COMPUTER SIMULATIONS OF PREBIOTIC EVOLUTION

V.I. ABKEVICH, A.M. GUTIN, and E.I. SHAKHNOVICH
Harvard University, Department of Chemistry
12 Oxford Street, Cambridge MA 02138

This paper is a review of our previous work on the field of possible ways of prebiotic evolution[1,2]. We propose an algorithm providing sequences of model proteins with rapid folding into a given native conformation. Thermodynamical analysis shows that the increase in speed is matched by an increase in stability: the evolved sequences are much more stable in their native conformation than the initial random sequence. We discuss a possible origin of the first biopolymers, having stable unique structure. We suggest that at the prebiotic stage of evolution, long organic polymers had to be compact in order to avoid hydrolysis and had to be soluble and thus must not be exceedingly hydrophobic. We present an algorithm that generates such sequences of model proteins. The evolved sequences turn out to have a stable unique structure, into which they quickly fold. This result illustrates the idea that the unique three-dimensional native structure of first biopolymers could have evolved as a side effect of a nonspecific physico-chemical factors acting at the prebiotic stage of evolution.

The problem of how first biopolymers could have evolved prebiotically is one of the most fundamental in modern science. The first macromolecules which served as enzymes must satisfy two conditions: (i) the structure must be thermodynamically stable and (ii) it must be kinetically reachable in a biologically reasonable time. The important question, from the evolutionary point of view, is how likely that such sequences, having unique 3-dimensional structure have appeared in the primordial soup. Assuming that initial polymerization of organic molecules at the prebiological stage resulted in the set of random sequences, it can be be formulated as what fraction of all sequences can fold to and be stable in a unique structure? This question concerns the very essence of the concept of prebiological evolution. Indeed, if the fraction of foldable sequences among all random sequences is large, then there may be no need in prebiotic selection, and protein-like sequences could have been selected randomly out of multitude of randomly synthesized biopolymers. Alternatively, if the set of foldable sequences constitutes vanishing fraction of all sequences, then some selection was likely to had been acting even at the prebiotic stage. In this paper we show that this is in fact the case. However, the main problem here is that at the prebiotic stage such property as unique structure (and ability to perform catalytic functions) was not likely to be a factor in selection (that is why this stage is *prebiotic*). Therefore in this case some other factors could have been guiding synthesis of prebiological polymers, and sequences with unique structure could have evolved as a byproduct of other

physico-chemical processes. In this paper we suggest a principle possibility of a scenario of this type.

It was shown analytically[3,4] and numerically[4−7] for different models of proteins that the condition of the thermodynamic stability of the native state is not restrictive. It was conjectured[3,4] that there exists a critical temperature T_c such that a significant fraction of random sequences have a stable unique structure at $T < T_c$. It was also pointed out[1] that the probability for a random sequence to have a stable unique structure at $T < T_c$ does not depend on its length N (this result was confirmed in a more recent numerical study[7]). However, at low temperatures folding rate decreases and the native conformation becomes kinetically less accessible. The implication is that random sequences mostly are not able to fold into a unique structure[8].

The principal way out is to find special sequences which have their native structure stable at temperatures when the folding is fast, resolving therefore the contradiction between the thermodynamic and kinetic requirements characteristic for random sequences[9]. The requirement that a sequence has a native structure stable at higher temperatures imposes the necessary condition that it is a pronounced energy minimum separated by a large energy gap from the set of non-native conformations[8−10].

The probability of finding such sequence by simply pulling it out at random from the "soup" of all possible sequences is very small for chains of realistic length. However, this difficulty may be overcome by a simple sequence design algorithm suggested[10,11]. This algorithm generates sequences which have sufficiently low energy in a chosen target conformation. The energy gap in designed sequences was indeed large enough to enable them to fold fast to the stable native conformation[12] which coincided with the target conformation used at the design stage. The design algorithm is based on the idea of Monte Carlo (MC) search in sequence space. It proceeds by making mutations in a sequence, biasing them to ones which decrease the relative energy of the native state at given amino acid composition, or relative energy. The algorithm made it possible to find sequences having relative energies sufficient for folding in spite of the fact that the fraction of such sequences among all possible ones is small.

The results of refs. 10 and 12 demonstrate that sequence design aimed at generating thermodynamically stable sequences makes this conformation accessible - i.e., that thermodynamic stability is a sufficient condition for fast folding. We may ask now, is it also a necessary condition for fast folding? This question is exceptionally important for our understanding of prebiotic evolution of any biopolymers with a unique native state. If thermodynamic stability is a sufficient condition for fast folding then two conditions to which enzymes must satisfy (see above) are equivalent to each other and one could easier imagine prebiotic process in which the first enzymes evolved.

To address this question, we use a simple algorithm of sequence selection

based on the principles similar to biological evolution, i.e. random substitutions and selection pressure. The requirement of rapid folding into one specific conformation modeled the selection pressure via acceptance of substitutions which result in faster folding and rejection of substitutions which slow down the folding rate.

Due to the fact that the mean first passage (to the native conformation) time (MFPT) is determined only approximately we developed the following 3-step algorithm for evaluation of the kinetic consequence of a point mutation. In step 1, we perform 2 folding runs and estimate very roughly the new MFPT. If it is longer than the original one, then the mutation is rejected; if the new MFPT is shorter, then we estimate the new MFPT more precisely in step 2. The purpose of the step 1 is to reject outright obviously poor mutations, which constitute the majority of all mutations. In step 2 we perform 10 folding simulations and therefore get much more precise estimate for the MFPT. If it is less than the original MFPT by 20%, then the mutation is accepted; otherwise it is rejected. If the mutation is accepted, then an additional 100 folding runs are performed (step 3) to get a reasonably good estimate for the new MFPT. This estimate is used as a current MFPT for comparisons with the MFPT of the next mutations.

Due to computer time limitations we restricted our study to chains of 27 monomers on a cubic lattice. The energy of a conformation of the chain is the sum of energies of pairwise contacts:

$$E = \sum_{1 \le i < j \le N} (B_0 + B(\xi_i, \xi_j))\Delta_{ij} \qquad (1)$$

where $\Delta_{ij} = 1$ if monomers i and j are lattice neighbors and $\Delta_{ij} = 0$ otherwise. ξ_i defines the type of amino acid residue in position i. $B(\xi, \eta)$ is a magnitude of contact interaction between amino acids of types ξ and η. We expect that the choice of the specific set of parameters for our study is not very essential (for more detailed discussion of this problem, see ref. 13). We used parameter set published[14]. This set of parameters was derived from statistics of contacts in proteins. B_0 is an energetic parameter having the meaning of overall attraction, it was introduced to bias conformations toward more compact ones making the native state belong to the set of maximally compact conformations. The motion of the chain is modeled by the standard cubic lattice MC algorithm[15,16]. Simulations were performed at temperature $T = 0.32$ and with $B_0 = -T = -0.32$.

We started our selection algorithm with random sequence which folded in its native conformation at temperature $T = 0.32$ in about $5 \cdot 10^6$ MC steps. The result of selection is presented in Fig.1. which shows the MFPT as a function of the number of accepted mutations. It can be seen that for about 100 accepted mutations the MFPT decreased almost two orders of magnitude.

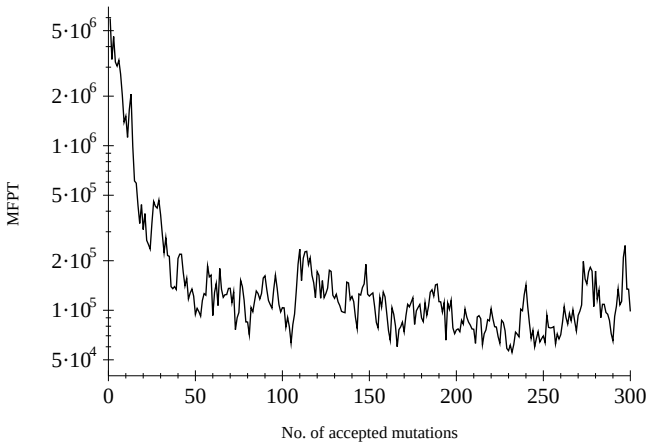

FIG. 1. Evolution of the MFPT (in MC steps) under the requirement of fast folding. This figure is taken from the ref. 1.

As we mentioned before, one way to make a chain fold rapidly is to make the energy E_{nat} of its native conformation as low as possible. The absolute value of the energy itself is not directly related to stability; what is important is the relative value of energy, or Z-score introduced by Eisenberg and coauthors[17]:

$$Z = \frac{E_{nat} - E_{av}}{\sigma} \tag{2}$$

where E_{av} is the average energy of compact non-native conformations. It could be estimated as $E_{av} = K \cdot e_{av}$, where $K = 28$ is the number of contacts in a maximally compact conformation and e_{av} is an average energy of all possible contacts with a corresponding dispersion σ.

The evolution of the relative energy of the native conformation Z is shown on Fig.2. It is seen that for the first 50 accepted mutations, when the MFPT decreases, Z decreases noticeably too, though because of strong fluctuations of Z the effect is not so pronounced. To get a clear impression of the properties of the ensemble of selected fast-folding sequences, we calculated the distribution of Z over 250 sequences starting from 50th accepted mutation. The corresponding histogram is shown on Fig.3. For comparison we also plotted a similar histogram for 250 random sequences with the same amino acid composition. It is clearly seen that the distribution of the parameter Z for fast-folding

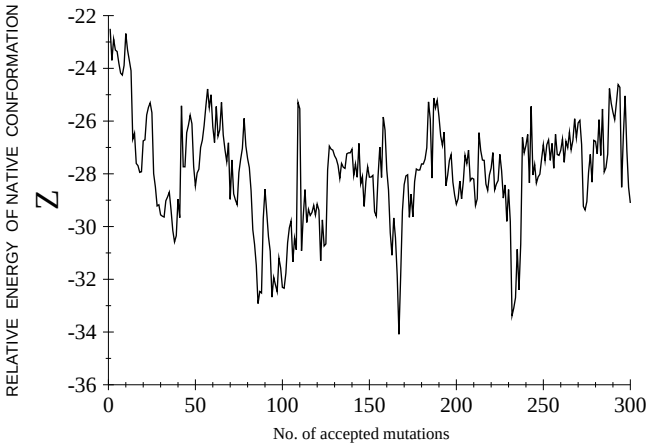

FIG. 2. Evolution of the relative energy of the native conformation Z under the requirement of fast folding. This figure is taken from the ref. 1.

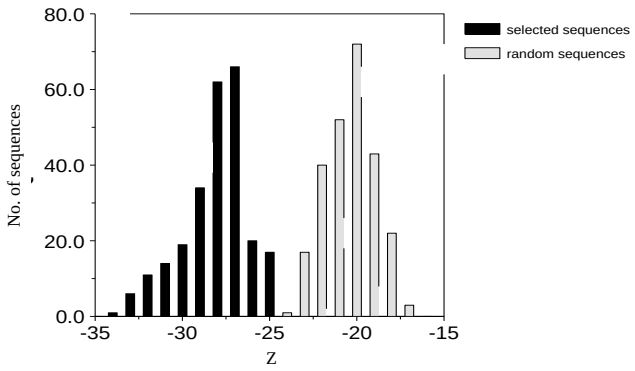

FIG. 3. Distribution of the relative energy of the native conformation Z for random sequences (white bars) and for the fast-folding sequences generated by the selection algorithm under the requirement of fast folding (filled bars). This figure is taken from the ref. 1.

sequences generated by the selection algorithm is clearly shifted to smaller values as compared to random sequences. The most probable value of Z for fast-folding sequences is about -27.5; at the same time for random sequences this number is close to -20. The distribution of Z for random sequences can be fit very well by a Gaussian. It is possible to determine from such a fit the probability to have a random sequence with $Z = -27.5$ corresponding to the median of the Z distribution for selected sequences. The elementary estimate gives $P(Z = -27.5) \approx 10^{-6}$. Therefore one out of million random sequences for 27-mer will behave like average fast-folding sequence from the pool of "evolutionary selected" ones. Even though thermodynamical stability is matched with increase in speed it is not the only factor which determines a folding rate. Two equally stable sequences might have very different folding time. We have compared folding time of evolved sequences and equally stable sequences obtained by a simple sequence design algorithm[10,11]. We found that evolutionary designed sequences fold in average ten times faster. However we were not yet able to find what makes evolutionary designed sequences so exceptionally fast-folding.

It was shown in a number of previous calculations that thermodynamic stabilization results in rapid folding. In this work, we have shown that the opposite is true as well: optimization of folding rate results in a pronounced thermodynamic stabilization of the native state. This result show that both key conditions for polypeptide sequences to be protein-like (i.e., thermodynamic stability and fast folding), can be achieved when one parameter, the relative energy *or* folding time, is minimized under the evolutionary pressure. This is due to the fact that one parameter, the relative energy, governs, to a large extend, both quantities.

Such schemes may model some aspects of protein evolution only at the biological stage where sequences, which reliably fold into their native conformations, have already evolved, and the evolutionary process presses for their improvement as more stable and fast-folding proteins. This fact manifests in both algorithms where the concept of unique native structure is introduced from the beginning, at the stage of formulation of selection rules.

This implies that such evolutionary improvement schemes must be "seeded" in the sense that first protein-like sequences, which are able to fold into their unique native conformations must have evolved at the stage of prebiological evolution.

It is quite clear that at the prebiological stage of evolution the notion of unique structure could hardly exist, and instead of "selection pressure", at which specific biological advantages were memorized and inherited, there could be only relatively nonspecific physico-chemical factors which governed synthesis and hydrolysis of the first organic polymers. Then question arises, how such nonspecific physico-chemical factors could have selected first protein-

like sequences?

In this paper we give a possible example of such "pre-biological" selection. While being unlikely to reproduce the actual process of prebiological evolution, it may illustrate an important idea of how biopolymers might have evolved as a result of nonspecific physico-chemical factors.

We consider a model where "physico-chemical" pressure consists of two factors: possibility of aggregation and hydrolysis[18]. The latter could be catalyzed by other existing low- and high molecular weight compounds. It is clear that a simple way to protect polypeptide from hydrolysis would be to make chains compact. This biases equilibrium distribution over sequences towards more hydrophobic ones, with prevailing attraction between residues. However, hydrophobic sequences tend to aggregate.

The real prebiotic selection of biopolymers precursors was likely to involve synthesis of large number of sequences over long period of time ($\sim 10^9$ years). The polymers which are less compact had higher probability to get hydrolyzed. This factor gradually shifts the sequence distribution towards more compact and, as will be shown in this work, more stable sequences with unique structure.

In this work we propose a simple algorithm which is similar in spirit to the one we described above and biases sequence sampling towards compact and water-soluble sequences. While being unlikely to reproduce the details of the actual process of pre-biological evolution, it generates in a computationally reasonable time the ensemble of sequences which meet these two criteria of compactness and solubility. Compactness can be measured directly in the simulation through the number of intramolecular contacts which chain forms: the greater this number, the more compact the chain is. As regards to solubility, it is computationally very costly to faithfully reproduce interactions of macromolecules with solvent while making the search in sequence space. The unrestricted condition of compactization would generate mainly hydrophobic sequences with tendency to aggregate. To prevent this we will impose the simple condition of solubility as a requirement that sequence search proceeds over sequences with constant amino acid composition. To model the requirement of solubility further, we introduced slight overall repulsion between molecules by taking $B_0 = 0.05$. Simulations were performed at temperature $T = 0.15$.

Now the main idea of the algorithm can be formulated: beginning with random sequences, we make substitutions preserving the overall composition of amino acids. If the chain becomes more compact, or gets compact faster, the substitution is accepted, otherwise it is discarded. After a number of cycles the procedure will generate sequences which will be compact and still not exceedingly hydrophobic. The ensemble of such sequences can be then studied to see whether their properties deviate from such of random sequences.

We changed the selection procedure in the following a way. Instead of optimization of the folding time, we optimized the average compactness over

a half of million MC steps starting with a random coil conformation. In other words, the requirement now is not only to reach a compact state rapidly, but also to keep the compactness over the course of simulation. This is a more faithful representation of the idea that compactness should protect the chains from hydrolysis. This combines two requirements: that compactization will be fast and that it will lead to persistent compact conformations.

Our selection mechanism is close to the described above. In step 1, 50 folding runs were performed to get a reasonably good estimate of average compactness during the first $5 \cdot 10^5$ MC steps. In step 2, an attempt of pairwise substitution is made. This corresponds to picking randomly two sites and "swapping" amino acids between them. This is the natural way of changing sequence without changing amino acid composition. If the new average compactness (number of intrachain contacts) during the first $5 \cdot 10^5$ MC steps is smaller, then the substitution is rejected. If it is larger, then we perform ten folding simulations and get a more precise estimate of the new average compactness during the first $5 \cdot 10^5$ MC steps in step 3. If it is at least 1 % higher than the original compactness, then the substitution is accepted; otherwise, the substitution is rejected.

Thirty two substitutions were accepted. For all of the generated sequences, we determined conformations (not necessarily maximally compact) with the lowest energy. This was done by means of extensive MC simulations. These conformations were identified as the native ones. Some of them are shown on Fig.4. After analyzing native conformations, we found that most of them differ one from another. Although for the first 7 substitutions, native conformations changed dramatically, for the next substitutions, the number of common contacts between native conformations for different sequences became large ($> 80\%$) so that sequences which evolved at the later stage of the procedure had structurally similar native conformations.The evolution of average compactness of the native conformations, with substitutions is shown on Fig.5. It is clear that the compactness of the native conformations grows rapidly and reaches its plateau value corresponding to conformations having 25 contacts.

Most importantly, the relative energy of the native conformation (Z) decreases with substitutions (Fig.5b). It implies that as a result of selection aimed at large average compactness the procedure, after some substitutions, decreases the relative energy of the native conformation to as low value as -33. One might see on Fig.3 that probability to choose such sequence randomly is vanishingly small. In other words, the requirement of a rapid compactization can lead to a result different from the compactization itself, that is, to the stable native conformation with very low energy. This conclusion is supported by the next plot on the Fig.5c which shows evolution of the thermodynamic probability of the native conformation. One can see that the native conformation of the starting randomly chosen sequence is extremely unstable: its

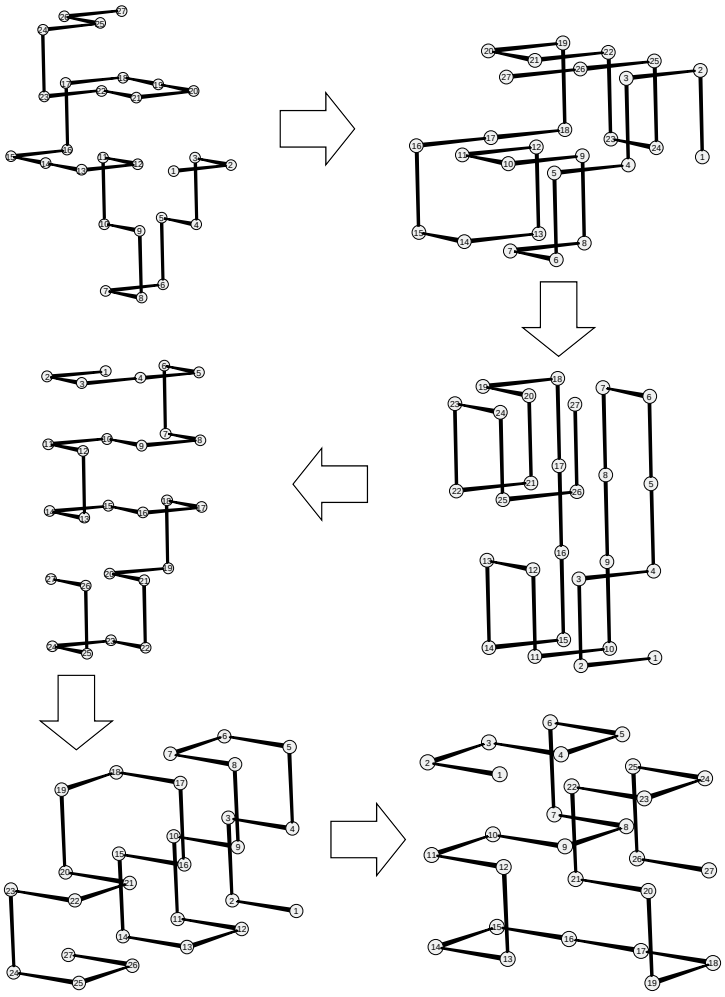

FIG. 4. Selection of sequences under the requirement of large average compactness over a half of million MC steps starting with a random coil conformation. Evolution of the native structure is shown. On this figure shown the native conformations of the original random sequence, sequence after the 5th substitution, sequence after the 10th substitution, sequence after the 20th substitution, sequence after the 25th substitution and sequence after the 32nd (last) substitution. This figure is taken from the ref. 2.

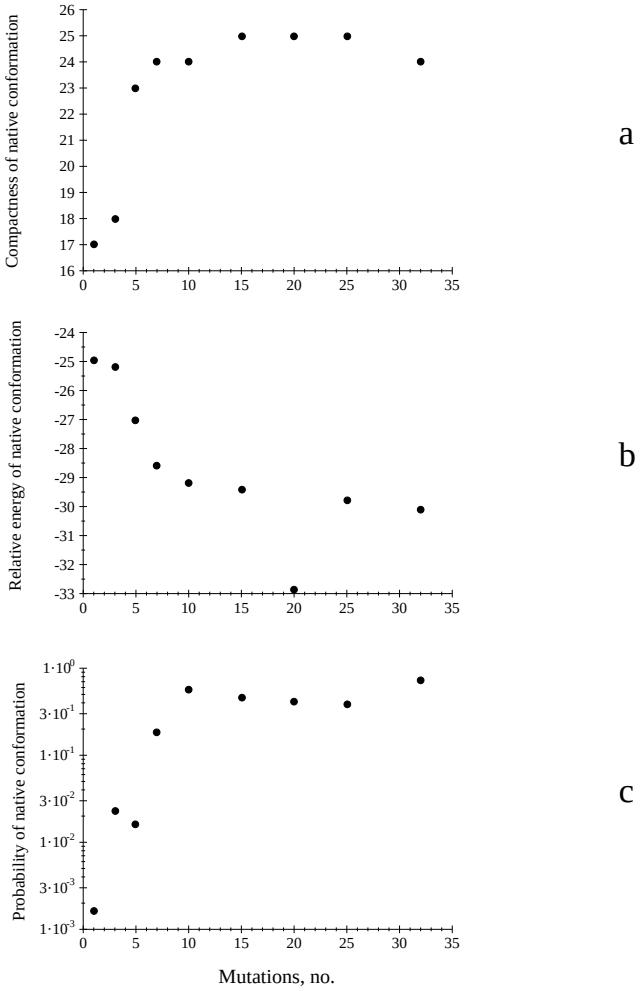

FIG. 5. Selection of sequences with 27 residues under the requirement of large average compactness over a half of million MC steps starting with a random coil conformation. (a) Evolution of compactness (the number of contacts in the native structure). (b) Evolution of the relative energy of the native conformation. (c) Evolution of the stability of the native conformation. This figure is taken from the ref. 2.

Boltzmann probability is only about 0.1%. At the same time for the selected sequences the Boltzmann probability of the native conformation is close to 50% after about 10 substitutions.

This suggests an idea of how proteins with stable native structures might have evolved. The main point here is that such stable structures could be a side effect of a different requirement, for example, the requirement on compactization. In the primordial "soup" different species with different sequences were in equilibrium. Some fractions aggregated; some were soluble but noncompact and were hydrolized. Our results suggest that a small fraction of all sequences were able to satisfy mutually conflicting non-specific requirements, and these sequences could have had a unique structure.

It is commonly believed that in the earliest cells, both the genetic and enzymatic components were RNA molecules[19−21]. If this hypothesis is correct, proteins evolved during biological evolution. However, the question remains as how ribozymes with unique native structure have evolved? The model described above can be used to understand better the driving forces of prebiological creation of any biopolymers with stable native structure.

The selection algorithm that we used herein may resemble rather biological selection than prebiological one since some favorable substitutions were accepted and passed by to further "generations" of sequences. This resemblance is superficial. The possible feedback in prebiological evolutionary scenario could be that more compact polypeptides were more likely to avoid hydrolysis. In the situation when polypeptides got synthesized randomly, this factor gradually shifted the sequence distribution to increase probability of polypeptides with stable native state. The possible physical mechanism of it is very simple: in spite of the fact that probability to randomly synthesize polypeptides with stable native state is very small, these sequences were not hydrolyzed and accumulated over many cycles of hydrolysis and synthesis; whereas other sequences underwent many turns of recycling.

Synthesis of random sequences represents a random walk in sequence space. Within the framework of this analogy, the stability of folded structures, which makes sequences nonhydrolyzable, is equivalent to absorbing walls; it is clear that the concentration of species at or near adsorbing boundaries is increased. This gives an equivalent explanation of the observed phenomenon and suggests the way how to construct analytical model of prebiological evolution.

Another striking result of this study is that, though the fraction of sequences with a stable unique structure is very small, the algorithm generated them after a small number of mutations. This is the common feature of evolution-like algorithms: selection pressure generates a directed drift in sequence space that makes generation of desirable (in our case compact) chains much more feasible than if such sequences were searched for randomly.

Our model does not consider such important properties of proteins as chi-

rality and enzymatic activity. But we found that the requirement of compactness without aggregation can cause a stable three-dimensional native conformation, without which enzymatic activity is unlikely to be possible. The model presented in this work, while being unlikely to reproduce the details of prebiotic evolution, illustrates a general idea that a stable native structure could have evolved as a side effect of equilibrium conditions that selected sequences satisfying simple physicochemical requirements.

Acknowledgments

This work was supported by a Packard Fellowship.

References

1. A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **92**, 1282 (1995).
2. V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **93**, 839 (1996).
3. E.I. Shakhnovich and A.M. Gutin, *Nature* **346**, 773 (1990).
4. A.M. Gutin and E.I. Shakhnovich, *J. Chem. Phys.* **98**, 8174 (1993).
5. E.I. Shakhnovich and A.M. Gutin, *J. Chem. Phys.* **93**, 5967 (1990).
6. H.S. Chan and K.A. Dill, *J. Chem. Phys.* **95**, 3775 (1991).
7. C. Camacho and D. Thirumalai, *Phys. Rev. Lett.* **71**, 2505 (1993).
8. A. Sali, E.I. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235**, 1614 (1994).
9. R. Goldstein, Z.A. Luthey-Schulten, and P. Wolynes, *Proc. Natl. Acad. Sci. USA* **89**, 9029 (1992).
10. E.I. Shakhnovich and A.M. Gutin, *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).
11. E.I. Shakhnovich and A.M. Gutin, *Protein Engineering* **6**, 793 (1993).
12. E.I. Shakhnovich, *Phys. Rev. Lett.* **72**, 3907 (1994).
13. V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, *Protein Science* **4**, 1167 (1995).
14. S. Myazawa and R. Jernigan, *Macromolecules* **18**, 534 (1985).
15. P.H. Verdier, *J. Chem. Phys.* **59**, 6119 (1973).
16. H.J. Hilhorst and J.M. Deutch, *J. Chem. Phys.* **63**, 5153 (1975).
17. J.U. Bowie, R. Luthy, and D. Eisenberg, *Science* **253**, 164 (1991).
18. A. Grosberg and A. Khohlov, *Unsolved Problems in Physiccs of Polymers*, preprint NCBI, Puschino (in Russian) (1981).
19. G.F. Joyce, *Nature* **338**, 217 (1989).
20. S.A. Benner, A.D. Ellington, and A. Tauer, *Proc. Natl. Acad. Sci. USA* **86**, 7054 (1989).
21. C. Wilson and J.W. Szostak, *Nature* **374**, 777 (1995).