Local Interactions Dominate Folding in a Simple Protein Model

Ron Unger$^{1,2,*}$ and John Moult$^2$

$^1$Department of Life Sciences
Bar-Ilan University
Ramat-Gan, 52900, Israel

$^2$Center for Advanced Research in Biotechnology
University of Maryland
Biotechnology Institute, 9600 Gudelsky Drive, Rockville MD 20850, USA

Recent computational studies of simple models of protein folding have concluded that a pronounced energy minimum (i.e. large gap in energy between low-energy states of the model) is a necessary and sufficient condition to ensure folding of a sequence to its lowest-energy conformation. Here, we show that this conclusion strongly depends on the particular temperature scheme selected to govern the simulations. On the other hand, we show that there is a dominant factor determining if a sequence is foldable. That is, the strength of possible interactions between residues close in the sequence. We show that sequences with many possible strong local interactions (either favorable or, more surprisingly, a mixture of strong favorable and unfavorable ones) are easy to fold. Progressively increasing the strength of such local interactions makes sequences easier and easier to fold. These results support the idea that initial formation of local substructures is important to the foldability of real proteins.

Introduction

What are the dominant contributions guiding the process of protein folding? The short life of folding intermediates makes experimental analysis very difficult. Theoretical analysis of the folding of polypeptide chains has been hindered by the large size and complexity of the systems to be considered. Simple models of protein folding, while losing some of the details, may offer broad insights. For example, the importance of the statistical characterization of energy landscapes (Bryngelson & Wolynes, 1987), and the importance of hydrophobicity in folding (Dill, 1990) have been highlighted by simple model studies.

Shakhnovich & Gutin (1990a) introduced a very simple and elegant model of hetropolymer folding that makes possible studies that are not feasible on a more complete representation of the problem. In a recent pair of papers Sali et al. (1994a,b) used this simple model to advance a specific view of the folding process. Their surprising conclusion was that a wide separation of energy levels among the lowest-energy states of the system is crucial to folding. Specifically, they found that if the gap in energy between the lowest and the second lowest compact conformations was large, then a Monte Carlo folding simulation (considered to reflect the kinetics of folding) was much more likely to find the conformation with the global minimum of energy. These findings led them to state (Sali et al., 1994a,b) that “the necessary and sufficient condition for folding in this model is that the native state is a pronounced energy minimum”. Failing to see how the gap between the two lowest-energy states of the system could have such a kinetic effect, we adopted the model system suggested by Sali et al. (1994a,b), and further analyzed its behavior. We noticed that by changing the temperature scheme, folding is no longer related to the gap size. We further analyzed the system and found that the locality of the interactions is a much more important factor in determining the kinetic folding process. Native conformations that include many favorable short-range interactions (between residues close in the sequence) are folded readily, while structures with a high proportion of favorable remote (long-range in the sequence) interactions fold very rarely. Similar results were obtained for sequences in which local interactions are dominant (i.e. have larger magnitude), regardless of whether they are favorable or unfavorable. These results are robust over a large range of parameter selection. This view, that local interactions are dominant in the folding process, is hardly revolutionary. Yet, it is reassuring that the
importance of locality, which has much experimental and theoretical support, can be invoked to explain the behavior of simple model systems.

The Model

The Shakhnovich & Gutin model (1990a,b) uses a chain of 27 amino acid residues folding onto a three-dimensional cubic lattice. Each residue is represented by a single bead, which occupies a lattice point. The model is contiguous (i.e. two neighboring residues in the sequence must occupy neighboring lattice points) and self-avoiding (i.e. only one bead can occupy one lattice point). Two residues interact when they occupy neighboring lattice points. A novel and powerful feature is the manner in which the effective sequence is determined. A sequence is represented by a pairwise energy function in which the strength of a possible interaction between each pair of residues along the chain is drawn randomly from a normal distribution with a mean of $B_0$ and standard deviation of $\sigma_B$. On a cubic lattice, a sequence 27 residues long has 156 possible interactions between pairs of residues. Interactions between bonded, i.e. contiguous, residues are not counted and, on a cubic lattice, odd-numbered residues can interact only with even-numbered residues. Thus, residue 1 can interact only with residues 4, 6, 8, . . ., 26, residue 2 with residues 5, 7, . . ., 27, residue 3 with residues 6, 8, . . ., 26, etc., a total of 156 interactions. A sequence is thus a set of all 156 potential interactions between pairs of residues along the sequence. The energy of a given conformation is the sum of interactions between the residues that occupy neighboring lattice points. A given conformation realizes only a subset of these interactions, e.g. a perfect $3 \times 3 \times 3$ cube contains 28 interactions (six interactions in each of the nine planar sections through the cube minus 26 bonded interactions.) Note that the sequences (i.e. the interaction set) determine the potential for interactions, while the actual energy of a conformation is determined only by the interactions that are realized.

Šali et al. (1994a,b) established that with a normal distribution of possible interactions where $B_0 = -2$ and $\sigma_B = 1$, the conformation with the lowest energy for a given sequence tends, with very high probability, to be contained within a perfect $3 \times 3 \times 3$ cube. That is, because on average interactions in this model tend to be attractive, the lowest energy conformation is likely to be a fully compact state. Thus one can find the conformation with the lowest energy simply by enumerating all of the 103,346 unique (non-symmetric) fully compact conformations (Shakhnovich & Gutin, 1990b). The simplicity of the model, and especially the elegance with which the global minimum of this non-trivial system can be determined, enabled Šali et al. (1994a,b) to conduct their study. For similar reasons we decided to use the same model. Sets of sequences were created as described (Šali et al., 1994b). Conformations were changed using a mixture of 50% pivot moves (where a residue is randomly chosen, and the rest of the chain is rotated randomly around it), 25% corner flipping (where a residue is flipped to an opposite corner), and 25% crankshaft moves (where a pair of contiguous residues rotate together around an axle defined by the two adjacent residues). Various mixtures of move types were tried and all yielded similar results. Conformations are considered only if they are self-avoiding. A new conformation is accepted if its energy is lower than the current one or if it passes the Metropolis test (Metropolis et al., 1953). That is, $\text{RND} < e^{-\Delta E/k_B T}$, where $\text{RND}$ is a random number chosen uniformly from the interval $[0, 1]$, $\Delta E$ is the energy change, $T$ is the temperature of the simulation, and $k_B$ is Boltzmann’s constant. Since the energy units are arbitrary, the Boltzmann constant in all the experiments is set to 1. Each simulation was run for 75,000,000 steps. Five runs were performed on each sequence. In accord with Šali et al. (1994a,b), sequences for which the simulation found the global minimum (known from the full enumeration on the cube) at least twice in the five runs are considered foldable.

The Dependence of Foldability on Gap Size and Temperature

The main conclusion of Šali et al. (1994a,b) was that the fate of a Monte Carlo simulation trying to fold such random sequences is largely dependent on the gap in energy, for a given sequence, between the lowest and the second-lowest fully compact conformations. When this gap is relatively large, then there is a higher chance that the simulation will find the conformation with the lowest energy, otherwise finding the global minimum is rare. This conclusion suggests a surprising link between a static parameter (i.e. the gap in energy between two fully compact conformations) and the dynamic behavior (i.e. the ability of a Monte Carlo simulation to reach the lowest-energy conformation). An intriguing part of their scheme is the way in which the temperature for the simulations was selected. The temperature is a crucial parameter for a Monte Carlo folding simulation. It provides the impetus to make up-hill energy moves and thus to escape from local minima. Studies on simple folding models have shown that the efficiency of folding varies with temperature (Miller et al., 1992; Chan & Dill, 1994; Socci & Onuchic, 1994; Camacho & Thirumalai, 1995).

Šali et al. (1994a,b) selected the temperature according to the following criteria: it is the temperature $T$ for which the expression:

$$X(T) = 1 - \sum_{i} \left[ \frac{e^{-E_i/B_0}}{\sum_j e^{-E_j/B_0}} \right]^3$$

has the value of 0.8.
The summation is done over all of the 103,346 compact conformations. Since $e^{-E_i/k_BT}/\sum_i e^{-E_i/k_BT}$ is the fractional occupancy of conformation $i$ at equilibrium, the order parameter $X$ is close to 1 when the system has many states with comparable Boltzmann occupancy, and is close to 0 when there is one dominant low-energy conformation. Although this parameter is a sum over all the compact conformations of the system, under these conditions it is typically dominated by the energy levels of the very few lowest-energy conformations. This can be demonstrated by calculating $X(T)$ including only a few terms corresponding to the lowest-energy conformations in the summation over $i$ in equation (1). Figure 1(a) shows the value of $X$ as the function of the number of conformations included, starting from the lowest in energy. Including only the two lowest-energy conformations yields a value of 97.7% of the full $X(T)$. As a consequence, $T$ is highly correlated with the gap between the two lowest energy conformations, as can be seen in Figure 1(b).

Table 1 summarizes the results of all the experiments described here. In experiment I, we repeated the experiments reported by Šali et al. (1994a,b) using 99 sequences, and obtained similar results: 23% of the sequences are foldable (i.e. find the global optimal solution in at least two of five runs) compared with 15% by Šali et al. (1994b). Sequences with large gaps are somewhat more likely to fold: amongst the 11 sequences with the largest gaps, four are foldable, while only one of the 11 sequences with the smaller gaps is foldable. For intermediate gap sizes no correlation with foldability is observed (Figure 2A). On the whole, there is indeed a correspondence between gap size and foldability, but it does not seem to be strong enough to support the claim made by Šali et al. (1994a,b) that a pronounced energy gap is a necessary and sufficient condition for folding in this model. Furthermore, foldability of the same sequences shows a stronger correlation with simulation temperature than with the gap size (Figure 2B). In order to further examine whether the foldability is really related to the gap or is a consequence of the particular scheme in which the temperature for each simulation was selected, the following experiments were performed: In experiment II, simulations were run at a fixed temperature, independent of gap size. We investigated three different values of temperature, $T = 0.7, 1$ and $1.3$ as low, middle, and high values from the range of temperatures used by Šali et al. (1994b). In this case 16%, 30% and 65% of the sequences were foldable, respectively. Thus, higher temperature yields higher success in folding these sequences, with no correlation with the gap size (results for $T = 1$ are shown in Figure 2C).

A further demonstration that the foldability as measured by Šali et al. (1994a,b) is a consequence of the temperature scheme and is not determined by the actual gap size came from experiment III: The same set of temperatures as in experiment I were used.
experiment I was used to obtain $T$ (i.e. using expression (1) as suggested by Šali et al. (1994b)), but these temperatures were shuffled randomly between the 99 sequences. Each one of the 99 sequences was randomly matched with a temperature originated from another sequence in the set, regardless of the gap size. For example, sequence 89, with a very small gap (0.29), was matched with $T = 1.8$, a very high temperature that originated from sequence 85 with gap of 5.02. Sequence 85 was assigned $T = 0.8$, originated from another sequence, and so on. Thus we have a system using an identical temperature distribution to that used by Šali et al. (1994b), but this time without any correlation with gap sizes. If the underlying gaps were the important factor, we would expect to see each sequence perform very similarly to its performance in experiment I, since the gap sizes are the same in both experiments. If, on the other hand, the temperatures themselves determine the outcome, then we expect foldability still to correlate with temperature. Because the distribution of the temperatures is identical, the overall performance of the set of sequences as a function of temperature is similar to experiment I (Figure 2B versus E). Overall, 29% of the sequences are now foldable, compared with 23% in experiment I. But sorted by gap size, the behavior is very different (Figure 2A versus D). For example, four of the 11 sequences with the smallest gap size are foldable, while only two of the 11 sequences with the largest gap size are foldable. Again, these results demonstrate that the temperatures and not the gap sizes are driving the simulations.

### Locality of Interactions

If the gap size is not the crucial factor in determining if a sequence is foldable, what is the important factor? We now set out to show that the answer is the locality of interactions. The Šali et al. scheme (1994a,b) for assigning pairwise interactions lends itself well to testing this hypothesis. For a given sequence, we define a favorable interaction locality index as:

$$ F = \sum_{ij} -E_{ij}(27 - |i-j|) $$

where $E_{ij}$ is the energy of an interaction between residue $i$ and residue $j$ in the sequence.

In this model a sequence is defined by a set of independent possible pairwise interactions. Thus, the possible interaction between a specific pair of residues may be changed without affecting other interactions. In particular, the locality of a sequence can be changed by swapping interactions between pairs of pairs of residues. The change can be made as desired to decrease or increase the locality index. For example, if the strength of the interaction between residues 4 and 9 (i.e. sequence distance of 5) is $-1.0$ and the strength of the interaction between residue 3 and 20 (sequence distance 17) is $-3.0$, and a more local sequence is required, then these interactions are swapped. For each original interaction set, one can repeat this process of swapping interactions exhaustively, in order to get the most local sequence. By swapping the interactions in the other direction, the most remote

### Table 1. Summary of folding simulations

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Purpose</th>
<th>Number of sequences</th>
<th>Folding simulation temperature scheme</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Reproduce Šali et al. experiments</td>
<td>99</td>
<td>Sali et al.</td>
<td>Large energy gaps imply high simulation temperatures, which yield foldability (Figure 2A and B)</td>
</tr>
<tr>
<td>II</td>
<td>Folding with fixed temperatures</td>
<td>99</td>
<td>Fixed $T = 0.7, 1, 1.3$</td>
<td>No correlation between gap size and foldability is seen (Figure 2C)</td>
</tr>
<tr>
<td>III</td>
<td>Uncouple lowest-energy gap size and simulation temperatures</td>
<td>99</td>
<td>Randomly assigning Šali et al. temperatures regardless of gaps</td>
<td>Foldability is seen to depend on temperature, not on gap size (Figure 2D and E)</td>
</tr>
<tr>
<td>IV</td>
<td>Role of locality</td>
<td>299</td>
<td>Fixed $T = 0.7, 1, 1.3$</td>
<td>Sequences with many possible favorable local interactions are much more foldable than sequences with remote favorable interactions</td>
</tr>
<tr>
<td>V</td>
<td>The effect of varying locality</td>
<td>192</td>
<td>Fixed: $T = 1$</td>
<td>Foldability increases gradually with sequence locality (Figure 2F)</td>
</tr>
<tr>
<td>VI</td>
<td>The effect of favorable and unfavorable strong local interactions</td>
<td>64</td>
<td>Fixed: $T = 1$</td>
<td>Foldability is facilitated by the presence of many possible strong local interactions both favorable and unfavorable (Figure 2G)</td>
</tr>
</tbody>
</table>

The Table is a quick guide to the six simulation experiments analyzed in this study. For each experiment we report the number of sequences tested, and the temperature scheme used in the simulations. The purpose of each experiment and its main conclusions are stated. For each sequence, the simulation was run five times, for 75,000,000 steps each. Sequences for which the simulation found the global minimum at least twice out of the five runs are considered foldable. The temperatures used in experiments I and III were according to the scheme of Šali et al. (i.e. $T$ such that $X(T) = 0.8$, see the text). In the other experiments the temperatures were fixed to the values mentioned in the Table.
sequence can be generated (i.e. where the most favorable interactions are between residues far away in the sequence). In this way, 99 pairs of most local and most remote sequences were created from the original 99 sequences of experiment I. Since this procedure creates new sequences, the cube conformations were enumerated for each one of these.

In experiment IV we analyzed the behavior of the most local and most remote sequences, and found that folding local sequences to their lowest-energy conformation is much easier than folding remote sequences. We noted that for the local sequences, the Monte Carlo simulations find extended, out of cube, conformations that have lower energy than the best conformations on the cube. This is obviously due to the fact that strong favorable $i \leftrightarrow i + 3$ interactions that are common in the local sequences tend to enforce elongated helix-like conformations. Since the best conformations found are not on the $3 \times 3 \times 3$ cube, we cannot prove directly that they are the global minima. Nevertheless, the fact that for 91% of the sequences the same conformation was obtained in at least three out of the five runs supports the assertion that the

![Figure 2](image)

**Figure 2.** The parameters affecting foldability. A to E, The dependence of foldability on gap size and temperature. The foldability of 99 sequences is shown in nine groups, sorted by either the size of the energy gap between the two lowest-energy conformations available to each sequence, or by the simulation temperature. A, Using the same temperature scheme as Sali et al. (1994b), sorted by gap size. Correlation between foldability and gap size can be seen. B, Using the same temperature scheme as Sali et al. (1994b), sorted by simulation temperature. The correlation between foldability and the simulation temperature is even stronger. C, Using a fixed temperature of $T = 1$, sorted by gap size. The link between gap size and foldability is no longer apparent. D, Randomly assigning temperatures from the same distribution as Sali et al. (1994b), sorted by gap size. No link between foldability and gap size is seen. E, Randomly assigning temperatures from the same distribution as Sali et al. (1994b), sorted by simulation temperature. Higher temperatures yield higher foldability. F and G, The foldability of sequences as a function of their potential to form interactions locally in the sequence. F, Distribution of foldability as a function of the favorable interaction locality index, which reflects the concentration of possible favorable interactions between residues close together in the sequence: 192 sequences sorted into eight bins by their locality index were used. The significant and progressive effect of locality on foldability is evident. G, Distribution of foldability as a function of the strong interaction locality index, which reflects the concentration of possible strong interactions (both favorable and unfavorable) between residues close together in the sequence: 64 sequences sorted into eight bins by their locality index were used. Again, the significant and progressive effect of the strong interaction locality index on foldability is evident.
simulations usually found the optimal conformation. Again, we tried a few fixed temperature values: For an intermediate temperature of \( T = 1 \), the local sequences show a very high foldability of 95%, while only 6% of remote sequences are foldable. At low temperature (\( T = 0.7 \)), 79% of local sequences are foldable but only 2% of remote sequences are. A similar trend is seen for a high temperature (\( T = 1.3 \)), where local sequences are 96% foldable, compared with 9% of remote sequences.

In experiment V, we explored the effects of gradually varying the locality by creating, for each interaction set, sequences with varying values of the locality index. We split the range between the locality indices of the most local and the most remote sequences into eight locality subranges. For each interaction set, we start with the most local sequence, and pick randomly a pair of interactions. The interactions are swapped if by doing so a sequence with lower locality index is created. Repeating this operation exhaustively yields a long series of sequences with gradually decreasing locality index, spanning all locality subranges. The first such sequence to have a locality index within a locality subrange was chosen to represent the subrange. In this manner, from each interaction set we obtained eight sequences with varying degrees of locality.

The first 24 of the 99 interaction sets were used (i.e. 192 sequences). The foldability of each sequence was probed in the usual way, i.e. comparing the energy of the best cube enumeration (done separately for each sequence) to the result of five Monte-Carlo runs with fixed temperature of \( T = 1 \). The results are shown in Figure 2F, where foldability is presented as a histogram of eight bins of locality subranges. It is clear that locality has a strong and progressive effect on foldability. The more local the sequence is, the more likely it is to fold.

So far, we have manipulated the sequence towards more local or remote favorable interactions. In experiment VI, we analyzed the affect of the locality of strong interactions, regardless of whether they are favorable or unfavorable. To do so, we used the generating procedure of experiment V, but now swapping interactions based on their deviation from the mean strength of all interactions. For example, with the mean strength of interaction of \( B_0 = -2 \), if the strength of the interaction between residues 4 and 9 is \(-3.0\) and the strength of the interaction between residue 3 and 20 is \(-0.5\), and a sequence with strong local interactions is to be obtained, then the interactions are swapped, since \(-0.5\) deviates more from the mean (\(-2\)) than does \(-3\). In this sense, the \(-0.5\) interaction is “stronger” than the \(-3.0\) interaction, although it is of course less favorable. The strong interaction locality index is defined as:

\[
S = \sum B_{ij}(27 - |i-j|) \tag{3}
\]

Using the same procedure as in experiment V, each interaction set was used as the source for eight sequences with varying strength index. Figure 2G shows the results for Monte Carlo simulations on the first eight interaction sets (i.e. 64 sequences).

Under this scheme the on-cube minimum does appear to be the global minimum for all sequences. This is probably due to the fact that favorable and unfavorable interactions are mixed along the sequence, so that there is no drive towards extended conformations. The importance of locality is again clear. Locally dominant sequences are much easier to fold than sequences in which the dominant interactions are remote. Once more the effect is progressive with increasing locality.

**Discussion**

To conclude, we have shown that the dependence of foldability on the size of the gap in energy between the two lowest energy conformations found by Šali *et al.* (1994a,b) is a direct consequence of their method of choosing a simulation temperature. Therefore, it is a property of the model and not of real proteins. The original work was subject to criticism on other grounds (Chan, 1995). One of them is that under the temperature scheme employed by Šali *et al.* (1994a,b), the number of steps is a measure only of the first passage time through the native conformation rather than of actual folding times.

Nevertheless, we find that the elegance of the original model makes it a powerful tool in understanding the basic processes involved in real protein folding. We have used it to establish that locality of strong interactions plays a crucial role in determining foldability. The idea of small substructures local in the sequence (so-called nucleation or initiation sites) forming early in the folding process and driving the subsequent pathway has been put forward many times in the past (e.g. see Wettlaufer, 1973; Moult & Unger, 1991; Dill *et al.*, 1993; Abkevich *et al.*, 1994; Guo & Thirumalai, 1995). Experimental data on both the earliest folding events (e.g. see Serrano *et al.*, 1992; Valey *et al.*, 1993) and the partial stability of appropriately chosen protein fragments (e.g. see Dyson *et al.*, 1992; Cox *et al.*, 1993) support this view. The availability of favorable local interactions in the model ensure such nucleation properties, and demonstrate their effectiveness in increasing overall foldability. It is reassuring that the importance of locality can be invoked to explain the detailed behavior of simple model systems. More surprising is the finding that a combination of strong favorable and unfavorable local interactions facilitate folding. A recent lattice study (Shrivastava *et al.*, 1995), in which sequences were designed according to the principle of minimal frustration (Bryngelson & Wolynes, 1987), has similarly demonstrated that sequences with repulsive interactions facilitate folding. Based on these results, one might speculate that the need to avoid non-native local interactions
in the folding process is equally important for achieving foldability. These observations may have practical implications by emphasizing the importance of including possible unfavorable interactions (so called “negative design”) in protein design, as a means of providing a strong and effective guide towards the desired conformation.

Acknowledgements

This work was supported by NIH grant GM 31034 to J.M. and by an Alon fellowship to R.U.

References


Edited by F. E. Cohen

(Received 28 December 1995; received in revised form 4 April 1996; accepted 15 April 1996)