Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8798–8802, September 1995
Biophysics

On constructing folding heteropolymers

(protein folding/sequence optimization/evolutionary algorithms/helix-coil transition/random energy model)

MARTIN EBELING AND WALTER NADLER

Institut für Theoretische Chemie, Universität Tübingen, Auf der Morgenstelle 8, D-72076 Tübingen, Federal Republic of Germany

Communicated by Rudolph A. Marcus, California Institute of Technology, Pasadena, CA, May 22, 1995 (received for review November 14, 1994)

ABSTRACT **Simplified models of the protein-folding process have led to valuable insights into the generic properties of the folding of heteropolymers. On the basis of theoretical arguments, Shakhnovich and Gutin [(1993) *Proc. Natl. Acad. Sci. USA* 90, 7195–7199] have proposed a specific method to generate folding sequences for one of these. Here we present a model of folding in heteropolymers that is comparable in simplicity but different in spirit to the one studied by Shakhnovich and Gutin. In our model, the proposed recipe for constructing folding sequences fails. We find that, as a rule, the construction of folding sequences is impossible to achieve by looking at the native conformation only. Rather, competing conformations have to be taken into account too. An evolutionary algorithm that generates folding sequences by optimizing both stability of the native state and folding time is described. Remarkably, this algorithm produces, among others, sequences that fold reproducibly to metastable states.**

In a recent paper, we have presented a simplified model of secondary structure formation in polypeptides (1) that combines ideas from the treatment of helix–coil transitions (2) and of two-dimensional polymer crystallization (3, 4). An essential feature of this model is a strongly simplified form of tertiary interactions between elements of secondary structure that allows one to study the influence of such interactions on the secondary and tertiary structure-formation process. Due to its essentially two-dimensional character, the model exhibits a strong neighborhood correlation between structural elements. However, it is well-known that neighborhood correlations in three-dimensional protein structure (5) and in collapsed polymers (6) are also stronger than expected from the analysis of three-dimensional self-avoiding random walks. Models of the type presented here have so far not been put to use in the study of protein folding to our knowledge.* This offers a chance to reexamine results obtained with other simplified models (8–20), especially since the model is easily implemented and already can be run effectively on personal computers. In the foregoing paper, we studied the homopolymer case (1). We showed that it is possible to obtain compact, mostly α-helical structures that resemble globular proteins in helix number and average helix length. The transition from the random coil to compact states was found to be essentially glass-like. In this paper, we present results for the heteropolymer case and use our model to investigate recently published notions on the construction of folding heteropolymers.

We represent the conformation of a polypeptide of length L by a string of labels $\sigma_i = h$, c^+, or c^0, where i ranges from 1 to L. The conformation h corresponds to residues with dihedral angles characteristic of α-helices, whereas c^+ and c^0 represent random coil residues. We assume that c^0 residues do not contribute to the distance between adjacent helices (i.e., two helices separated solely by c^0 residues are taken to be in contact), whereas helices with at least one residue with conformation other than c^0 between them are not in contact. Thus, the interconversion $c^0 \leftrightarrow c^+$ allows one to model the formation and disruption of tertiary contacts between helices. The free energy of a conformation $\{\sigma_i\}$ with sequence $\{A_i\}$ is given by

$$\begin{aligned} F(\{\sigma_i\}, \{A_i\}) = &\sum_{n=2}^{L-1} H(\sigma_{n-1}, \sigma_n, \sigma_{n+1})[\Delta E(A_{n-2}) + \Delta E(A_{n+2})]/2 \\ &+ \sum_{n=1}^{L-1} \sum_{m=n+1}^{L} C_{n,m}(\{\sigma_i\})[k(A_n) + k(A_m)]/2 \\ &- T\sum_{n=1}^{L} \Delta S(\sigma_n, A_n). \end{aligned} \quad [1]$$

The three terms in Eq. 1 describe the contribution of hydrogen bonds, tertiary interactions, and entropic contributions due to local conformation space restrictions, respectively (compare ref. 1).

Following the formalism by Lifson and Roig (2), three successive monomers must be in helical conformation to be spanned by a hydrogen bond. Therefore, if $\sigma_{n-1} = \sigma_n = \sigma_{n+1} = h$, we have a hydrogen bond linking residues $n - 2$ and $n + 2$. We describe this by defining $H(\sigma_{n-1}, \sigma_n, \sigma_{n+1}) = 1$ in this case and 0 otherwise. The strength of the hydrogen bond between two monomers $n - 2$ and $n + 2$ is determined by the mean of their respective ΔE parameters. Note that the two monomers forming the hydrogen bond need not be in the helical conformation, as opposed to the three monomers in between them. For consistency, two dummy coil residues A_0 and A_{L+1}, with $\Delta E(A_0) = \Delta E(A_{L+1}) = 0$, are added, which allow the formation of hydrogen bonds bridging the first three or last three residues, respectively, but which do not otherwise contribute to F (21).

Any two helices separated solely by c^0 residues are considered to be in contact with each other. In our simplified treatment of the tertiary interactions, we assume helices to be arranged in parallel and in register (Fig. 1). All of the residues of the shorter helix are then taken to be in contact with their counterparts on the longer helix. This can be formulated as $C_{n,m}(\{\sigma_i\}) = 1$ when residues n and m are in contact in chain conformation $\{\sigma_i\}$ and 0 otherwise. The contact energy is simply taken as the mean of the respective contact parameters k of the two residues in contact. This interaction scheme does not take α-helix topology into account and, of course, could be modeled more realistically, but it serves the purpose of introducing an additional sequence-dependent type of interaction into the helix–coil transition model.

Finally, the entropic term represents only contributions to the system's entropy that arise from local conformation space restrictions. Since the conformation space volume $V(h)$ acces-

Abbreviations: REM, random energy model; MC, Monte Carlo.

*The only other model that we are aware of that includes both tertiary interactions (in that case, hydrophobic interactions) and secondary structure formation is by Thomas and Dill (7); however, their model is quite unrelated to the one presented here.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

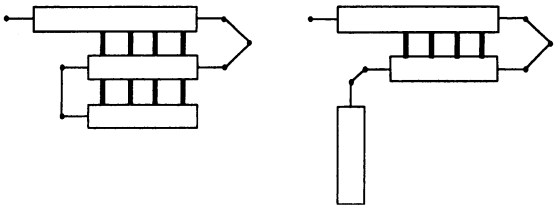

FIG. 1. Schematic representation of polymer chain conformation, $L = 20$, in which three distinct α-helical stretches have formed. (*Left*) Three helices interact with each other (indicated here by black bars between the helices). The chain conformation might be $c^+h_6c_3^0h_4c_2^0h_4$. As a shorthand notation, we describe this conformation as $_1\mathbf{6}_3\mathbf{4}_2\mathbf{4}$, where numbers in boldface stand for helix lengths, and subscripts indicate loop regions between helices. (*Right*) Contact between two of the helices has been broken, resulting in a chain conformation of $c^+h_6c_3^0h_4c^0c^+h_4$, or $_1\mathbf{6}_3\mathbf{4}_{[2]}\mathbf{4}$, where square brackets indicate that the helices on either side of the loop region are not in contact.

sible to h residues is smaller than that for non-h residues, as mirrored in a Ramachandran plot (22), the conversion $c^+ \leftrightarrow h$, for example, is accompanied by a loss in conformational entropy. We define $\Delta S(\sigma_i, A_i) = k_B \ln[V(\sigma_i, A_i)/V(c^+, A_i)]$. Since other possible entropic terms for the chain as a whole (namely, those due to degeneracy of random coil sections) are not represented, the role of F is equivalent to that of a Hamiltonian in other models.

Any sequence of residues $\{A_i\}$, $i = 1 \ldots L$, will be completely characterized by the three parameter sets $\{\Delta E(A_i)\}$, $\{k(A_i)\}$, and $\{\Delta S(\sigma_i, A_i)\}$. We set $\Delta S(c^0, A_i) = 0$, allow two values for each of the three parameters $\Delta E(A_i)$, $k(A_i)$, and $\Delta S(h, A_i)$, and consider the $2^3 = 8$ residue types resulting from the combination of the possible parameter values as described in Table 1. They have been chosen in such a way that $\Delta S(h) < 0$† for all residue types and that both the hydrogen bond and contact energy parameters can be favorable as well as unfavorable to structure formation. All the results presented here have been obtained for a constant temperature with $k_BT = 0.108\ |\Delta E(\mathbf{A})|$.

When looking for the optimal conformation of a given sequence (in the low temperature limit), one has to take into account the fact that residues in different positions have to fulfill different requirements. For example, the residues situated in the loops between helices do not establish any tertiary contacts. In addition, only the first and last residues of any loop take part in hydrogen bonds, whereas central residues of short helices may contribute to tertiary contacts but not to hydrogen bonds. Therefore, it would be locally desirable to place residue types **D** and **G**, which contribute favorably to tertiary interactions but unfavorably to hydrogen bonds, in the center of short helices. Likewise, it would be locally desirable to put residue types **B** and **E** at the first or last position in loops, etc. However, since correlations between residue conformations, as expressed by the terms $H(\sigma_{n-1}, \sigma_n, \sigma_{n+1})$ and $C_{n,m}(\{\sigma_i\})$ in Eq. **1**, are relevant for determining F, it will usually be impossible to devise a chain conformation so that all of the residues are put in a local context that makes optimal use of their potential interactions. This phenomenon is known from spin glasses and is called "frustration" (24, 25). It is understood to occur in proteins too (26, 27). Consequently, given any particular sequence, it will in general be difficult to determine its ground-state conformation. In essence, this is what makes determining protein structure from amino acid sequence a problem.

†The homopolymer studied in ref. 1 corresponds to the parameter values $k/\Delta E = 0.6$, $\Delta S(c^0) = \Delta S(c^+) = 0$, and $\Delta S(h)/k_B = -4.26 + \ln(2) \approx -3.57$, the latter value derived from experimental data first discussed by Zimm and Bragg (23).

Table 1. Characterization of residue types

Type	ΔE	$\Delta S(h)$	k	p
A	−1.0	−2.0	−0.6	1/6
B	−1.0	−2.0	+0.3	1/6
C	−1.0	−3.57	−0.6	1/8
D	+0.5	−2.0	−0.6	1/8
E	−1.0	−3.57	+0.3	1/8
F	+0.5	−2.0	+0.3	1/8
G	+0.5	−3.57	−0.6	1/12
H	+0.5	−3.57	+0.3	1/12

When looking for the optimal sequence for a given target conformation, sequence optimization procedures will select appropriate residue types according to the local requirements of a given target conformation. For example, for a helix of length five that establishes a tertiary contact in the conformation, residue types **A** and **D** in the central position will both minimize F. However, it will be impossible to choose between them simply because, in this local context, the capacity to form hydrogen bonds is not probed. For loop regions, the situation is even worse. Thus, the F function for any target conformation in general will be highly degenerate in sequence space. However, the properties of residues in neutral or "undecidable" positions may play a significant role in other possible chain conformations. In this context, we note that substitutions of single amino acids in proteins have been extensively studied. While little is known about the energetic consequences of such substitutions, it is well appreciated that the resulting structural changes are sometimes only minor local rearrangements but sometimes are extensive perturbations of the overall protein structure (28). From the foregoing it follows that holding any conformation constant and minimizing F by adjusting the sequence in sequence space may not suffice to make that conformation the global minimum space or even a local minimum in conformation space, let alone to establish a folding sequence. Such a procedure might even arrive at the type **A** homopolymer: for any conformation, this sequence is a global minimum for F in sequence space but obviously will not fold into that conformation. Therefore, when constructing a folding sequence for a particular target conformation, one has to consider other possible conformations and to introduce residues that not only favor the target conformation but also suppress others. This is reminiscent of "negative design" strategies used by researchers in the field of *de novo* protein design (29).

Shakhnovich and coworkers (15–19) have found in their model that the energy spectrum of a folding sequence exhibits a pronounced energy gap between the native state and all of the other conformations—a feature that is not typically found in the spectra of random sequences. Shakhnovich and Gutin (16) have demonstrated as a consequence that folding sequences for their model can be engineered by minimizing the energy of the native state with respect to sequence while keeping the amino acid composition constant. They argue that the constraint of constant composition not only prevents convergence to a homopolymer but also ensures that minimizing the energy of the target conformation in sequence space produces an energy gap between the target and all other conformations. They support this claim by using results derived from Derrida's random energy model (REM) (30, 31) onto which the ensemble of all possible sequences can be mapped under certain conditions (26, 32). When applied to proteins or models thereof, the self-averaging properties of the REM (27) let one expect that rearranging a sequence to minimize the target conformation's energy does not affect the statistical properties of the energy spectrum of all other possible conformations. It is concluded that use of the proposed optimization procedure lowers the target conformation energy se-

lectively, while all other conformation energies on average remain unchanged.

The REM does not make any structural statements about conformations at all but simply treats their respective energies as statistically independent—hence, the name. Proteins differ from the REM in that energies of related conformations are strongly correlated. Therefore, minimizing the target's energy will affect other conformations with similar interactions. To apply the concepts of the REM to the construction of folding sequences in a protein model, one has to postulate that conformations different from but with interactions similar to that of the target will not interfere with the folding process. Ideally, they should belong to the target's potential well in the energy landscape and should be higher in energy than the target itself—otherwise, the chain could fold to either of these conformations also and no folding sequence would emerge. The question arises whether validity of these assumptions is a generic property of protein folding models and, thus, of proteins. For the model presented here, the approach suggested by Shakhnovich and Gutin (16) fails to produce folding sequences reliably.

To arrive at test conformations and test compositions, we first generated some random heteropolymer sequences and tried to establish their ground state conformations. Ten random sequences RS_1 to RS_{10} for $L = 100$ were obtained by using the *a priori* probabilities $p(A_i)$ given in Table 1. These probabilities have been chosen so that residue types favoring helical conformations due to their ΔS or ΔE values are a little more abundant. Thereby, the existence of mostly α-helical ground-state conformations is ensured. Table 2 shows the resulting conformations and their sequence compositions obtained by Metropolis Monte-Carlo (MC) (33) simulations and/or Genetic Algorithms (34). After their respective ground states were established, each of the sequences was subjected to 100 runs of Metropolis MC simulation in conformation space, with 5000 MC steps each, starting in the all-c^+ state, a procedure that will be referred to as a "folding experiment." Folding performance was assessed by determining the number of MC runs (out of 100) in which the sequences encountered their respective ground states (successful MC runs), the time (number of MC steps) after which they first encountered it, and how much of the remaining simulation time they spent in the ground state once they had reached it. Our results, given in Table 3, indicate that although random heteropolymer sequences in general have a nondegenerate "native" state corresponding to the global minimum in F, only few of them will actually fold reproducibly and stably to this state. This phenomenon has been observed with other models of protein folding too (15).

With the ground-state conformations and compositions of Table 2 as a starting point, application of the procedure proposed by Shakhnovich and Gutin (16) is straightforward. Holding the conformation constant, we performed for each random heteropolymer 10 runs of Metropolis MC optimization in sequence space, with the constraint of constant composition. The runs were stopped after 10^4 MC steps each, and the optimum sequence encountered so far was chosen from every run. The 100 sequences S_1 to S_{100} thus obtained were further analyzed. First, for all of these sequences, we found that the value of F for the target conformation had been significantly lowered in all cases (data not shown). In a second step, each of the sequences was subjected to 10 folding experiments. Inspection of the results showed that 78 of the 100 sequences during the simulations had produced conformations lower in F than the target conformation and thus could already be dismissed. The 22 remaining optimized sequences again were selected for 100 folding experiments each. Six of 22 sequences failed to reach the target even once. The results for the remaining 16 sequences are given in Table 4.

Table 2. Test conformations and compositions, $L = 100$

	Composition	Conformation*
C_1	$A_{22}B_{16}C_{12}D_8E_{14}F_{10}G_6H_{12}$	$2_3 1_9 1_7 1_7 1_4 3_1 0_7 6_8 4_1 6_1 6$
C_2	$A_{18}B_{12}C_{13}D_{14}E_{20}F_{10}G_4H_9$	$11_1 8_1 8_1 5_2 9_1 9_5 13_1 13_5 5_2$
C_3	$A_{18}B_{14}C_{11}D_{16}E_{14}F_{12}G_7H_8$	$3_1 9_3 3_3 4_1 9_5 10_4 13_6 14_1 10_1$
C_4	$A_{20}B_{17}C_{10}D_8E_{13}F_9G_{11}H_{12}$	$1_3 1_3 1_4 5_5 5_5 1_9 1_1 5_1 5_2 1_5 1_7$
C_5	$A_{19}B_{13}C_{16}D_{10}E_{12}F_{12}G_7H_{11}$	$6_8 8_1 8_3 8_2 7_1 1_4 2_6 4_8 1_7 1_5$
C_6	$A_{18}B_{16}C_{13}D_9E_{10}F_{16}G_7H_{11}$	$1_1 0_3 1_2 1_1 4_6 7_5 7_1 3_7 4_2 4_1 4_1 7$
C_7	$A_{14}B_{17}C_7D_{10}E_{16}F_{13}G_{10}H_{13}$	$1_5 1_5 1_5 1_3 1_7 1_1 2_1 1_3 4_8 1_4 2_1 2_1 4_7$
C_8	$A_{18}B_{21}C_{13}D_{10}E_{14}F_{16}G_1H_7$	$1_1 0_8 1_1 1_3 6_1 3_1 1_8 2_9 1_1 4_2$
C_9	$A_{15}B_{21}C_{15}D_{10}E_{14}F_{10}G_{10}H_5$	$20_1 4_3 1_3 1_1 4_1 1_0 3_1 0_1 8_2 8_1$
C_{10}	$A_{21}B_{16}C_{12}D_9E_{15}F_{12}G_8H_7$	$15_1 20_3 23_1 1_1 3_6 1_5 2_9$

*For a description of the shorthand notation, see Fig. 1.

Table 3. Folding performance of random sequences

Conformation	Random sequence	Successful MC runs, no. of 100	Folding time (MC steps), mean ± SD	Average stability, %
C_1	RS_1	62	1795 ± 1224	7.5 ± 5.8
C_2	RS_2	46	1153 ± 909	4.1 ± 1.1
C_3	RS_3	0	—	25.8 ± 12.3*
C_4	RS_4	7	2890 ± 961	0.1 ± 0.05
C_5	RS_5	12	3187 ± 969	4.8 ± 6.1
C_6	RS_6	21	3263 ± 1128	1.7 ± 0.9
C_7	RS_7	48	2565 ± 1172	6.9 ± 10.0
C_8	RS_8	19	3139 ± 1277	25.9 ± 20.9
C_9	RS_9	2	5314 ± 74	10.4 ± 0.3
C_{10}	RS_{10}	21	2665 ± 1313	15.7 ± 4.2

*Determined from simulations starting in the target conformation.

While quantitative criteria for folding sequences are difficult to state, one should demand that a sequence folds rapidly and reproducibly to its target conformation, which in turn should correspond to a relatively stable structure. The degree to which this combined kinetic and thermodynamic criterion is met in repeated folding experiments can be measured simply by the average simulation time spent in the native conformation. Only 3 of 100 optimized sequences seem to meet this criterion—namely, sequences S_{80}, S_{81}, and, with reservations, S_{76}. Thus, the described procedure results in folding sequences for only 2 of the original 10 conformations.

Given a criterion of folding performance, there are a variety of stochastic optimization procedures one can choose

Table 4. Folding performance of optimized sequences

Conformation	Optimized sequence	Successful MC runs, no. of 100	Folding time (MC steps), mean ± SD	Average stability, %
C_1	S_6	85	1672 ± 1207	29.4 ± 8.2
C_2	S_{12}	96	899 ± 957	20.9 ± 3.5
	S_{13}	63	531 ± 579	37.9 ± 3.1
	S_{16}	68	1234 ± 981	28.9 ± 7.5
	S_{19}	80	1577 ± 1237	10.4 ± 3.9
C_3	S_{23}	19	1993 ± 1264	69.8 ± 20.5
C_4	—	—	—	—
C_5	—	—	—	—
C_6	S_{52}	59	1284 ± 690	15.7 ± 3.5
	S_{54}	17	1689 ± 1191	34.3 ± 16.2
	S_{59}	89	665 ± 688	41.4 ± 2.4
C_7	—	—	—	—
C_8	S_{76}	73	395 ± 522	58.1 ± 3.5
	S_{80}	74	480 ± 524	74.2 ± 4.7
C_9	S_{81}	97	1065 ± 853	71.5 ± 2.7
	S_{88}	99	1152 ± 725	33.8 ± 1.5
	S_{89}	52	1264 ± 1089	59.9 ± 2.1
	S_{90}	32	793 ± 681	25.9 ± 1.1
C_{10}	S_{100}	89	1338 ± 1009	7.5 ± 2.7

to optimize sequences. To construct folding sequences for our model, we have experimented with an optimization procedure that relies heavily on evolutionary processes and will be described in more detail elsewhere. It consists basically of two separate evolutionary optimization steps. In the first step, only stability of the target conformation is selected for. To this end, repeated Metropolis MC simulation runs, starting in the desired target conformation, are performed for a population of sequences, with the random sequences of Tables 2 and 3 as initial population. The total time spent in the target conformation is taken as the fitness of each sequence, according to which it is represented in the following simulation round. Random mutations and crossing-over allow for changes in the competing sequences, notably without any constraints on composition or even average composition. This optimization procedure is stopped when average stabilities of about 80% have been reached for the target conformation.

The second optimization step then starts with the population of sequences obtained in the first one. Now, the sequences are subjected to repeated folding experiments—i.e., simulation runs starting in the all-c^+ state. Again, the time spent in the target conformation is taken as their respective fitness. Thus, in this second optimization step, sequences are selected that fold rapidly and reproducibly to the target conformation and stay there for as long as possible.

It is important to note that the second optimization step alone is unlikely to produce folding sequences, since the probability for most random sequences to hit the desired target conformation is vanishingly small. Only after the first optimization step, which ensures that the target conformation is at least a relatively deep local minimum in conformation space, is there a sufficiently high probability for the sequences of the population to reach the target conformation when starting from the random coil state. Further details of the described procedure will be presented elsewhere.

In comparison to Tables 3 and 4, we present in Table 5 data for the folding sequences designed to fold to the conformations C_1 to C_{10} of Table 2. Success rates and folding times of all of these sequences are comparable to those of the best ones in Table 4, and their average stability is significantly higher. Note the large variances in the folding times given in Tables 4 and 5; this illustrates the necessity to study the folding behavior of test sequences by a sufficiently large number of folding experiments.

In addition to the results presented so far, we have carried through other lines of sequence analysis. It has been repeatedly stated that a necessary and sufficient condition for folding sequences in their model is that the native state is a pronounced energy minimum—i.e., that there is a large gap between the lowest and second lowest states in the energy spectrum (18, 19). Since a complete enumeration of states is not possible with our model, we have developed other strategies of obtaining the relevant data. Starting in the all-c^+ conformation, we performed 50 folding experiments for each sequence under study. During each simulation run, the 100 conformations lowest in F were stored so that, after 50 runs, somewhere between 100 and 5000 different conformations were obtained for each sequence. The spectra of Fig. 2 show the F values of these conformations up to a boundary of $F - F_{min} = 1$.

Looking at Fig. 2 *Top*, one first notes that the random sequences RS_1 to RS_{10} do indeed lack a large gap in their spectra. The little variation there is does not seem to be correlated with folding performance. Fig. 2 *Middle* and Table 4 show that the procedure of lowering the ground state in F with the constraint of constant composition does not always succeed in producing a notable gap in the spectrum, as one

Table 5. Folding performance of evolved sequences

Conformation	Optimized sequence	Successful MC runs, no. of 100	Folding time (MC steps), mean ± SD	Average stability, %
C_1	F_1	99	446 ± 414	90.1 ± 2.6
C_2	F_2	93	326 ± 379	70.9 ± 2.3
C_3	F_3	93	1513 ± 1005	84.1 ± 9.1
C_4	F_4	88	956 ± 828	82.8 ± 9.4
C_5	F_5	80	1521 ± 760	89.2 ± 2.2
C_6	F_6	88	1627 ± 841	64.5 ± 8.3
C_7	F_7	93	1172 ± 612	65.7 ± 5.9
C_8	F_8	83	909 ± 647	73.3 ± 4.4
C_9	F_9	94	1083 ± 920	75.9 ± 9.1
C_{10}	F_{10}	96	1183 ± 698	77.8 ± 4.2

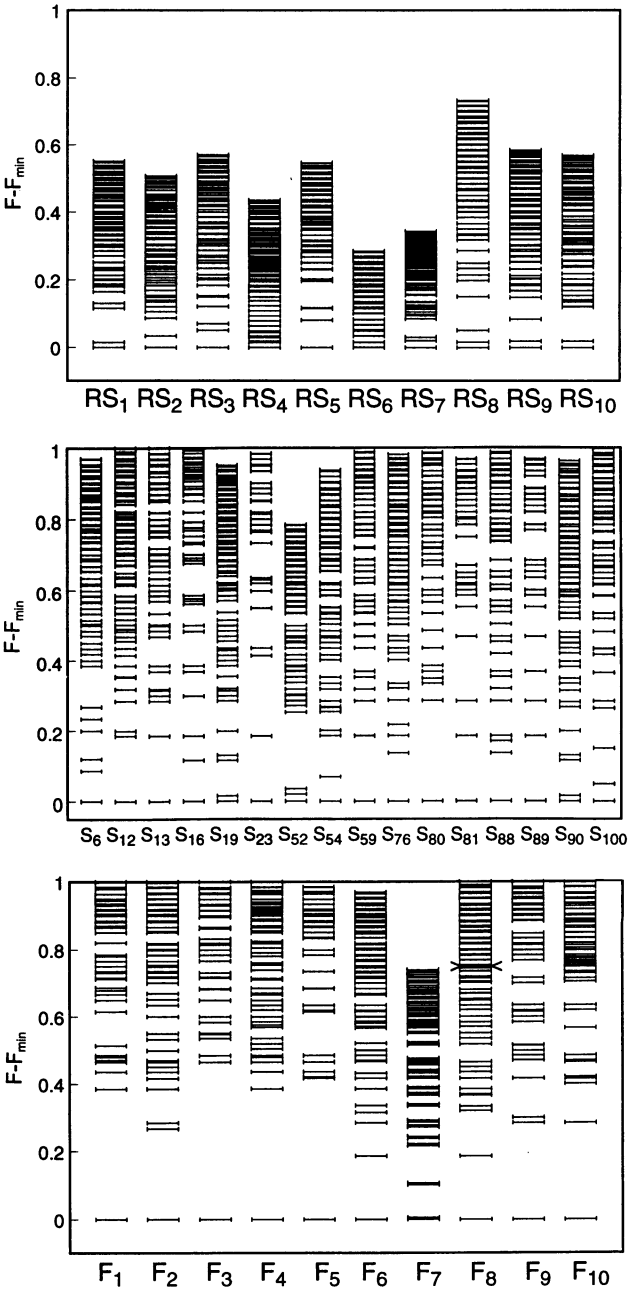

FIG. 2. Spectra of various sequences discussed in the text. Spectra are plotted relative to their lowest states and are not supposed to be complete in their upper parts. For RS_3, F_{min} does not correspond to the ground state given in Table 2 but does correspond to the lowest state found during folding experiments.

would expect from the analysis by Shakhnovich and Gutin (16) based on the REM. This indicates that the model presented here does not belong to the REM universality class.‡ In addition, the spectra of some nonfolding sequences (for example, sequences S_{13}, S_{23}, and S_{59}) exhibit gaps comparable to or wider than those of the folding sequences. Thus, a gap of the size found for those sequences is not a sufficient condition for a folding sequence. Most of the spectra of the folding sequences (Fig. 2 *Bottom*) do indeed show the expected large energy gap, which ensures thermodynamic stability of the ground state. Remarkably, however, it turns out that this is not a necessary condition for a folding sequence either. Sequence F_7 has two nearly degenerate states at the bottom of its spectrum, of which the only marginally lower one is stably attained in well over 90% of the folding experiments, obviously because of kinetic preferences. Even more interesting is F_8 which as a product of our simulated evolutionary process folds reproducibly and stably (on the timescale of millions of MC steps; data not shown) to its target conformation, although this is only a local minimum and lies well above the ground state in the spectrum (marked by arrows in Fig. 2 *Bottom*). Once again, kinetic factors must be responsible for this behavior. It is interesting to note that similar results have recently been reported for some proteins too (36). Therefore, we emphasize that to identify a folding sequence, one has to consider not only spectra of conformation energies but also possible routes connecting the corresponding conformations in conformation space. For example, one has to demand that there be routes via only moderate barriers of free energy from all of the other kinetically accessible local minima to the ground state. Results pointing in this direction have been obtained for other simplified protein folding models, too (9–12).

In ref. 17 Shakhnovich proposes a method, also based on the REM, to obtain a threshold value F_c for any given random sequence and target conformation. Folding sequences in their target conformation should lie well below this threshold value. We found, however, that none of the folding sequences presented in this paper comes even close to satisfying this criterion.

To conclude, the REM has been applied in several studies of protein structure and folding behavior with interesting results (26, 32, 37–40). However, our data indicate that some results derived from it do not hold for the model presented. Therefore, we suggest that its concepts be applied only carefully when studying the folding properties of protein-like heteropolymers. In particular, as the main result of this work, we object to the notion that folding sequences in general can be constructed by looking at the target conformation only, as may be possible for models belonging to the REM universality class. Rather, a host of other conformations that are different from the target may have to be taken into consideration. We suppose that for more realistic models, only optimization approaches that probe dynamic properties are suited to achieve this goal.

It is a pleasure to thank T. Krausche for continuing and stimulating discussions. W.N. also thanks A. Gutin and E. Shakhnovich for interesting and enjoyable discussions at an Aspen Center for Physics workshop. M.E. gratefully acknowledges support by a stipend from the Studienstiftung des Deutschen Volkes.

‡In ref. 35 it is argued on the basis of various approximations that heteropolymers in $d < 2$ dimensions do not belong to the REM universality class. However, the arguments presented there do not hold for $d = 2$. Therefore, we believe that essentially two-dimensional models like ours cannot be excluded *a priori* from that universality class.

1. Ebeling, M. & Nadler, W. (1993) *J. Chem. Phys.* **99,** 6865–6875.
2. Lifson, S. & Roig, A. (1961) *J. Chem. Phys.* **40,** 1963–1974.
3. Zwanzig, R. & Lauritzen, J. I. (1968) *J. Chem. Phys.* **48,** 3351–3360.
4. Lauritzen, J. I. & Zwanzig, R. (1970) *J. Chem. Phys.* **52,** 3740–3751.
5. Schulz, G. E. & Schirmer, R. H. (1979) *Principles of Protein Structure* (Springer, New York), Chaps. 5.3 and 8.3.
6. Grosberg, A. Y., Nechaev, S. K. & Shakhnovich, E. I. (1988) *J. Phys. (Paris)* **49,** 2095–2100.
7. Thomas, P. D. & Dill, K. A. (1993) *Prot. Sci.* **2,** 2050–2065.
8. Chan, H. S. & Dill, K. A. (1991) *J. Chem. Phys.* **95,** 3775–3787.
9. Miller, R., Danko, C., Fasolka, M. J., Balazs, A. C., Chan, H. S. & Dill, K. A. (1992) *J. Chem. Phys.* **96,** 768–780.
10. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.
11. Chan, H. S. & Dill, K. A. (1993) *J. Chem. Phys.* **99,** 2116–2127.
12. Camacho, C. J. & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6369–6372.
13. Yue, K. & Dill, K. A. (1993) *Phys. Rev. E* **48,** 2267–2278.
14. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 1942–1946.
15. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67,** 1665–1668.
16. Shakhnovich, E. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
17. Shakhnovich, E. (1994) *Phys. Rev. Lett.* **72,** 3907–3910.
18. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
19. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369,** 248–251.
20. Socci, N. D. & Onuchic, J. N. (1994) *J. Chem. Phys.* **101,** 1519–1528.
21. Poland, D. & Scheraga, H. A. (1970) *Theory of Helix–Coil Transitions in Biopolymers* (Academic, London, New York).
22. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963) *J. Mol. Biol.* **7,** 95–99.
23. Zimm, B. H. & Bragg, J. K. (1959) *J. Chem. Phys.* **31,** 526–535.
24. Toulouse, G. (1977) *Commun. Phys. (London)* **2,** 115–119.
25. Anderson, P. W. (1978) *J. Less Common Met.* **62,** 291–294.
26. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
27. Frauenfelder, H. & Wolynes, P. G. (1994) *Phys. Today* **47,** 58–64.
28. Goldenberg, D. P. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 353–404.
29. Richardson, J. S. & Richardson, D. C. (1989) *Trends Biochem. Sci.* **14,** 304–309.
30. Derrida, B. (1980) *Phys. Rev. Lett.* **45,** 79–82.
31. Derrida, B. (1981) *Phys. Rev. B* **24,** 2613–2626.
32. Shakhnovich, E. I. & Gutin, A. M. (1989) *Biophys. Chem.* **34,** 187–199.
33. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
34. Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems* (Univ. of Michigan Press, Ann Arbor).
35. Shakhnovich, E. I. & Gutin, A. M. (1989) *J. Phys. A* **22,** 1647–1659.
36. Sinclair, J. F., Ziegler, M. M. & Baldwin, T. O. (1994) *Nat. Struct. Biol.* **1,** 320–326.
37. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6,** 307–310.
38. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6,** 597–601.
39. Goldstein, R., Luthey-Schulten, Z. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4918–4922.
40. Goldstein, R., Luthey-Schulten, Z. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.