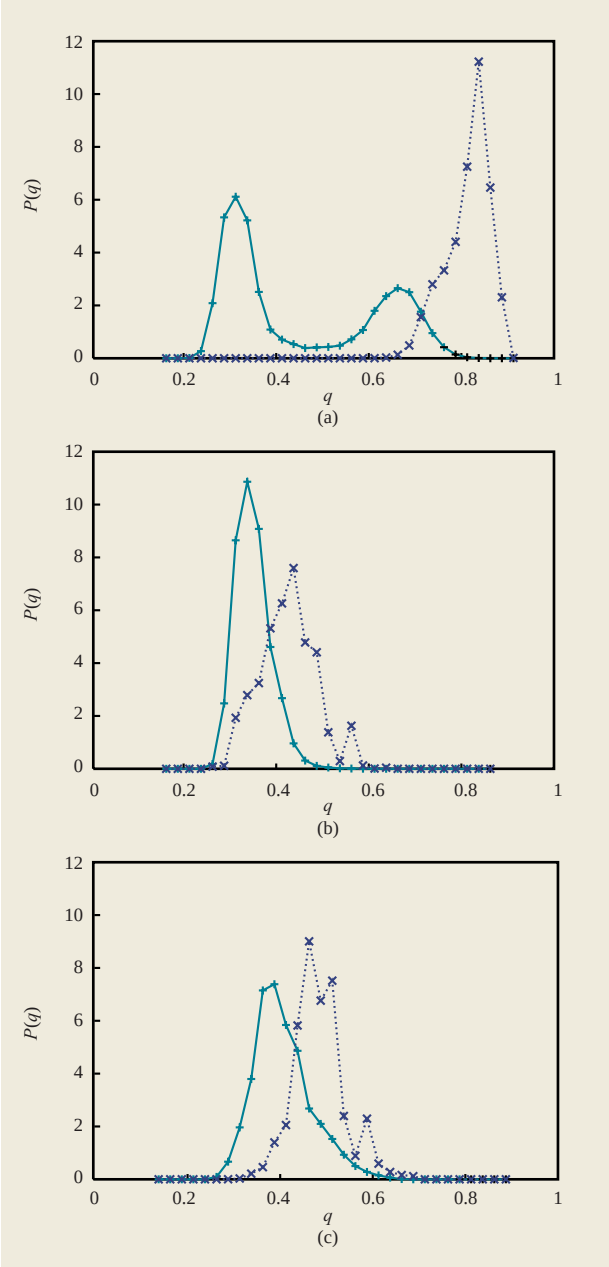

Figure 8

Equilibrium probability distribution, $P(q)$, of observing a similarity q between two independently selected structures at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. $P(q)$ is normalized such that $\int_0^1 P(q)\,dq = 1$.

nearly linear, so a thermodynamic glass transition in the region $\tilde{T} = 0.6$–0.8 appears likely.

Further investigation of the onset of glassy behavior is facilitated by considering order parameters other than those quantifying similarity to the native state. For example, **Figure 8** shows the probability distribution $P(q)$ of overlaps q obtained between members of the equilibrium ensemble at two different temperatures for the three Hamiltonians studied. Although, naturally, the Gō model is the antithesis of glassy behavior, it is nonetheless helpful to consider $P(q)$ for this case first; see Figure 8(a). At the Gō folding temperature ($\tilde{T} = 1.06$), $P(q)$ has two distinct peaks: one at $q \approx 0.31$ corresponding to the globule minimum, and one at $q \approx 0.66$ corresponding to the native basin (the structures with strong similarity to a particular structure—in this case the native—are also similar to one another). At the lower temperature $\tilde{T} = 0.9$, only the native-like peak survives, and it moves to higher q, since at this temperature the native basin is more tightly focused on the native structure [see Figure 5(a[i])] and the equilibrium structures are consequently more similar to one another. This situation contrasts sharply with that found for the prediction Hamiltonian. Here, the maximum in $P(q)$ moves only from $q = 0.34$ to $q = 0.41$ (1r69) and $q = 0.36$ to $q = 0.44$ (1bg8a) between $\tilde{T} = 1.06$ and 0.9. This movement reflects the loss in entropy of the ensemble as the temperature is lowered; however, in this temperature range there is no compelling evidence of the existence of a small number of strongly localized basins (traps); i.e., there is no second peak at large values of q. There could be several reasons for the absence of such a signal. Most obviously, the temperature range investigated could still be above the kinetic glass transition temperature (T_A) where distinct traps separated by free-energy barriers form [44, 48]. Alternatively, the system could be below T_A, but with the globule basin still thermodynamically dominant because of its larger entropy. Also, problems with sampling can never be entirely ruled out: Free-energy barriers may make some basins kinetically inaccessible on the simulation time scale, or there could simply be so many traps that the probability of return to any one is low on the simulation time scale. In this regard, note that the presence of a large number of dissimilar traps will also reduce the expected magnitude of any second peak in $P(q)$, which is approximately inversely proportional to the number of traps.

It is interesting to probe an individual trap in detail. **Figure 9** shows the free energy, energy, and entropy as a function of similarity Q_t to a local minimum of the energy function found—as discussed in the previous section—from an annealing run with the prediction Hamiltonian for protein 1r69. Figure 9(a) indicates that in the temperature range studied, there is only a single minimum in $F(Q_t)$ (lying close to $Q_t = 0.4$). Thus, the local minimum found from annealing is not the center of a localized basin at $\tilde{T} = 0.9$. However, it is clear that at some lower temperature this situation must change, since the state is a local minimum of the energy function, but not

a unique one. There are some indications from Figure 9 that at a temperature not far below $\tilde{T} = 0.9$, $F(Q_t)$ will develop a double-minimum structure. In particular, in comparison with $F(Q)$ [Figure 5(b[i])], $F(Q_t)$ is significantly flatter in the region $0.6 < Q_t < 0.8$, and becomes flatter with decreasing temperature. Furthermore, $E(Q_t)$ [Figure 9(b)] remains funneled at $\tilde{T} = 0.9$, indicating that below this temperature the trend in F is continued (i.e., high-Q_t states are reduced in free energy relative to those at low Q_t); and that a second minimum may appear at high Q_t. The trend in F suggests that this occurs at roughly $\tilde{T} = 0.7$–0.8. Indeed, direct calculation of $F(Q_t)$ at $\tilde{T} = 0.7$, although an extrapolation of 30% in temperature and outside the range for which the multiple histogram method is quantitatively reliable, does indeed suggest a double-minimum structure for $F(Q_t)$, with a second minimum occurring in the range $Q_t = 0.7$–0.8. Note that this implies a basin size somewhat smaller than the $Q = 0.55$ used to estimate the thermodynamic glass transition temperature above; if it is typical, that estimated temperature should be revised downward a little. Of course, all traps will to some extent have different $F(Q_t)$, and analysis of a single one does not necessarily give the full picture. However, the main point of our analysis here is to illustrate techniques that may be used to investigate trapped states, rather than to provide an exhaustive study of a single Hamiltonian.

We now consider how to identify where in the energy function improvements might be made. In this regard, it is often useful to consider the contribution of different components of the energy to the slope of the funnel. A natural first step is to consider separately the associative memory and backbone contributions. In **Figure 10** the backbone energy is shown as a function of Q for two different temperatures for the three Hamiltonians considered here, and in addition it is shown as a function of similarity to the trapped state just discussed. For the case of protein 1r69 [Figure 10(b)], the backbone provides a significant contribution to the funneling, roughly 35ε between $Q = 0.3$ and $Q = 0.8$ at $\tilde{T} = 1.06$, and 25ε at $\tilde{T} = 0.9$. Unsurprisingly, this is similar to the amount of backbone funneling provided in the Gō case [Figure 10(a)]. The slope in the case of protein 1bg8a is less marked, and almost disappears at $\tilde{T} = 0.9$. Further analysis (not illustrated) shows that the major contribution to the variation of E_{back} with Q is from the Ramachandran potential. Figure 10 thus immediately implies that an increase in the weight of the Ramachandran potential will lead to a shift in the minimum of $F(Q)$ toward the native state for protein 1r69. Increasing the weight of this term will lead to a depopulation of those levels with unfavorable ϕ–ψ angles, which lie at intermediate values of Q; i.e., the entropy of the globule basin will be reduced. There is naturally a limit beyond which further increase of the Ramachandran weight will not improve prediction results, as would be indicated by a flattening out of a plot of Ramachandran energy against Q (much as seen for protein 1bg8 at $\tilde{T} = 0.9$). Further increase of the weight of an unbiased Ramachandran energy in the

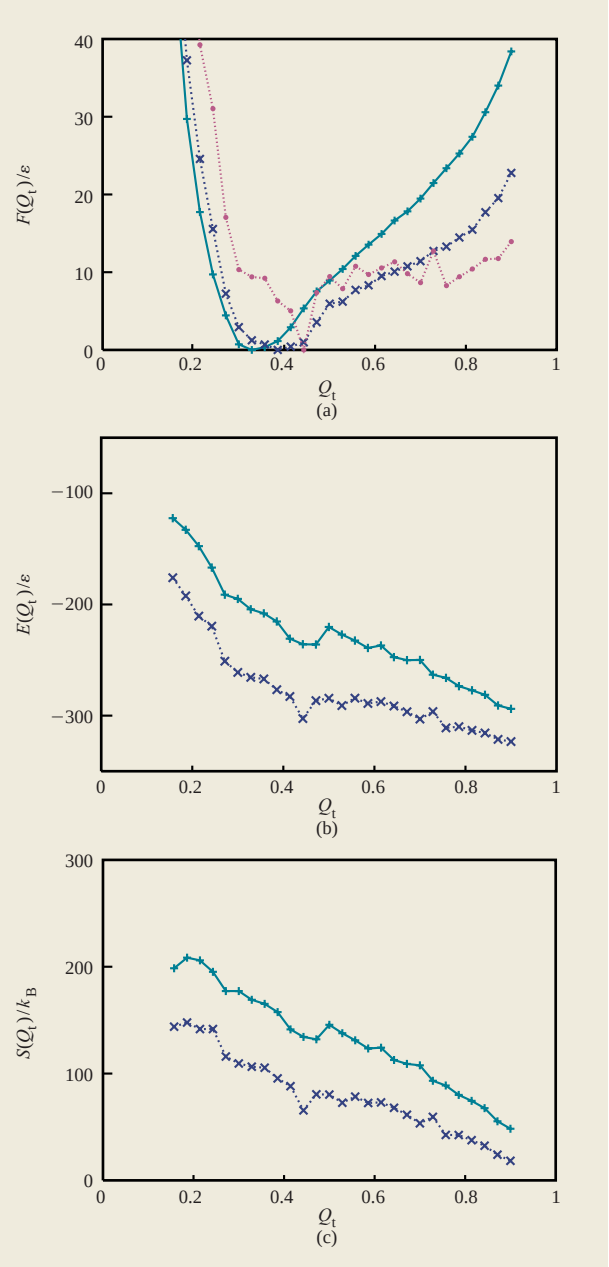

Figure 9

(a) Free energy, (b) energy, and (c) entropy vs. similarity to a low-energy non-native "trapped" state (Q_t), at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×). In (a), the free energy is also shown at $\tilde{T} = 0.7$, where a double-minimum structure is apparent.

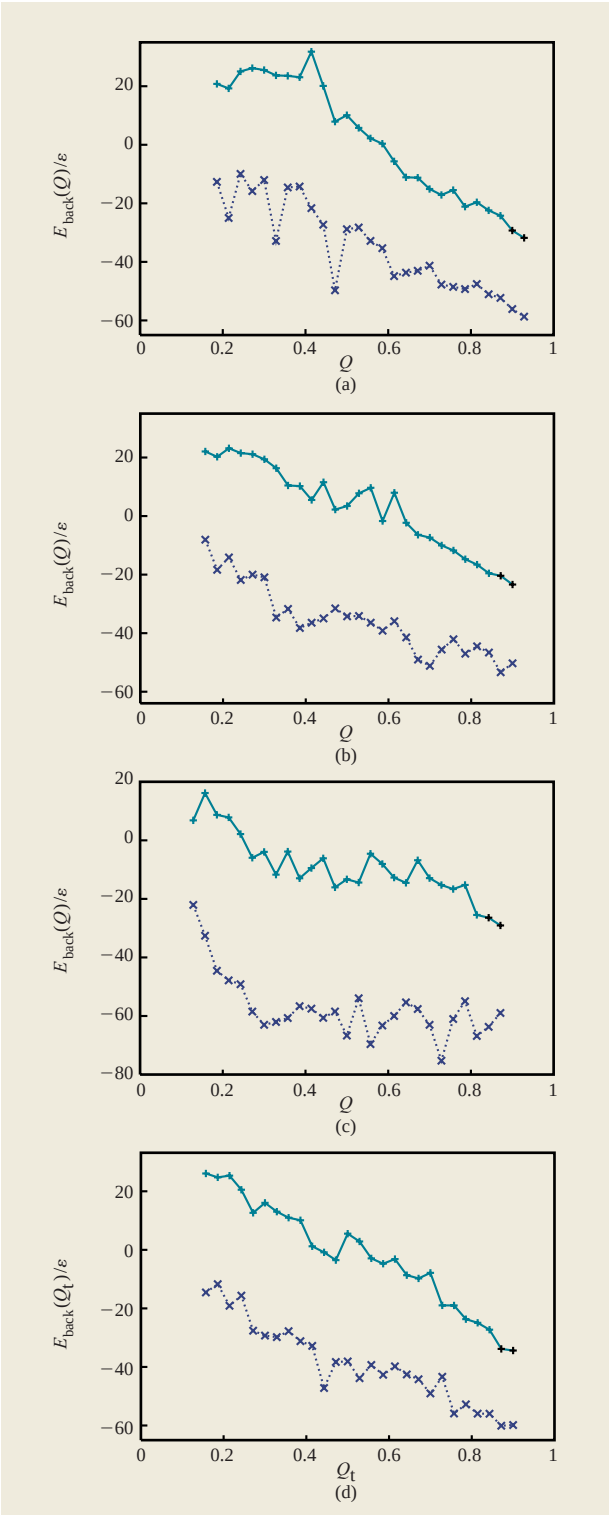

Figure 10

Backbone contribution to the energy as a function of similarity to the native state at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. In (d) the backbone energy is shown for protein 1r69 as a function of similarity to a trapped state.

Hamiltonian will lead to emphasized glassy behavior with basins centered on relatively few states with near-perfect ϕ–ψ angles. [It is worth noting that the trapped state found with the current Hamiltonian already has a slightly more funneled $E_{back}(Q)$ plot than the native; see Figure 10(d).] Therefore, the weight of the backbone should be changed cautiously, after considering all of the proteins in the training set as well as any kinetic consequences of such a change. The use of secondary-structure predictions to produce a Ramachandran potential biased to predicted secondary structure could also help.

Figure 11 shows the combined contribution of associative memory and contact (i.e., non-backbone) terms to the energy as a function of Q. The behavior of $E_{amc}(Q)$ largely mirrors that of $E(Q)$ shown in Figure 6. In particular, $E_{amc}(Q)$ is funneled to the native state in the Gō case, and also, though less strongly, for protein 1r69 at $\tilde{T} = 1.06$, while at $\tilde{T} = 0.9$, $E_{amc}(Q)$ is flat beyond about $Q = 0.4$ for protein 1r69, and slopes away from the native state for protein 1bg8a in the temperature range shown. Thus, at $\tilde{T} = 0.9$, the weak residual funneling beyond $Q = 0.4$, apparent in $E(Q)$ for protein 1r69 [Figure 6(b)], arises entirely from the backbone contribution (Figure 10) as opposed to E_{amc}. Note also that the difference between E_{amc} as a function of similarity to the native and E_{amc} as a function of similarity to the trapped state is a late funneling worth about 20ε between $Q_t = 0.5$ and 0.9 [compare Figures 11(b) and 11(d)]. This is much greater than the corresponding difference arising from backbone contributions, indicating that the trapped state is largely stabilized by associative memory or contact contributions.

It would naturally be desirable to improve the quality of the non-backbone terms. To investigate them in more detail, we break up their contributions according to sequence separation. Specifically, E_{amc} is separated into three terms $\{E_x\}$ for which x = short, medium, or long, corresponding to the three proximity classes of the Hamiltonian. These are shown as a function of Q in **Figure 12**. The individual components of E_{amc} are found to follow trends similar to those of E_{amc} (or even the total energy E). For example, all components $\{E_x\}$ contribute similarly to the funneling in the Gō model, which is funneled down to the native state [Figure 12(a)], while in the prediction case the funneling is weaker or absent after a certain value of Q [Figures 12(b, c)]. However, it is more profitable to focus on the difference between the different classes of interaction. For example, Figure 12(b) shows that the contribution to the funneling for protein 1r69 above $Q = 0.4$ comes almost entirely from the short-range interactions at $\tilde{T} = 1.06$; at $\tilde{T} = 0.9$, there remains weak funneling in this class of interactions, while the other two classes give a small funneling away from the native state [leading to a flat total $E_{amc}(Q > 0.4)$; see Figure 11(b)]. This situation is essentially repeated for

protein 1bg8a, with the contribution to funneling being largest for E_{short}, followed by that for E_{medium}. This implies that a reweighting of the overall contributions of short-, medium-, and long-range interactions to increase the weight of short-range ones will move the minimum in $F(Q)$ toward the native state (at least in the temperature range of Figure 12). The balance among short-, medium-, and long-range interactions is optimal when all three curves level off at the same temperature.

While Figure 12 aids assessment of the balance among the various terms, it is difficult to use it alone to assess the relative merits of the energy function used in the three classes. The additional information required is a measure of the amount of ordering by class, i.e., a knowledge of the thermodynamics as a function of $\{Q_x\}$. For example, the main difference between $\{E_x\}$ as a function of Q and as a function of Q_t is an additional funneling that occurs near the trapped state in the short- and medium-range classes of interactions. This does not imply that there is a flaw in these particular classes, especially given that the trap has $Q_{short} = 0.8$, $Q_{medium} = 0.56$, but $Q_{long} = 0.29$. A more plausible explanation is that improvement in the long-range potential would help discriminate more against such traps. Another example for which $\{Q_x\}$ information is required is shown in Figure 12(b[i]). Here, the caldera-like nature of the long- and medium-range classes compared to the funneled short-range energy could in principle be explained by a scenario in which the medium-to-long-range potential is so good that Q_{medium} and Q_{long} are close to unity for all structures encountered in the simulation. That this is not in fact the case can be seen in **Figure 13** [part (b[i]) in particular], in which the total free energy of the proteins is shown as a function of Q_x. The minimum of $F(Q_x)$ for protein 1r69 at $\tilde{T} = 1.06$ lies at 0.62 (short), 0.44 (medium), and 0.27 (long). Thus, not only is the contribution of the short-range potential to the overall funnel largest in this case, but the short-range native order is at the same time greater than that in the other two classes. When the temperature is reduced to $\tilde{T} = 0.9$, the minima in $F(Q_{short})$ move further toward the native state than in the other two classes. Protein 1bg8a shows behavior similar to that of 1r69 (and there is also no significant difference using similarity to the trapped state by class as order parameters). The short-range native order for protein 1r69 is comparable to that of Gō [compare Figures 13(b[ii]) and (a[ii])], while the native order in the medium-range class is considerably less than that in the Gō folded minimum, but is more than in the Gō globules. The order in the long-range class is similar to that in the Gō globules.

The above suggests that improvements should be directed at the long-range (contact) potential, which does not appear to provide sufficient discrimination. This is also indicated by **Figure 14**, in which the components of the

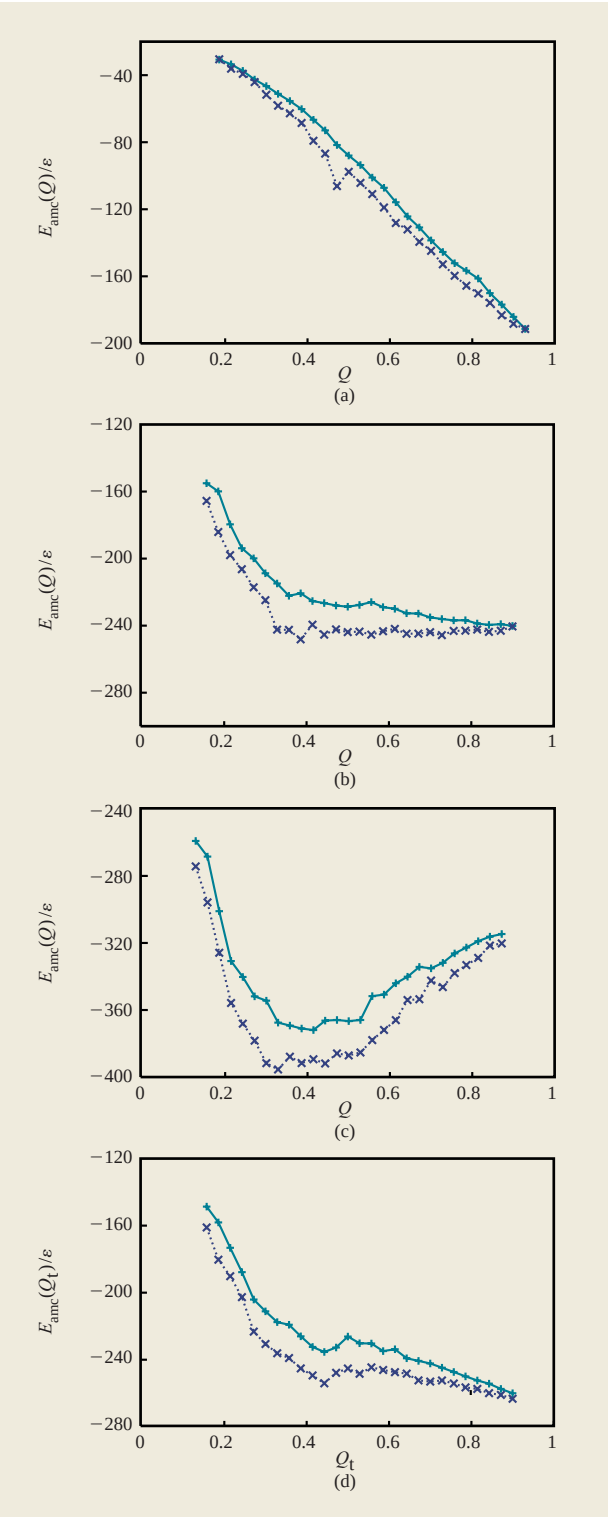

Figure 11

Total associative memory and contact contribution to the energy as a function of similarity to the native state at $\tilde{T} = 1.06$ (+) and $\tilde{T} = 0.90$ (×) for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. In (d) the AMC energy is shown for protein 1r69 as a function of similarity to a trapped state.

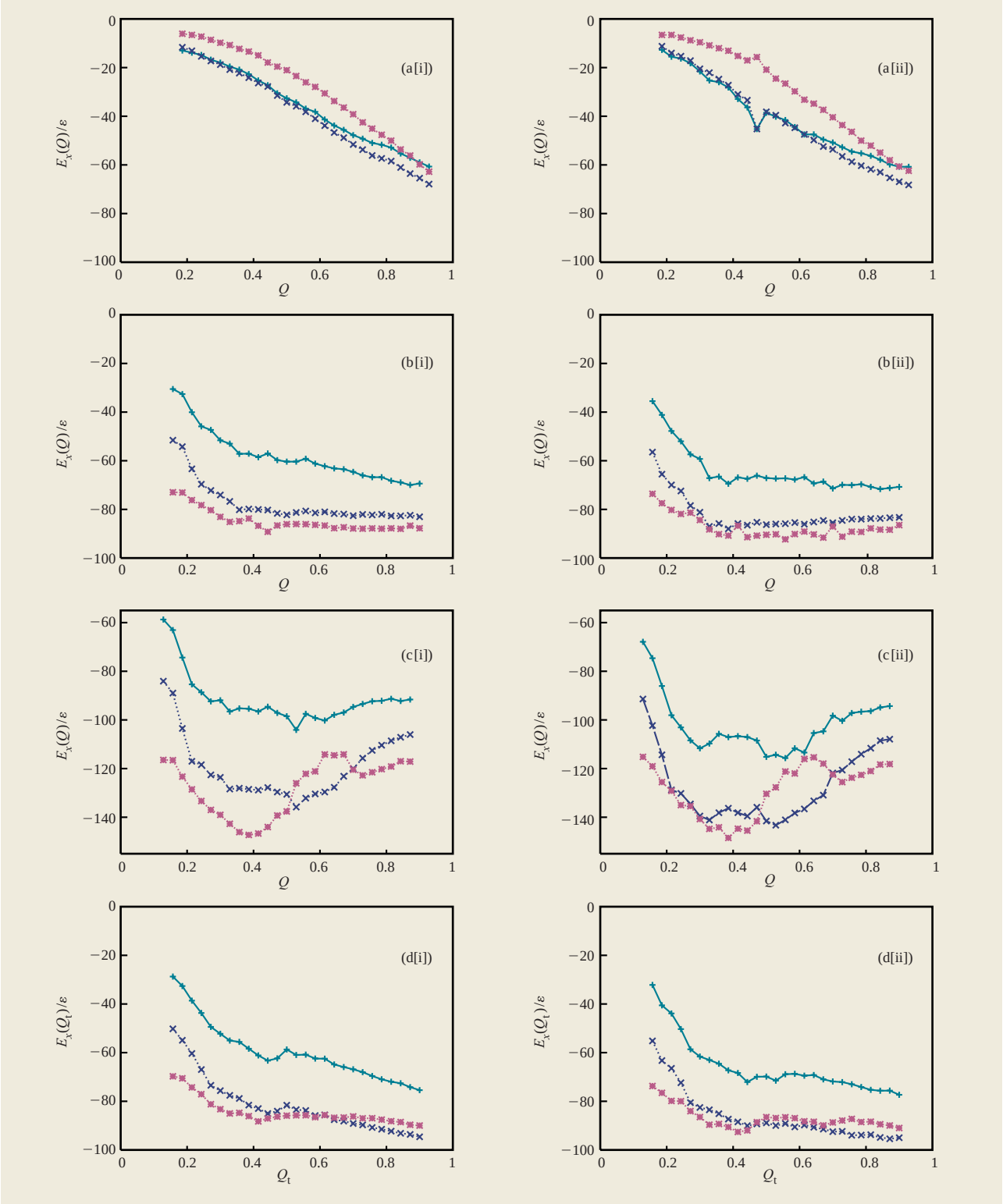

Figure 12

Components of the AMC contribution to the energy E_x [for x = short (+), medium (×), and long (*)] as a function of similarity to the native state at (i) $\tilde{T} = 1.06$ and (ii) $\tilde{T} = 0.90$. The plots are for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. In (d) the same components of the AMC energy are shown for protein 1r69 as a function of similarity to a trapped state.

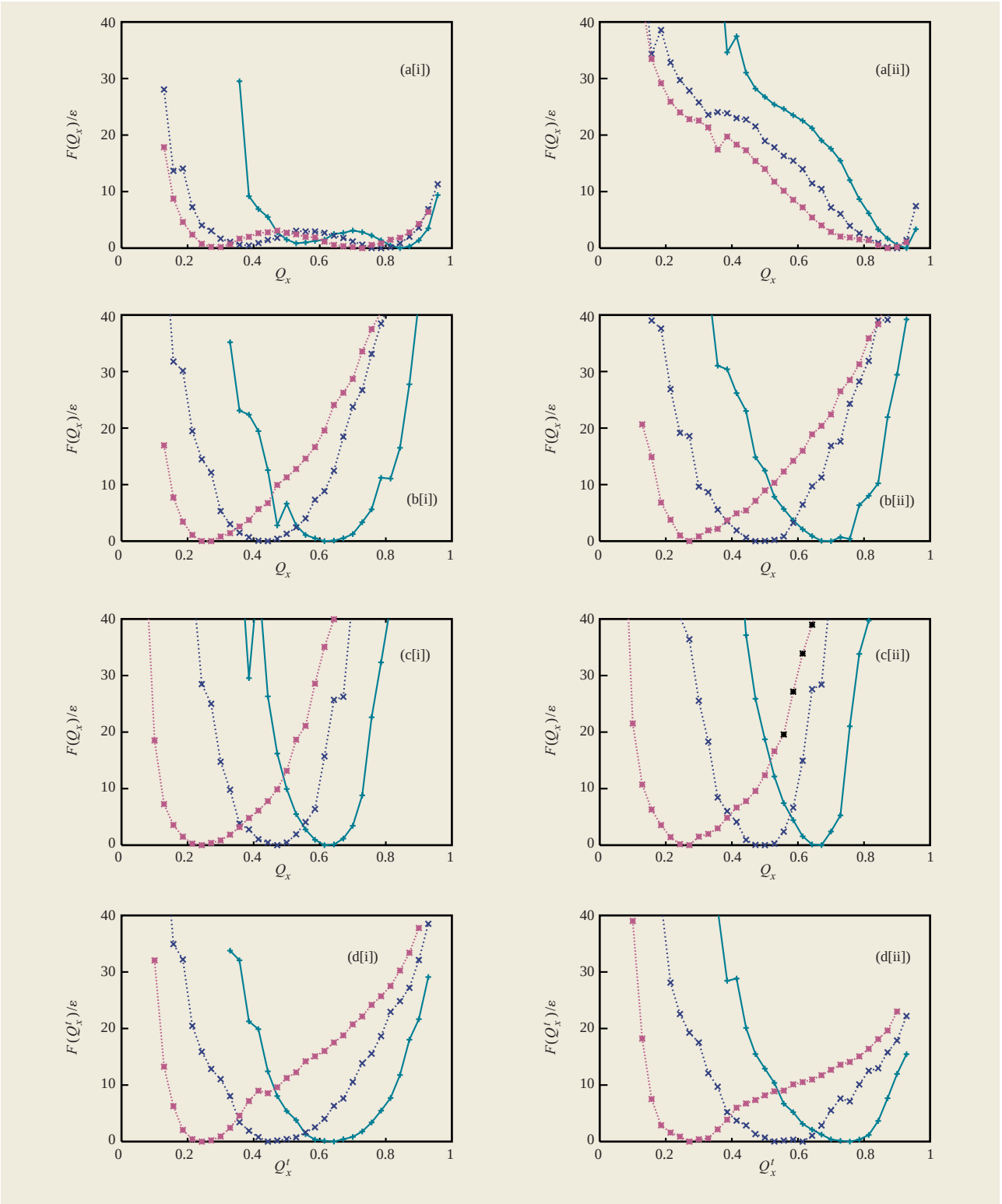

Figure 13

Free energy as a function of the sequence-dependent native similarity parameters Q_x [for x = short (+), medium (×), and long (*)] at (i) $\tilde{T} = 1.06$ and (ii) $\tilde{T} = 0.90$. The plots are for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. In (d) the free energy is shown analogously for protein 1r69 as a function of similarity to a trapped state Q_x^t.

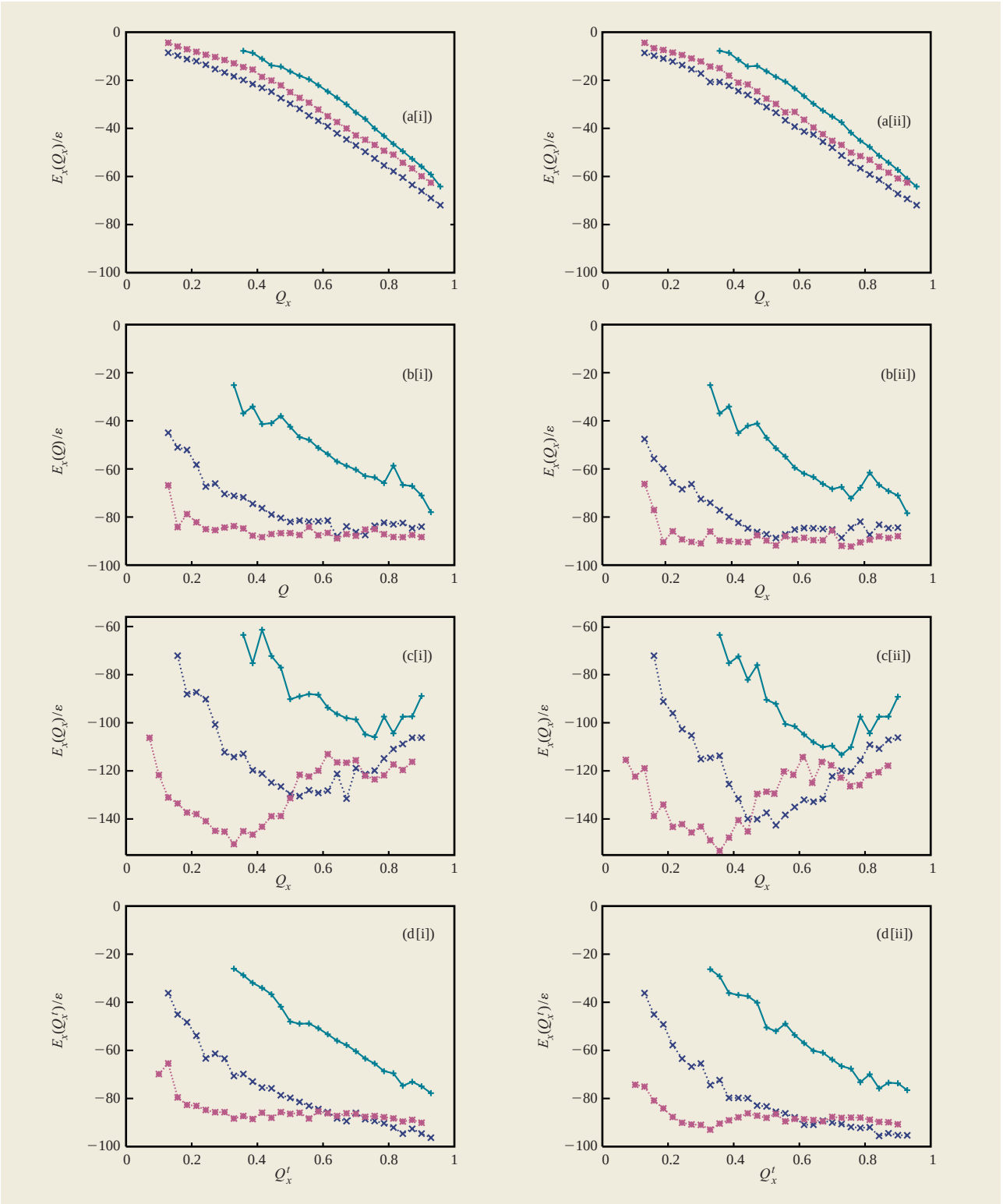

Figure 14

Components of the AMC contribution to the energy [for x = short (+), medium (×), and long (*)] as a function of the corresponding component of similarity to the native state (Q_x) at (i) $\tilde{T} = 1.06$ and (ii) $\tilde{T} = 0.90$. The plots are for (a) the Gō model, (b) protein 1r69, and (c) protein 1bg8. In (d), the same components of the AMC energy are shown for protein 1r69 as a function of similarity to a trapped state.

energy, E_x, are plotted as functions of their corresponding Q_x values. This method of displaying the data is probably the clearest way of comparing the quality of different terms, although, as discussed above, that used in Figure 12 makes possible a more ready assessment of the balance among terms. It is immediately apparent from Figure 14(b) that the funneling in the short-range class is excellent (to $Q_{\text{short}} = 1$), while in the medium range it is good (to $Q_{\text{medium}} = 0.7$ at $\tilde{T} = 1.06$ and to $Q_{\text{medium}} = 0.55$ at $\tilde{T} = 0.9$); however, in the long-range class it is poor (weakly to $Q_{\text{long}} = 0.4$ at $\tilde{T} = 1.06$, and not beyond $Q_{\text{long}} = 0.2$ at $\tilde{T} = 0.9$). The situation with regard to the long-range class is somewhat better for protein 1bg8, but the discrimination in this class is clearly worse than in the other two classes. This is not very surprising, since there exist good local signals for secondary structure, so design of good potentials is easier in the short-range (local) as opposed to long-range classes. The discrimination achieved in the medium-range class is probably responsible for the relative success of the prediction Hamiltonian.

Although we have concentrated on describing how changes of the potential affect free-energy profiles and funnel characteristics, free-energy profiling can be used to investigate other aspects of prediction schemes. We end with an example using a consensus sequence scheme. One might argue that fluctuations in sequence away from a consensus are tolerated and are either evolutionally neutral or represent adaptations for function. Thus, a consensus sequence might have a more pronounced funnel. To study this we prepared a consensus sequence for the training protein 1r69. This was done by identifying the 14 most similar sequences via a BLAST search [49], and then identifying the most common class of residue at each position in the sequence after performing a multiple-sequence alignment. This is a simplified version of the multiple-sequence averaging scheme of Keasar et al. [25, 26]. In our four-letter code (H = hydrophobic, P = hydrophilic, A = acidic, B = basic), the actual protein 1r69 sequence and the consensus sequence are, respectively, PHP**P**BHB**P**BB**H**AHPH**A**APAHPABHP**P**PAAPHAA-HAAPBPB**B**PBHH**P**AHPPPHPHP**H**AHHHAPP and PHP-**A**BHB**A**BB**P**AHPH**P**APAHPABHP**H**PAAPHAAHAAPBP-B**P**PBHH**H**AHPPPHPHP**P**AHHHAPP. The eight positions where the sequences differ are shown in bold. The free energy and the energy of the consensus sequence were calculated in the key range $Q = 0.2$ to $Q = 0.6$ and are compared to the results for protein 1r69 in **Figures 15(a)** and **15(b),** respectively. There is a clear but modest shift in the position of the free-energy minimum from $Q = 0.36$ to $Q = 0.39$. This is accompanied by a greater funneling in the medium-Q range ($Q = 0.3$–0.5). There is also evidence that the energy function in the long-range proximity class, in particular, has better discrimination when the consensus sequence is used; see **Figure 15(c)**. This is despite the fact that it was the protein 1r69 sequence, and not the consensus sequence, that was used in the optimization procedure. Thus, while the caldera picture still applies and the resolution of predictions is limited, using the consensus

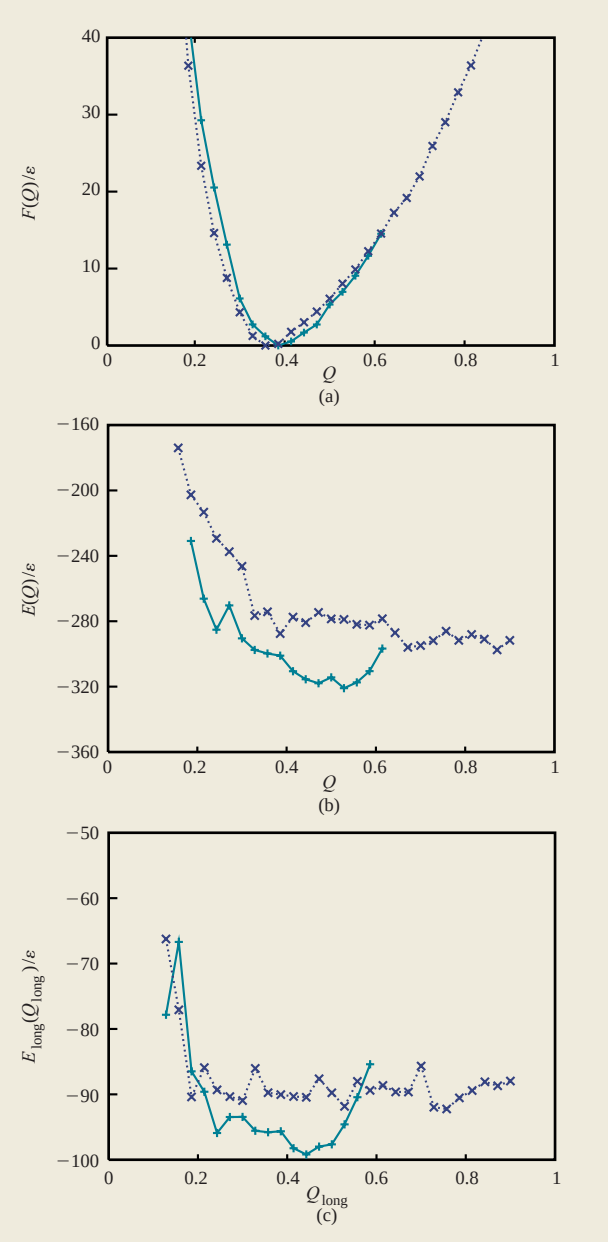

Figure 15

Thermodynamic profiles of the consensus sequence (+) compared to the actual protein 1r69 sequence (×) at $\tilde{T} = 0.9$. Part (a) shows the free energy as a function of Q; part (b) shows the total energy as a function of Q; part (c) shows the long-range (contact) energy as a function of the corresponding order parameter Q_{long}.

sequence is seen to have a small beneficial effect in smoothing out the energy landscape. More sophisticated ways of using multiple sequence information such as that described by Keasar et al. may have a greater effect.

5. Concluding remarks

The protein structure prediction problem is essentially one of finding an energy function with a funnel-shaped landscape for naturally occurring protein sequences. Additionally, to make the problem computationally feasible it is naturally desirable that the energy function be as simple as possible; i.e., that a reduced description of the protein be used, omitting explicit solvent molecules and possibly some side-chain atoms, and containing few expensive many-body interactions. It is not obvious that such a reduced description can in fact produce a funnel-shaped landscape in the absence of homology information, even for a limited subset of proteins (such as alpha-helical proteins). However, while current *ab initio* prediction schemes are far from perfect, the best are certainly better than random, indicating that their landscapes are funneled to some extent.

In this paper, we have shown, using well-known methods, how the shape of the energy landscape may be quantified for a structure-prediction Hamiltonian. While the method was illustrated for an associative-memory Hamiltonian with no homology information, it is generally applicable to other energy functions. There are two main reasons to analyze an energy function as described here. First is the wealth of information obtainable from such an analysis, allowing the quality of the energy function to be objectively quantified according to a number of measures. For example, knowledge of $F(Q, T)$ trivially allows estimation of the amount of sampling time required to find a structure of a specified quality; by measuring the free energy as a function of similarity to trapped states, estimates of the kinetic glass transition temperature and basin size can be made. Knowledge of $S(Q, T)$ gives the total number of thermally occupied states as a function of temperature, and leads to an estimate of the thermodynamic glass transition temperature. Perhaps most significantly, $E(Q, T)$ gives a direct and quantitative measure of the shape of the landscape. For an ideal energy function, $E(Q, T)$ is funneled down to the native state at all temperatures, and so the position of the global free-energy minimum is controlled by the exchange of energy for entropy: At low enough temperature, a good "prediction" will be obtained. For the prediction energy function, $E(Q, T)$ is funneled to the native state at high temperature (where folding is entropically prevented), but when the temperature is reduced, the slope in the funnel decreases until a situation is reached in which the energy landscape resembles a caldera; i.e., it is only funneled up to a certain value of nativeness, beyond which it is flat. This reflects the presence of competing (non-native) interactions; i.e., folding is energetically prevented by the presence of low-energy non-native structures (traps).

The information from thermodynamic analysis appears to come at a large computational cost: For the example studied here, the analysis requires 30 times the computing time of a single prediction. This cost may be reduced by focusing on a smaller region of phase space, dictated in part by expectations for the energy function; if, for example, a prediction of nativeness of $Q = 0.55$ is required, there is much to be learned just from studying the thermodynamics in the range $Q = 0.3$ to $Q = 0.55$. Additionally, the comparison of computing time with that of a single prediction is not a fair one, since for a partially funneled landscape the quality of predictions will vary between individual annealing runs, and several are required to gauge the quality of the energy function. While the thermodynamic analysis can never completely replace judging the quality of the predictions as a means of evaluation (the predictions are, after all, the "final product"), there is a second key reason to perform it. That is, analysis of the thermodynamics as a function of various order parameters points the way to where improvements can and should be made to the prediction scheme. On a straightforward level, it can guide the design of an annealing schedule. Examining components of the energy as a function of overall nativeness allows the balance between various terms to be assessed and gives an indication how they should be reweighted. Examining components of the energy as a function of the corresponding nativeness [e.g., $E_{\text{short}}(Q_{\text{short}})$], gives a better indication of which components individually require improvement. Once the optimum parameterization has been obtained for a given energy function, further improvement requires refining the form of the energy function. Thermodynamic knowledge can in principle be used to help design two-stage prediction schemes in which a reduced description is used to quickly narrow the search problem to a relatively small basin, after which more detail (e.g., extra atoms or many-body interactions) can be "switched on" to narrow the ensemble of predicted structures still further.

Although we have not discussed it here, the use of sampled structures with a range of nativeness, combined with a knowledge of the $E(Q)$ and $F(Q)$ in that range, also opens up the possibility of alternative optimization criteria. A simple example is to maximize the stability gap between molten globule structures and an ensemble of structures constrained to lie within, say, 4 Å RMSD of the native state (rather than the native state itself), with the objective of increasing the funneled character of the landscape in the region where it is most needed. Such schemes could potentially be enhanced if combined with the histogram analysis, which allows calculation of energy

and free energy as a function of the parameters in the energy function without additional molecular dynamics simulations. We hope to return to these issues in the near future.

Appendix A: Backbone Hamiltonian

The same backbone Hamiltonian as used here, with minor differences, has been discussed more expansively elsewhere; see for example [35]. For completeness, the main points of the backbone energy function used in this paper are summarized below. $E_{\rm back}$ is composed of several terms:

$$E_{\rm back} = E_{\rm SHAKE} + E_{\rm ev} + E_{\rm chain} + E_{\rm chi} + E_{\rm rama}\,. \tag{A1}$$

Each term depends only on the positions of C^α, C^β, and backbone O atoms (all assumed to have unit mass), which are thus the only atoms to enter the dynamics. It aids visualization, however, to express some of the forces between these atoms in terms of the variables

$$\mathbf{r}_{N_i} = 0.483\mathbf{r}_{C^\alpha_{i-1}} + 0.703\mathbf{r}_{C^\alpha_i} - 0.186\mathbf{r}_{O_{i-1}} \tag{A2}$$

and

$$\mathbf{r}_{C'_i} = 0.444\mathbf{r}_{C^\alpha_i} + 0.235\mathbf{r}_{C^\alpha_{i+1}} - 0.321\mathbf{r}_{O_i}\,, \tag{A3}$$

where, in an obvious notation, $\mathbf{r}_{N_i}$ and $\mathbf{r}_{C'_i}$ are the positions the nitrogen and C′ carbons of the protein backbone would respectively assume, given ideal amino acid geometry.

The chain connectivity is maintained by the SHAKE algorithm [50], which constrains the neighboring C^α–C^α distance, as well as the C^α–C^β bond and distances from the oxygens to the neighboring two C^α atoms at their ideal values. Excluded volume effects are included via a harmonic interaction between C^α–C^α, C^α–C^β, C^β–C^β, and O–O pairs of atoms separated by less than $r_{\rm ev}$:

$$E_{\rm ev} = \varepsilon\lambda^{\rm C}_{\rm ev}\sum_{x,y}\sum_{i<j}\theta[r^{\rm C}_{\rm ev}(j-i) - r_{C^x_iC^y_j}][r^{\rm C}_{\rm ev}(j-i) - r_{C^x_iC^y_j}]^2 + \varepsilon\lambda^{\rm O}_{\rm ev}\sum_{i<j}\theta(r^{\rm O}_{\rm ev} - r_{O_iO_j})(r^{\rm O}_{\rm ev} - r_{O_iO_j})^2\,, \tag{A4}$$

where x and y can each take the values α and β. Note that $r^{\rm C}_{\rm ev}$ has a sequence dependence; its actual values are given below. The remaining terms act to maintain chain geometry close to that of an ideal peptide chain (with some flexibility). First, the correct bond angles at C^α are maintained by a combination of the SHAKE algorithm and harmonic potentials of the form

$$E_{\rm chain} = \varepsilon\lambda_{\rm chain}\sum_i\{(\bar r_{N_iC^\beta_i} - 2.46)^2 + (\bar r_{C'_iC^\beta_i} - 2.51)^2 + (\bar r_{N_iC'_i} - 2.45)^2\}. \tag{A5}$$

Second, chirality at the C^α atoms is maintained using the potential

$$E_{\rm chi} = \varepsilon\lambda_{\rm chi}\sum_i(\chi_i - \chi_0)^2\,, \tag{A6}$$

where

$$\chi_i = (\bar{\mathbf{r}}_{C'_iC^\beta_i}\times\bar{\mathbf{r}}_{C^\alpha_iN_i})\cdot\bar{\mathbf{r}}_{C^\alpha_iC^\beta_i} \tag{A7}$$

and $\chi_0 = -0.83$. Finally, there is a term designed to produce a distribution of dihedral (ϕ, ψ) angles roughly reflecting that found in real proteins, and commonly displayed in the form of Ramachandran plots,

$$\begin{aligned}E_{\rm rama} = -\varepsilon\lambda_{\rm rama}\sum_{i=2}^{N-1}\;& 1.3149e^{-15.398\{[\cos(\phi+2.051)-1]^2+[\cos(\psi-2.138)-1]^2\}}\\ &+1.17016e^{-100.521\{[\cos(\phi+1.353)-1]^2+[\cos(\psi-2.4)-1]^2\}}\\ &+1.29264e^{-49.0954\{[\cos(\phi+1.265)-1]^2+[\cos(\psi+0.218)-1]^2\}}\\ &+1.78596e^{-419.123\{[\cos(\phi+1.265)-1]^2+[\cos(\psi+0.929)-1]^2\}}.\end{aligned}$$

This term has been somewhat modified from that given in Reference [35] to better fit the backbone torsional angles observed in the standard Ramachandran map for non-glycine residues [51]. Nevertheless, the Ramachandran potential used here is similar to that used previously in that barriers between the minima are deliberately constructed to be lower than in reality in order to promote more facile chain dynamics.

The parameters chosen are $\lambda^{\rm C}_{\rm EV} = 15.0$, $\lambda^{\rm O}_{\rm EV} = 15.0$, $\lambda_{\rm chain} = 30.0$, $\lambda_{\rm rama} = 1.0$, $\lambda_{\rm chi} = 20.0$, $r^{\rm C}_{\rm ev}(j-i<5) = 3.5$ Å, $r^{\rm C}_{\rm ev}(j-i\geq 5) = 4.5$ Å, and $r^{\rm O}_{\rm ev} = 3.5$ Å.

Finally, we note that while the backbone given above was used in generating structures for the optimization procedure outlined in Section 2, the backbone used in the thermodynamic analysis described in the results section in addition included a term dependent on the radius of gyration of the protein, $R_{\rm g}$ (calculated from C^α positions). This is given by a quadratic well centered at $R^{\rm pred}_{\rm g}(N) = 2.2N^{0.38}$ (which is an estimate of the radius of gyration of single-domain proteins [52]). Specifically,

$$E_{\rm radius} = \varepsilon\lambda_{\rm radius}(\bar R_{\rm g} - \bar R^{\rm pred}_{\rm g})^2\,, \tag{A8}$$

for $0.75 < R_{\rm g}/R^{\rm pred}_{\rm g} < 1.5$, and $E_{\rm radius}$ is constant outside this range. We take $\lambda_{\rm radius} = 10.0$. This term has little effect on the thermodynamics for the prediction Hamiltonian, and is not required for collapse in this case. It is merely included to facilitate collapse in the case of the Gō model, which otherwise does not collapse before folding.

Appendix B: Training proteins and memories

The PDB codes of the ten training proteins are 1r69, 1utg, 3icb, 256b, 4cpv, 1ccr, 2mhr, 1mba, 2fha, and 1rgp. The 36 memory proteins are 1jhg, 256b, 1bgf, 5icb, 1ah7, 2a0b, 1tx4, 1avs, 1c3d, 1a28, 1ak0, 2abk, 1ail, 1lis, 1b4f, 1pbv, 1huw, 1lki, 2lbd, 1vin, 1aa7, 1bja, 1nsg, 1beo, 1au1, 1rcb, 1e2a, 1b10, 1hiw, 1col, 1szt, 1hul, 1a17, 1axd, 1baj, and 1kxu. For training protein 3icb, memory proteins 5icb and 1avs are replaced with 1cf7 and 1aep; for training protein 1rgp, memory protein 1tx4 is replaced with 1cf7. This ensures that Q for training proteins aligned with memories is $Q < 0.4$ in all cases, and typically $Q \approx 0.2$. Similarly for protein 1bg8, examined in this paper as an "unknown" protein outside the training set, memories 1a28 and 1a17 are replaced by 1cf7 and 1aep.

Appendix C: Gō Hamiltonian

The Gō model used in this paper is defined by

$$E_{\text{G}\bar{\text{o}}} = E_{\text{G}\bar{\text{o}}}^{\text{AM}} + E_{\text{back}} . \tag{A9}$$

The backbone part, E_{back}, is identical to that used in the prediction Hamiltonian and described in Appendix A. The other term is an associative memory term with its minimum at the native structure of protein 1r69,

$$E_{\text{G}\bar{\text{o}}}^{\text{AM}} = -\frac{\varepsilon}{a_{\text{G}\bar{\text{o}}}} \sum_{i \le j-3} \gamma_{\text{G}\bar{\text{o}}}[x(|i-j|)] \exp\left[-\frac{(r_{ij} - r_{ij}^{N})^2}{2\sigma_{ij}^2}\right] . \tag{A10}$$

The sum over i and j runs over all unique pairs of atoms (C^{α}–C^{α}, C^{α}–C^{β}, C^{β}–C^{α}, C^{β}–C^{β}) with sequence separation of at least three residues. The interaction between $C^{\alpha}(i)$ and $C^{\beta}(j)$ atoms is thus a Gaussian well centered at their native separation r_{ij}^{N}. The well widths are given by the same formula as in the prediction Hamiltonian, $\sigma_{ij} = |i-j|^{0.15}$ Å. The weights $\gamma_{\text{G}\bar{\text{o}}}(x)$ given to interactions in each proximity class are in the ratio 4.9:1.35:1.0 (short:medium:long). This is chosen so that energy is evenly distributed among the three proximity classes, as was also the case for the prediction Hamiltonian. The dimensionless constant $a_{\text{G}\bar{\text{o}}}$ is chosen so that Equation (2) is satisfied, which ensures that the energy difference between the PDB structure and a completely unfolded structure is the same for the Gō model and the prediction Hamiltonian for protein 1r69.

Acknowledgments

We wish to thank Hans Frauenfelder for helpful discussions. We thank Mike Prentiss for preparing the multiple sequence alignment. Computations were carried out principally at the National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign. This work was supported by NIH Grant No. PHS 2 R01 GM44557-10.

References

1. C. B. Anfinsen, E. Haber, M. Sela, and F. White, Jr., *Proc. Natl. Acad. Sci. USA* **47,** 1309 (1961).
2. C. B. Anfinsen, *Science* **181,** 223 (1973).
3. J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84,** 7524 (1987).
4. P. E. Leopold, M. Montal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **89,** 8721 (1992).
5. J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Ann. Rev. Phys. Chem.* **48,** 545 (1997).
6. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins* **21,** 167 (1995).
7. L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.* **254,** 260 (1995).
8. G. S. Huang and T. G. Oas, *Proc. Natl. Acad. Sci. USA* **92,** 6878 (1995).
9. J. Sabelko, J. Ervin, and M. Grubele, *Proc. Natl. Acad. Sci. USA* **96,** 6031 (1999).
10. B. A. Shoemaker and P. G. Wolynes, *J. Mol. Biol.* **287,** 657 (1999).
11. B. A. Shoemaker, J. Wang, and P. G. Wolynes, *J. Mol. Biol.* **287,** 675 (1999).
12. J. J. Portman, S. Takada, and P. G. Wolynes, *Phys. Rev. Lett.* **81,** 5237 (1998).
13. E. Alm and D. Baker, *Proc. Natl. Acad. Sci. USA* **96,** 11305 (1999).
14. V. Muñoz and W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **96,** 11311 (1999).
15. O. V. Galzitskaya and A. V. Finkelstein, *Proc. Natl. Acad. Sci. USA* **96,** 11299 (1999).
16. J. D. Bryngelson, *J. Chem. Phys.* **100,** 6038 (1994).
17. V. S. Pande, A. Y. Grosberg, and T. Tanaka, *J. Chem. Phys.* **103,** 9482 (1995).
18. A. Šali and T. L. Blundell, *J. Mol. Biol.* **212,** 403 (1990).
19. J. Bowie, R. Luthy, and D. Eisenberg, *Science* **253,** 164 (1991).
20. A. V. Finkelstein and B. Reva, *Nature* **351,** 497 (1991).
21. A. Godzik, A. Kolinski, and J. Skolnick, *J. Mol. Biol.* **227,** 227 (1992).
22. R. A. Goldstein, Z. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **89,** 4918 (1992).
23. D. T. Jones, W. R. Taylor, and J. M. Thornton, *Nature* **358,** 86 (1992).
24. V. N. Maiorov and G. M. Crippen, *J. Mol. Biol.* **227,** 876 (1992).
25. C. Keasar, R. Elber, and J. Skolnick, *Fold. Des.* **2,** 247 (1997).
26. C. Keasar, D. Tobi, R. Elber, and J. Skolnick, *Proc. Natl. Acad. Sci. USA* **95,** 5880 (1998).
27. A. Y. Badretinov and A. V. Finkelstein, *J. Comp. Biol.* **5,** 369 (1998).
28. A. V. Finkelstein, *Fold. Des.* **2,** 115 (1997).
29. K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **95,** 2932 (1998).
30. Special issue of *Proteins*, Volume 37(S3) (1997).
31. Special issue of *Proteins*, Volume 29(S1) (1999).
32. C. Hardin, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Associative Memory Hamiltonians for Structure Prediction Without Homology: Alpha-Helical Proteins," *Proc. Natl. Acad. Sci. USA* **97,** 14235 (2000).
33. M. S. Friedrichs and P. G. Wolynes, *Science* **246,** 371 (1989).
34. M. S. Friedrichs, R. A. Goldstein, and P. G. Wolynes, *J. Mol. Biol.* **222,** 1013 (1991).
35. C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, *Proteins* **34,** 281 (1999).
36. K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes,

Prot. Sci. **5,** 1043 (1996).
37. R. A. Goldstein, Z. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **89,** 9029 (1992).
38. J. G. Saven and P. G. Wolynes, *J. Mol. Biol.* **257,** 199 (1996).
39. N. Gō, *Ann. Rev. Biophys. Bioeng.* **12,** 183 (1983).
40. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61,** 2635 (1988).
41. N. D. Socci and J. N. Onuchic, *J. Chem. Phys.* **103,** 4732 (1995).
42. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **63,** 1195 (1989).
43. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comp. Chem.* **13,** 1011 (1992).
44. M. Sasai and P. G. Wolynes, *Phys. Rev. Lett.* **65,** 2740 (1990).
45. M. Sasai and P. G. Wolynes, *Phys. Rev. A* **46,** 7979 (1992).
46. S. S. Plotkin, J. Wang, and P. G. Wolynes, *Phys. Rev. E* **53,** 6271 (1996).
47. S. S. Plotkin, J. Wang, and P. G. Wolynes, *J. Chem. Phys.* **106,** 2932 (1997).
48. S. Takada and P. G. Wolynes, *Phys. Rev. E* **55,** 4562 (1997).
49. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215,** 403 (1990).
50. J. Ryckaert, G. Ciccotti, and H. Berendsen, *J. Comp. Phys.* **23,** 327 (1977).
51. P. A. Karplus, *Prot. Sci.* **5,** 1406 (1996).
52. A. Kolinski, J. Skolnick, A. Godzik, and W.-P. Hu, *Proteins* **27,** 290 (1997).

Received June 30, 2000; accepted for publication November 17, 2000

Michael P. Eastwood *Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093 (eastwood@chem.ucsd.edu).* Dr. Eastwood is a Postdoctoral Research Chemist at the University of California at San Diego, where he has investigated aspects of protein folding. He received a B.A. degree in chemistry from Oxford University in 1994, and a D.Phil. degree from their Physical and Theoretical Chemistry Department in 1998. Before joining the Department of Chemistry and Biochemistry at UC San Diego, he was a Postdoctoral Research Associate in the Department of Chemistry at the University of Illinois at Urbana–Champaign.

Corey Hardin *Department of Chemistry, University of Illinois, 600 S. Mathews Avenue, Urbana, Illinois 61801 (chardin@verdandi.scs.uiuc).* Mr. Hardin received his B.S. degree in chemistry and his B.A. degree in biology from Drake University. He is currently pursuing both a Ph.D. in biophysics and an M.D. degree as part of the Medical Scholars Program at the University of Illinois at Urbana–Champaign. He joined the laboratory of Dr. Wolynes in 1995.

Zaida Luthey-Schulten *Department of Chemistry, University of Illinois, 600 S. Mathews Avenue, Urbana, Illinois 61801 (zan@uiuc.edu).* Dr. Schulten is a Professor of Chemistry and Biophysics at the University of Illinois at Urbana–Champaign. Her main research interests involve the theoretical and computational study of biomolecules and the statistical mechanics of the genome. She received a B.S. degree in chemistry from the University of Southern California in 1969, an M.S. degree in chemistry from Harvard University in 1971, and a Ph.D. degree in applied mathematics from Harvard University in 1974. From 1975 to 1980, she was a Research Fellow at the Max-Planck Institute for Biophysical Chemistry in Goettingen, and from 1980 to 1985 a Research Fellow in the Department of Theoretical Physics at the Technical University of Munich.

Peter G. Wolynes *Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093 (pwolynes@ucsd.edu).* Dr. Wolynes has broad research interests in chemical physics, statistical mechanics, and biomolecular physics. He has made crucial contributions to understanding localization phenomena, glassy behavior, and electron transfer reactions. He pioneered the application of ideas from the statistical mechanics of spin glasses to the protein folding problem, thereby opening up a very active area of current research. He earned a bachelor's degree from Indiana University in 1971 and a master's degree from Harvard University in 1972, both in chemistry. He received a doctorate in chemical physics at Harvard in 1976. He was a faculty member at Harvard before joining the chemistry department of the University of Illinois at Urbana–Champaign in 1980. He was made a full professor of chemistry in 1983, of physics in 1985, and of biophysics in 1989. In 2000, he joined the faculty of the University of California at San Diego, where he holds the Francis H. C. Crick chair. He was elected a member of the National Academy of Sciences in 1991. Dr. Wolynes has received many other honors, including the Peter Debye Award in 1999.