Stability Threshold as a Selection Principle for Protein Design

Michele Vendruscolo\textsuperscript{1}, Amos Maritan\textsuperscript{1}, and Jayanth R. Banavar\textsuperscript{2}

\textsuperscript{1} Istituto Nazionale per la Fisica della Materia (I.N.F.M.) and International School for Advanced Studies (S.I.S.S.A.), Via Beirut 2-4, 34014 Trieste, Italy
\textsuperscript{2} Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802

The sensitivity of the native states of protein-like heteropolymers to mutations modelled as perturbations in the interaction potential between amino acids is studied. The stability threshold against mutations is shown to be zero for random heteropolymers on a lattice in two dimensions, whereas a design procedure modelling evolution produces a non-zero threshold. We introduce an evolution-like protein design procedure based on an optimization of the stability threshold that is shown to naturally ensure thermodynamic stability as well.

PACS numbers: 87.15.Bv, 87.10.+e

Natural proteins are made up of sequences of amino acids that fold rapidly into specific compact structures corresponding to minimum free energy states called native states \([1,1,2]\). The structure of the native state conformation controls the functionality of the protein. Because the number of possible random amino acid sequences and the number of possible conformations are gigantic, an important issue is an understanding of the selection principles that apply to protein sequences and/or native state structures.

The two key ingredients of evolution are diversity (afforded by the availability of 20 amino acids) and stability (a functionally useful sequence should not be mutated away). The stability of the occupancy of the native state on increasing the temperature (i.e. the thermodynamic stability) has been argued to be a characteristic of a good folder \([14,5,11,12]\). Here, we consider a different kind of stability against mutations modelled as structure selection. The effects of destabilizing factors such as variations in the denaturant concentration and site directed mutagenesis on the kinetics of folding have been recently investigated in Refs. \([14]\). A careful analysis of such perturbations has proven helpful in the clarification of the folding funnel in terms of collective coordinates \([14,16,17]\).

We consider self-avoiding chains of \(N\) monomers on a 2D square lattice. The Hamiltonian is

\[
H_\delta(\Gamma) = \sum_{ij} B_{i,j} \delta(r_i - r_j),
\]

where \(B_{i,j}\) is the coupling between monomer \(i\) and \(j\) and \(\delta\) is nonzero (and equal to 1) only if \(r_i\) and \(r_j\) are adjacent sites on the lattice and \(i\) and \(j\) are not next to each other in sequence. This Hamiltonian is well known in protein modelling \([16,17]\). For a given sequence (and \(B_{i,j}\) with \(N \leq 25\)), we enumerate the energies of all possible conformations and are able to determine the native state (ground state) conformation exactly.

We consider two versions of the model. In the first (which we will call the \(B_{i,j}\) model), the \(N\) monomers are assumed to be distinct. The \(B_{i,j}\) matrix is symmetric and has \(N(N + 1)/2\) elements. In order to obtain a random heteropolymer, these elements are drawn from a Gaussian distribution with mean value -2 and variance 1. Effectively, the matrix \(B\) represents a certain sequence. The model is identical to that studied by Dinner et al. \([16]\). The random
contact energies are in approximate correspondence to a more realistic parametrization of the contact energies by Miyazawa and Jernigan [20,21] or by Kolinski, Godzik and Skolnik [22]. In order to model evolved sequences with a large stability gap [1], we follow the rank-ordered procedure outlined by Shrivastava et al. [23] of shuffling the $B_{i,j}$ entries to assign the most favorable attractive interactions to the native contacts of the maximally compact native fold chosen as fixed target.

In the second model (denoted as the MJ model), each monomer is chosen to represent one of the twenty amino acids with the interactions determined by Miyazawa and Jernigan [20,21]. A random sequence would correspond to a random choice of the amino acids. In order to mimic the relevant feature of sequences selected by evolution it is no longer possible to follow the rank ordering procedure because the $B_{i,j}$ entries cannot be shuffled at will. Instead, after having fixed a target fold [3], we have used a recently proposed protein design procedure [12] entailing an optimization scheme in sequence space which allows one to obtain sequences with a desired native state conformation and a required measure of thermodynamic stability, enforced by fixing a “design temperature” $T_d$ and selecting those sequences with $T_j > T_d$.

Our calculations begin with the selection of two distinct interaction matrices which we shall call $B$ and $C$. We shall consider 4 choices for $B$: the random and the evolved $B_{i,j}$ and MJ models. $C$ is chosen randomly. The ground states of the $B$ and $C$ sequences are generally distinct. We now consider mutations of the sequence along a trajectory parametrized by a mixing coefficient $a \in [0,1]$ that changes the interaction matrix from $B$ to $C$ [17]:

$$B_a = (1-a)B + aC.$$  \hspace{1cm} (2)

The coefficient $a$ is a measure of the distance in sequence space between $B$ and $B_a$. This is a quite general perturbation of which the natural occurring ones are a subset. The structural similarity of the ground state conformations of these two sequences is given by the normalized distance $\Delta(a) = d(B_a,B)/d(C,B)$, where the distance $d(X,Y)$ is defined by

$$d(X,Y) = \sqrt{\sum_{i,j=1}^{N} (r_{i,j} - r'_{i,j})^2}.$$  \hspace{1cm} (3)

where $r_{i,j}$ and $r'_{i,j}$ are the Euclidean distances between amino acids $i$ and $j$ in the the two native states of sequences $X$ and $Y$ respectively. Note that $\Delta$ has been normalized so that it is 1 when $a = 1$, as long as the ground states of $B$ and $C$ are distinct.

Our primary probe of the stability to mutations is via a study of the dependence of $\Delta$ on $a$. Qualitatively similar trends are found for both the models – the signature of the selection in sequence space is in the quite distinct behavior of random and evolved sequences. A summary of our results for the behavior of the average $\Delta$ as a function of $a$ for $N = 16$ is shown in Fig. 1. The curves have been obtained as an average over 1000 realizations of independently chosen $B$ and for each of them over 1000 realizations of $C$ for the $B_{i,j}$ model and over 10 realizations of $B$ and for each of them over 1000 realizations of $C$ for the MJ model. The average stability threshold is zero for random heteropolymers [24] and is distinctly non-zero for the evolved cases.

Furthermore, the stability threshold goes up with the overall thermodynamic stability – in Figure 1, for the MJ model, the region of stability against mutations increases with the design temperature $T_d$. The threshold is somewhat reduced but is clearly non-zero when one considers rank ordered $B_{i,j}$ sequences that have native states in conformations that are not maximally compact. One increasing $N$, the number of monomers, comprising the evolved sequences, the stable phase along the $a$-axis increases in size along with a sharpening of the $\Delta - a$ curve suggestive of a sharp phase transition at the onset of instability in the thermodynamic limit.

One may also define an individual stability threshold $a_k(B,C)$ in the strength $a$ of the perturbation above which $\Delta$ becomes non-zero for the first time – the native structure of sequence $B$ is destabilized. Normalized probability distributions $P(a_k)$ of the individual stability thresholds for the random and evolved $B_{i,j}$ models are shown in Fig. 2. They underscore the different behaviors in the two cases. Our results, in the random case, are marginally related to a study of Bryngelson [25] and in the evolved case, to a recent study of Pande et al. [1], where the authors addressed the issue of stability of the ground state against inaccuracies in the potential. Bryngelson [25] used a mean field theory to estimate the probability of predicting the correct structure of a sequence of monomers if the interatomic potential is known only to an accuracy of $\eta$. A non-zero $\eta$ could arise from variations in the solvent properties or due to the imperfect parametrization or determination of the potential between amino acids or, as in our case, from mutations in the sequences. Pande et al. [1] showed analytically that the ground state
of designed sequences is significantly robust against
the introduction of random noise in the interaction
matrix.

Imperfectly folded proteins within the cell are de-
molished by proteolytic enzymes [24]. Unfolded pro-
teins can be present either as a product of a desta-
abilizing mutation or due to a variation of the solvent
properties. Thus the folded structure must be robust
against perturbations in order to survive in the cell.
The role of sequence selection is to produce robust
sequences which rapidly fold into a target confor-
mation with a specific function. Can one develop a
design scheme starting from existing functional se-
quen ces and produce artificial homologues with bet-
ter functionality?

The design scheme works as follows. 1) Select an
initial random sequence. 2) Compute its MJ matrix
B and its stability threshold by extracting a set of
100 realizations of the perturbation C. 3) Subject
the sequence to Monte Carlo optimization proce-
dure: monomers are swapped and the new sequence
is accepted if its stability threshold is increased. 4)
After 1000 such Monte Carlo steps, stop the opti-
mization and compute the folding transition tem-
perature $T_f$ of the resulting sequence. As shown in
Fig. 3, $T_f$ averaged over sequences correlates well
with the threshold.

We turn now to a recent study of Li et al. [15] who
suggested that certain highly designable structures
that are the unique native states of a large number
of sequences are special in that they are thermo-
dynamically more stable than other structures and
are stable against mutations in the sequence. Their
study was of chains comprising 27 monomers made
up of two kinds of amino acids, hydrophobic (H) and
polar (P), on a simple cubic lattice. The bare values
of their interaction parameters were expressed (af-
after scaling and shifting) in convenient units so that
$E(P,P)=0$, $E(H,P)=-1$ and $E(H,H)=-2.3$. Our own
studies of the identical model (without an overall
attractive shift in the bare interaction parameters
that would promote maximally compact structures)
necessitate the consideration of non-maximally com-
 pact. These findings show that, at least
for interactions that do not always lead to maxi-
 mally compact native states, the selection process
primarily involves the sequences and not the struc-
ture. The present study has concentrated on the
thermodynamic effects of mutations. The two-way
link between resistance against mutations and ther-
modynamic stability demonstrated here could also
have ramifications on the dynamics of folding, for
example, in the key role that specific amino acids
have in nucleation mechanisms of folding [16,25].

We are grateful to Hao Li and Chao Tang for use-
ful discussions. This work was supported by NASA,
NATO, NSF, The Petroleum Research Fund admin-
istered by the American Chemical Society and the
Center for Academic Computing at Penn State.

References

[1] C. Anfinsen, Science 181, 223 (1973).
[2] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. 34, 187 (1989).
[3] H. S. Chan and K. A. Dill, Physics Today, 46, 24 (1993).
[4] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. 90, 7195 (1993).
[5] V. S. Pande, Yu. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. 91, 12972 (1994).
[6] A. Sali, E. I. Shakhnovich, and M. Karplus, Nature 369, 248 (1994).
[7] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. 103, 9482 (1995).
[8] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science 267, 1619 (1995)
[9] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and
P. G. Wolynes, Proteins Struct. Funct. Genet. 21, 167 (1995).
[10] K. D. Klimov and D. Thirumalai, Phys. Rev. Lett. 76, 4070 (1996).
[11] J. Deutsch and T. Kurosky, Phys. Rev. Lett. 76, 323 (1996).
[12] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. 77, 1901 (1996).
[13] For both the $B_{i,j}$ and MJ models, we have repeated the analysis for three different target folds, with no significant difference in the results.
[14] C. A. Orengo, D. T. Jones, and J. M. Thornton, Nature 372, 631 (1994).
[15] H. Li, R. Hellig, C. TANG, and N. Wingreen, Science 273, 666 (1996).
[16] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. 105, 5860 (1996). J. N. Onuchic, N. D. Socci, Z. Luthey-Shulten, and P. G. Wolynes, Folding and Design 1, 441 (1996).
[17] In the case of the $B_{i,j}$ model, a more appropriate definition would be $B_a = \sqrt{1 - a^2 B + a C}$. With such a definition, the variance of $B_a$ would be unaffected and therefore the scale of the folding transition temperature would be unchanged. See Refs. [12, 16].
[18] A. Dinner, A. Sali, M. Karplus, and E. I. Shakhnovich J. Chem. Phys. 101, 1444 (1994).
[19] M. Cieplak, S. Vishveshwara, and J. R. Banavar, Phys. Rev. Lett. 77, 3681 (1996).
[20] S. Miyazawa and R. Jernigan, Macromolecules 18, 534 (1985).
[21] S. Miyazawa and R. Jernigan, J. Mol. Biol. 256, 623 (1996).
[22] A. Kolinski, A. Godzik, and J. Skolnik, J. Chem. Phys. 98, 7420 (1993).
[23] I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan, and J. R. Banavar, Proc. Natl. Acad. Sci. U.S.A. 92, 9206 (1995).
[24] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, cond-mat/9606130 preprint.
[25] J. D. Bryngelson, J. Chem. Phys. 100, 6038 (1994).
[26] T. E. Creighton, Proteins. Structures and Molecular Properties. W. H. Freeman & Company (New York), (1993).
[27] J. U. Bowie, R. L"othy, and D. Eisenberg, Science 253, 164 (1991).
[28] E. I. Shakhnovich, V. I. Abkevich, and O. B. Ptitsyn, Nature 379, 96 (1996).

FIG. 1. Average structural similarity $\Delta$ as a function of the perturbation $\alpha$. The non evolved and evolved cases for both the models used are shown. For the $B_{i,j}$ model the averages have been taken over 1000 realizations of $B$ and for each of them over 1000 realizations of $C$. For the MJ model we averaged over 10 realizations of $B$ and for each of them over 1000 realizations of $C$. In our design procedure for the MJ model, only $B$ sequences with a folding transition temperature $T_f$ greater than the indicated value of the design temperature $T_d$ (0.0, 1.0, and 1.2, respectively) were considered. $T_f > 0$ indicates no selection.

FIG. 2. The normalized probability distributions $P(\alpha_t)$ of the individual stability thresholds $\alpha_t$ for the non evolved and evolved $B_{i,j}$ case. The statistics are the same as in Fig. 1.

FIG. 3. Folding transition temperature $T_f$ as a function of the stability threshold $\alpha$ in the MJ model. The sequences have been designed through threshold optimization.

FIG. 4. The $Z$ score plotted against $N_S$ for the 2D version of the 3D model considered by Li et al. [15]. The $Z$ score is the difference between the average energy $\langle E \rangle$ of the compact conformations and the ground state energy, $E$, scaled by the dispersion $\sigma$, $Z = (\langle E \rangle - E)/\sigma$. $\langle E \rangle$ and $\sigma$ were calculated as averages over all configurations with 7 or more contacts, which are those competing to be the native state. Maximally compact configurations have nine contacts. The $Z$ score is a measure of the thermodynamic stability of the native state. Similar trends are observed on considering the energy gap instead of the $Z$ score.
Bij non evolved
Bij evolved
MJ Tf > 0.0
MJ Tf > 1.0
MJ Tf > 1.2
