# Statistical Theory of Protein Combinatorial Libraries

Hidetoshi Kono, Jinming Zou and Jeffery G. Saven

Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104, USA

## ABSTRACT

Combinatorial experiments provide new ways to probe the determinants of protein folding and to identify novel folding amino acid sequences. These types of experiments, however, are complicated by both enormous conformational complexity and by large numbers of possible sequences. We present and apply a statistically based, computational approach for identifying the properties of sequences compatible with a given main chain structure. The method yields the likelihood of each of the amino acids at preselected positions in a given protein structure. The theory may be used to quantify the characteristics of sequence space for a chosen structure without explicitly tabulating sequences. We apply the method to consider the energetic separation of a target structure from other possible structures and to identity the monomer probabilities at selected positions of the immunoglobulin light chain-binding domain of protein L, for which many variant folding sequences are available.

***Keywords***: protein engineering, structural biology, bioinformatics, combinatorial chemistry

## 1 INTRODUCTION

Protein folding spans biology, physics, and chemistry and has applications to biomedicine and biomaterials. Since proteins are the direct products of genes, folding is fundamental to the expression of genetic information in the cell. A quantitative and predictive understanding of protein folding will accelerate the interpretation of genomic information. Folding is also of fundamental physical interest, since it involves spontaneous ordering at the molecular scale. With few exceptions, proteins fold reversibly to unique structures. The three-dimensional folded structure of a protein is encoded in its sequence of amino acids. Thus we may be able to predict structure from sequence alone and to design desired folded structures through careful choice of sequence. Using synthetic sequences, features important in protein stability and folding kinetics may be probed via selective mutations. Once particular structures can successfully designed, the opportunity then exists for the design of novel functional proteins. Potentially, these ideas can be expanded beyond the naturally occurring biopolymers. Folding polymers, both biological and synthetic, could yield new types of structures and properties and lead to novel pharmaceuticals, catalysts, and materials.

Combinatorial experiments provide new ways to probe the determinants of protein folding and to identify novel folding sequences [1]–[5]. In these experiments, large numbers of distinct sequences are created and assayed for predetermined structural or functional properties. Combinatorial experiments may be used to discover novel sequences that fold to a particular structure and to provide a broader picture of the determinants of protein folding than simple site directed modifications alone. Combinatorial techniques are complicated, however, by both the enormous conformational complexity of proteins and by the large numbers of possible sequences. For example, the number of possible protein sequences having N amino acids is $20^N$. A characterization of the ensemble using conventional molecular modeling is not feasible, but surveying large portions of the library will be useful, however, for revealing trends concerning the interactions that stabilize a particular structure or for identifying sequences with a desired structure and function. While a number of powerful computational techniques for protein design have been developed [6]–[12], a quantitative computational theory that avoids explicit generation of particular sequences will be helpful in designing and interpreting combinatorial experiments. Many such experiments have been guided by qualitative chemical considerations, but a quantitative theory will yield a more efficient means of harnessing the diversity of the molecules studied.

We present and apply a statistically based, computational approach for identifying the properties of amino acid sequences compatible with a given main chain structure. The theory may be adapted to include a number of important features of protein folding, including the effects of protein side chain conformations which may be included in an atom-based fashion. Calculations may be performed for a variety of similar backbone structures to identify sequence properties that are robust with respect to minor changes in main chain structure. Rather than specific sequences, the method yields the likelihood of each of the amino acids at preselected positions in a given protein structure. The theory may be used to quantify the characteristics of sequence space for a cho-

sen structure without explicitly tabulating sequences. The accuracy of the theory has been verified using small exactly solvable model systems, for which explicit enumeration of sequences is feasible [13], [14]. We also apply the method to calculate the identity probabilities of selected positions of the immunoglobulin light chain-binding domain of protein L, for which many variant folding sequences have been examined experimentally. While the experimental sampling of possible sequences is necessarily sparse, the calculations compare favorably with the experimentally observed amino acid probabilities [5], [15].

## 2 THEORY

The statistical, entropy-based theory has a structure very similar to statistical thermodynamics and also draws upon contemporary molecular modeling techniques to estimate the number and composition of sequences that are likely to fold to a given three dimensional structure. In the theory, constraints can be introduced to focus combinatorial libraries. Such constraints can be physical (e.g., the overall energy of sequences) or synthetic (e.g., the patterning of amino acid properties). This theory yields the number and composition of sequences likely to be compatible with a particular structure and a chosen set of constraints [14]. The theory takes as input a given target structure and a many-body energy (or scoring) function. Importantly, because explicit enumeration is avoided, the properties of an exponentially large number of possible protein sequences can be addressed.

Theis statistical approach addresses the number and composition of sequences compatible with a particular folded protein structure. Let $S$ be the logarithm of the number of sequences for a target structure. If the energy of the sequences is fixed, $S$ is equivalent to a microcanonical entropy and may be termed the *sequence entropy.* As in statistical thermodynamics, $S$ is maximized with respect to any unconstrained internal parameters. Here the internal parameters include the probabilities $w_i(\alpha)$ that each site $i$ in the structure is occupied by monomer type $\alpha$.

$$S = -\sum_{i=1}^{N}\sum_{\alpha=1}^{m} w_i(\alpha)\ln w_i(\alpha) \qquad (1)$$

$N$ is the chain length and $m$ is the number of possible monomer types. $S$ is maximized subject to constraints on the sequences. The constraints need only be functions of the monomer probabilities. The constraints may specify values of such global quantities as the folded state energy $E_f$. "Patterning" constraints can also be incorporated, wherein certain amino acids are precluded from occupying particular sites [1]. Composition constraints may be included to specify the number of each type of monomer used in making the sequences in the library. A constrained maximization of $S$ yields the $w_i(\alpha)$. In this way, number and composition of sequences having particular values of $E_f$ and any other constraint conditions can be determined.

The theory is practical and extendable. Other constraints may be easily included. Upon introducing constraints, the number of different molecules in a particular chemical ensemble decreases. This reduction in library size is due to a fundamental concept in statistical thermodynamics: the imposition of any internal constraints in a system usually decreases the overall entropy. Thus the theory may be used to design and focus combinatorial experiments. For a given target structure, we can quickly investigate the ramifications of such constraints on the number and identities of allowed sequences. For example, correlations between monomers may be identified by constraining the identity of one position and examining how this affects the identities of nearby residues. In implementing the theory, a set of self-consistent nonlinear equations is solved, but the computational time necessary to solve these equations goes as $N^a$, where $N$ is the number of residues and $a = 1 \sim 2$. In contrast, the time required for explicit tabulation is exponentially dependent on $N$. Thus the theory provides a tractable method of characterizing and designing sequence ensembles of folding macromolecules, where typically $N = 10^1 \sim 10^3$.

## 3 APPLICATIONS

### 3.1 Lattice Model and Stability Gap

We have tested the theory by comparing its results with those of an exactly solvable system, a cubic lattice polymer ($N = 27$ and $m = 2$), where the exact enumeration of all $2^{27}$ sequences is computationally feasible [13]. For a "protein-like" energy function [16], the theory is in excellent agreement with the exact results for both $S(E_f)$ and the sequence identity probabilities $w_i(\alpha)$. The theory may also be used to focus libraries on regions having lower values of the target state energy $E_f$, by fixing the hydrophobicity of buried residues [13].

A sequence with a low energy in the target structure need not necessarily fold to that structure. Such a sequence may possess other conformations that are of comparable energy; such sequences will not fold to unique structures. Researchers have addressed these issues by using design criteria that account for structures other than the target [17], [18]. We are incorporating such criteria into the statistical theory of protein sequences. A simple measure of sequence-structure compatibility that takes into account non-target structures is the "stability gap," which we define as $\Delta = E_f - \langle E \rangle_u$ [14]. Here, $E_f$ is the energy of a sequence in the target

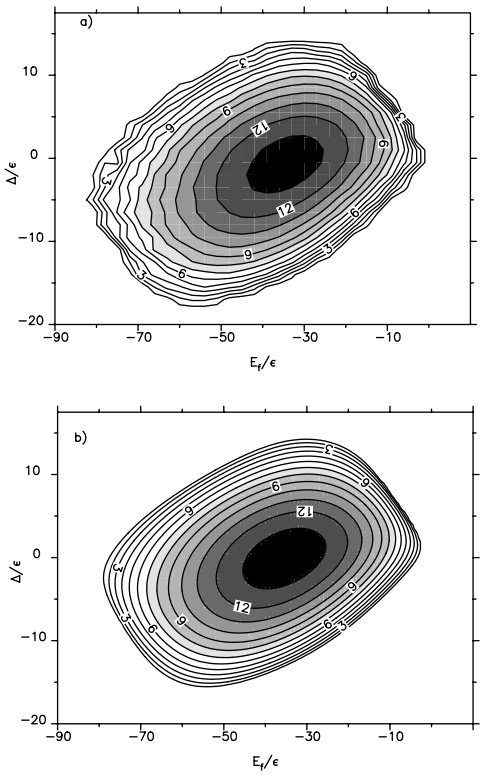


Figure 1: $S(E_f, \Delta)$ for 27-mer lattice model. The target structure is the most "designable" structure of Li et al [16]. Two types of monomer are used. The energy function is described in [14]. (a) Exact enumeration result. (b) Theoretical result. The spacing between contours is 1, and the range extends from $S = 1$ to $S = 13$.

structure, and $\langle E \rangle_u$ is the average energy of a set of non-target (or unfolded) conformations.

We have extended these methods so that the distribution of the stability energy gap $\Delta$ may also be estimated. In order to specify the stability gap $\Delta$, an ensemble of non-folded states is required. We choose the set of all 103,346 compact, cubic conformations [19]. Using the theory, we can examine the number and composition of sequences in a library as functions of both $E_f$ and $\Delta$ (Figure 1). As can be seen, the theory is in excellent agreement with the results of the exact enumeration. The range and shape of the sequence entropy are recovered quantitatively by the theory. Note that there is only a weak correlation between $E_f$ and $\Delta$, in agreement with the notion that energy minimization alone may be insufficient for sequence design. The theoretical estimates for $S$ and $w_i(\alpha)$ are in excellent agreement with the exact results for different values of $E_f$ and $\Delta$ [14].

These lattice studies illustrate a number of key features of the method. The comparison with the exact results in each case confirms that we have an accurate computational tool for estimating the number of sequences and monomer probabilities for arbitrary structures. The theory can be used to identify monomers that are important to stabilizing particular structures. Local and global constraints on the sequences may be imposed to address particular questions and to focus a combinatorial library. We have also shown how the stability of sequences may be tuned by either stabilizing the folded state (decreasing $E_f$) or by destabilizing unfolded structures (increasing $\langle E \rangle_c$).

## 3.2 Side chain packing

We have extended the statistically based, computational approach to identify the properties of sequences compatible with a given main chain structure. The theory includes a number of important features of protein folding. Protein side chain conformations are included in an atom-based fashion. As a result, the complexity of the problem is greatly increased, since not only are multiple amino acids possible, but each amino acid has multiple side chain conformational (rotamer) states. Calculations are performed for a variety of similar backbone structures to identify sequence properties that are robust with respect to minor changes in main chain structure. Rather than specific sequences, the method yields the likelihood of each of the amino acids at preselected positions in a given protein structure. To account for hydrophobic effects, we have developed an environmental energy that it is consistent with other simple hydrophobicity scales and have shown that it is effective for side chain modeling. We calculate the amino acid probabilities at selected positions of the immunoglobulin light chain-binding domain of protein L, for which many variant folding sequences are available [5]. The calculations compare favorably with the experimentally observed identity probabilities, especially for buried residues (see Fig. 2).

# REFERENCES

[1] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685, 1993.

[2] R. T. Sauer. Protein folding from a combinatorial perspective. *Folding & Design*, 1:R27–R30, 1996.

[3] D. D. Axe, N. W. Foster, and A. R. Fersht. Active barnase variants with completely random hydrophobic cores. *Proc Natl Acad Sci USA*, 93:5590–5594, 1996.

[4] Biao Ruan, Joel Hoskins, Lan Wang, and Philip N. Bryan. Stabilizing the subtilisin bpn pro-domain by phage display selection, how restrictive is the amino acid code for maximum protein stability? *Protein Sci*, 7:2345–2353, 1998.

[5] D. E. Kim, H. D. Gu, and D. Baker. The sequences of small proteins are not extensively optimized for

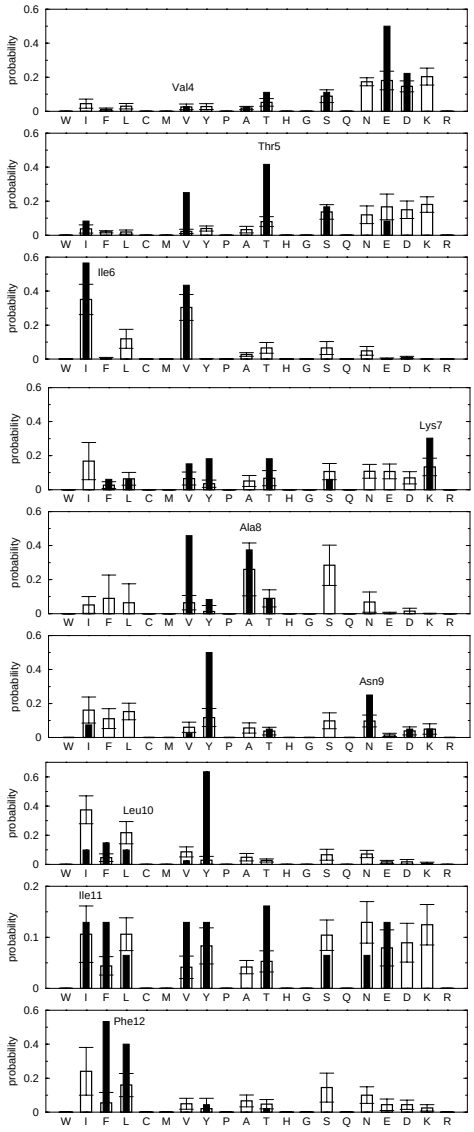


Figure 2: Calculated amino acid probabilities for strand 1 of protein L (shown by open bars with error bar). Filled bars are the probabilities based on the amino acid frequency obtained from a phage-display selection experiment [5]. The sequence position and amino acid of the wild type are indicated on each plot. Even though the experiment involves a sparse sampling of only 22 sequences, the calculated and observed probabilities are in good agreement in many cases and trends with regard to hydrophobic identity are maintained.

rapid folding by natural selection. *Proc. Natl. Acad. Sci. U. S. A.*, 95:4982–4986, 1998.

[6] E. I. Shakhnovich and A. M. Gutin. A new approach to the design of stable proteins. *Protein Eng*, 6:793–800, 1993.

[7] D. T. Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci*, 3:567–574, 1994.

[8] H. W. Hellinga and F. M. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA*, 91:5803–5807, 1994.

[9] H. Kono and J. Doi. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins*, 19:244–255, 1994.

[10] John R. Desjarlais and Tracy M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci*, 4:2006–2018, 1995.

[11] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1462–1467, 1998.

[12] X. Jiang, E. J. Bishop, and R. S. Farid. A de novo designed protein with properties that characterize natural hyperthermophilic proteins. *J. Am. Chem. Soc.*, 119:838–839, 1997.

[13] J. G. Saven and P. G. Wolynes. Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. Phys. Chem. B*, 101:8375–8389, 1997.

[14] J. Zou and J. G. Saven. Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. *J Mol Biol*, 296:281–294, 2000.

[15] H. Gu, N. Doshi, K. T. Kim, K. T. Simons, J. V. Santiago, S. Nauli, and D. Baker. Robustness of protein folding kinetics to surface hydrophobic substitutions. *Protein Science*, 8:2734–2741, 1999.

[16] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.

[17] R.A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc Natl Acad Sci USA*, 89:9029–9033, 1992.

[18] E. I. Shakhnovich. Protein design: a perspective from simple tractable models. *Folding & Design*, 3:R45–R58, 1998.

[19] A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding. a lattice model study of the requirements for folding to the native state. *J Mol Biol*, 235:1614–1636, 1994.