Stability of Designed Proteins against Mutations

R. A. Broglia,[1,2,3] G. Tiana,[1,3] H. E. Roman,[1,2] E. Vigezzi,[1,2] and E. Shakhnovich[4]
[1]*Dipartimento di Fisica, Università di Milano, Via Celoria 16, I-20133 Milano, Italy*
[2]*INFN, Sezione di Milano, Via Celoria 16, I-20133 Milano, Italy*
[3]*The Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark*
[4]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138*

(Received 28 September 1998)

The stability of model proteins with designed sequences is assessed in terms of the number of sequences, obtained from the designed sequence through mutations, which fold into the "native" conformation. By a complete enumeration of the sequences obtained by introducing up to four point mutations and up to seven composition-conserving mutations (swapping of amino acids) in a 36mers chain and by running dynamic simulations on the mutated sequences, it is found that this number depends on the gap between the native conformation and the bulk of misfolded conformations, but not on the particular designed sequence, provided its associated energy gap is large. [S0031-9007(99)09339-4]

PACS numbers: 87.15.By

A number of previous analyses [1–6] (reviewed in [7,8]) have provided arguments and supporting evidence for the deep connection existing between the energetic properties of protein sequences and their ability to fold fast into their native conformations. In particular, it was found [2–4] that the presence of a large (compared to the dispersion of interaction energies) energy gap between the native state and the bulk of misfolded conformations that are structurally dissimilar to the native state is an important factor that ensures fast folding into the native conformation (foldability requirement). A number of observations support the notion that sequences of natural proteins have been optimized to satisfy the foldability requirement.

(1) Random sequences undergo noncooperative folding transition [9,10] while designed sequences and proteins fold cooperatively [1,3,4,11,12].

(2) The native state of random sequences is very unstable even to small changes in potential function [13], while the ones that have larger gaps are much more robust with respect to changes in the energy function [14,15]. The latter behavior is characteristic of real proteins that exhibit the remarkable ability to maintain their native structure intact in a wide range of conditions including variation of temperature, pH, solvent composition, etc.

(3) It was shown theoretically that ground (native) states of random sequences are very unstable with respect to point mutations: The probability that a mutated sequence has the same native state scales as γ^{-8}, where γ is the number of conformations per one residue in the chain [16]. In contrast, real proteins are able to accommodate numerous mutations that are neutral with respect to structure changes [17] . This fact has obvious implication for the molecular evolution of proteins: It accounts for the existence of large families of proteins that may have diverged from a common root. Proteins belonging to a family have similar sequences and their native states are also structurally similar.

The stability of designed sequences with respect to point mutations has been demonstrated [18] in simulations, and the fact that larger energy gaps imply a greater ability of the designed sequence to accommodate many neutral mutations, that is, mutations with a nondegenerate ground state lying inside the gap, has been acknowledged [1–4,19–21]. In particular, Finkelstein *et al.* [22] studied the fold stability to mutations (in this case, to mutations in the potential), depending on the energy gap between the native and the bulk of misfolded conformations. The work by Govrindarajan and Goldstein [23,24] concentrated on the so-called neutral networks, that is, sets of mutations that preserve the native structure of a model protein determining its diffusion in the space of interaction parameters (cf. also Ref. [25]). While these analyses provided a number of important insights, in particular showing how the plasticity of protein sequences can exist with remarkable robustness of the resulting structure, they did not address the fundamentally important question of how many mutations can a sequence undergo without losing its ability to fold into the native state. In this paper, we address this question and provide, for the first time, the actual numbers obtained by exhaustive enumeration of mutations. We furthermore demonstrate, by running Monte Carlo (MC) dynamics, that starting from a designed sequence which displays a large gap, all mutated sequences which preserve (to some extent) the gap fold into the native conformation.

For the analysis we use a lattice model of a protein that has been used earlier by us [18,26,27] and others [28,29]. The model sequences are composed of amino acids of 20 types and contain 36 monomers. Two amino acids are considered interacting if they occupy neighboring positions on the lattice but are not sequence neighbors. The energy of the interaction depends on the identity of the amino acids involved, so that there is a 20 × 20 parameter matrix that describes the energetics in the model. We used the set of parameters suggested by Miyazawa and Jernigan (Table 6 of Ref. [30]). The associated standard deviation of the interaction energies between different amino acid types is $\sigma = 0.3$.

Our approach to protein simulations is based on the idea of designing sequences having a large energy gap in the target conformation chosen to serve as a native state for simulations [4,31]. A sequence that has sufficiently low energy in a conformation chosen as native is denoted as S_{36} (cf. caption to Fig. 1). This sequence is the same as was studied in previous publications [18,26,27]. In the units we are considering ($RT_{\text{room}} = 0.6$ kcal/mol), the energy of S_{36} in its native conformation [cf. Fig. 1(a)] is $E_N = -16.5$. Starting from a random configuration, the sequence S_{36} always reaches the native configuration, and it does it in a rather short "time," of the order of 10^6 MC steps. This is a consequence of the fact that the value of the energy gap δ ($= 2.5$), that is, the energy difference $E_N - E_C$ between the energy of the native and the lowest dissimilar configuration (configuration with a similarity parameter q [32] much smaller than one), is large, much larger than the variance of the contact energies. The value of E_C was calculated through a set of low temperature MC samplings in conformational space.

The goal of the present analysis is to characterize quantitatively how many mutations can S_{36} tolerate without losing the ability to fold into its native conformation. In other words, our study aims at providing an estimate of the number of sequences, having a certain degree of sequence similarity to S_{36}, that fold into its native structure. To characterize quantitatively single or multiple mutations, we ascribe to them a value ΔE [18], defined as the difference between the energies of the altered sequence (S'_{36}) and of the intact chain (S_{36}), both calculated in the native configuration [Fig. 1(a)]. The quantity ΔE is a measure of how the energy gap changes upon a mutation provided that the distribution of energies of conformations that are dissimilar to the native state remains unaffected by the mutation [18]. This was shown to be the case when mutations do not change the amino acid composition [3,4,18]. In this paper we have analyzed mutations which conserve and which do not conserve the composition of the protein. This gives a lower and upper limit for the number of mutations which the "wild-type" sequence can tolerate.

A complete enumeration of all sequences S'_{36} has been done up to seven mutations keeping fixed the amino acid composition of the chain (swapping), and up to four mutations without this constraint (pointlike). By simulating the dynamics [4] of sequences chosen among the mutated sequences, with the same composition of S_{36} (all sequences with two and three mutations and 10^3 randomly chosen sequences with seven mutations) and with $\Delta E < \delta$, it turned out that, in 98% of the cases, they can reach the native conformation in a time comparable to the folding time of S_{36}. Repeating the same analysis on 10^3 sequences with up to four pointlike mutations, we observed that only in 57 cases did the chain find conformations dissimilar from the native one, with lower energy, and it was not able to find the native conformation within the simulation time.

Furthermore, we studied the impact of pointlike mutations on sequences having different degrees of design. To this end we calculated the distributions $n_2(\Delta E)$ associated with two pointlike mutations for the case of three sequences designed to fold into the structure shown in Fig. 1(a) with different energy gaps, which in all cases

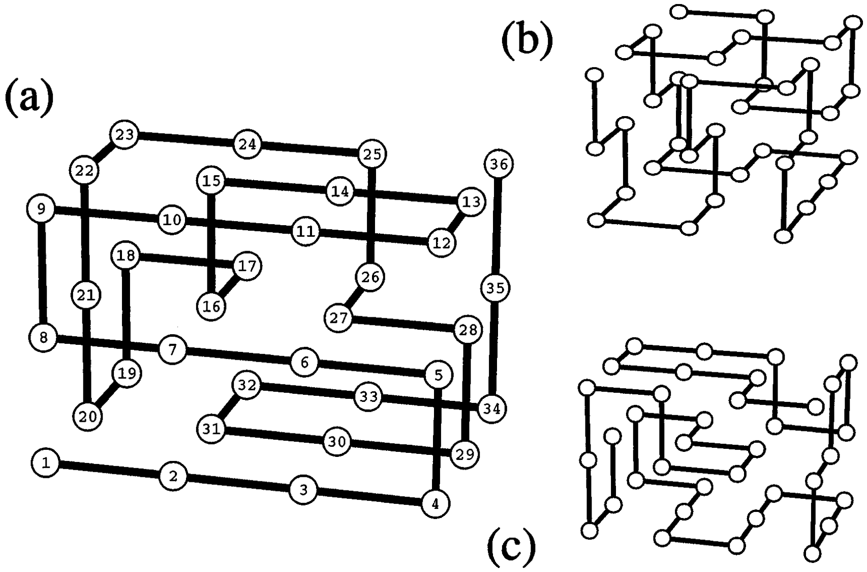

FIG. 1. (a) Conformation onto which the sequence $S_{36} \equiv$ SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA has been designed. In (b) and (c) we display two other fully compact target structures, used also as natives (for other sequences). These conformations were generated by collapsing a homopolymeric chain at low temperature.

fulfill the condition $\delta \gg \sigma$. The distributions for all three sequences appear to be very similar to each other (Fig. 2). This is also true for composition-conserving mutations and for the different numbers of mutations we have analyzed (data not shown). These results suggest that the distribution $n_m(\Delta E)$ associated with m mutations has some degree of universality, and applies to all designed sequences regardless of the value of δ (provided $\gg \sigma$). This is a nontrivial result since in the case under study, unlike the HP model of Ref. [25], there are unconnected regions of sequence space folding into the same native structure [33].

Given a sequence characterized by an energy gap δ, it is then possible to calculate the number of sequences which folds into the same native structure (i.e., for which there is still some energy gap between the native structure and the bulk of decoys) and which differs from the designed sequence by m mutations. To do this, one has to calculate the quantity

$$N_m(\delta) = \int_{-\infty}^{\delta} dE\, n_m(E)\,. \qquad (1)$$

As an example, we provide the function $N_m(\delta)$ for $m = 4$ and for the case of pointlike mutations (Fig. 3). An interesting dependence of the number of fold-preserving mutations on the initial fold stability (value of the energy gap δ) is observed, and can be parametrized in terms of a superposition of error functions. The curve increases smoothly as a function of δ, reaching a plateau at about $\delta \approx 8$, in keeping with the fact that the average value of $\overline{\Delta E}$ associated with single mutations on S_{36} is $\approx$0.9 [18], i.e., a value much smaller than δ.

The values of N_m for the sequence S_{36} for up to four pointlike and seven swap mutations are shown in Table I. The calculation of the total number of sequences $N(\delta) = \sum_m N_m(\delta)$ is beyond our calculational power, and can be established only with approximate methods [34]. In any case, the numbers shown in Table I, in particular that there are 2.78×10^8 sequences, i.e., 1.9% of all of the sequences generated by seven composition conserving mutations which preserve (to any extent) the gap, provide a lower limit to the total number of sequences folding into the same native structure. This fact is a result of our study rather than an assumption, obtained by running extensive dynamics simulations on the mutated sequences. To be noted is that the native fold stability is preserved by a sizable fraction of all of the investigated mutations, and that this fraction depends on the number of mutated residues rather than on whether the mutations conserve composition or not. In fact, the ratio of the number of mutations with $\Delta E < \delta$ to the total number of mutations are accurately parametrized by the Gaussian function $\exp[-0.5(m/m_0)^2]$ with $m_0 = 2.46$ for both composition conserving and pointlike mutations. The same study has been repeated using two other fully compact target structures [Figs. 1(b) and 1(c)], generated by the collapse of a 36 monomer homopolymeric chain at low temperature (below the θ point, see, e.g., [35]). The results are virtually identical to the ones shown in Fig. 3.

Previous studies have shown that the normalized gap $\xi = \delta/\sigma$ (or the closely related z score [1,5,32,36]) is a major determinant of the ability of sequences to fold. The results presented above further demonstrate that the stability of a designed sequence to mutations, and thus the number of mutations which preserve folding to the same native structure, is also controlled by the energy gap. To this end, the "resilience" of sequences against point mutations is directly related to their energetic impact: If the cumulative effect of mutations on the energy of the native state is weak enough so that the energy gap for

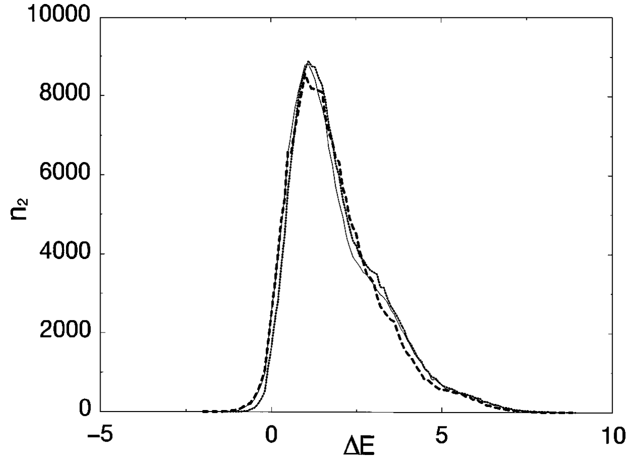

FIG. 2. Distribution $n_2(\Delta E)$ associated with two pointlike mutations, carried out in three different sequences, with gaps $\delta = 1.3$ (dashed curve), $\delta = 1.6$ (solid curve), and $\delta = 2.5$ (dotted curve), respectively. The three "root" sequences display no appreciable similarity.

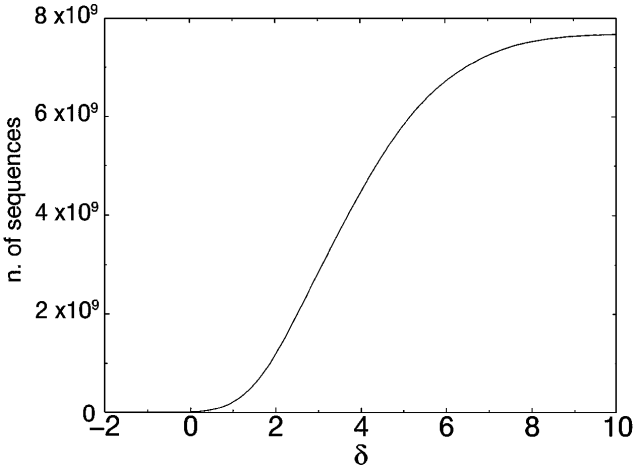

FIG. 3. Number of sequences which fold into the conformation shown in Fig. 1(a) and obtained from all possible four amino acid pointlike mutations of S_{36} that still preserve the gap.

TABLE I. The number of mutated sequences S'_{36} which fold into the native conformation shown in Fig. 1(a). In column one the number of mutations m is shown. Columns 2 and 3 are associated with composition-conserving results (c.), while columns 4 and 5 correspond to pointlike mutations (n.c.). Columns 2 and 4 display the number of sequences associated with a change in energy ΔE smaller than the gap δ, while columns 3 and 5 display the total number of sequences associated with the number of mutations m.

m	$\Delta E < \delta$ (c.)	Tot (c.)	$\Delta E < \delta$ (n.c.)	Tot (n.c.)
1			613	684
2	447	630	1.59×10^5	2.27×10^5
3	6518	14 280	2.30×10^7	4.89×10^7
4	1.37×10^5	5.30×10^5	1.99×10^9	7.68×10^9
5	2.10×10^6	1.66×10^7		
6	2.53×10^7	5.16×10^8		
7	2.78×10^8	1.55×10^{10}		

the native state remains, the mutations are neutral and the mutated sequences will still fold into the native state, albeit at a decreased stability.

In summary, in this paper we demonstrate, by running dynamical simulations on mutated sequences, that the whole issue of estimating the number of mutations that are tolerated by a designed sequence (and hence the number of sequences that fold to the same conformation) is reduced to enumerating mutations that keep the energy gap. Furthermore, we provide (by complete enumeration) a quantitative estimate of this number.

This work was partially supported by NIH Grant No. RO1 GM52126 (to E. S.) and by the School of Biophysics of the University of Copenhagen (to G. T.). Financial support by NATO under Grant No. CRG 940231 is gratefully acknowledged.

[1] R. Goldstein, Z. A. Luthey-Schulten, and P. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918–4922 (1992).

[2] A. Sali, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614–1636 (1994).

[3] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195–7199 (1993).

[4] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907–3910 (1994).

[5] S. Govindarajan and R. A. Goldstein, Biopolymers **36**, 43–51 (1995).

[6] A. Gutin, V. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).

[7] J. Bryngelson, J. N. Onuchic, N. D. Socci, and P. Wolynes, Proteins: Struct. Funct. Genetics **21**, 167–195 (1995).

[8] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29–40 (1997).

[9] K. Gast, A. F. Chafotte, D. Zirwer, Y. Guillou, M. Mueller-Fromme, C. Cadieux, M. Hodges, G. Damaschun, and M. Goldberg, Protein Sci. **6**, 2578–2588 (1997).

[10] A. Davidson and R. Sauer, Proc. Natl. Acad. Sci. U.S.A. **91**, 2146–2150 (1994).

[11] G. Makhatadze and P. L. Privalov, Adv. Protein Chem. **47**, 307–425 (1995).

[12] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10 428–10 435 (1991).

[13] J. D. Bryngelson, J. Chem. Phys. **103**, 6038–6045 (1994).

[14] V. Pande, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **103**, 1–10 (1995).

[15] A. P. de Araujo and T. Pochapsky, Folding Des. **1**, 299–314 (1996).

[16] E. I. Shakhnovich and A. M. Gutin, J. Theor. Biol. **149**, 537–546 (1991).

[17] T. Creighton, *Proteins Structure and Molecular Properties* (Freeman, New York, 1992).

[18] G. Tiana, R. A. Broglia, H. E. Roman, E. Vigezzi, and E. I. Shakhnovich, J. Chem. Phys. **108**, 757–761 (1998).

[19] S. Govindarajan and R. Goldstein, Proc. Natl. Acad. Sci. U.S.A. **93**, 3341–3345 (1995).

[20] E. I. Shakhnovich, Folding Des. **3**, R45–R58 (1998).

[21] M. Vendruscolo, Physica (Amsterdam) **249A**, 576–580 (1998).

[22] A. V. Finkelstein, A. Gutin, and A. Badretdinov, Proteins: Struct. Funct. Genetics **23**, 142–149 (1995).

[23] S. Govindarajan and R. Goldstein, Biopolymers **42**, 427 (1997).

[24] S. Govindarajan and R. Goldstein, Proteins **29**, 461 (1997).

[25] E. Bornberg-Bauer, Biophys. J. **73**, 2392 (1997).

[26] V. Abkevich, A. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10 026–10 036 (1994).

[27] V. Abkevich, A. Gutin, and E. I. Shakhnovich, J. Chem. Phys. **101**, 6052–6062 (1994).

[28] N. Socci, W. Bialek, and J. Onuchic, Phys. Rev. E **49**, 3440–3443 (1994).

[29] D. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070–4073 (1996).

[30] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534–552 (1985).

[31] E. I. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96–98 (1996).

[32] A. Gutin, V. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282–1286 (1995).

[33] G. Tiana, E. I. Shakhnovich, and R. A. Broglia (to be published).

[34] R. A. Broglia, D. Provasi, H. E. Roman, and G. Tiana (to be published).

[35] I. Lifshitz, A. Grosberg, and A. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978).

[36] J. U. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164–169 (1991).