

Analytical Approach to the Protein Design Problem

Edo L. Kussell and Eugene I. Shakhnovich

Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, Massachusetts 02138
(Received 22 April 1999)

We present an analytical method for determining the designability of protein structures. We apply our method to the case of two-dimensional lattice structures, and give a systematic solution for the spectrum of any structure. Using this spectrum, the designability of a structure can be estimated. We outline a hierarchy of structures, from most to least designable, and show that this hierarchy depends on the potential that is used.

PACS numbers: 87.10.+e

The protein design problem asks which and how many amino-acid sequences fold into a given protein structure. This problem, having obvious practical and evolutionary significance, attracted considerable attention and effort of experimentalists [1] and theorists [2,3]. In particular, Tang and co-workers carried a thorough study of the design problem for a simple square and cubic lattice model of proteins [3]. These authors enumerated all sequences made of two types of monomers and all conformations for two-dimensional 36-mers and three-dimensional 27-mers. This calculation provided a density of states in sequence space for the studied models, i.e., how many sequences have a given energy in a given conformation. For each conformation, the “designability,” which is defined as the number of sequences that have this conformation as a ground state, was determined in [3]. Interestingly, it was found that designability varied from conformation to conformation. Furthermore, the structures that possess certain degrees of symmetry turned out to be most designable [3]. This result tempts one to speculate that the factor of designability may be the cause of symmetries observed in real proteins.

The physical reason for the observed relationship between geometric properties of conformations and their designability remained unclear. In particular, it is important to evaluate how robust this relationship is with respect to choice of lattice model, model protein “alphabet,” and interaction parameters between model aminoacids. Recent numeric studies [4] and [5] showed that the set of the most designable structures does indeed depend on the potential and amino-acid alphabet used.

A deeper understanding of the designability problem requires an analytical solution that may shed light on the general geometric features that make structures designable. Such a solution will facilitate an informed connection between model results and real proteins.

In this Letter, we present a statistical mechanical analysis of designability, and give an analytical solution to the problem on a two-dimensional lattice. Our analysis is based on the analogy between protein design and a class of statistical-mechanical spin models established in earlier works [2,6,7].

We use the standard lattice model to represent protein conformations. On a lattice, a structure is defined as a self-avoiding walk of length N . The energy of a sequence of residues $\{\sigma_i\}$ in a structure with coordinates of the monomers $\{r_i\}$ is

$$H(\{\sigma_i\}, \{r_i\}) = \sum_{i < j}^N U(\sigma_i, \sigma_j) \Delta(r_i, r_j). \quad (1)$$

$\Delta(r_i, r_j)$ is defined to be 1 when r_i and r_j are nearest neighbors on the lattice and not sequence neighbors along the chain. For real proteins the definition of a contact requires a choice of a cutoff distance [8]. The folding potential is given by a symmetric matrix $U(\alpha, \beta)$ whose entries are the pairwise interaction energy between amino-acid residues of types α and β .

In any instance of the design problem, the structure $\{r_i^0\}$ is given (we will refer to this structure as the “target conformation” for design), and we are asked to find all sequences whose native (folded) state is $\{r_i^0\}$. The number \mathcal{N} of such sequences is the *designability* of the structure. Since the residues of a protein are tightly packed in space, only maximally compact lattice structures are generally studied. On a square lattice these correspond to self-avoiding walks that completely fill a square. The folding potential H is therefore assumed to have a nonspecific attractive term (which we have omitted), that favors compact conformations.

In order for a sequence to fold and to be stable in the desired target conformation, it must have lowest energy in this conformation compared to its energy in all alternative conformations. The spectrum of energies of alternative conformations is qualitatively similar to that of a random heteropolymer, i.e., it consists of a continuum part, with a high density of states whose lowest energy is E_c , and of a few discrete energy levels lying below E_c [9,10]. Importantly, the value of E_c is self-averaging, i.e., it depends on amino-acid composition rather than on a particular sequence [2,11].

Thus the issue of designability is reduced to the question of how many sequences with a given amino-acid composition can fold into a given conformation with an energy $E \ll E_c$. Each structure has a sequence-space

energy spectrum given by the energy density

$$n(E, \Delta) = \sum_{\{\sigma_i\}} \delta[H(\{\sigma_i\}, \Delta) - E]. \quad (2)$$

The above expression corresponds to a system of N spins $\{\sigma_i\}$ confined to an interaction geometry dictated by $\Delta(r_i, r_j)$, with $\sigma_i \in 1, \dots, q$, where q is the number of possible residue types. In this language the amino-acid composition of a sequence is equivalent to total magnetization and is denoted by the letter C , which is in general a q -dimensional vector. The probability $p(E, C)$ that a sequence $\{\sigma_i\}$ of composition C , will fold into a structure Δ , where $E = H(\{\sigma_i\}, \Delta)$, is [12]

$$p(E, C) = \exp\left[\frac{-1}{\sqrt{4\pi N \log \gamma}} \exp\left(\frac{E - E_c}{T_c}\right)\right], \quad (3)$$

where T_c is the temperature of the thermodynamic “freezing” transition in random heteropolymers having composition C , and γ is the effective number of conformations per monomer [13]. These properties of E_c render it a valuable link between sequence and structure space. If we know the spectrum $n(E, \Delta, C)$ of a structure Δ over all sequences of fixed composition C , we can compute its expected designability as follows:

$$\mathcal{N}(\Delta) = \sum_{E, C} P(E, C) N(E, \Delta, C). \quad (4)$$

On a two-dimensional square lattice, maximally compact structures have a property that makes the spectrum calculation $n(E, \Delta, C)$ analytically tractable. If the ends of the chain are prescribed to be neutral—they do not interact with any monomers—then the entire structure can be decomposed into a direct product of one-dimensional loops and strands. A loop is a collection of S monomers in which monomer i is in contact with monomer $i + 1$, for $1 \leq i \leq S - 1$, and monomer 1 is in contact with monomer S ; if monomer 1 and monomer S are not in contact, we call the collection a strand. Figure 1 shows the decomposition of a particular 36-mer into strands and loops (ends have been allowed to interact in this figure). This decomposition derives from the restriction that nearest neighbors along the chain are not in contact, leaving each monomer in the interior of a structure exactly 2 contacts to make. Since this excludes branched systems of contacts, the only possibilities left are loops and strands. Monomers lying on the square border of a structure can make only one contact because the chain must occupy two of the three neighboring sites. Strands, therefore, must begin and end on the border of a structure, and loops must lie entirely within it. Since the corners of a structure cannot interact, a square structure of length N has $4\sqrt{N} - 8$ interacting border sites. Thus, there must be $2\sqrt{N} - 4$ strands in each structure. Including the ends of the chain would add at most 2 branched systems, or would change the number of strands by at most 2. While the number of strands is largely the same for all structures, the distribution of lengths of strands is structure dependent, and will prove to be the determinant of designability.

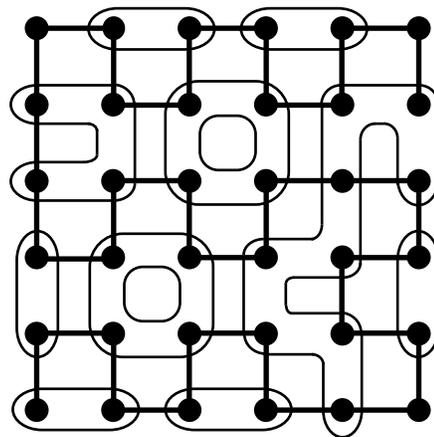


FIG. 1. Strand/loop decomposition for a 36-mer.

We restrict this discussion to systems with two types of amino acids, which we denote $+1$ and -1 , although the general methodology can certainly be applied to larger alphabets. Following the analogy between protein design and spin models [2,7], amino-acid types will be called “spins” in what follows. The interaction energies are taken to be $U(1, 1) = a$, $U(-1, -1) = b$, and $U(1, -1) = U(-1, 1) = d$. The general aim is to compute $n(E, \Delta, C)$, which can be deduced from the partition function at composition C , over all sequences $\{\sigma_i\}$:

$$Z(C) = \sum_{\{\sigma_i\}} e^{-\beta H\{\sigma_i\}} \delta\left(C - \sum_i \sigma_i\right).$$

Note that composition in the 2-spin case is a scalar corresponding to the net magnetization. The delta function can be expressed as a Fourier integral, and if we define the partition function for the j th loop or strand to be

$$Z_j(\omega) = \sum_{\{\sigma_i\}_j} e^{-\beta H\{\sigma_i\} - i\omega \sum \sigma_i},$$

where $\{\sigma_i\}_j$ is the set of all spins in the j th loop or strand, we can write

$$Z(C) = \int e^{i\omega C} d\omega \prod_{j=1}^{s+l} Z_j(\omega), \quad (5)$$

where s and l are the number of strands and loops, respectively. $Z_j(\omega)$ can be computed by the transfer matrix method; we define the transfer matrix T as

$$T = \begin{pmatrix} A e^{-i\omega} & D \\ D & B e^{i\omega} \end{pmatrix},$$

where $A = e^{-\beta a}$, $B = e^{-\beta b}$, and $D = e^{-\beta d}$. If λ_+ and λ_- are the eigenvalues of T , we have for a loop, $Z_j(\omega) = \lambda_+^{n_j} + \lambda_-^{n_j}$, while for a strand

$$Z_j(\omega) = \sum_{\sigma_1, \sigma_{n_j}} e^{-(\sigma_1 + \sigma_{n_j})i\omega/2} \langle \sigma_1 | T^{n_j-1} | \sigma_{n_j} \rangle, \quad (6)$$

where n_j is the length of the j th loop or strand. The expression for strands is more complicated than that for loops because the first and last monomers in a strand have

to be summed over separately. By diagonalizing T , the following expression for strands can be derived:

$$Z_j(\omega) = [(2D - A - B)(\lambda_+^{n_j-1} - \lambda_-^{n_j-1}) + (e^{i\omega} + e^{-i\omega})(\lambda_+^{n_j} - \lambda_-^{n_j})]/(\lambda_+ - \lambda_-). \quad (7)$$

The eigenvalues can be calculated, and one observes that if $D^2 = AB$, we get $\lambda_+ = Ae^{-i\omega} + Be^{i\omega}$, while $\lambda_- = 0$. This situation corresponds to any potential for which $d = (a + b)/2$. We will call this the symmetric potential, because any such potential is the same as $a = 1, b = -1, d = 0$. Under the symmetric potential, $Z_j(\omega) = \lambda_+^{n_j}$ for loops, and $Z_j(\omega)$ for strands simplifies considerably to

$$Z_j(\omega) = \lambda_+^{n_j-2}(\sqrt{A}e^{-i\omega} + \sqrt{B}e^{i\omega})^2. \quad (8)$$

If N is the chain length of a structure, the total partition function (5) for the symmetric potential now becomes

$$Z(C) = \int e^{i\omega C} d\omega \lambda_+^{N-2s}(\sqrt{A}e^{-i\omega} + \sqrt{B}e^{i\omega})^{2s}.$$

Using the binomial expansion, the integral can be calculated exactly and the density of states, $n(E, C)$, obtained. Figure 2 shows a plot of $n(E, C)$ using $a = -1, b = 1, N = 36$, and $C = 0$.

Under the symmetric potential, a structure's spectrum depends only on its length, composition, and the number of strands. Since these three are the same for each structure, the spectra of all structures are identical (up to end effects). This result implies that all structures have the same expected designability, with fluctuations only due to the effects of interacting ends and to any sequence-independent terms of E_c . For large N these terms will be insignificant.

The symmetric potential is a singular case in that structural features play no role in designability. It has previously been identified in [5] as an important special case for design, where designabilities of structures were obtained computationally. The narrow range of designabilities observed in [5] shows that end effects are indeed small even for small N . We now show how designability becomes sensitive to structure under perturbations of this potential.

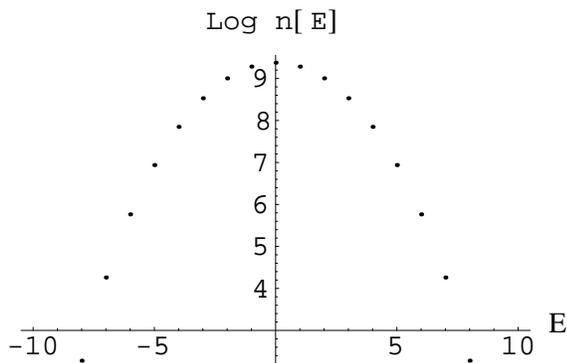


FIG. 2. Symmetric potential spectrum of a 36-mer.

Under the symmetric potential, one eigenvalue is zero. If we add a small perturbation u to the potential, so that $a' = a + u, b' = b$, and $c' = (a + b)/2$, we would expect one eigenvalue still to be small compared to the other. Under a perturbed potential, then, $\lambda_-/\lambda_+ \ll 1$. Starting from the general expression for strands (7), we observe that if $\lambda_- < \lambda_+$, we have two geometrical series in $R = \lambda_-/\lambda_+$, and we can rewrite it as

$$Z_j(\omega) = \lambda_+^{n_j-2} \Theta \left[1 + \sum_{k=1}^{n_j-2} R^k - (e^{i\omega} + e^{-i\omega}) \frac{\lambda_-}{\Theta} R^{k-1} \right],$$

where $\Theta = Ae^{-2i\omega} + Be^{2i\omega} + 2D$. The corresponding expression for loops is

$$Z_j(\omega) = \lambda_+^{n_j}(1 + R^{n_j}).$$

Using these expressions in Eq. (5) for $Z(C)$ results in a product of the loop and strand Z_j 's. We can expand this product in R , and obtain the following kinds of terms: The principal contribution $\lambda_+^{N-2s} \Theta^s$ (zero order in R) is a structure-independent term corresponding to the symmetric potential. Terms coming from a single loop of length n_j look like $\lambda_+^{N-2s} \Theta^s R^{n_j}$. Terms coming from a single strand of length n_j are

$$\lambda_+^{N-2s} \Theta^s \left[R^k - (e^{i\omega} + e^{-i\omega}) \frac{\lambda_-}{\Theta} R^{k-1} \right], \quad (9)$$

where $1 \leq k \leq n_j - 2$. Strands of length 2 contribute no terms. Thus, strands have terms at all powers of R less than $n_j - 1$, while loops only perturb at R^{n_j} . Since the smallest possible loop is of length 4, strands of lengths greater than 2 are much more important to the structure of the spectrum than loops of any length.

To calculate a perturbed spectrum, one still has to back transform the Fourier variable ω in each term. Setting $U = e^{-B\omega}$, the perturbed eigenvalues must satisfy $\lambda_+ + \lambda_- = AUe^{-i\omega} + Be^{i\omega}$ and $\lambda_+\lambda_- = AB(U - 1)$. The principal contribution to the partition function is

$$Z_{\text{principal}} = \int \lambda_+^{N-2s} \Theta^s e^{i\omega C} d\omega,$$

which can be computed by binomial expansions of λ_+^{N-2s} . The only difficulty in doing so comes from odd powers of a square-root term in the expression for λ_+ , which would normally lead to a power series expansion rather than a finite binomial expansion. The approximation we make in our computation is to replace the odd powers of the square root with the next even number, so that we have terminating binomial expansions at all powers. To compute a perturbation of the form $\lambda_+^{N-2s} \Theta^s R^k$, we use $R = AB(U - 1)/\lambda_+^2$ and obtain

$$Z_{\text{perturb}} = \int \lambda_+^{N-2s-2k} \Theta^s A^k B^k (U - 1)^k e^{i\omega C} d\omega,$$

which is similar to $Z_{\text{principal}}$ except for another binomial that can be expanded. Computing perturbations of any order is therefore relatively straightforward.

We performed computations to check our theory with a full enumeration of sequences and structures that was published by Tang and coauthors [3]. These authors used

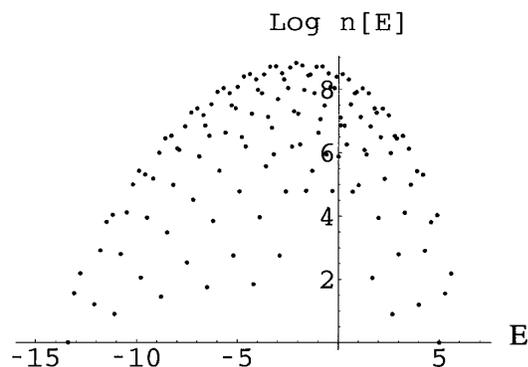


FIG. 3. Principal contribution to 36-mer spectrum with $u = -0.3$.

a potential ($a = -2.3$, $b = 0$, $d = -1$) which can be rescaled to the perturbed potential $a = -1$, $b = 1$, $d = 0$, $u = -0.3$. Figure 3 shows a log plot of the principal (structure-independent) contribution to the spectrum of a 36-mer. This lowest-order approximation of the density of states bears a striking resemblance to the spectrum calculated by full enumeration for 2D structures (Ejtehadi [14]). A full, perturbative spectrum calculation is in good agreement with the enumerated spectrum. Qualitative features such as the shape and the anomalous speckled pattern emerge from our theory. The pattern of the spectrum is essentially a splitting of the states of the symmetric potential due to the perturbation u .

The expected designability (4) depends on the value of E_c . In Fig. 4 is a plot of the change in expected designability due to the first-order strand perturbation as a function of E_c . Note that the logarithm of negative numbers was taken in absolute value and plotted as a negative value.

The change in the expected designability of the structure due to this perturbation is negative for a large range of E_c in the lower part of the spectrum. Since strands of lengths greater than 2 will contribute a first-order perturbation, the more such strands in a structure, the less designable it will be. The most designable structures will, therefore, contain the maximum number of strands of length 2. This is the major effect determining designability; adding higher-order terms does not change this result. The effect which

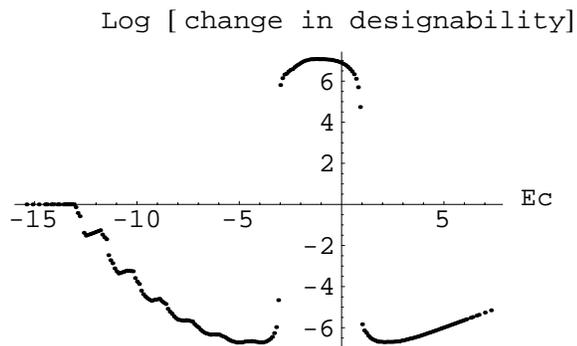


FIG. 4. Change in expected designability due to first-order perturbation.

2-strands have on designability is due to the condition $u < 0$. One can also perturb with $u > 0$ and obtain the opposite result: 2-strands are not desirable. All of the plots shown in Fig. 4 correspond to composition $C = 0$. Similar plots are obtained over a large range of C , with significant qualitative differences for $|C|$ close to N , i.e., as the sequence approaches a homopolymer. Our conclusion remains unchanged by these differences because at extreme compositions we expect very few sequences to design any given structure.

Deciding whether or not a strand of length n is better for designability than a strand of length $n + 1$ requires looking at the change in expected designability due to higher order perturbations. We find that up to strands of length 5, the shorter the strand, the more designable the structure. For longer lengths, differences in designability become very small, and resolution of these differences requires precise knowledge of E_c . Researchers have conjectured that minimal degeneracy [15] and high designability [3] lead to highly symmetrical structures. With $u < 0$, it seems plausible that the constraint of having many 2-strands necessitates symmetry in the most designable structures.

We have presented a general, analytical approach to the protein design problem, and by applying it to the case of two-dimensional structures, we have shown how it effectively recognizes features that make some structures much more designable than others. This allowed us to identify a hierarchy of designability based on strand lengths. Future work should generalize these results to larger alphabets (which leads to a q th degree polynomial in λ), and to lattices of dimension three (which leads to highly branched systems of contacts).

We thank Dr. Grosberg and Dr. Gutman for interesting discussions, W. Ejtehadi and co-workers for sharing their data, and NIH for support.

- [1] B. I. Dahiyat and S. L. Mayo, *Science* **278**, 82 (1997).
- [2] E. I. Shakhnovich, *Fold. Des.* **3**, R45 (1998).
- [3] H. Li *et al.*, *Science* **273**, 666 (1996).
- [4] N. E. G. Buchler and R. A. Goldstein, *Proteins* **34**, 113 (1999).
- [5] M. R. Ejtehadi *et al.*, *J. Phys. A* **31**, 6141 (1998).
- [6] E. I. Shakhnovich and A. Gutin, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7195 (1993).
- [7] E. I. Shakhnovich and A. Gutin, *Protein Eng.* **6**, 793 (1993).
- [8] S. Miyazawa and R. Jernigan, *Macromolecules* **18**, 534 (1985).
- [9] E. Shakhnovich and A. Gutin, *Nature (London)* **346**, 773 (1990).
- [10] J. Bryngelson and P. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524 (1987).
- [11] E. Shakhnovich, *Phys. Rev. Lett.* **72**, 3907 (1994).
- [12] A. Sali *et al.*, *J. Mol. Biol.* **235**, 1614 (1994).
- [13] A. Gutin *et al.*, *J. Chem. Phys.* **108**, 6466 (1998).
- [14] M. Ejtehadi (private communication).
- [15] K. Yue and K. Dill, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 146 (1995).