The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models

Jon M Sorenson[1] and Teresa Head-Gordon[2]

Background: Recent studies have proposed various sources for the origin of cooperativity in simplified protein folding models. Important contributions to cooperativity that have been discussed include backbone hydrogen bonding, sidechain packing and hydrophobic interactions. Related work has also focused on which interactions are responsible for making the free energy of the native structure a pronounced global minimum in the free energy landscape. In addition, two-flavor bead models have been found to exhibit poor folding cooperativity and often lack unique native structures. We propose a simple multibody description of hydration with expectations that it might modify the free energy surface in such a way as to increase the cooperativity of folding and improve the performance of two-flavor models.

Results: We study the thermodynamics and kinetics of folding for designed 36-mer sequences on a cubic lattice using both our solvation model and the corresponding model without solvation terms. Degeneracies of the native states are studied by enumerating the maximally compact states. The histogram Monte Carlo method is used to obtain folding temperatures, densities of states and heat capacity curves. Folding kinetics are examined by accumulating mean first-passage times versus temperature. Sequences in the proposed solvation model are found to have more unique ground states, fold faster and fold with more cooperativity than sequences in the nonsolvation model.

Conclusions: We find that the addition of a multibody description of solvation can improve the poor performance of two-flavor lattice models and provide an additional source for more cooperative folding. Our results suggest that a better description of solvation will be important for future theoretical protein folding studies.

Addresses: [1]Department of Chemistry, University of California, Berkeley, CA 94720, USA. [2]Physical Biosciences and Life Sciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

Correspondence: Teresa Head-Gordon
E-mail: TLHead-Gordon@lbl.gov

Key words: hydration, hydrophobic interaction, lattice models, protein folding, two-state kinetics

Received: **18 August 1998**
Revisions requested: **29 September 1998**
Revisions received: **21 October 1998**
Accepted: **16 November 1998**

Published: **11 December 1998**
http://biomednet.com/elecref/1359027800300523

Folding & Design 11 December 1998, **3**:523–534

Introduction

What are the forces that guide a polypeptide chain to fold both quickly and correctly to its native state? In the past decade, theoretical [1–3] and experimental studies [3–6] have made important progress towards piecing together this complex story. It is now appreciated that the resolution of Levinthal's well-known paradox lies in the existence of biases in the free energy landscape [2,7,8] guiding the unfolded chain to the native state. These biases are often described as creating a funnel [9,10] in the energy landscape. The theoretical backing for this picture has been developed from extensive studies and is now well understood from the standpoint of simplified lattice and off-lattice models. We are now at a position where we can change the lattice model description in a simple and desirably realistic manner and investigate the effects of such a perturbation on the resulting thermodynamics and kinetics. In this paper, we investigate the effect of adding features of hydration forces to a simple lattice model of protein folding.

Nearly all lattice and many off-lattice studies designed to investigate protein folding have concentrated on residue–residue interactions on the protein chain and do not include explicit residue–water and water–water interactions. A full study of protein folding with explicit solvent and realistic atomistic potentials will probably remain infeasible for years to come. Some notable exceptions do exist [11–13], primarily focusing on unfolding studies; however, the multiple studies desired for good statistics and eliciting general folding principles are still only manageable with simplified models [14]. Solvation forces are not completely left out in simplified models; their effects are partially accounted for in the residue–residue interactions.

We chose lattice models for our first study of adding more realistic solvation forces to simplified protein folding models because they have a long history [15] and have now been well characterized. The energy terms in lattice models are typically specified only for residues in

nearest-neighbor contact on a cubic lattice. The energy of the chain is given by the expression:

$$E = \sum_{i<j}^{N} B_{ij}\Delta_{ij} \quad (1)$$

where the double sum is over the N residues of the chain, B_{ij} is the contact energy between residues i and j, and Δ_{ij} is 1 if residues i and j are nearest neighbors and not contiguous on the chain and 0 otherwise. The contact energy terms are taken from statistical studies of the distribution of residue–residue contacts in real proteins [16–18], are drawn randomly from a statistical distribution [19], or are motivated from physical pictures [20]. It has been argued that statistical potentials derived from protein crystal structures in the PDB suffer from the fact that they are neither potentials of mean force [21] nor the correct potentials necessary to recover the desired native states from which they were drawn [22]. A further criticism of the popular Miyazawa–Jernigan (MJ) contact potentials [16] used in many lattice model studies of protein folding [23–28] is that contacts made between hydrophilic residues are predicted to be as much as twice as favorable as contacts between hydrophobic residues. This prediction is counter to experimental, simulation and theoretical studies [29]. The result is that sequences designed with the MJ parameters favor a core of hydrophilic residues [25], in sharp contrast to the hydrophobic core known to be a key element of protein structure [30]. Recently derived contact potentials [18] that account better for the effects of chain connectivity on the distribution of residue–residue contacts seem to correct this aberrant prediction; however, the use of pairwise contact potentials neglects two prominent features of hydration forces: their many-body nature and their potentially long-range effects.

Recent simulations have made it increasingly apparent that hydrophobic forces are strongly nonpairwise additive [31–34]. The free energy of forming a cluster of hydrophobic solutes can differ by over 50% from the predicted free energy based on the assumption of pairwise additivity [31]. The multibody nature of free energy potentials has been repeatedly emphasized [35,36], but most studies to date have employed strictly pair potentials. This is somewhat justified in that we hardly know the exact nature of the pairwise interaction between residues on the protein chain, let alone their multibody interactions. One aim of this present study is to examine the effect of adding a simple multibody potential on the previous conclusions drawn from studies of lattice models with pairwise-additive energies. The simple model we propose below also incorporates some of the long-range nature of hydration forces. The long-range forces we investigate are intimately tied to the multibody nature of our description; this is similar to the long-range solvent effects found in simulations [34,37] arising from solvent-mediated solute–solute interactions [38,39].

A further aspect of hydration forces that has been noted several times [3,40], but not addressed in the context of lattice models, is their temperature dependence [41]. The contact potentials used in lattice models are free energies, and as such they will depend on temperature. Molecular dynamics simulations have indicated that the contact minimum in the hydrophobic potential of mean force deepens with increasing temperature [42]. This same study has shown that the free energy of the solvent-separated minimum is relatively temperature independent. These results indicate that incorporating temperature dependence into the potential parameters of a lattice model might involve not only a temperature-dependent well-depth but also a temperature-dependent length scale of interaction. Because of this added and only partly understood complexity, we do not address this aspect here and instead leave the issue of incorporating the temperature dependence of hydration forces for future studies.

There have been several previous attempts to add various features of solvation to lattice models of protein folding. Perhaps the approach most similar to our present study is the recent lattice model study by Hao and Scheraga [43] where solvent-accessible surface area terms were added to the free energy of the chain. Their energy function has the form:

$$E = \sum_{i<j}^{N} B_{ij}\Delta_{ij} + \sum_{i}^{N} u_i |s_i - s_i^0| \quad (2)$$

where the new term on the right represents a free energy that is dependent on the solvent-accessible surface area of each monomer. The solvation state s_i counts the number of monomers neighboring monomer i, s_i^0 is the optimal number of neighbors for monomer i and $u_i > 0$ biases each monomer towards its optimum solvation state. In their study, the preferred solvation states were selected to represent a variety of hydrophilic and hydrophobic residues, and the unknown parameters u_i and B_{ij} were optimized to produce a good foldable model. They found that the inclusion of the solvation terms produced a model that folded more quickly to the native state with less chance of being stuck in energetically low-lying misfolded states compared to the same model with $u_i = 0$.

Other solvation-motivated lattice model studies have studied the effect of making contacts more repulsive [27,40,44]. They have found that more repulsive terms in the energy function force the protein to fold in a more all-or-none transition, collapsing and folding to the native state in a concerted manner. Misfolded compact states were also found to be destabilized, in agreement with the conclusions of Hao and Scheraga [43]. Onuchic *et al.* [45] have added a solvent-accessible surface area term to only the core monomer of the native structure of the chain in an attempt to model the presence of denaturant. They found that by increasingly favoring desolvation of the core

Table 1

Foldable sequences in the solvation model.

Sequence	Hydrophobic/polar composition
1	HHPHPPHHHHPPPPHHPHPHHHHHPPPPHPHPPHPP
3	HPHPPHHPPHPHPHPPHHHPPHHHPPPPHPHPHPHH
6	HHPHHPPHPPPHPPHPHHHHPPPPHHPPHPHHHHPP
20	PPHPHHPPHHPPPHPPPHPHHPPHHHHPHPHPHHPH
26	PPHPHPPHPHHPHPPHPHPPHHHPHPPPHHHHPPHH
29	HHHPPHHPHHHHPHHPPHPHHPHPPHPPPHHPPPPP
30	HPHPHHPPHHPPHHHHPHPHPPPPHPHHPHPPPHHP
35	HHHPHPPPPHHPHPHHPPPHPHPHHHHHHPPHPPPP

monomer, the barrier to folding to the native state was progressively increased.

A further motivation of our current study was to attempt to 'rescue' the status of two-flavor lattice models. Two-flavor models are those in which residues are allowed to be only one of two types, traditionally hydrophobic (H) or polar (P). From a computational and theoretical point of view, they are some of the simplest models that display important features of real proteins — unique ground states and folding to these ground states, overcoming a Levinthalian search. Whereas the early HP model proposed by Lau and Dill [20] has been criticized for its lack of nondegenerate ground states and an energy gap [23,46,47], other two-flavor models do seem to possess unique ground states and foldable sequences [44,48]; however, two-flavor models have been routinely criticized for not possessing the proper energy gap or T_f/T_g ratio predicted for real proteins [1,49]. It was hoped in our current study that by making the interaction between two flavors of monomers more complex, we might regain some of the desirable features present in multiflavor models such as more nondegenerate ground states and a larger energy gap [23].

Results and discussion

Our proposed description of solvation in a lattice model and our methods for simulation and sequence selection are detailed in the Materials and methods section. We found eight foldable sequences for study with the solvation model; their sequence compositions are listed in Table 1. Some of the folding properties of these sequences and properties of their native structures are given in Table 2. The folding temperature listed in this table is the thermodynamic folding temperature T_f^{SO}, defined as the temperature where the relative population of the native state is 50% [44,48]. The relative population of the native state, $P_n(T)$, is defined as:

$$P_n(T) = \frac{e^{-E_{nat}/T}}{\sum_E \Omega(E) e^{-E/T}} \quad (3)$$

Table 2

Properties of foldable sequences studied with the solvation model.

	Solvation		Nonsolvation		
Sequence	E_{min}	T_f^{SO}	E_{min}	T_f^{SO}	RCO (%)
1	−35.17	0.46	−36	0.42	30.1
3	−34.50	0.39	−36	0.30	31.1
6	−36.00	0.50	−34	0.39	27.6
20	−36.50	0.45	−36	na	26.3
26	−35.92	0.43	−36	na	28.9
29	−35.83	0.34	−36	0.31	28.3
30	−36.00	0.36	−36	na	29.0
35	−35.17	0.39	−36	na	32.1

Sequence numbers correspond to the sequences in Table 1. E_{min} is the native state energy, T_f^{SO} is the thermodynamic folding temperature and RCO is the relative contact order for each structure [51].

where $\Omega(E)$ is the density of states for energy E. $\Omega(E)$ was calculated with the histogram Monte Carlo method [44,50]. The accuracy of the calculated $\Omega(E)$ was confirmed by calculating E versus T and C_v versus T curves and comparing these to the values found by simple averaging from Monte Carlo simulations at various temperatures.

Also listed in Table 2 is the relative contact order (RCO) for each native state structure — a measure of how many local versus nonlocal interactions are present for a given structure [51]. Such considerations have been proposed to correlate with the folding kinetics of simplified models [26,52] and real proteins [51]. In this work, we found that the RCO has utility as a simple topological descriptor for aiding structure selection (see the Materials and methods).

Each of these eight sequences was also studied with the nonsolvation model. Sequences 20, 26, 30 and 35 were found to have degenerate ground states without solvation and consequently could not be used for folding studies. Table 2 shows that the folding temperature is consistently higher for the sequences under the solvation model than with the nonsolvation model. To validate studying the same sequence in both models, we verified that the four sequences studied with and without solvation were optimally designed sequences in both models. This confirms that the same sequences would have been arrived at had we followed the six selection steps as in the Materials and methods section but instead used the nonsolvation model in the design and enumeration steps. Our two models are energetically similar enough to allow the same sequence to fold to the same native structure in both models.

Degeneracy

It was initially hoped that the introduction of the solvation model would lift degeneracy and produce more unique ground states than the corresponding two-flavor nonsolvation model. The observation above that of the eight

foldable sequences only four had nondegenerate native states in the nonsolvation model indicates that this lifting of degeneracy was partially achieved.

This conclusion rests on sequences that passed all six selection steps (see the Materials and methods). To show its validity for more sequences, we took the structures that did not produce foldable sequences in the solvation model and followed the first four selection steps in the nonsolvation model to study the degeneracy of the resulting native states. Every sequence/structure examined in this way had a degenerate ground state in the nonsolvation model. Four of these structures produced sequences with nondegenerate ground states in the solvation model (sequences that failed selection step five). We see again that a solvation component has partly lifted the degeneracy problem that plagues two-flavor models [46,47].

This observation is not surprising when we recast our model as a multiflavor model. Table 3 shows how the contact energies given by Equations 5, 6, 10 and 11 can be reformulated in a multiflavor fashion. It is known that multiflavor models have more sequences with nondegenerate ground states [1,47]. The difference between our solvation model and a true multiflavor model is that the flavors of each monomer are environment dependent and are able to change over the course of the simulation. In effect, the protein is given some freedom to redesign itself as it folds.

Table 3

Representation of the solvation model as a multiflavor model.

	H0	H1	H2	H3	P0	P1	P2
H0	−1	−1	−1	−1	0	0.25	0.5
H1	−1	−1	−1	−1	0.167	0.417	0.667
H2	−1	−1	−1	−1	0.333	0.583	0.833
H3	−1	−1	−1	−1	0.5	0.75	1.0
P0	0	0.167	0.333	0.5	1	0.5	0
P1	0.25	0.417	0.583	0.75	0.5	0	−0.5
P2	0.5	0.667	0.833	1.0	0	−0.5	−1

The number after the residue type is the solvation state s_i (Equation 9). Flavors H4, H5, P3, P4 and P5 are not shown because by Equation 8 they are equivalent to flavors with lower solvation states. H0 and P0 do not actually occur in simulation because energies are present only between residues in contact and the presence of a single contact would necessarily raise the solvation state above zero.

Kinetics

The folding kinetics were explored for each sequence by varying temperature and collecting statistics on mean first-passage times for folding to a collapsed state (≥ 36 contacts), folding to a compact state (40 contacts) and folding to the native state. Following previous kinetics studies [44,48,53], if in a particular run a sequence was found not to fold within the maximum simulation time of 10^9 steps, we averaged the maximum time into the mean. As such, the reported times are all lower bounds to the true mean first-passage times; the associated standard deviations should give a sense of how much this averaging has affected the reported times. Table 4 shows the temperature dependence of the mean first-passage times for folding for the eight sequences. The fastest folding times and temperature of fastest folding are shown in Table 5. Although the thermodynamic folding temperatures listed in Table 2 vary by up to 40%, the temperature of fastest folding appears more sequence independent; in the language of past studies, it appears to be a self-averaging property [54]. Each sequence folds faster under the solvation model, although the extent of this varies from a factor of 5.3 for sequence 3 to near equality for sequence 6. On the basis of the four sequences studied in both models, it appears that the solvation model has modified the topology of the energy surface in such a way as to better guide the search for the native state.

T_f/T_g

Of particular interest for comparing minimalist protein folding models with experiment is the ratio of the folding temperature to the glass transition temperature, T_f/T_g [49]. This ratio gives a simple characterization of the steepness of the protein folding funnel for theoretical and real

Table 4

Mean first-passage times for folding versus temperature for the foldable sequences in the solvation model.

	Temperature								
Sequence	0.45	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.1
1	9.7(3)	8.3(5)	7.0(7)	5.6(5)	7.1(5)	7.2(7)	9.6(3)	9.6(3)	10.0(0)
3	8.8(7)	5.8(7)	2.5(6)	1.3(2)	3.0(5)	4.9(7)	9.5(3)	9.6(5)	10.0(0)
6	6.8(7)	3.7(6)	1.3(2)	1.0(1)	1.1(1)	2.2(3)	7.0(6)	10.0(0)	10.0(0)
20	8.9(5)	6.2(6)	3.1(4)	1.8(3)	1.5(1)	1.8(3)	5.0(7)	9.2(4)	10.0(0)
26	7.6(7)	2.2(5)	0.9(2)	0.6(1)	0.7(1)	1.5(3)	6.5(7)	9.7(4)	10.0(0)
29	8.5(7)	5.1(7)	3.0(5)	1.4(2)	1.2(1)	1.3(2)	4.4(7)	9.0(6)	10.0(0)
30	9.7(4)	9.3(5)	5.0(7)	3.7(5)	1.8(3)	3.9(7)	8.4(5)	9.6(4)	10.0(0)
35	9.1(6)	7.7(7)	5.1(7)	1.9(3)	2.3(4)	4.1(7)	8.0(7)	9.5(6)	10.0(0)

Times are in 10^8 Monte Carlo moves. The uncertainty in the last digit is given in parentheses. The temperature is in units of E/k_B.

Table 5

Fastest folding times for foldable sequences.

	Solvation		Nonsolvation	
Sequence	τ_{MFPT}	T_f^{kin}	τ_{MFPT}	T_f^{kin}
1	5.2(8)	0.63	6.2(8)	0.61
3	1.3(2)	0.60	6.9(6)	0.62
6	1.0(1)	0.60	1.0(3)	0.63
20	0.9(2)	0.66	na	–
26	0.5(1)	0.61	na	–
29	1.0(1)	0.67	1.2(2)	0.61
30	1.5(4)	0.63	na	–
35	1.9(3)	0.60	na	–

Times are measured in 10^8 Monte Carlo moves. The uncertainty in the last digit is given in parentheses. T_f^{kin} is the kinetic folding temperature, corresponding to the temperature of fastest folding.

Table 6

T_f/T_g for the foldable sequences.

	Solvation			Nonsolvation		
Sequence	T_g	T_f^{SO}/T_g	T_f^{HS}/T_g	T_g	T_f^{SO}/T_g	T_f^{HS}/T_g
1	0.50	0.92	1.15	0.55	0.76	0.87
3	0.51	0.75	1.09	0.51	0.59	0.67
6	0.49	1.02	1.30	0.50	0.78	0.99
20	0.51	0.87	1.23	–	–	–
26	0.49	0.88	1.30	–	–	–
29	0.53	0.64	1.02	0.51	0.61	0.79
30	0.55	0.65	1.00	–	–	–
35	0.53	0.74	1.05	–	–	–

T_g is the kinetic glass temperature, T_f^{SO} is the thermodynamic folding temperature found from $P_n(T_f^{SO}) = 0.5$, and T_f^{HS} is the folding temperature found from the tangent construction with the density of states.

proteins. It has been found that T_f/T_g is about 1.3 for two-flavor models whereas the experimental ratio is expected to be approximately 1.6 from comparison to predictions of the random energy model [10,49]. Lattice models with more flavors have been found to possess a T_f/T_g ratio closer to that predicted for real proteins [49,55].

There are several possible definitions of the folding and glass transition temperatures for use in calculating this ratio. For the folding temperature, we used the thermodynamic folding temperature given above (T_f^{SO}) and the folding temperature where the free energy of the native state is equal to the free energy minimum of the unfolded states [43,56,57]. This second definition of a thermodynamic folding temperature, denoted as T_f^{HS}, is similar to the previously given definition, but it produces a different temperature. It can either be found from calculating the free energy:

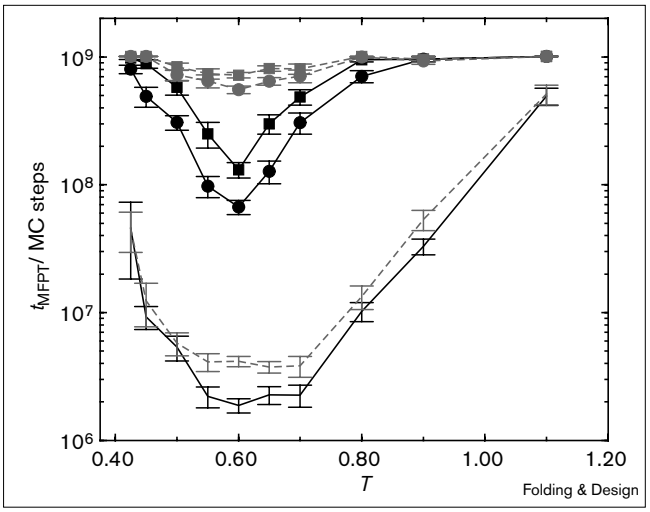
$$F(E)=E-TS(E) \quad (4)$$

using the histogram method and finding the temperature that equates the free energies of the native state and the minimum free energy of the unfolded states, or from a tangent construction using the density of states [43]. The form of $F(E)$ for sequences in the solvation and nonsolvation model is shown in the next section. For the glass transition temperature we used a kinetic definition proposed by Socci and Onuchic [48]. The kinetic T_g is defined as the temperature below the folding temperature at which the folding time is half way between the maximum simulation time, τ_{max}, and the fastest folding time for that sequence (given in Table 5).

Table 6 shows the result of this calculation for sequences in the solvation and nonsolvation models. Depending on the definition of folding temperature used, we see that the folding temperature is either below or above the glass transition temperature for our model. Good folding sequences should have folding temperatures above the glass temperature [48]. Seven of our eight sequences do not meet this requirement with the above definition of a kinetic glass temperature and T_f^{SO} for the folding temperature, but all pass with the second definition of folding temperature. In contrast, the four sequences in the nonsolvation model are bad folders using either definition. We expect that the poor values of T_f^{SO}/T_g arise from the unfolding and folding interactions chosen for this study (Equations 10 and 11). These matrices are not optimized, and we would anticipate optimized interactions to produce better ratios. In all cases, the T_f/T_g ratio is higher for sequences under the solvation model; this indicates that the addition of solvation terms has shaped a better free energy surface for folding.

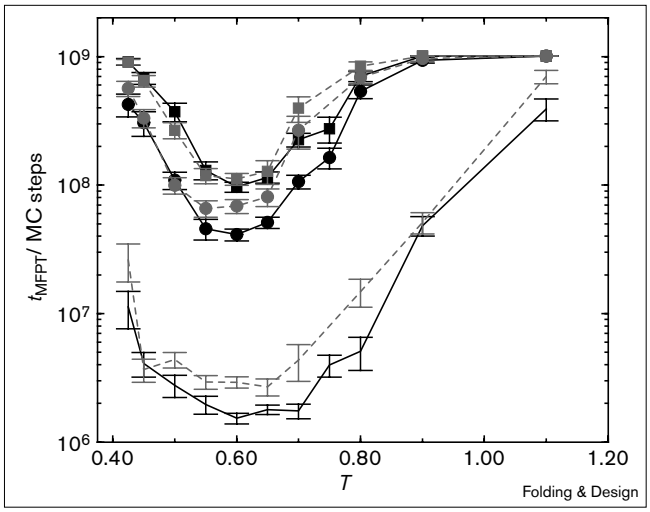
Since the T_f/T_g ratios are noticeably dependent on the definition of a folding temperature, we should make a few remarks about the various definitions of T_f. The concept of a folding transition is borrowed from the theory of phase transitions in bulk systems, and as such is not a precise description for a finite-sized system such as our 36-mer polymer chain; thus, several definitions have been put forward for use. Perhaps the best definition of T_f would be the one that is most similar to the definition used in experimental thermodynamic studies — defining T_f as the temperature of the maximum in the heat capacity versus temperature curve [58]. While this definition might be problematic for some protein folding models that exhibit a strong collapse transition as well as a folding transition [59,60], we found this not to be a problem in this work. We have evaluated this temperature for our sequences and found that it defines a folding temperature that is closer to T_f^{HS} than T_f^{SO} (JM Sorenson, T Head-Gordon, unpublished data). Examining the values of T_f^{HS}/T_g, Table 6 shows that sequences clearly have a more favorable ratio under the solvation model. This definition also appears more useful for discriminating between the solvation and nonsolvation models because it correlates better with the folding speed of the sequence, as seen in

Figure 1



Mean first-passage times for collapse (no symbols), compaction (circles) and folding (squares) for sequence 3 in the solvation model (solid lines) and nonsolvation model (dashed lines).

Figure 2



Mean first-passage times for collapse (no symbols), compaction (circles) and folding (squares) for sequence 6 in the solvation model (solid lines) and nonsolvation model (dashed lines).

Table 5. We show in the next section that T_f^{HS} is closely related to the degree of two-state kinetics present in the folding transition, and the higher T_f^{HS}/T_g ratios for the solvation model indicate a more cooperative folding process.

Part of determining T_f/T_g also depends on how we define the glass temperature, T_g. (It has been recently suggested that a true glass transition does not exist for lattice models [61]; as such, the definition of a glass transition temperature becomes more complicated.) In previous work [48], it has been shown that T_g has some dependence on τ_{max}. In our present work, we might expect a much greater dependence on τ_{max} since, unlike the previous studies, the fastest folding times of many of our sequences are less than an order of magnitude different from τ_{max}. This is especially a concern for calculating T_g for the slower folding sequences in the nonsolvation model. Because of this dependence on τ_{max}, the reported kinetic T_g is an upper bound on the true kinetic glass transition temperature [48]. The extent of this was tested for sequence 6 by running simulations at low temperatures for ten times τ_{max} (10^{10} steps). The resulting prediction for the kinetic T_g was shifted to lower temperature by 9% and appears converged. This is similar to the shift found by Socci and Onuchic [48], and would increase our T_f/T_g ratios by about 10%.

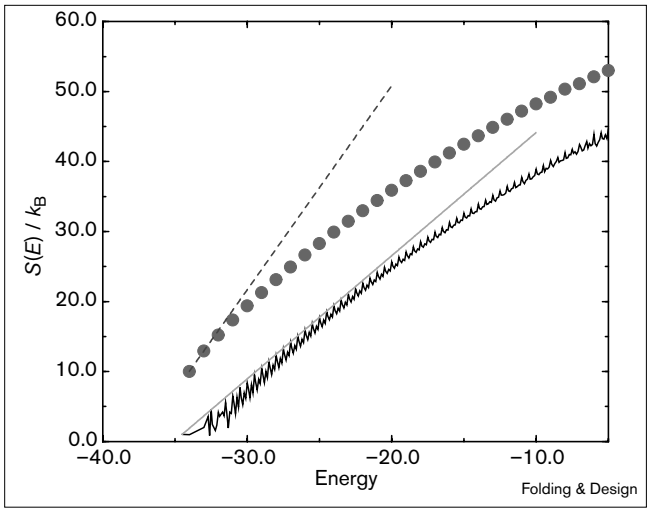
Sequences 3 and 6

The folding kinetics in Table 5 show that sequences fold faster in the model with solvation terms present. To investigate what might be underlying this difference between models, we further examined sequences 3 and 6 in the solvation and nonsolvation models. We chose these sequences to be representative of the trends found in the kinetics; sequence 3 showed the greatest speed-up in folding in the solvation model and sequence 6 showed the least effect. The temperature dependence for the mean first-passage times for these two sequences in the solvation and nonsolvation models are shown in Figures 1 and2. At any given temperature, the collapse, compaction and folding times are faster in the solvation model, with the only exception being at lower temperatures for sequence 6 where the times are equal within the associated uncertainties. The close coincidence of the times for compaction and folding show that folding in both models is an all-or-none transition, similar to that seen in other studies with repulsive potentials [27,40,44]. The convergence of these two times has been observed before in the context of 27-mer folding in the nonsolvation model [44].

Figures 3 and 4 show the density of states computed with the histogram Monte Carlo method for sequences 3 and 6 in the two models. We can readily see that the logarithm of the density of states, $S(E) = \ln\Omega(E)$, is more concave for sequences under the solvation model. This has been noted before as a criterion for two-state kinetics and fast folding [56]. The fine structure of the density of states for the solvation model is not noise; it is a consequence of the distribution of fractional contact energies given in Table 3. For example, integral and half-integral values of the energy are more likely to occur than other fractions.

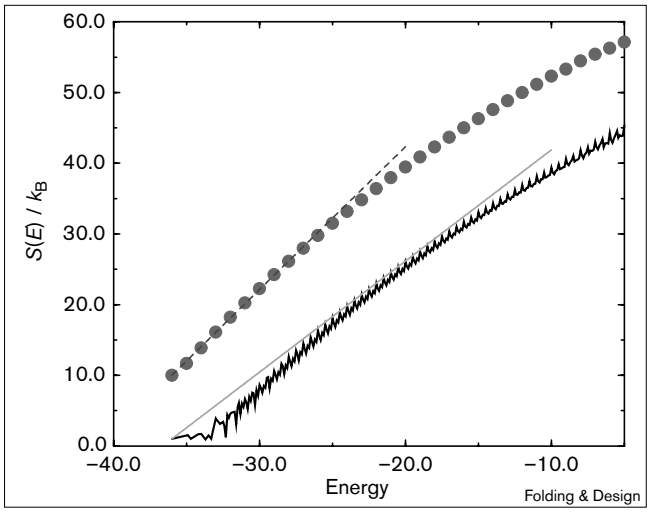
The corresponding $P_n(T)$ versus T curves are shown in Figure 5. This figure shows well the higher thermodynamic folding temperatures found for sequences in the solvation model. The folding curves are shifted to the right by ≈0.1 in the solvation model. (Boltzmann's constant k_B

Figure 3



Logarithm of the density of states for sequence 3 in the solvation model (solid line) versus the nonsolvation model (circles). Values have been offset by one for the solvation model and ten for the nonsolvation model for better clarity. The tangent lines are drawn to illustrate the concavity of $S(E)$; the slopes of the lines are the inverse folding temperature ($1/T_f^{HS}$).

Figure 4



Logarithm of the density of states for sequence 6 in the solvation model (solid line) versus the nonsolvation model (circles). See the legend to Figure 3 for details.

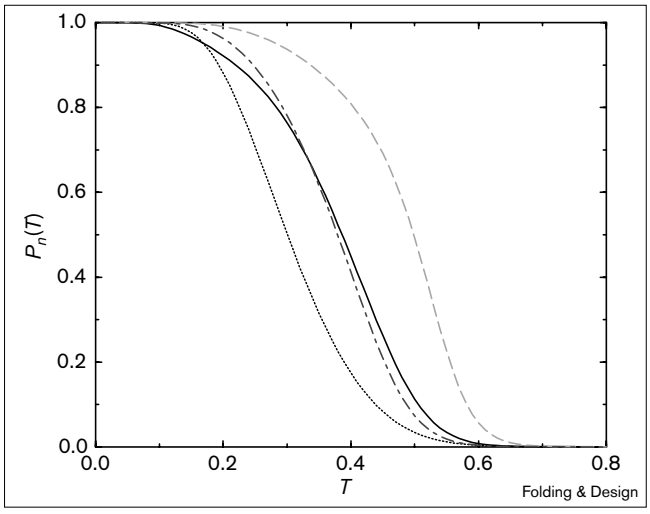
was set to 1 in our simulations so the unit of temperature is equivalent to the unit of energy.)

The origin of the observed differences in kinetics becomes clearer when we examine the free energy of folding versus energy, $F(E)$, at the temperatures where the free energy of the folded state equals that of the minimum free energy of the unfolded states. Figure 6 shows this comparison for sequence 3 in the solvation and nonsolvation models. The first set of curves corresponds to $T = 0.57$, where the free energy of the folded state is equal to the minimum free energy of the unfolded states in the solvation model. The second set of curves are for $T = 0.345$, which is the equivalent temperature in the nonsolvation model. Figure 7 is the corresponding figure for sequence 6. The scale for the free energies is a relative scale; for comparison, the curves shown here were offset to make $F(-5) = 0$.

The free energy curves are dramatically different from one another in the solvation and nonsolvation models. In particular, the curves in the solvation model look similar to curves in previous studies that found good two-state kinetics [43,56,57]. The corresponding curves for the nonsolvation model barely exhibit two minima, a prerequisite for two-state kinetics. This is equivalent to the observation above that $S(E)$ is not very concave in the nonsolvation model [57]. These plots give a graphical illustration of the T_f/T_g ratios summarized in Table 6. The extremely low value of T_f^{HS}/T_g for sequence 3 in the nonsolvation model helps explain why its kinetics are so slow. We see here that this arises from a small value of T_f^{HS} because of the lack of a well-defined cooperative transition. Sequence 6 in the nonsolvation model shows some evidence for two-state kinetics and consequently has a more favorable T_f^{HS}/T_g ratio in Table 6.

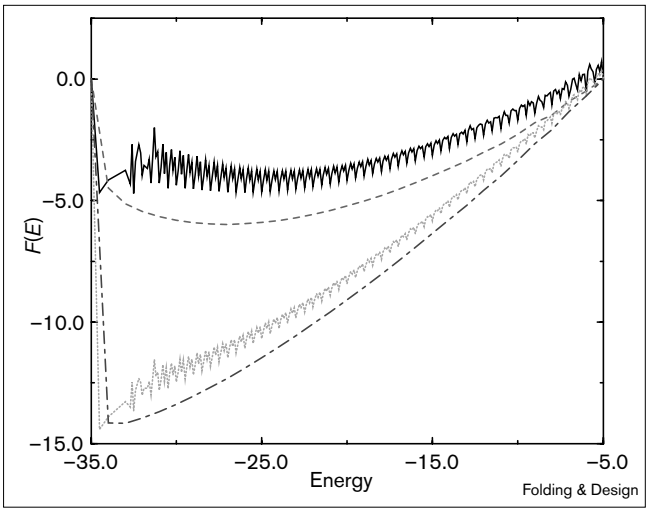
A close examination of the free energy curve in Figure 6 for $T = 0.57$ and sequence 3 in the solvation model also

Figure 5



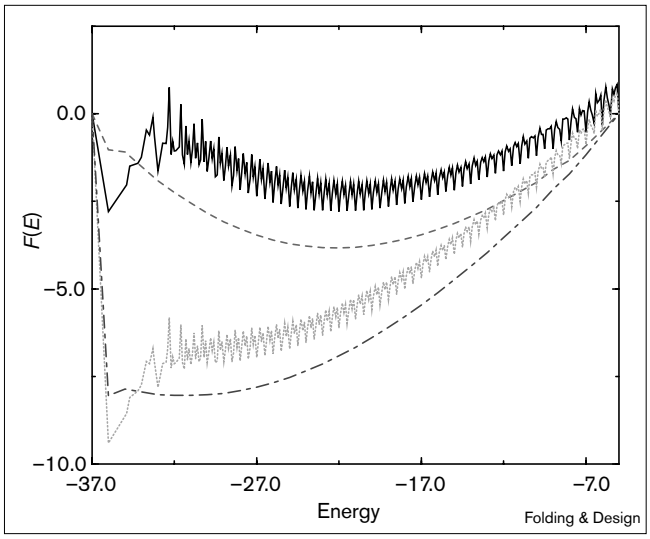
Relative population of the native state versus temperature for sequences 3 and 6 in the solvation and nonsolvation models. Sequence 3, solvation model (solid line); sequence 3, nonsolvation model (dotted line); sequence 6, solvation model (dashed line); sequence 6, nonsolvation model (dotted-dashed line).

Figure 6



Free energy versus energy for sequence 3 in the solvation model and nonsolvation model. $T = 0.57$, solvation model (solid line); $T = 0.57$, nonsolvation model (dashed line); $T = 0.345$, solvation model (dotted line); $T = 0.345$, nonsolvation model (dotted-dashed line).

Figure 7



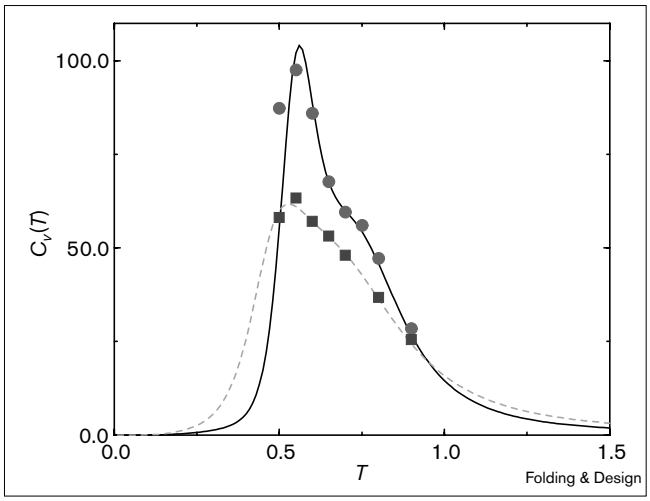
Free energy versus energy for sequence 6 in the solvation model and nonsolvation model. $T = 0.636$, solvation model (solid line); $T = 0.636$, nonsolvation model (dashed line); $T = 0.487$, solvation model (dotted line); $T = 0.487$, nonsolvation model (dotted-dashed line).

offers a possible explanation why the folding temperature ($T_f^{HS} = 0.57$) predicted from this plot is so high whereas the folding temperature predicted from $P_n(T) = 0.5$ is relatively low ($T_f^{SO} = 0.39$). We see the existence of an intermediate at $E = -32.5$ with free energy lower than both the folded and unfolded state at this temperature. It is likely that the presence of this energetically favorable intermediate slows down the search for the ground state and is responsible for the low thermodynamic folding temperature. At lower temperatures, the intermediate is destabilized relative to the ground state as seen in Figure 6. We would expect that destabilizing this intermediate with sequence design might make sequence 3 a faster folding sequence and increase its thermodynamic folding temperature.

A final example of the sharper two-state kinetics in the solvation model can be seen in the heat capacity curves shown in Figure 8 for sequence 6. From the figure, we can see that the heat capacity curve is much sharper and more peaked for the solvation model, characteristic of a first-order-like transition. The heat capacity in both models also exhibits a shoulder at higher temperatures indicating a weak pre-folding collapse transition. It is curious that the folding of sequence 6 in the solvation model is more cooperative, yet the sequence folds at comparable speeds in both models. It would appear that the rougher free energy curve in the solvation model, seen in Figure 7, works against this sequence, possessing many stable traps in the region of the transition ensemble. Similar to sequence 3, sequence design could be used to destabilize these obstacles to faster folding.

These thermodynamic results combined with our kinetic data indicate that the addition of solvation terms to our model has changed the underlying free energy landscape. The differences in the free energy surfaces for the folding of sequence 6 are illustrated in Figure 9. In the solvation model at this temperature, the native state is favorable enough to create a marked depression at the center of the

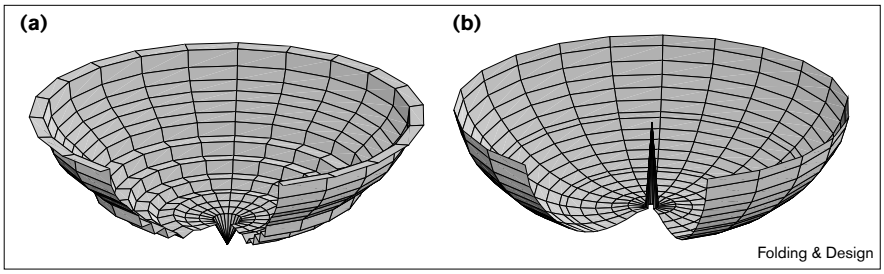
Figure 8



Heat capacity (C_v) versus temperature for sequence 6 in the solvation (solid line) and nonsolvation models (dashed line). The curves were generated with the histogram Monte Carlo method; the points are taken from Monte Carlo simulations at those temperatures.

Figure 9

Free energy versus entropy for sequence 6 in the **(a)** solvation and **(b)** nonsolvation models at $T = 0.5$. The depth of the funnel corresponds to the free energy and the radial coordinate is the entropy. $T = 0.5$ is below T_f^{HS} for sequence 6 in the solvation model but not in the non-solvation model.



funnel. The slow steps of folding will be in searching through the plateau of partially collapsed states just above the native state in free energy. Such an entropic bottleneck in the energy landscape has been described by Dill and Chan as a champagne glass landscape [8]. We see that in the nonsolvation model, the free energy minimum favors a multitude of partially collapsed states, and the native state is decidedly unfavorable. Similar to the free energy curves in Figures 7 and 8, the energy landscape is much smoother in the nonsolvation model. The addition of multibody terms roughens the free energy surface in the solvation model allowing for a higher likelihood of traps that hamper fast folding, but also making possible discrete-state kinetics and a more pronounced global minimum, important for good folding.

The combined evidence points to the conclusion that folding is a more cooperative, two-state process in our solvation model. Higher cooperativity and faster folding have been noted before in the context of models that incorporate multibody structure biases [56,57,62]. Here, we see better two-state folding in a model that incorporates a different form of multibody interaction, one arising from the many-body nature of solvation. Our model gives a simple demonstration of the increased cooperativity associated with multibody interactions in the context of lattice models.

Conclusions

It is widely appreciated that water plays an important role in governing the forces that control protein structure and stability [63,64]. The strong hydration forces that are responsible for hydrophobic attraction and stabilization of a protein's native core are expected to also play an important role in governing how the protein folds quickly to the proper folded state. To address the issue of how solvation forces might influence the kinetics of protein folding, we have examined the addition of simple features of solvation forces to a minimalist model of protein folding. The 36-mer lattice model examined is far from the complex reality of genuine proteins in aqueous solvent, but it possesses some of the essential features of the protein folding problem such as a unique ground state and a large set of possible conformations ($\approx 10^{24}$) [65]. More importantly, we studied a lattice model that is closely related to many previous models with well-characterized kinetics and thermodynamics from over 20 years of studies [1,3,15].

To address the issue of solvation in lattice models we proposed a simple model that incorporates many-body and long-ranged forces while retaining a simple form. In particular, we were able to examine the effect of the breakdown of the pairwise additivity assumption often made when using potentials of mean force. As noted above, recent simulation results have emphasized the acknowledgement of this departure from traditional approximations [31–34].

We have found that the adoption of a model that incorporates some of these features of solvation forces leads to faster folding, unique native states, and a more cooperative, two-state folding transition. Of particular interest is the fact that these properties are not typically found in traditional two-flavor models. We find it encouraging that such a simple model can recover some of these important properties observed in the folding of small lattice proteins. We have observed that the inclusion of multibody hydration forces leads to a more cooperative folding transition, similar to the effect of multibody internal interactions on other protein folding models [57,62]. This lends support to the view that hydration forces are an important source of cooperativity in the protein folding transition [66].

Our conclusion that a simple lattice model of protein folding can be improved with a description of solvation forces motivates further research into the experimental characterization of the solvation forces present between residues on the protein chain [67–69]. Now that the addition of solvation-like terms has been shown to affect lattice protein studies, it will be important to better understand the nature of the solvation terms necessary in more detailed theoretical studies of protein folding.

Materials and methods

The model

The use of lattice models for protein folding has now been described numerous times in the literature [1,3]. For our studies, we modeled 36-mers as self-avoiding walks on a cubic lattice with each residue represented by a single interaction site.

The energy of a particular protein configuration is given by an energy function depending on the contacts between interaction sites, similar to Equation 1. In proposing a new solvation model for lattice model studies, we desired to find a form for the energy which, while remaining relatively simple, captures several aspects of hydration: different free energies of solvation for hydrophobic and polar residues, multibody effects and long-range effects. Our proposed form is that of Equation 1, where the double sum is over the N residues of the chain and Δ_{ij} is the same contact function as defined above. Our approach differs from previous work in the definition of the contact energy matrix:

$$B_{ij} = \left(1-\lambda_{ij}\right)B_{ij}^{u} + \lambda_{ij}B_{ij}^{f} \tag{5}$$

where B_{ij}^{u} represents the contact energy matrix for the unfolded chain and B_{ij}^{f} is the contact energy matrix for the folded chain. We let the energy of contacts interpolate between a matrix of unfolded contact energies and a matrix of folded contact energies, with $0 \le \lambda_{ij} \le 1$, the interpolation parameter, representing the degree to which a particular contact is solvated.

This form is motivated by the breakdown of the pairwise additivity assumption noted above. Many recent simulations have investigated the multibody nature of hydration forces [31,33,34,70]. The bulk of evidence from these studies suggests that the nature of the pairwise hydrophobic interaction changes strongly depending on the surrounding concentration of additional solutes. Other hydration forces such as hydrophilic–hydrophilic interactions would be expected to also display some multibody character. By making the contact energy of a pair of residues dependent on the solvation state of the pair, we can incorporate these kind of effects into a lattice model.

We have many options for how to choose λ_{ij}, the pair solvation state parameter. In this paper we chose the following form:

$$\lambda_{ij} = \frac{\lambda_i + \lambda_j}{2} \tag{6}$$

where $0 \le \lambda_{ij} \le 1$ is a parameter dependent on the solvation state of residue i. An example of a related form, not investigated here, is:

$$\lambda_{ij} = \lambda_i \lambda_j \tag{7}$$

Some of the properties of the form in Equation 7 and the relative merits of the first form versus the second are explored in the Supplementary material (published with this paper on the internet). The individual monomer solvation parameter λ_i is defined as:

$$\lambda_i = \begin{cases} s_i / s_i^0 \text{ if } s_i / s_i^0 < 1 \\ 1 \quad \text{otherwise} \end{cases} \tag{8}$$

where s_i is a measure of the solvent-accessible surface area of monomer i:

$$s_i = \sum_{j}^{N} \Delta_{ij} \tag{9}$$

and s_i^0 is a measure of the optimal solvation state for residue i. We chose $s_i^0 = 3$ for hydrophobic residues and $s_i^0 = 2$ for polar residues, representing the tendency for hydrophobic residues to bury themselves in the protein interior, away from solvent. Our definition of s_i is similar to that used by Hao and Scheraga [43]. It differs in that we do not count residues adjacent on the chain in determining the solvation states of monomers. Thus $0 \le s_i \le 4$ for monomers 2 through N–1 and $0 \le s_i \le 5$ for the two end monomers.

Our current study is a two-flavor model; the type of each residue is restricted to be either hydrophobic (H) or polar (P). For the unfolded chain contact energy matrix we chose:

$$\mathbf{B}^{u} = \begin{array}{c} \\ H \\ P \end{array} \begin{array}{c} \begin{array}{cc} H & P \end{array} \\ \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \tag{10}$$

For the folded contact energy matrix we chose:

$$\mathbf{B}^{f} = \begin{array}{c} \\ H \\ P \end{array} \begin{array}{c} \begin{array}{cc} H & P \end{array} \\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \end{array} \tag{11}$$

The form of the unfolded matrix is motivated by the observation that hydrophobic surfaces are attracted to each other in water, and the interaction between hydrophilic surfaces is more repulsive [29]. Our own experimental work [67–69] is focused on probing the extent to which these observations remain true on the molecular scale of residue–residue interactions. We have recently found evidence from neutron scattering studies that hydrophobic amino acids are indeed positively correlated with each other in solution [69]. The folded matrix in Equation 11 is similar to a form studied in previous theoretical [71,72], design [73,74] and simulation [44,48] studies. Our matrix differs from this previous work in that the average interaction energy is more repulsive; most of the past studies with this form have added a background attractive field [44,48,71,73].

The form of the matrix encourages compact ground states. We wanted to encourage folding to maximally compact states because this allows full enumeration [75] as a check that the ground state is indeed nondegenerate and the minimum energy structure of the compact states. Because of the repulsive interactions both in the unfolded matrix and the folded matrix, all ground states in our model possess some degree of frustration and we are not guaranteed that a maximally compact state gives the lowest energy structure. However, in the comparative folding studies reported above, all of the evaluated sequences fold to compact structures, with no lower energy structure found in the course of numerous long simulations (more than 100 simulations of 10^9 Monte Carlo moves each for each sequence).

Simulations

Our Monte Carlo simulations used the standard move set of one- and two-monomer moves employed in past studies [19,44,76]. The relative probabilities of one-monomer versus two-monomer moves were taken from Sali *et al.* [19]. We used the Metropolis energy criterion for acceptance of a new move [77]. Moves that were rejected because they caused chain overlap were still counted as steps for the purpose of calculating elapsed 'time'. This agrees with the definition used by Socci and Onuchic [48] and Pande *et al.* [74] but differs from that used by Shakhnovich and coworkers [19]. Folding studies were performed by starting with an arbitrary random coil configuration from a high temperature simulation, and allowing the chain to equilibrate for 20×10^6 moves. Chains were studied for an additional 10^9 moves. Energies were binned every 20,000 moves for use with the Monte Carlo histogram method [50] and mean first-passage times were calculated for collapse, compaction and folding to the native state [48].

As this is a comparative study, we also performed simulations using the folding matrix alone for interactions; that is, $B_{ij} = B_{ij}^{f}$ in Equation 5. This model is referred to as the nonsolvation model in the text. The nonsolvation study can be compared to a similar 27-mer study by Socci and Onuchic [44], where part of their study looked at folding using the same matrix (E_{avg} in their terminology).

Sequence selection

We were interested in comparing the thermodynamics and kinetics of protein folding in our proposed solvation model with that in the nonsolvation model. For this purpose, we needed to find sequences that folded in both our solvation model and the nonsolvation model. Sequence selection was performed by the following set of steps:

1. First we chose a compact $3 \times 3 \times 4$ structure. We used several different approaches, described below, for finding appropriate structures.

2. Next, a sequence was designed within the solvation model that should fold to this structure. Sequences were designed at low temperature

using the constant composition design algorithm proposed by Shakhnovich and Gutin [73]. We fixed the sequence composition at 50% H, 50% P.

3. If the energy of the best designed sequence was not sufficiently low, we went back to step 1. This was done because early studies showed that when frustration is high enough, poorly optimized sequences tend not to fold to compact states but collapse instead to degenerate non-compact states. We found that for many structures, the distribution of contacts was such that a sufficiently low energy sequence could not be designed (Sorenson JM, Head-Gordon T, unpublished data).

4. We next enumerated the 84,731,192 maximally compact structures for a 36-mer [75] to find if the lowest energy for this sequence corresponded to a nondegenerate compact structure.

5. If so, folding studies were conducted to verify that the sequence could fold to this compact state within a reasonable number of moves (10^9) and the putative native state was the lowest energy state found in the simulation.

6. Sequences that passed this last criterion were considered foldable and their thermodynamics and kinetics were examined in the nonsolvation model as well.

Each step reduced the number of sequences available for study. Initially, 60 structures were looked at in step 1. Of these, we enumerated compact conformations for 35 sequences in step 4. From this, twelve sequences were evaluated in step 5 and finally eight sequences were found to pass to step 6.

Trial structures for step 1 were chosen from several sources. The first sequence was designed on a 36-mer structure used by Shakhnovich and coworkers in many previous studies [24,25,27,78]. Other structures were generated randomly from enumerating the maximally compact structures. Several structures came from step 4 when enumeration would show that a sequence had a lower energy structure than the one it had been originally designed for.

Unbiased random selection of structures from the 84,731,192 possible maximally compact structures can lead to many structures that are not able to pass all six steps, and it is useful to identify simple topological descriptors of structures that are more likely to produce foldable sequences. To improve our random selection of maximally compact structures, we first grouped the compact structures into subsets based on their relative contact order (RCO). The relative contact order is a measure of how many local versus nonlocal interactions are formed for a given structure [51,79]. It is defined as the average sequence distance between contacting residues:

$$RCO = \frac{1}{N_c N} \sum_{i<j}^{N} |i-j| \Delta_{ij} \qquad (12)$$

where N_c is the total number of contacts, N is the total number of residues and Δ_{ij} is as defined above. Many kinetics studies have tried to correlate this kind of order parameter with the folding time, although there remains a debate whether local interactions or nonlocal interactions are more important for determining fast folding [26,51,52]. We found that structures with an RCO of around 27–30% led to designed sequences that were more likely to pass all six steps, so many of the randomly selected compact structures were drawn from this subset. For comparison, we also randomly picked structures with an RCO 21–26% or an RCO of 30–40% but found that the resulting sequences were less likely to pass step 4. From full enumeration, we found that the lowest possible RCO for a maximally compact 36-mer is 21.11% and the highest possible RCO is 54.58%. The average RCO is 33.95% with a standard deviation of 4.12%. Future studies might also benefit from this grouping of structures into simple topological categories for the purposes of random structure selection.

Supplementary material

An analysis of the solvation model is available as Supplementary material published with this paper on the internet.

Acknowledgements

THG would like to acknowledge financial support from Air Force Office of Sponsored Research Grant #FQ8671-9601129 and US Department of Energy Contract #DEAC-03-76SF00098. JMS thanks the National Science Foundation for a Graduate Research Fellowship. We would also like to thank V. Pande and G. Hummer for encouraging conversations and R.M. Glaeser for pointing out reference [79].

References

1. Shakhnovich, E. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
2. Onuchic, J., Luthey-Schulten, Z. & Wolynes, P. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545-600.
3. Dobson, C., Sali, A. & Karplus, M. (1998). Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* **37**, 868-893.
4. Englander, S. & Mayne, L. (1992). Protein folding studied using hydrogen-exchange labeling and 2-dimensional NMR. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 243-265.
5. Baldwin, R. (1996). On-pathway versus off-pathway folding intermediates. *Fold. Des.* **1**, R1-R8.
6. Fersht, A. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
7. Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc. Natl Acad. Sci. USA* **89**, 20-22.
8. Dill, K. & Chan, H. (1997). From Levinthal to pathways and funnels. *Nat. Struct. Biol.* **4**, 10-19.
9. Bryngelson, J., Onuchic, J., Socci, N. & Wolynes, P. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195.
10. Wolynes, P., Onuchic, J. & Thirumalai, D. (1995). Navigating the folding routes. *Science* **267**, 1619-1620.
11. Daggett, V. & Levitt, M. (1993). Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 353-380.
12. Boczko, E. & Brooks, C. (1995). First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* **269**, 393-396.
13. Brooks, C. (1998). Simulations of protein folding and unfolding. *Curr. Opin. Struct. Biol.* **8**, 222-226.
14. Shakhnovich, E. (1996). Modeling protein folding: the beauty and power of simplicity. *Fold. Des.* **1**, R50-R54.
15. Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183-210.
16. Miyazawa, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.
17. Miyazawa, S. & Jernigan, R. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
18. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci.* **6**, 676-688.
19. Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636.
20. Lau, K. & Dill, K. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986-3997.
21. Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: are these meaningful potentials? *J. Chem. Phys.* **107**, 3698-3706.
22. Thomas, P. & Dill, K. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457-469.
23. Shakhnovich, E. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907-3910.
24. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
25. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6062.

26. Abkevich, V., Gutin, A. & Shakhnovich, E. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460-471.
27. Gutin, A., Abkevich, V. & Shakhnovich, E. (1995). Is burst hydrophobic collapse necessary for protein folding? *Biochemistry* **34**, 3066-3076.
28. Govindarajan, S. & Goldstein, R. (1997). The foldability landscape of model proteins. *Biopolymers* **42**, 427-438.
29. Israelachvili, J. (1992). *Intermolecular and Surface Forces.* Academic Press Limited, London.
30. Creighton, T. (1993). *Proteins: Structures and Molecular Properties.* W.H. Freeman and Company, New York 2nd edition.
31. Rank, J. & Baker, D. (1997). A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding. *Protein Sci.* **6**, 347-354.
32. Tsai, J., Gerstein, M. & Levitt, M. (1997). Simulating the minimum core for hydrophobic collapse in globular proteins. *Protein Sci.* **6**, 2606-2616.
33. Brugè, F., Fornili, S., Malenkov, G., Palma-Vittorelli, M. & Palma, M. (1996). Solvent-induced forces on a molecular scale: non-additivity, modulation and causal relation to hydration. *Chem. Phys. Lett.* **254**, 283-291.
34. Martorana, V., Bulone, D., Biagio, P. S., Palma-Vittorelli, M. & Palma, M. (1997). Collective properties of hydration: long range and specificity of hydrophobic interactions. *Biophys. J.* **73**, 31-37.
35. Ben-Naim, A. (1980). *Hydrophobic Interactions.* Plenum Press, New York.
36. Dill, K. (1997). Additivity principles in biochemistry. *J. Biol. Chem.* **272**, 701-704.
37. Grayce, C. & dePablo, J. (1994). The effect of solvation on the conformation of freely jointed repulsive trimers. *J. Chem. Phys.* **101**, 6013-6023.
38. Grayce, C. & Schweizer, K. (1994). Solvation potentials for macromolecules. *J. Chem. Phys.* **100**, 6846-6856.
39. Grayce, C. (1996). The conformation of hard-sphere polymers in hard-sphere solution calculated by single-chain simulation in a many-body solvent influence functional. *J. Chem. Phys.* **106**, 5171-5180.
40. Chan, H. & Dill, K. (1998). Protein folding in the landscape perspective: chevron plots and non-arrhenius kinetics. *Proteins* **30**, 2-33.
41. Baldwin, R. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl Acad. Sci. USA* **83**, 8069-8072.
42. Lüdemann, S., Schreiber, H., Abseher, R. & Steinhauser, O. (1996). The influence of temperature on pairwise hydrophobic interactions of methane-like particles: a molecular dynamics study of free energy. *J. Chem. Phys.* **104**, 286-295.
43. Hao, M.-H. & Scheraga, H. (1997). On foldable protein-like models: a statistical-mechanical study with Monte Carlo simulations. *Physica A* **244**, 124-146.
44. Socci, N. & Onuchic, J. (1995). Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732-4744.
45. Onuchic, J., Socci, N., Luthey-Schulten, Z. & Wolynes, P. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441-450.
46. Yue, K., Fiebig, K., Thomas, P., Chan, H., Shakhnovich, E. & Dill, K. (1995). A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA* **92**, 325-329.
47. Wolynes, P. (1997). As simple as can be? *Nat. Struct. Biol.* **4**, 871-874.
48. Socci, N. & Onuchic, J. (1994). Folding kinetics of proteinlike heteropolymers. *J. Chem. Phys.* **101**, 1519-1528.
49. Onuchic, J., Wolynes, P., Luthey-Schulten, Z. & Socci, N. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl Acad. Sci. USA* **92**, 3626-3630.
50. Ferrenberg, A. & Swendsen, R. (1989). Optimized Monte Carlo data analyis. *Phys. Rev. Lett.* **63**, 1195-1198.
51. Plaxco, K., Simons, K. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.
52. Unger, R. & Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988-994.
53. Socci, N., Onuchic, J. & Wolynes, P. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868.
54. Goldenfeld, N. (1992). *Lectures on Phase Transitions and the Renormalization Group.* Addison-Wesley Publishing Company, Reading, Massachusetts.
55. Socci, N., Nymeyer, H. & Onuchic, J. (1997). Exploring the protein folding funnel landscape. *Physica D* **107**, 366-382.
56. Hao, M.-H. & Scheraga, H. (1997). Characterization of foldable protein models: thermodynamics, folding kinetics and force field. *J. Chem. Phys.* **107**, 8089-8102.
57. Hao, M.-H. & Scheraga, H. (1998). Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.* **277**, 973-983.
58. Atkins, P. (1990). *Physical Chemistry.* W.H. Freeman and Company, New York fourth edition.
59. Guo, Z. & Thirumalai, D. (1994). Kinetics of protein folding: nucleation mechanism, time scales, and pathways. *Biopolymers* **36**, 83-102.
60. Guo, Z. & Brooks, C. (1997). Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein. *Biopolymers* **42**, 745-757.
61. Gutin, A., Sali, A., Abkevich, V., Karplus, M. & Shakhnovich, E. (1998). Temperature dependence of the folding rate in a simple protein model: search for a 'glass' transition. *J. Chem. Phys.* **108**, 6466-6483.
62. Kolinski, A., Galazka, W. & Skolnick, J. (1996). On the origin of the cooperativity of protein folding: implications from model simulations. *Proteins* **26**, 271-287.
63. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1-63.
64. Dill, K. (1990). Dominant forces in protein folding. *Biochemistry* **31**, 7133-7155.
65. deGennes, P.-G. (1979). *Scaling Concepts in Polymer Physics.* Cornell University Press, Ithaca, N.Y.
66. Chan, H., Bromberg, S. & Dill, K. (1995). Models of cooperativity in protein folding. *Phil. Trans. R. Soc. Lond. B* **348**, 61-70.
67. Pertsemlidis, A., Saxena, A., Soper, A., Head-Gordon, T. & Glaeser, R. (1996). Direct evidence for modified solvent structure within the hydration shell of a hydrophobic amino acid. *Proc. Natl Acad. Sci. USA* **93**, 10769-10774.
68. Head-Gordon, T., Sorenson, J., Pertsemlidis, A. & Glaeser, R. (1997). Differences in hydration structure near hydrophobic and hydrophilic amino acids. *Biophys. J.* **73**, 2106-2115.
69. Pertsemlidis, A., Soper, A., Sorenson, J. & Head-Gordon, T. (1998). Evidence for microscopic, long-range hydration forces for a hydrophobic amino acid. *Proc. Natl Acad. Sci. USA* in press.
70. Hummer, G., Garde, S., García, A., Paulaitis, M. & Pratt, L. (1998). The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Natl Acad. Sci. USA* **95**, 1552-1555.
71. Sfatos, C., Gutin, A. & Shakhnovich, E. (1993). Phase diagram of random copolymers. *Phys. Rev. E* **48**, 465-475.
72. Gutin, A. & Shakhnovich, E. (1993). Ground state of random copolymers and the discrete random energy model. *J. Chem. Phys.* **98**, 8174-8177.
73. Shakhnovich, E. & Gutin, A. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
74. Pande, V., Grosberg, A. & Tanaka, T. (1994). Folding thermodynamics and kinetics of imprinted renaturable heteropolymers. *J. Chem. Phys.* **101**, 8246-8257.
75. Pande, V., Grosberg, A., Joerg, C. & Tanaka, T. (1994). Enumeration of the hamiltonian walks on a cubic sublattice. *J. Phys. A: Math. Gen.* **27**, 6231-6236.
76. Hilhorst, H. & Deutch, J. (1975). Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. *J. Chem. Phys.* **63**, 5153-5161.
77. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1099.
78. Mirny, L., Abkevich, V. & Shakhnovich, E. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103-116.
79. Chan, H. (1998). Matching speed and locality. *Nature* **392**, 761-763.

Supplementary material

The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models

Jon M Sorenson and Teresa Head-Gordon

Folding & Design 11 December 1998, **3:**523–534

Analysis of the solvation model

In this appendix, we make explicit the long-range and multibody nature of the solvation model proposed in Equations 1, 5, 6, and 8. We start with Equations 1 and 5 for the energy of the chain:

$$E = \frac{1}{2}\sum_{i,j}^{N}\left[\left(1-\lambda_{ij}\right)B_{ij}^{u} + \lambda_{ij}B_{ij}^{f}\right]\Delta_{ij} \tag{13}$$

where the factor of 1/2 is now in place to account for double counting. The form of λ_{ij} described in Equations 6 and 8 was used in our study, but for the purposes of the following analytical treatment it is useful to redefine λ_i as

$$\lambda_i = \frac{s_i}{s_i^0} \tag{14}$$

that is, we no longer restrict λ_i and λ_{ij} to be ≤ 1. For our studies, we enforced this restriction to keep the λs as interpolative parameters. This minor change for the purposes of the following development does not affect the resulting conclusions about the range and multibody nature of the model.

Equation 13 is first rewritten as:

$$E = \frac{1}{2}\sum_{i,j}^{N} B_{ij}^{u}\Delta_{ij} + \frac{1}{2}\sum_{i,j}^{N}\lambda_{ij}B'_{ij}\Delta_{ij} \tag{15}$$

where $B'_{ij} = B^{f}_{ij} - B^{u}_{ij}$. Inserting the definitions of s_i in Equation 9 and λ_{ij} from Equation 6 we have:

$$E = \frac{1}{2}\sum_{i,j}^{N} B_{ij}^{u}\Delta_{ij} + \frac{1}{4}\sum_{i,j}^{N}\left(\sum_{k}\frac{1}{s_i^0}\Delta_{ik} + \sum_{k}\frac{1}{s_j^0}\Delta_{jk}\right)B'_{ij}\Delta \tag{16}$$

This can be further simplified by introducing the definition:

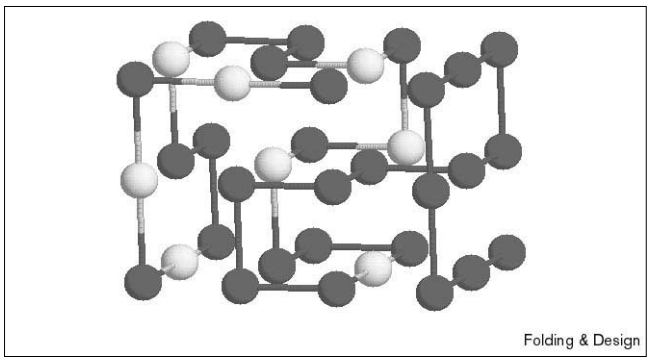
$$\Delta^2_{ik} = \sum_{j}\frac{1}{s_j^0}\Delta_{ij}\Delta_{jk} \tag{17}$$

Δ^2_{ik} is a similar operator to Δ_{ij} but its effect is longer ranged in that it connects sites that are next-nearest neighbors on the lattice. Δ^2_{ik} is nonzero if a path of two nonbonded interactions can be drawn between residues *i* and *k*. The action of Δ^2_{ik} is depicted in Figure S1. Like Δ_{ij}, Δ^2_{ik} ranges from 0 to 1 (as long as $s_j^0 \geq 2$), although it can take on fractional values because of the $1/s_j^0$ weight in the sum.

With this definition, we can perform the sum over *i* for the first term in the second sum and sum over *j* for the second term in the second sum to arrive at:

$$E - \frac{1}{2}\left(\sum_{i,j}B_{ij}^{u}\Delta_{ij} + \sum_{i,j}B'_{ij}\Delta^2_{ij}\right) \tag{18}$$

Figure S1



The action of the operator Δ^2_{ik}. Sites that are connected by this operator to the core site on the left are shown in white. Note that this site is connected to itself ($\Delta^2_{ii} \neq 0$, see text).

The first term on the right is of the same form as Equation 1, the pairwise contact energy used in traditional lattice models. The new term connects sites that are further away, up to next-nearest neighbors. We see that by making the strength of a residue–residue interaction dependent on the relative solvent accessibility of the pair, we have incorporated a dependence of the energy on interactions that are further apart than nearest neighbor.

From the definition of Δ^2_{ij}, Equation 17, we have:

$$\Delta^2_{ii} = \sum_{j}\frac{1}{s_j^0}\Delta_{ij} = s'_i \tag{19}$$

Thus Δ^2_{ii} is a measure of the solvent accessibility of site *i*, similar to the definition of s_i (Equation 10), but differing in that it is weighted differently. With this identification, the final form of the energy becomes:

$$E = \frac{1}{2}\left(\sum_{i}B'_{ii}s'_i + \sum_{i,j}B_{ij}^{u}\Delta_{ij} + \sum_{i\neq j}B'_{ij}\Delta^2_{ij}\right) \tag{20}$$

where now the first term is a Scheraga-like (see Equation 2) solvent accessibility term, the second term is a nearest-neighbor term and the third term links next-nearest neighbors.

This completes the explicit demonstration of the longer ranged aspects of the proposed solvation model. From this development, it is not difficult to show that the multiplicative definition of λ_{ij} given in Equation 7 requires the definition of an analogous Δ^3_{ij} operator, and the energy now depends on interactions between residues that are next-next-nearest neighbors as well as nearest-neighbor contacts.