Martin Ebeling
Walter Nadler
Institut für Theoretische Chemie
Universität Tübingen
Auf der Morgenstelle 8
D-72076 Tübingen, Germany

Protein Folding: Optimized Sequences Obtained by Simulated Breeding in a Minimalist Model

For a minimalist model of protein folding, which we introduced recently, we investigate various methods to obtain folding sequences. A detailed study of random sequences shows that, for this model, such sequences usually do not fold to their ground states during simulations. Straightforward techniques for the construction of folding sequences, based solely on the target structure, fail. We describe in detail an optimization algorithm, based on genetic algorithms, for the "simulated breeding" of folding sequences in this model. We find that, for any target structure studied, there is not only a single folding sequence but a patch of sequences in sequence space that fold to this structure. In addition, we show that, much as in real proteins, nonhomologous sequences may fold to the same target structure. © *1997 John Wiley & Sons, Inc.*

INTRODUCTION

While a detailed, microscopic analysis of the protein folding process remains impossible using currently available experimental or computational techniques, simplified models of protein folding can be studied in order to learn about the generic properties of folding processes in heteropolymers. Common simplifications include the representation of amino acid residues by one or a few effective atoms, the analysis of chain dynamics in lattice spaces, the use of spaces with reduced dimensionality, and the substitution of the various different modes of interaction between residues by a simple contact potential. The aim of such minimalist models is not to model particular proteins of known amino acid sequence but, rather, to elucidate the necessary conditions for folding processes in linear heteropolymers. Such insights may also be useful for the future development of more realistic model treatments, since they might point out which features are crucial to the folding process and thus have to be modeled carefully.

Recently, we have introduced a simple model of secondary and tertiary structure formation in polypeptides,[1-3] based on the Lifson–Roig model for the helix–coil transition in amino acid homopolymers,[4] and on models for polymer crystallization.[5,6] The essential feature of the model is the combination of the helix–coil dynamics with a simplified form of tertiary interactions between α helices. For the homopolymer case, we have shown that the helix–coil transition with its continuous mode of transition and its tendency to produce only a single, long helix in the low temperature regime is modified to exhibit a first-order phase transition, as to be expected for proteins, too, and to establish structures that resemble globular proteins in average helix number and average helix length.[1,2,7] Since our model is easily implemented and does not require excessive computational effort, it can already be studied on personal computers. On the other hand, given access to modern workstations, it becomes possible to study large ensembles of sequences and to obtain reliable statistics derived from many simulation runs. This

Received March 4, 1996; accepted May 22, 1996.
Biopolymers, Vol. 41, 165–180 (1997)
© 1997 John Wiley & Sons, Inc.
CCC 0006-3525/97/020165-16

seems especially important in the light of recent findings that suggest proteins do not usually fold by a single, specified pathway, but use a multitude of possible routes to reach their target structures.[8,9] We note that our model emphasizes secondary structure dynamics, and therefore differs from most other model systems currently under study.[10–20] Since generic properties of folding are expected to hold for all model types, this offers the change to reexamine results obtained with other models.

When studying a minimalist model, one has to demonstrate, first of all, that the model shows dynamic processes resembling protein folding, i.e., that there are folding sequences. In some cases, such sequences can be obtained from simple inspection by introducing interaction patterns that favor native substructures.[21–23] However, this is not likely to hold for all protein models, let alone for proteins. Therefore, in recent years, there have been attempts to devise general algorithms for obtaining folding sequences in simple protein models. For the heteropolymer case of our model, we have shown that random sequences usually do not show folding-like behavior.[3] Folding sequences could not be obtained using methods described for other simplified models of protein folding.[3,17] However, we were able to devise an optimization procedure that yielded folding heteropolymer sequences for all target conformations studied.[3] Here, we describe and discuss our results in more detail.

DESCRIPTION OF THE MODEL

The conformation of a polypeptide of length L is represented 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. Two helices separated solely by c^0 residues are assumed 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 to model the formation and disruption of tertiary contacts between helices. The free energy of a sequence $\{A_i\}$ in conformation $\{\sigma_i\}$ is given by

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

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

Three successive monomers must be in helical conformation in order to be spanned by a hydrogen bond. Therefore, if $\sigma_{n-1} = \sigma_n = \sigma_{n+1} = h$, there is 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 zero 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. 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 contribute otherwise to F.[24] Note that in random coil stretches, identified by the absence of hydrogen bonds, some monomers (or even pairs of monomers) may attain the helical conformation and will then be labeled h, accordingly. However, these monomers will not contribute to the hydrogen bond term in Eq. (1) and will not be considered part of helices.

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 (Figure 1). All 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 $C_{n,m}(\{\sigma_i\}) = 1$ if residues n and m are in contact in chain conformation $\{\sigma_i\}$, and zero otherwise. The contact energy is simply taken as the mean of the respective contact parameters k of the two residues in contact.

Finally, the entropic term represents contributions to the system's entropy which arise from *local* conformation space restrictions. We denote by $V(\sigma_i, A)$ the conformation space volume accessible to a monomer of type A with local conformation σ_i. We then define the local conformational entropy by $\Delta S(\sigma_i, A) = k_B \ln[V(\sigma_i, A)/V(c^+, A)]$. By this definition, the conformational entropy of the local conformation c^+ is arbitrarily set to

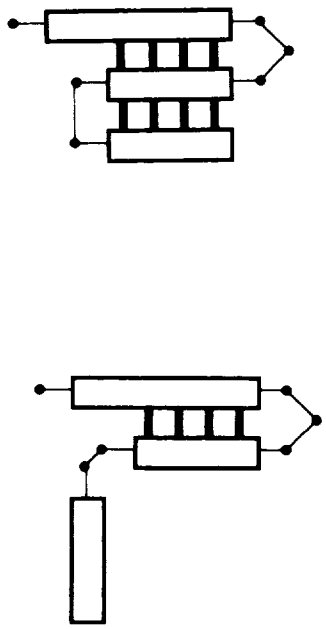

FIGURE 1 Tertiary interactions between α helices, as described in the model. (Top) Three helices of lengths 6, 4, and 4 monomers, respectively (open rectangles), with two interhelical contacts (black bars between helices). (Bottom) The contact between the second and third helices has been broken.

zero for all monomer types A, which is convenient since we are only concerned with relative entropies in the following. For simplicity, we also set $\Delta S(c^0, A) = 0$ for all residue types A, i.e., we assume equal a priori probabilities for the occurrence of the two different random coil conformations. [Note that Eq. (1) covers the more general case of $\Delta S(c^0, A) \neq 0$.] Since the conformation space volume $V(h, A)$ accessible to monomers in h conformation is smaller than that for monomers in random coil conformations, as mirrored in a Ramachandran plot,[25] the conversion $c^+ \leftrightarrow h$, for example, is accompanied by a change in conformational entropy.

Chain conformations can often be characterized by what will be called the chain *structure* $\mathcal{C}$, specified by the positions and lengths of helices and the positions of interhelical contacts. Chain structures will be described by the following shorthand notation: (1) lengths of helices are set in boldface, (2) lengths of interhelical loops are set as subscripts between the helix lengths, and (3) loop lengths are set in brackets when there is no contact between the two adjacent helices. The chain structures in Figure 1 are $\mathcal{C}_{\text{top}} = {}_1\mathbf{6}_3\mathbf{4}_2\mathbf{4}$, and $\mathcal{C}_{\text{bottom}} = {}_1\mathbf{6}_3\mathbf{4}_{[2]}\mathbf{4}$. Since the local conformation of the first monomer in structure $\mathcal{C}_{\text{top}}$ may be either c^0 or c^+, *two* different chain conformations correspond to this structure. For chain structure $\mathcal{C}_{\text{bottom}}$ in the figure, in addition, the local conformations of the two monomers between the second and third helices may be either one of the three possibilities c^0c^+, c^+c^+, or c^+c^0 (but not c^0c^0, as no contact is established), yielding a total number of *six* different chain conformations for this structure. For a more detailed description of the relation between chain conformations and chain structures, see Appendix A.

Any sequence of residues $\{A_i\}$, $i = 1, \ldots, L$, is completely characterized by the parameter sets $\{\Delta E(A_i)\}$, $\{k(A_i)\}$, and $\{\Delta S(h, A_i)\}$. We allow two values for each of these parameters and consider the $2^3 = 8$ monomer types resulting from the combination of the possible parameter values as described in Table I. All the results presented in the following have been obtained at a constant temperature with $k_BT = 0.108|\Delta E(\mathbf{A})|$. The system size is $L = 100$ monomers for the results presented here.

Since $\Delta S(h) < 0$ for all monomer types (Table I), the entropic term in Eq. (1) will dominate the others in the high temperature range, and the random coil conformations c^+ and c^0 will prevail, independently of the polymer sequence. For low temperatures, however, the energetic contributions from hydrogen bonds and tertiary contacts will cause the formation of helices and interhelical contacts. Precisely which chain structures will be established depends on the polymer sequence. In the following, the chain structure corresponding to the chain conformation with the lowest value of F at the above temperature will be referred to as its *ground state*.

CHARACTERIZATION OF RANDOM SEQUENCES

Ground States

In Ref. 3, we have discussed the phenomenon of frustration in the model presented here. In frus-

Table I Description of Monomer Types

	Parameter Values			
Monomer	ΔE	$\Delta S(h)/k_B$	k	p
A	−1.0	−2.0	−0.6	1/6
B	−1.0	−2.0	+0.3	1/6
C	−1.0	−3.567	−0.6	1/8
D	+0.5	−2.0	−0.6	1/8
E	−1.0	−3.567	+0.3	1/8
F	+0.5	−2.0	+0.3	1/8
G	+0.5	−3.567	−0.6	1/12
H	+0.5	−3.567	+0.3	1/12

trated systems, the various *local* conditions for energy minimization cannot all be satisfied simultaneously, i.e., the system state with minimum *global* energy will necessarily contain local substructures with unfavorable energy. The phenomenon of frustration was first described for spin glasses[26,27] but is known to occur in proteins, too.[28,29] Frustrated systems are characterized, in general, by complicated energy landscapes with various structurally different local minima separated from each other by high barriers. The ground state has to be sought among all these minima, and will therefore be difficult to obtain. Deterministic algorithms, like the method of steepest descent, will only find *local* minima in the vicinity of the starting point of the optimization. The *global* energy minimum cannot be found reliably with such methods. In principle, one would have to search the whole conformation space systematically to make sure that a given minimum is indeed the global one.

In recent years, various stochastic optimization algorithms have been developed for studying such complicated energy landscapes.[30-32] These algorithms do not search the whole state space of the system, and thus are not guaranteed to find the global energy minimum either. However, they usually produce system states with very low energy, i.e., at least close to the global minimum, within short times. Stochastic optimization algorithms are usually applied several times to a problem; the lowest state found in any run is then taken to be the ground state.

Since we expect frustration to occur in random sequences for the model presented here,[3] we have followed two different stochastic optimization strategies to identify structures low in F. The first procedure is the Metropolis–Monte Carlo (MC) simulation scheme. As an elementary step, a monomer is chosen randomly, and an attempt is made at altering its local conformation. Such an attempt is accepted or rejected according to the Metropolis criterion.[33] This criterion ensures an overall tendency to go downhill in F yet, at the same time, enables the system to go uphill occasionally, so that it is not necessarily trapped in local minima. Therefore, the Metropolis algorithm is suited to explore rough energy landscapes. A Monte Carlo step consists of L elementary steps, during which, on the average, each monomer is chosen once for an attempt at changing its conformation. The optimization procedure employed here consists of Metropolis–MC computer simulations of 5000 MC steps length, starting in the $(c^+)_L$ chain conformation. Such a simulation will be called a "folding experiment" in the following.

As a second, completely different approach to structural optimization, we have used a genetic algorithm[30,34] specified as follows. Chain conformations are represented as chromosomes with L genes. (A) As a starting point, a population of $N = 4000$ chromosomes is generated by randomly assigning the alleles h and c^0, with $p(h) = 0.8$ and $p(c^0) = 0.2$, respectively, to their genes. For convenience, the local conformation c^+ is not considered here, because compact chain structures, where all possible interhelical contacts are established, are expected for most of the ground states. (B) As the target function to be optimized, we choose $\mathcal{F}(\{\sigma_i\}) = \ln[-1000(F(\{\sigma_i\}) - 1)]$, which increases monotonically with decreasing F. (C) Chromosomes are perpetuated to the next generation following the "remainder stochastic sampling with replacement" algorithm[34] (see Appendix B for details). (D) We choose a mutation rate of $p_{mut} = 0.005$ per gene per generation, and simple crossing-over with $p_c = 0.25$ per chromosome per generation. This procedure is repeated over 500 generations, and the chromosome, i.e., the chain conformation, lowest in F is retained. It is then subjected to 500 MC steps of the Metropolis–MC simulation as described above, in order to check for even better conformations in its vicinity in conformation space. (In addition, during these short MC simulation runs, monomers in c^+ conformation may be introduced to allow for noncompact ground states.)

The above described methods were applied to a set of 110 random sequences that were generated from the monomer types of Table I, using the a priori probabilities p given in the table. For each random sequence, 50 folding experiments and 50 structural optimizations by genetic algorithm were done. For two of the 110 sequences, the lowest value of F found during these optimization procedures was fourfold degenerate, i.e., represented by four different chain structures; 16 other sequences had two different chain structures with minimum F. (We note that degenerate chain structures for any particular sequence resemble each other closely in all the cases observed.) All in all, the structural optimization procedures yielded a total of 122 chain structures for the 110 random sequences studied. Fifteen of these 122 chain structures were found to be noncompact, i.e., not all possible interhelical contacts were established due to unfavorable interactions between some of the helices. In 85 of these 122 cases, both optimization techniques yielded iden-

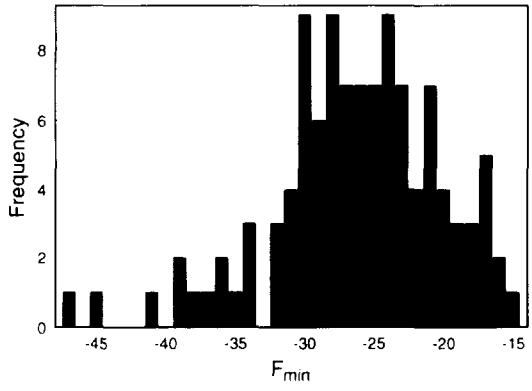

FIGURE 2 Distribution of lowest F values for 110 random sequences, as determined by folding experiments and/or genetic algorithm as described in the text.

tical results; in 7 cases, the genetic algorithm found the lowest conformation in F, in 30 others, the Metropolis algorithm (folding experiments) was more successful. The good agreement of two independent optimization procedures indicates that, in most cases, the true ground states have been identified. Figure 2 shows the distribution of the F values of the ground states for all 110 sequences.

Folding Behavior

Given the ground states of the random sequences under study, we next ask for their folding performance. As a *folding criterion*, we expect a folding sequence to reach its target structure reproducibly during repeated folding experiments, *and* to be stable in this structure. In addition, we require folding sequences to fold *rapidly* to their target structure. This third part of the criterion may appear somewhat artificial, since there is no evidence that particularly fast-folding sequences have actually been selected during natural evolution. However, proteins employ folding mechanisms that are rapid at least compared to an unbiased random search in conformation space, a feature known as Levinthal's paradox.[35] In addition, we include this criterion for practical reasons. Fast-folding sequences will allow us to study many folding trajectories, and thus to elucidate efficient folding strategies, which is an especially important task for a minimalist model of protein folding.

The folding experiments described in the foregoing section already reveal that, in general, random sequences do not fold reproducibly to their respective ground states: for more than half of the sequences, the ground state is reached in fewer than 10% of the folding experiments (data not shown). In addition, simulations were performed with the sequences starting in their ground state structures. It was determined how much simulation time was spent, on the average, in the respective chain structure. The results of the folding and stability studies are summarized in Figure 3. Given the above folding criterion, folding sequences are to be sought in the righthand lower back corner of the figure. Only very few of the studied 110 random sequences come even close to being folding sequences; most of them are clustered in the lefthand upper front corner of the figure. While several of the sequences reach their ground states reproducibly, they are usually not stable there. On the other hand, some sequences are comparatively stable in their target structures once they have reached them, but fail to do so in most of the folding experiments.

CONSTRUCTION OF FOLDING SEQUENCES

Straightforward Approaches to the Problem

In the foregoing section it was shown that folding sequences are not readily obtainable from a ran-

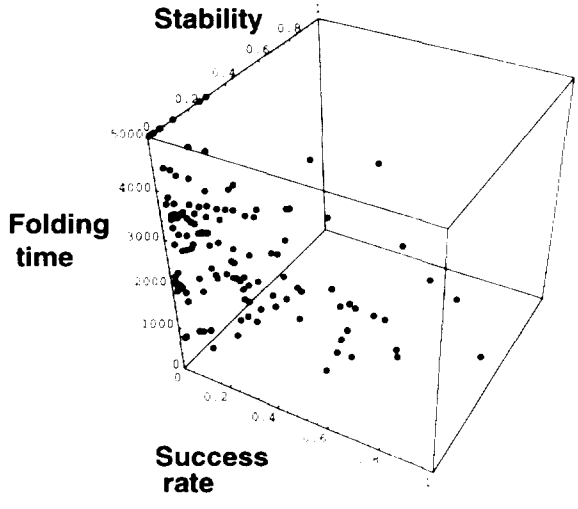

FIGURE 3 Folding performance as mirrored by average success rate and average folding time (from folding experiments), and average stability in the ground state, for 110 random sequences. Sequences for which the ground state was determined by genetic algorithm, i.e., which did not fold to the ground state, are shown with zero success rate and maximum folding time (upper left). Note that for sequences with low average success rate, the values of average folding time and average success rate are to be viewed as rough estimates only. Average stability values were determined separately in 50 simulations each, starting in the ground state.

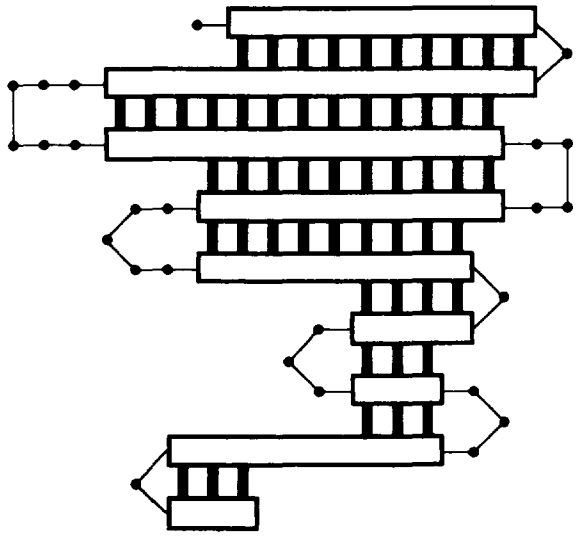

FIGURE 4 Chain structure $\mathcal{C}_3$ (Table III) shown as described in Figure 1.

dom search in sequence space. That leaves us with the task of demonstrating that there are sequences that fulfill the above folding criterion in the model presented here. In addition, one would like to be able to construct sequences that not only fold, but fold to a given specific structure. Thus, the question of sequence design arises.

In Ref. 2, it was shown that the construction of folding sequences is a nontrivial task in the model presented here. As a first approach, one might consider to look for a sequence that minimizes F for a given target structure. However, such a sequence optimization might even yield a type A homopolymer, since monomers of type A (Table I) optimize all possible interactions in any chain conformation. In fact, the minimum value of F for any chain conformation $\{\sigma_i\}$ is given by $F_{\min}(\{\sigma_i\}) = \min_{\{A_i\}} F(\{\sigma_i\}, \{A_i\}) = F(\{\sigma_i\}, \mathbf{A}_L)$, where L denotes the number of monomers in the chain. However, during folding experiments, the type A homopolymer will fold to various different chain structures (compare with Refs. 1 and 2); obviously, it cannot be a folding sequence for *any* given chain structure. Therefore, optimization with the constraint of constant composition has been proposed to prevent the optimization algorithms from converging to a homopolymer.[17] However, in Ref. 3 it was shown that, for the model studied here, this optimization scheme does not produce folding sequences either.

Another approach is to selectively stabilize native interactions.[21,22] Taking the chain structure $\mathcal{C}_3$ (Figure 4 and Table III below) as an example, one can choose for any position the monomer type that favors *only* those interactions occurring at this position, but disfavors all other interactions. By this procedure, one arrives at the sequence

$$\{A_i\}_3 = \mathbf{D_3EA_3B_3A_3EHED_3EHEAD_2AEA_9EH_3EA_{10}EH_2EA_{13}EH_4EA_{14}EA_{10}E}$$

which still satisfies $F(\mathcal{C}_3, \{A_i\}_3) = F_{\min}(\mathcal{C}_3) = -73.30$. However, in 10 folding experiments, the sequence $\{A_i\}_3$ invariably yielded chain structures *lower* in F, among them $\mathcal{C}_{3'} = \mathbf{3_1 15_3 14_1 17_1 13_6 14_1 11}$ with a value of $F(\mathcal{C}_{3'}, \{A_i\}_3) = -75.78$. Therefore, the sequence $\{A_i\}_3$ with optimized native interactions does neither fold to the target structure $\mathcal{C}_3$, nor have it as its ground state. Note that in $\mathcal{C}_{3'}$, some of the monomers are *locally frustrated* in order to enable the chain as a whole to form a structure lower in F. This frustration is the reason that $F(\mathcal{C}_{3'}, \{A_i\}_3) > F_{\min}(\mathcal{C}_{3'}) = -94.01$.

Figure 5 illustrates these findings schematically: straightforward attempts to construct folding sequences, which rely solely on the target structure as input information, fail for the model presented here. Rather, they usually produce sequences that either do not fold at all or fold to chain structures different from the target structure.

An Algorithm for "Simulated Breeding" of Folding Sequences

Given the above folding criterion, the search for a folding sequence can be viewed as an optimization

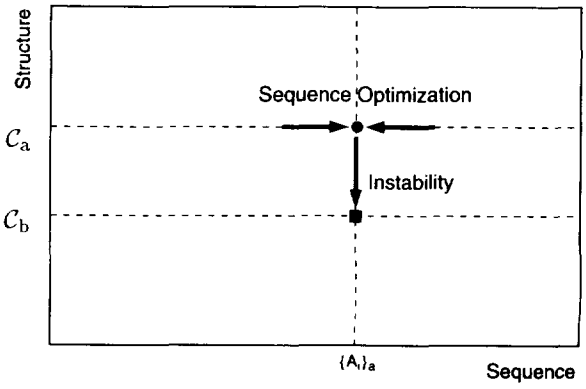

FIGURE 5 Schematic minimization of F in sequence and structure space. Holding the structure fixed at $\mathcal{C}_a$, one can always find a sequence $\{A_i\}_a$ with $F(\mathcal{C}_a, \{A_i\}_a) = F_{\min}(\mathcal{C}_a)$ by sequence optimization. However, for this sequence, there will usually be another chain structure $\mathcal{C}_b$ with $F(\mathcal{C}_b, \{A_i\}_a) < F(\mathcal{C}_a, \{A_i\}_a)$, resulting in structural instability.

Table II Random sequences, $L = 100$

	Composition	Sequence
RS_1	$A_{22}B_{16}C_{12}D_8E_{14}F_{10}G_6H_{12}$	GADFDBBAHBHEAAABFFDACCBCA AEHEEAHGBBCFBGACBECFCAAEH EEDCEHACDADHBGFFHEBEAEBEB FHFGBHBCDAAEHCDCAABGFAHAA
RS_2	$A_{18}B_{12}C_{13}D_{14}E_{20}F_{10}G_4H_9$	ACFBFABADBHECGECAADDHDABE ACHCEAHDADHBFEABABFFHECDD CAEEAAFFFGHDCCAHBDAEEEEAB GFCBECEBCEDDAHEEEEGCDDBEF
RS_3	$A_{18}B_{14}C_{11}D_{16}E_{14}F_{12}G_7H_8$	AADEDAEAEEDCBAHBDDADHCDED BEDADCEBFAFEEHGBGEBGAFFDA BBFHBCGFDFCHAFCEFGBGEFHAB FDECABHABACDECDHFCABCGDAA
RS_4	$A_{20}B_{17}C_{10}D_8E_{13}F_9G_{11}H_{12}$	AHADBBDDEDHEAEHGHGAEHBABH FEHACFGEEEGBEABAGGHHFCAAA ACHFEBBCHBAGDCCAFBEACGFGD EBBEFDBAFBGCBBDAAAGFBAHCC
RS_5	$A_{19}B_{13}C_{16}D_{10}E_{12}F_{12}G_7H_{11}$	GHCABAHDFGFDHBFCFCACBFBAD CCBCCDBHHGAEACGACFBAFDADA EEADFEACBBCGAHFBEEAHGEFAE DFEBGCAAAHEBHBDCAHCHCEFED
RS_6	$A_{18}B_{16}C_{13}D_9E_{10}F_{16}G_7H_{11}$	HBFAECEEHAFAFFAGCBAACCADF EBFDABHBHBBAADCBHFHBDEBDD BCAFEHHFGCBDBEGGCFDDBBFEG HBCGFFCHAGFFEACFCEHAAAAAC
RS_7	$A_{14}B_{17}C_7D_{10}E_{16}F_{13}G_{10}H_{13}$	HDBDBAEGDAACEFBFBCHDGDFGA EEEEFBGHDDBCABBFFAFHGEBEH HEFCGADEHHEEEABBCGGCAGFAE BFCHABHEEADBGBBFDAFBHHHAF
RS_8	$A_{18}B_{21}C_{13}D_{10}E_{14}F_{16}G_1H_7$	BEDAADBAEFAEDFFFCHBEBFECB FECGFBDFCHDFHCBFADEAABBBA ABBBAFFABEFCECAAAFEFEBEBD ADCBCHDABAHFCCHCABEEDCBBH
RS_9	$A_{15}B_{21}C_{15}D_{10}E_{14}F_{10}G_{10}H_5$	BCBEDEHEFBECFBBFEBCFCCGCA GHEADFAEEBBBACHCEBDDECCAC BFBBGAEDDBCGAFGHABFECBAAC DAEDGBDBFAGGGGBBACFEADBAH
RS_{10}	$A_{21}B_{16}C_{12}D_9E_{15}F_{12}G_8H_7$	AFBAFABBAEADBBBCDDFECCBAA DGEADACEFHACHECHFFAAEBGBH ADAEAEBCAFFBEFFBEAAEAEAGE HCFGDCDDBGGFGAECBGHBCEHCB

problem: For any given target structure, look for a sequence that spends as much simulation time as possible in the target structure during folding experiments. We have devised an optimization procedure, based on genetic algorithms, which achieves this goal.

Of the 110 random sequences described above, ten were chosen randomly. These sequences, labeled RS_1 to RS_{10}, are specified in Table II. In Table III, the ground states for these sequences are given. Figure 4 shows the chain structure $\mathcal{C}_3$ as an example. In the following, we adopt the chain structures of Table III as target structures and start our optimizations with the corresponding random sequences.

Since, for a given target structure, it will, in general, be impossible to construct a sequence that hits the target structure at all, we first look for sequences for which the target structure corresponds to a pronounced, albeit not necessarily *global*, minimum in F, i.e., we select for stability in the target structure. The genetic algorithm is formulated as follows: (A) For target structure $\mathcal{C}_i$ (Table III), we

Table III Ground States of Random Sequences, $L = 100$

Sequence	Chain Structure	F
RS_1	$\mathcal{C}_1 = {}_2\mathbf{3}_1\mathbf{9}_1\mathbf{17}_1\mathbf{14}_3\mathbf{10}_7\mathbf{6}_8\mathbf{4}_1\mathbf{6}_1\mathbf{6}$	−30.09
RS_2	$\mathcal{C}_2 = \mathbf{11}_1\mathbf{8}_1\mathbf{8}_1\mathbf{5}_2\mathbf{9}_1\mathbf{9}_5\mathbf{13}_1\mathbf{13}_5\mathbf{5}_2$	−30.12
RS_3	$\mathcal{C}_3 = \mathbf{3}_1\mathbf{9}_3\mathbf{3}_3\mathbf{4}_1\mathbf{9}_5\mathbf{10}_4\mathbf{13}_6\mathbf{14}_1\mathbf{10}_1$	−21.95
RS_4	$\mathcal{C}_4 = {}_1\mathbf{3}_1\mathbf{3}_1\mathbf{4}_5\mathbf{5}_5\mathbf{5}_1\mathbf{9}_1\mathbf{15}_1\mathbf{15}_2\mathbf{15}_1\mathbf{7}$	−24.65
RS_5	$\mathcal{C}_5 = \mathbf{6}_8\mathbf{8}_1\mathbf{8}_3\mathbf{8}_2\mathbf{7}_1\mathbf{14}_2\mathbf{6}_4\mathbf{8}_1\mathbf{7}_1\mathbf{5}$	−28.35
RS_6	$\mathcal{C}_6 = {}_1\mathbf{10}_3\mathbf{12}_1\mathbf{14}_6\mathbf{7}_5\mathbf{7}_1\mathbf{3}_7\mathbf{4}_2\mathbf{4}_1\mathbf{4}_1\mathbf{7}$	−24.52
RS_7	$\mathcal{C}_7 = {}_1\mathbf{5}_1\mathbf{5}_1\mathbf{5}_1\mathbf{3}_1\mathbf{7}_1\mathbf{12}_1{}_1\mathbf{3}_4\mathbf{8}_1\mathbf{4}_2\mathbf{12}_1\mathbf{4}_7$	−18.08
RS_8	$\mathcal{C}_8 = {}_1\mathbf{10}_8\mathbf{11}_1\mathbf{3}_6\mathbf{13}_1\mathbf{18}_2\mathbf{9}_1\mathbf{14}_2$	−31.33
RS_9	$\mathcal{C}_9 = \mathbf{20}_1\mathbf{4}_3\mathbf{13}_1\mathbf{14}_1\mathbf{10}_3\mathbf{10}_1\mathbf{8}_2\mathbf{8}_1$	−31.67
RS_{10}	$\mathcal{C}_{10} = \mathbf{15}_1\mathbf{20}_3\mathbf{23}_1\mathbf{11}_3\mathbf{6}_1\mathbf{5}_2\mathbf{9}$	−32.56

take $N = 10$ identical copies of sequence RS_i (Table II) as the starting population. (B) For every sequence, its fitness $\mathcal{F}$ is determined as the average relative amount of simulation time spent in the target structure during $n = 10$ short MC simulations (usually 200 MC steps), starting in the target structure. (C) The next generation is then built up following the "remainder stochastic sampling with replacement" algorithm[34] (Appendix B). (D) For every position in every sequence, the monomer type at that position is mutated with $p_{mut} = 0.005$ to *any other* monomer type with equal probability, i.e., neither the sequence compositions nor the a priori probabilities of Table I are conserved; every sequence is chosen with $p_c = 0.25$ for simple crossing-over at a random site with a random partner.

Figure 6 (left-hand part) illustrates the optimization for three of the studied cases. The *fitness* $\mathcal{F}$, i.e., the relative amount of simulation time spent in the target structure, is shown as a function of the generation number. In all of the 10 studied cases, we obtained sequences that spent well over 50% of the total simulation time in their respective target structures after 150 to 250 generation cycles.

We select for folding sequences in a second optimization step. The corresponding genetic algorithm runs as follows: (A) We use the resultant population of the first optimization step as a starting population. (B) For each sequence, its fitness $\mathcal{F}$ is determined as the average relative simulation time spent in the target structure during 10 folding experiments, i.e., simulations over 5000 MC steps, starting in the chain conformation $(c^+)_{100}$. (C) and (D) the selection mode for the next generation and all other parameters remain as described above.

Figure 6 (right-hand part) shows the time course of three folding optimizations. After 70–230 generations, we obtained sequences that, averaged over ten folding experiments, spent over 60% of the simulation time in the target. Note that this value is lower than the maximum attainable stability, since the folding process itself requires some time. For each of the ten optimization procedures performed, the sequence with the maximum observed fitness was retained. These folding sequences are labeled F_1 to F_{10} and are given in Table IV.

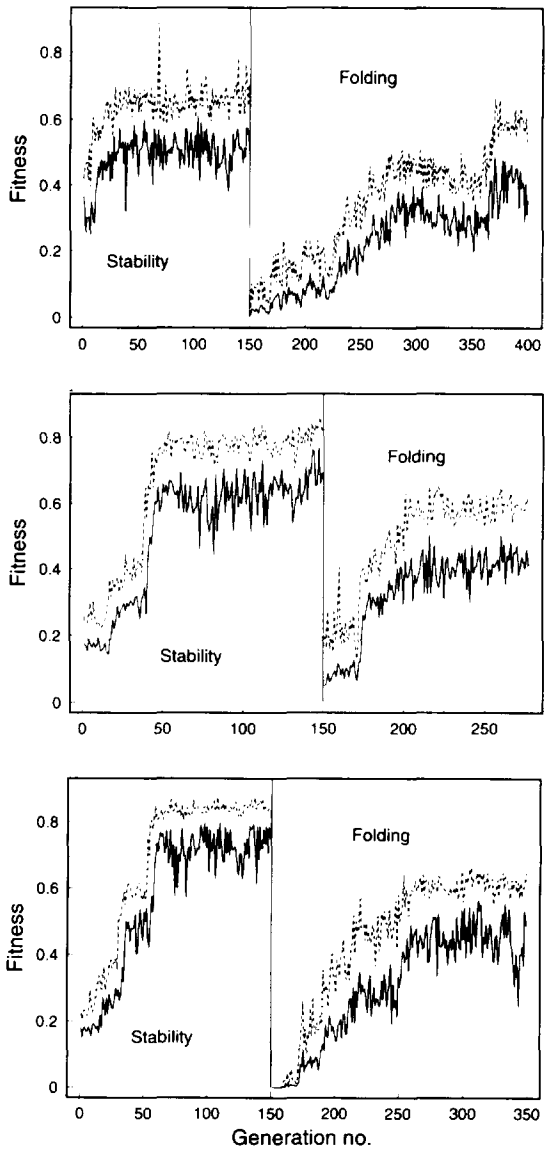

FIGURE 6 Fitness of sequences during stability optimization and folding optimization, respectively. (Solid line) Average fitness in the population. (Dashed line) Maximum fitness in the population. (Dotted line) Maximum fitness encountered so far during the optimization. (Top) Optimization for chain structure $\mathcal{C}_8$. (Middle) Optimization for chain structure $\mathcal{C}_9$. (Bottom) Optimization for chain structure $\mathcal{C}_{10}$.

Table IV Folding sequences, $L = 100$

	Composition	Sequence
F_1	$A_{21}B_{14}C_{11}D_9E_{16}F_9G_9H_{11}$	GEGCDBBCDAHEBAEBDFGAAEBAG BFBECAHGEFCDBGAEBECCFADEH EEDFBAACDEDHBGHGHEHEAABAB HHFFFFECDHAECACAAABGCAHAA
F_2	$A_{20}B_{19}C_{14}D_{10}E_{15}F_{12}G_5H_5$	AEFBBHCADABEAGBCAADGEDABB ACHCEFCDADHBGAABABBFEECBD AAEECABFFGHDCDCEAAAGEBEBB AHFBCCEBCBFDAFFFFECCDFBEE
F_3	$A_{17}B_{10}C_{11}D_{11}E_{16}F_{17}G_{10}H_8$	DHDECABEEEGEFBHEDDGFDEGFD CBDEHAEAFDFEHHGBCAFGGCADE FBFHBCFFDACAEAEFGBEHEFFAG FFECAAAAAAGDBACHFACBCGFCB
F_4	$A_{17}B_{12}C_{14}D_{14}E_{18}F_9G_{10}H_6$	EHADEBDCEDCCAEHGHGAEDAFBF FEHACFGDEDABEAHAEGEDGDAAD ECAFCBBCGEAGDDCABEEACFFDG EEDEFDBAEBGGCBAHABFCBBCCC
F_5	$A_{21}B_{10}C_{12}D_{11}E_{14}F_{12}G_{10}H_{10}$	AGBADAHHEDBHFBCAFGCCECEAH GEBDEABHHGAGACFDDFFADGAHA DEBDFACEEBCACFFAFEAFAHCAH HFECGGAAGBGEBDCAEFCEDBDEA
F_6	$A_{18}B_{14}C_{13}D_{10}E_{16}F_{12}G_8H_9$	HAAEEGECFBFEFEAGGBCEBCADE EBABAECBHABAADEDCHHFDCBDF CGADFGHFEEBDBFBACGDDFCHEH HBAAFFBHADFGEECACEGAACBCA
F_7	$A_{13}B_{20}C_{13}D_{13}E_{11}F_{10}G_8H_{12}$	ADBHBAECGFDCEFEDCCBGDDBCC EEHACBDHDCAEABBFAABHGGFHH BEHBDGDADHEDBFBBCCGBDGFAE BFCHABEBCAHBFCAFDABHGHFBE
F_8	$A_{13}B_{23}C_{13}D_{10}E_{13}F_{17}G_3H_8$	HCBABDAAAFAHDFFHHFCCFFBCB FECAFBDFCHBGHCBGADEBABEEC BFDBFEDABGBBFCCAABEFEFBED BDCCBHFAEBAFECDBEFBEDEBBH
F_9	$A_{17}B_{23}C_{13}D_6E_{15}F_8G_9H_9$	BABECFAGCBECBBBEEBCFCCGCA HHEDBFABEBBBAHHFEBDBEAHAC BAACHFBGDBCEAFGHAHFEBBHAC BEEDGBDCAAEGGGBBAAFEADCGE
F_{10}	$A_{22}B_{17}C_{10}D_{11}E_{12}F_5G_{11}H_{12}$	ADDAFACBAEBEDHACDDGACCAHD DAFAEABCGFHCHEADFHBACBBBH BGAEDEEBAHHGEBHAEADBGGAAE HBHGDBCABGGGGAHEBFACCEABB

With each of the folding sequences of Table IV, 100 additional folding experiments were performed. The results are summarized in Table V and illustrated in Figure 7. A comparison of the random sequences (spheres) and the optimized sequences F_1 to F_{10} (cubes) in the figure shows that folding sequences have, indeed, been generated by the described procedure. Given these folding sequences, it is now possible to study their thermodynamic and kinetic characteristics and to compare them to those of non-folding sequences. In Ref. 2, we have shown that the folding sequences obtained for our model do not necessarily have spectra with a wide gap between the lowest and second lowest possible values of F, which has been proposed as a general criterion for folding sequences.[18,19] In fact, we found that one sequence, F_8, folds fast and reproducibly to a target structure that is *not* the ground state for this sequence.[3] The target structure, $\mathcal{C}_8 = {}_1\mathbf{10}_8\mathbf{11}_1\mathbf{3}_6\mathbf{13}_1\mathbf{18}_2\mathbf{9}_1\mathbf{14}_2$, with $F(\mathcal{C}_8) = -26.573$, differs from the ground state structure, $\mathcal{C}_{8'} = {}_1\mathbf{10}_8\mathbf{11}_1\mathbf{3}_2\mathbf{3}_1\mathbf{13}_1\mathbf{18}_2\mathbf{9}_1\mathbf{14}_2$, with

Table V Folding Performance of Bred Sequences

Chain Structure	Sequence	Successful MC Runs, No. of 100	Folding Time (MC steps) Mean ± SD	Average Stability (%) Mean ± SD
$\mathcal{C}_1$	F_1	99	466 ± 414	90.1 ± 2.6
$\mathcal{C}_2$	F_2	93	326 ± 379	70.9 ± 2.3
$\mathcal{C}_3$	F_3	93	1513 ± 1005	84.1 ± 9.1
$\mathcal{C}_4$	F_4	88	956 ± 828	82.8 ± 9.4
$\mathcal{C}_5$	F_5	80	1521 ± 760	89.2 ± 2.2
$\mathcal{C}_6$	F_6	88	1627 ± 841	64.5 ± 8.3
$\mathcal{C}_7$	F_7	93	1172 ± 612	65.7 ± 5.9
$\mathcal{C}_8$	F_8	83	909 ± 647	73.3 ± 4.4
$\mathcal{C}_9$	F_9	94	1083 ± 920	75.9 ± 9.1
$\mathcal{C}_{10}$	F_{10}	96	1183 ± 698	77.8 ± 4.2

$F(\mathcal{C}_{8'}) = -27.339$, by lacking a single, short helix inserted in one of the loops. Another sequence, F_7, has two nearly degenerate structures at the bottom of its spectrum which, again, differ only by insertion of a short helix. However, $\mathcal{C}_7 = {}_1\mathbf{5}_1\mathbf{5}_1\mathbf{5}_1\mathbf{3}_1\mathbf{7}_1\mathbf{12}_{11}\mathbf{3}_4\mathbf{8}_1\mathbf{4}_2\mathbf{12}_1\mathbf{4}_7$, with $F(\mathcal{C}_7) = -27.344$, is reached in over 90% of the folding experiments, whereas $\mathcal{C}_{7'} = {}_1\mathbf{5}_1\mathbf{5}_1\mathbf{5}_1\mathbf{3}_1\mathbf{7}_1\mathbf{12}_1\mathbf{3}_7\mathbf{3}_4\mathbf{8}_1\mathbf{4}_2\mathbf{12}_1\mathbf{4}_7$, with $F(\mathcal{C}_{7'}) = -27.339$, is reached in less than 10% of folding experiments. As the examples of F_7 and F_8 demonstrate, kinetic preferences can contribute crucially to the folding performance of sequences. A detailed study of the folding kinetics of sequences F_1 to F_{10} is under way.

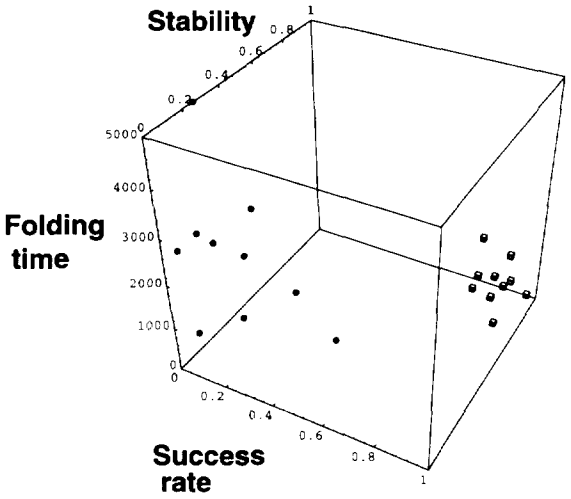

FIGURE 7 Folding performance, represented as in Figure 3. (Spheres) Results for random sequences RS_1 to RS_{10} and the chain structures $\mathcal{C}_1$ to $\mathcal{C}_{10}$, given in Tables II and III. (Cubes) Results for the folding sequences F_1 to F_{10} of Tables IV and V and chain structures $\mathcal{C}_1$ to $\mathcal{C}_{10}$ of Table III.

Stochastic optimization processes using genetic algorithms are often termed *evolutionary*,[30] which might suggest an analogy to natural evolutionary processes. However, the method described here, like most genetic algorithm approaches employed in optimization tasks, differs from natural evolutionary processes in various respects. The sequences have to adapt successively to two different requirements: (a) stability optimization, starting in the target structure, and (b) folding optimization, starting in the $(c^+)_L$ random coil. In addition, the population size is kept fixed throughout the optimization irrespective of the average fitness. An extinction of the population is thereby prevented artificially. For these reasons, the described method is more reminding of an artificial breeding process than of natural evolution. We term it "simulated breeding" to emphasize this point.

A Population of Folding Sequences

A closer look at the optimizations described above reveals that the algorithm does not converge to a single sequence in any of the cases studied. Even after the average fitness has reached a nearly constant high level, mutation and crossing-over events continue to generate new sequences, as is shown in Figure 8. There is only a very small fraction of back mutations, and the number of sequences generated increases linearly with generation number. On the average, each tested sequence was present in 2.5 copies and survived for less than two generations. The reason for this behavior is to be sought in the high dimensionality of sequence space. Even when there have already been generated several hundred different sequences, the probability to encounter one of them again in a random walk in a space of

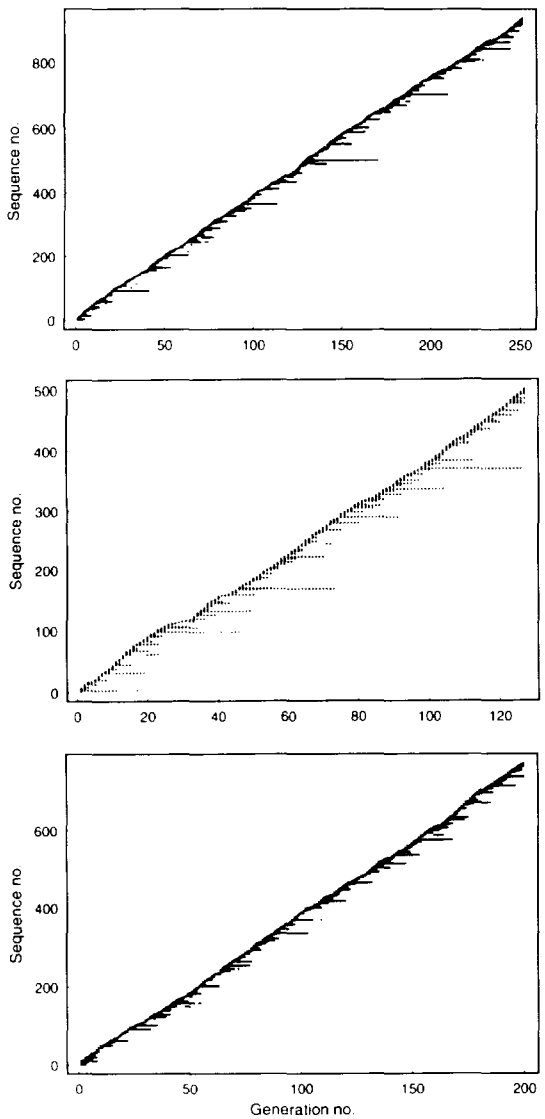

FIGURE 8 Evolution of sequence populations during the folding optimization procedures of Figure 6. The life span of each sequence is shown. (Top) Optimization for chain structure $\mathcal{C}_8$. (Middle) Optimization for chain structure $\mathcal{C}_9$. (Bottom) Optimization for chain structure $\mathcal{C}_{10}$.

100 dimensions is very small since each mutation step will usually lead into a new direction.

Since the mutation rate and the average lifetime of the sequences remain unchanged even late during the optimization procedures, i.e., when the fitness is generally high, one expects the algorithm to yield not only one, but several folding sequences during each run. To prove this, we have selected the 100 sequences with the highest average fitness values from the folding optimization in Figure 6 (bottom), i.e., for chain structure $\mathcal{C}_{10}$. Each of these sequences was subjected to 100 folding experiments to estimate their respective fitness more accurately. The average fitness values for the sequences lie in the ranges from 0.34 to 0.6; therefore, all these sequences are at least moderately good folding sequences.

As a distance measure between two sequences $\{A_i\}$ and $\{B_i\}$ in sequence space, we chose the *Hamming distance*

$$H(\{A_i\}, \{B_i\}) = \sum_i [1 - \delta(A_i, B_i)] \quad (2)$$

with δ being the Kronecker symbol. For a population P_N of N sequences $\{A_i\}_n$, $n = 1, \ldots, N$, we define the mean Hamming distance within that population by

$$\bar{H}(P_N) = \frac{2}{N(N-1)} \times \sum_{m=1}^{N} \sum_{n=m+1}^{N} H(\{A_i\}_m, \{A_i\}_n) \quad (3)$$

Given 8 different monomer types, the mean Hamming distance for the population of all 8^L possible sequences of length L is found to be $\bar{H}(P_{8^L}) = \frac{7}{8}L$. Now let $1 \le s \le 100$ denote the *rank* of the hundred best folding sequences, and $P_s = \{\{A_i\}_1, \ldots, \{A_i\}_s\}$, the population of the s best folding sequences. Figure 9 (top) shows the fitness $\mathcal{F}(\{A\}_s)$ as a function of $\bar{H}(P_s)$, for $s = 1$ to $s = 100$. The resulting graph decreases only slowly over a range of $\bar{H}$, indicating that several very good folding sequences can be found within a certain region in sequence space. This is in agreement with the observation, reported above, that the optimization procedure does not converge to a single sequence: it does not have to, since there are many different sequences that fold nearly equally well. Figure 9 (bottom), shows the mean Hamming distance of the populations P_s, for $s = 1$ to $s = 100$, as a function of s. $\bar{H}(P_s)$ reaches a value of about 7 for $s = 4$ and remains below 8.5 up to $s = 83$. This indicates that the folding sequences in the population P_{83} are distributed approximately uniformly over a patch in sequence space that can be characterized by a mean Hamming distance of about 8 between its individual sequences.

As to be expected from this value, the various folding sequences found are homologous. Figure 10 shows the probability distribution of the mutual overlap $Q(\{A_i\}, \{B_i\}) = L - H(\{A_i\}, \{B_i\})$ between any two sequences $\{A_i\}$, $\{B_i\}$ in P_{83}, the

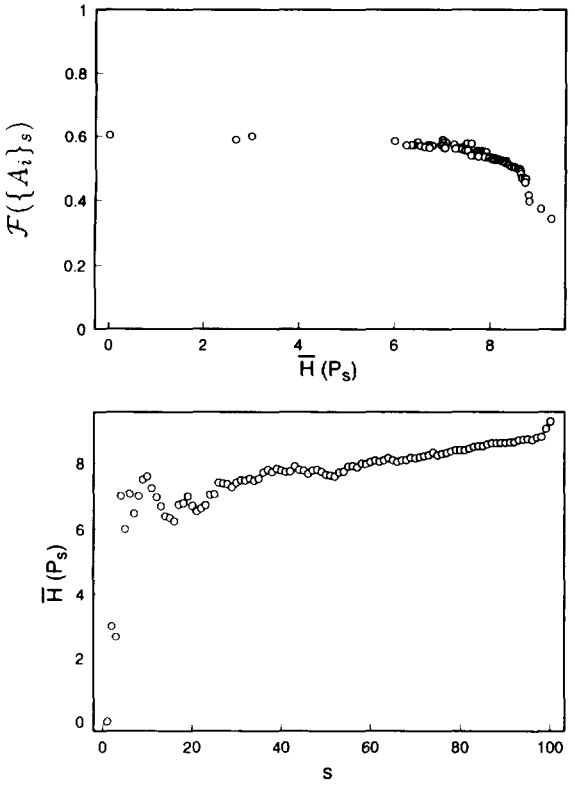

FIGURE 9 Population of the 100 best folding sequences from the folding optimization in Figure 6, (bottom). (Top) The fitness $\mathcal{F}(\{A_i\}_s)$ is shown as a function of the mean Hamming distance, $\bar{H}(P_s)$, for rank $s = 1, \ldots, 100$. Explanation in text. (Bottom) The mean Hamming distance, $\bar{H}(P_s)$, is shown as a function of rank s.

population of the 83 sequences with an average fitness higher than 0.5. Two sequences picked at random from this population will coincide, on the

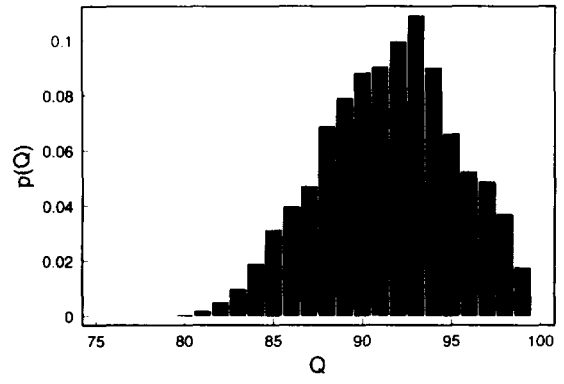

FIGURE 10 Distribution of mutual overlaps Q between sequences in the population P_{83}, discussed in the text. The distribution is centered around $Q = 91.5$.

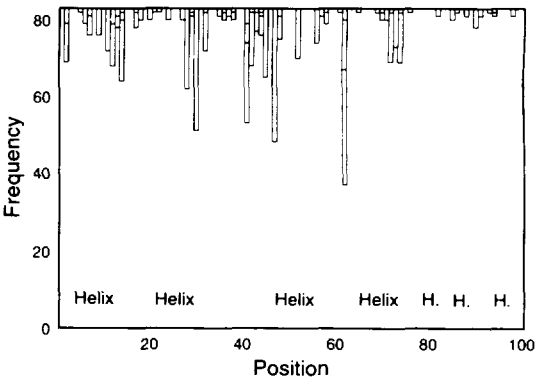

FIGURE 11 Comparison of the sequences in the population P_{83} discussed in the text. For each position, the frequency distribution of monomer types in the population is indicated; for clarity, individual monomer types are not given. For most of the positions, a single monomer type clearly dominates all others. Positions 30, 41, 47, and 62 are the most variable ones. A look at the target structure, indicated by dotted lines and "Helix" labels, does not give any clue as to why these positions should be allowed more variation than others.

average, at 91.5 out of their 100 monomers. Twenty-three out of 100 positions in the sequences of P_{83} are strictly conserved. In addition, at most of the remaining positions, a single monomer type clearly dominates in the population. Figure 11 shows the frequency distribution of monomer types over the 100 positions within the sequence population P_{83}. The folding performances found in P_{83} are presented in Figure 12. It is clearly seen that

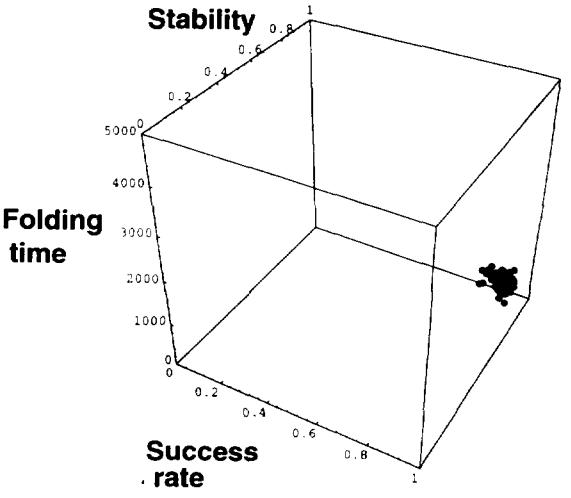

FIGURE 12 Folding performance, represented as in Figure 3, for the sequences of population P_{83} discussed in the text.

they form a cluster in the "folding corner" of the figure.

Supplement to the Described Method

When looking at Figure 6, one notices that the bottleneck of the described method for the breeding of folding sequences lies at the transition from stability optimization to folding optimization. It cannot be guaranteed that sequences that are stable in their target structure will reach this structure during folding experiments. In Figure 6 (bottom), the folding fitness remains zero for 8 generations after completion of the stability optimization. During this period, the algorithm performs an unbiased random walk in sequence space until it first hits a sequence that reaches the target structure. From then on, the algorithm proceeds to yield folding sequences within about 150 generation cycles. However, such a course of the optimization process is highly unlikely. Normally, a random walk in sequence space will not reach a sequence with fitness greater than zero in a reasonable amount of time.

This problem can be illustrated by looking at another interesting optimization course. As in Figure 6 (bottom), we chose $\mathcal{C}_{10}$ (Table III) as the target structure. However, this time we started the optimization algorithm not with ten copies of RS_{10}, but with ten copies of RS_1 (Table II). The sequence RS_1 is, not surprisingly, completely unstable in the chain structure $\mathcal{C}_{10}$. However, during the stability optimization, each newly generated sequence starts each simulation run in the target structure. Therefore, the optimization algorithm will, in a reasonable amount of time, be able to accumulate mutations that selectively stabilize the target structure. As the left-hand part of Figure 13 (top) shows, the stability optimization produced sequences that are extremely stable in the target structure after about 1200 generations. However, the population of sequences obtained after 1500 generations of stability optimization did not contain any folding sequences; even after 20 generations of folding optimization, no folding sequences were obtained (data not shown). To overcome this difficulty, we introduced a series of intermediate optimization steps during which we selected for tolerance toward disturbances of the target structure. The genetic algorithm runs as follows: (A) The sequences are set to a starting structure given by their target structure disturbed by changing D randomly chosen local conformations to c^+; (B) the relative amount of simulation time spent in the target structure during 10 simulation runs of 250 MC steps is taken as the fitness $\mathcal{F}$; (C) and (D) all other parameters remain as described above for stability and folding optimizations.

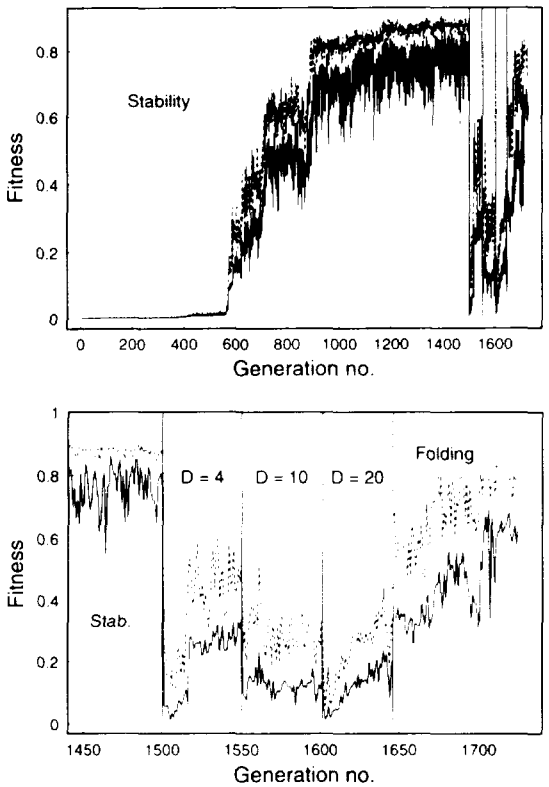

FIGURE 13 Optimization for chain structure $\mathcal{C}_{10}$, starting with random sequence RS_1. (Top) Overview: Starting from RS_1, a sequence completely incompatible with, i.e., unstable in, the target structure, the stability optimization nonetheless generates sequences stable in $\mathcal{C}_{10}$ in a reasonable amount of time (note that, due to the shorter simulation time employed, 25 generations of stability optimization take only as long as a single generation of folding optimization). (Bottom) Detail: Transition from stability optimization to folding optimization via tolerance optimization.

Figure 13 (bottom) shows a detail of Figure 13 (top). After completion of the stability optimization, 50 generations of tolerance optimization toward $D = 4$ disruptions of the target structure followed. The resulting sequences were able to cope with $D = 10$ disruptions (immediately after the stability optimization, none of the sequences could tolerate 10 random disruptions and still find back to the target structure; not shown). After 51 generations of optimization, D was raised to 20. After 46 generations at $D = 20$, moderately good folding sequences had already been obtained. One hundred fifty generations of folding optimization then

yielded excellent folding sequences; the best one found,

CBBAAEBCBGBBEBBEDDGDBBBBB

FAGEBCFBFCCEHFDDGABEEEBDB

BCBHAAFBABAHBDDCBFAACFDDE

FBBCDACDEEDDAAEEEFHACBBEB

had a fitness of 0.59, comparable to the best sequences discussed in the context of Figures 6 and 9. Surprisingly, however, this sequence does not resemble the sequences of the population P_{83} discussed above; at 61 of the total 100 positions, it carries monomer types that are not present at their respective positions in any of the sequences in P_{83}. Thus, the folding sequence obtained here belongs to a completely different region in sequence space. We have here the interesting result, known from real proteins, too, that nonhomologous sequences may adopt the same structure.

SUMMARY AND OUTLOOK

Simplified models have contributed much to the present views on the generic properties of protein folding. The question how to systematically obtain folding sequences in such minimalist protein folding models has only recently been addressed. For the model studied here, straightforward approaches, based on consideration of the target structure only, usually fail to produce folding sequences. Here, we presented an alternative approach to the problem. We opt for a folding criterion that requires sequences to fold fast and reproducibly to their target structure and to be stable in this structure. Such sequences will spend a maximum amount of simulation time in their target structures during simulated folding experiments. Given such a specific, quantitative folding criterion, the design of folding sequences amounts to an optimization problem on a fitness landscape over sequence space. However, this fitness landscape will, in general, be rugged and, therefore, difficult to study. Therefore, we have devised a stochastic optimization procedure based on genetic algorithms to select folding sequences. Sequences are first optimized with respect to their stability in the target structure. In a second step, sequences are selected that fold fast and reproducibly to the target structure and are stable there. Where necessary, an intermediate "tolerance optimization" step may be included during which sequences are selected that are able to return to their target structure after the structure has been randomly disturbed.

This algorithm of "simulated breeding" was applied successfully to obtain folding sequences for various target structures. This result puts us in a position to study the characteristics of folding sequences in our model, with the hope of gaining new insights into dynamical processes resembling protein folding. Already, there is evidence for a pivotal role of kinetic preferences for some of the sequences studied. In addition, we have shown that nonhomologous sequences may fold to the same target structure, as is known from proteins, too. More detailed studies of the kinetics of folding processes in the model are under way.

APPENDIX A: MICRO-, MESO-, AND MACROSTATES

The free energy f of a system is given by

$$\beta f = \ln \sum_{x} \exp[-\beta E(x)] \qquad \text{(A1)}$$

where $\beta = (k_B T)^{-1}$, with T temperature and k_B Boltzmann's constant. The summation ranges over all *microstates* x of the system, e.g., all possible combinations of atomic coordinates. In the model presented here, the system is described by *chain conformations* $\{\sigma_i\}$. Each of the three local conformations (h, c^0, c^+) is represented by numerous different microstates, and therefore, the chain conformations $\{\sigma_i\}$ can be considered *macrostates* of the system. These macrostates are assumed to cover all the possible microstates of the system. In addition, in the approximation leading to Eq. (1), all the microstates contributing to the same macrostate are assumed to have the same energy. Hence, letting $g(\{\sigma_i\})$ be the number of microstates belonging to the macrostate $\{\sigma_i\}$, we can write

$$\begin{aligned}\exp[\beta f] &= \sum_{x} \exp[-\beta E(x)] \\ &= \sum_{\{\sigma_i\}} g(\{\sigma_i\}) \exp[-\beta E(\{\sigma_i\})] \\ &= \sum_{\{\sigma_i\}} \exp[-\beta E(\{\sigma_i\}) + S(\{\sigma_i\})/k_B] \\ &= \sum_{\{\sigma_i\}} \exp[-\beta F(\{\sigma_i\})] \qquad \text{(A2)}\end{aligned}$$

Here, the entropy $S(\{\sigma_l\})$ of a macrostate $\{\sigma_l\}$ is given by $S(\{\sigma_l\}) = k_B \ln[g(\{\sigma_l\})]$, and we have $\beta F = \beta E - S/k_B$. In the model presented here, we use relative entropies

$$S_{\text{rel}}(\{\sigma_l\}) = S(\{\sigma_l\}) - S_0 = \sum_{l=1}^{L} \Delta S(\sigma_l, A_l) \qquad \text{(A3)}$$

S_0 is chosen so that the chain conformation $(c^+)_L$, in which all L monomers are in c^+ conformation, is assigned a relative entropy of $S_{\text{rel}}[(c^+)_L] = 0$. Since the relative entropy of any chain conformation is easily calculated [compare Eq. (A3)], the choice of chain conformations as system coordinates is a useful one. Note that, while f is the thermodynamic free energy of the system, F denotes the free energy of a specific macrostate. Comparing the first and last lines of Eq. (A2), one finds that the role of F, Eq. (1), in a model description using chain conformations is analoguous to that of the energy E, i.e., the Hamiltonian, in a model description on the basis of microstates.

The situation becomes more complicated with the introduction of *chain structures* as defined in the text. These are *macrostates* with respect to the chain *conformations*, which, in turn, can be considered *mesoscopic* states now. The number of chain conformations corresponding to a particular chain structure is not always easily determined. For example, the chain structure $\mathcal{C} = {}_4\mathbf{5}_2\mathbf{3}_{[3]}\mathbf{3}$ of a system with length $L = 20$ is found to represent 572 different chain conformations: the random coil section "[3]" corresponds to the 11 local conformations $c^+c^+c^+$, $c^+c^0c^0$, $c^0c^+c^0$, $c^0c^0c^+$, $c^+c^+c^0$, $c^+c^0c^+$, $c^0c^+c^+$, c^+hc^0, c^+hc^+, c^0hc^+, c^0hc^0; in addition, the random coil section of length 4 at the front end corresponds to another $3^3 \cdot 2 - 2 = 52$ local conformations, yielding a total of $11 \cdot 52 = 572$ chain conformations with identical chain structure. The different chain conformations corresponding to any chain structure may differ in their F values. Therefore, the Hamiltonian has no well-defined value for a chain *structure*. For an approximate characterization of a chain structure $\mathcal{C}$, we define the lowest F value among the contributing chain conformations as the chain structure's F value, $F(\mathcal{C})$.

APPENDIX B: REMAINDER STOCHASTIC SAMPLING WITH REPLACEMENT

When applying genetic algorithms, there are various ways of generating a new population of chromosomes once the fitness values for the current population have been determined.[34] The "remainder stochastic sampling with replacement" procedure consists of the following steps: (A) For any chromosome C_j, $j = 1, \ldots, N$, of the current population, the probability to appear in the new population is given by $p_j = \mathcal{F}(C_j)/\Sigma_j \mathcal{F}(C_j)$, where $\mathcal{F}(C_j)$ stands for the fitness value of chromosome C_j. (B) For each chromosome with $p_j \geq N^{-1}$, one copy is taken to the new population, and p_j is lowered by N^{-1}, until $p_j < N^{-1}$; thereby, the survival of at least one copy of the most successful chromosome is ensured. (Note, however, that this copy may be altered by subsequent mutation or crossing-over.) (C) The remaining vacancies in the new population are filled by choosing randomly among the current chromosomes, selecting chromosome C_j with probability p_j.

It is a pleasure to thank T. Krausche for interesting and stimulating discussions on protein folding and stochastic optimization. WN thanks W. Wolff for an interesting discussion on evolution and breeding. ME gratefully acknowledges a stipend from the Studienstiftung des deutschen Volkes.

REFERENCES

1. Ebeling, M. & Nadler, W. (1993) *J. Chem. Phys.* **99,** 6865–6875.
2. Ebeling, M. & Nadler, W. (1993) *J. Chem. Phys.* **100,** 4719E.
3. Ebeling, M. & Nadler, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8798–8802.
4. Lifson, S. & Roig, A. (1961) *J. Chem. Phys.* **40,** 1963–1974.
5. Zwanzig, R. & Lauritzen, J. I. (1968) *J. Chem. Phys.* **48,** 3351–3360.
6. Lauritzen, J. I. & Zwanzig, R. (1970) *J. Chem. Phys.* **52,** 3740–3751.
7. Ebeling, M. & Nadler, W. (1996) *Phys. Rev. E* **53,** 3365–3368.
8. Matthews, C. R. (1993) *Ann. Rev. Biochem.* **62,** 653–683.
9. Roder, E. & Elöve, G. A. (1995) in *Mechanisms of Protein Folding*, Pain, R. H., Ed., IRL Press, Oxford, pp. 26–54.
10. Chan, H. S. & Dill, K. A. (1991) *J. Chem. Phys.* **95,** 3775–3787.
11. Miller, R., Danko, C., Fasolka, M. J., Balazs, A. C., Chan, H. S. & Dill, K. A. (1992) *J. Chem. Phys.* **96,** 768–780.
12. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.

13. Chan, H. S. & Dill, K. A. (1993) *J. Chem. Phys.* **99,** 2116–2127.
14. Camacho, C. J. & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6369–6372.
15. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 1942–1946.
16. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67,** 1665–1668.
17. Shakhnovich, E. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
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. Go, N. & Abe, H. (1981) *Biopolymers* **20,** 1013–1031.
22. Shrivastava, I., Vishveshwara, S., Cieplak, M. & Maritan, A. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9206–9209.
23. Krausche, T. & Nadler, W., to be submitted.
24. Poland, D. & Scheraga, H. A. (1970) *Theory of Helix-Coil Transitions in Biopolymers,* Academic Press, London, New York.
25. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963) *J. Mol. Biol.* **7,** 95–99.
26. Toulouse, G. (1977) *Commun. Phys.* (London) **2,** 115–119.
27. Anderson, P. W. (1978) *J. Less Common Met.* **62,** 291–294.
28. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
29. Frauenfelder, H. & Wolynes, P. G. (1994) *Phys. Today* **47,** 58–64.
30. Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems,* University of Michigan Press, Ann Arbor.
31. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983) *Science* **220,** 671–680.
32. Dueck, G. (1993) *J. Comp. Phys.* **104,** 86–92.
33. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
34. Michalewicz, Z. (1994) *Genetic Algorithms + Data Structures = Evolution Programs,* 2nd ed., Springer, Berlin.
35. Levinthal, C. (1969) in *Mössbauer Spectroscopy in Biological Systems,* Debrunner, P., Tsibris, J. C. M. & Münck, E., Eds., University of Illinois Press, Urbana, pp. 22–24.