Computational Analysis of Sequence Selection Mechanisms

Leonid Meyerguz[1], Catherine Grasso[2], Jon Kleinberg[1], and Ron Elber[1*]

[1]Department of Computer Science

4130 Upson Hall

[2]Center of Applied Mathematics

Cornell University

Ithaca NY 14853

Phone: (607)255-7416
Fax: (607)255-4428
e-mail: ron@cs.cornell.edu

Summary:

Mechanisms leading to gene variations are responsible for the diversity of species, and are important components of the theory of evolution. One constraint on gene evolution is that of protein foldability; the three-dimensional shapes of proteins must be thermodynamically stable. We explore the impact of this constraint and calculate properties of foldable sequences using 3660 structures from the protein databank. We seek a selection function that receives sequences as input, and outputs survival probability based on sequence fitness to structure. We compute the number of sequences that match a particular protein structure with energy lower than the native sequence, the density of the number of sequences, the entropy, and the "selection" temperature. The mechanism of structure selection for sequences longer than two hundred amino acids is approximately universal. For shorter sequences, it is not. We speculate on concrete evolutionary mechanisms that show this behavior.

I. Introduction

An intriguing question in evolution is that of the mechanism of sequence selection; what is the means that selects the sequences we see today? The space of all possible sequences for a protein of length L is enormous (20^L possibilities). Only a tiny fraction of this space is realized in genomes, making the selection process particularly intriguing.

To model evolutionary processes we consider a selection function denoted by G that accepts as an input a protein sequence and returns a survival probability. The decision depends on a set of variables that may include (for example) protein function, protein flexibility, protein stability, and more. Here we focus on only one variable: the stability energy, E, of sequences embedded in 3660 alternative structures. The single variable, E, is clearly important (unstable proteins could not possibly work) and study of its impact will provide a reference framework for potential studies of other variables (e.g. biological activity, flexibility, etc.). The energy of a sequence is determined by fitting it into a structure. We examine the sequence-structure matches, one structure at a time.

To determine the space of selected sequences, we examine the number of alternative sequences as a function of energy. A rough estimate suggests that the number of sequences of a protein of length L with energy lower than E is growing exponentially in L. Since typical protein lengths (L) are between 100 and 1000 amino acids, this number (which we denote by $N(E)$) is enormous. While it is naïve to think that the sequence space has been explored extensively by nature, it is useful to examine known protein

sequences and shapes and to infer mechanisms that may have led to the current distribution of proteins.

We assume that native sequences were selected because they were highly probable as a function of energy. Hence, the probability of sampling a native sequence A_n with energy E_n is a maximum as a function of energy. This assumption is tested in the present manuscript by examining the energy distribution of homologous proteins. If the distribution of the energies of homologous proteins is narrow and centered near the energy of the native proteins then the above assumption is expected to be sound.
The probability of finding a sequence between energy E and $E+\Delta E$ is given by $P(E)\Delta E = G(E)\Omega(E)\Delta E$, where $P(E)$ is the probability density, ΔE is a small energy interval, and the number density as a function of energy - $\Omega(E)$ - is given by $\Omega(E)\Delta E = N(E+\Delta E) - N(E) \approx \frac{dN}{dE}\Delta E$. Since $\Omega(E)$ is a rapidly increasing function, $G(E)$ must be a rapidly decreasing function to keep the energies of probable sequences below the unfolded state. Here we use our assumption of highly probable native sequences and search for a maximum of the probability for the native sequence A_n and its corresponding native energy E_n, we have

$$\left.\frac{d\log[P(E)]}{dE}\right|_{E_n} = \left.\frac{d\Omega/dE}{\Omega(E)} + \frac{dG/dE}{G(E)}\right|_{E_n} = 0$$

As discussed below we can (and do) compute the function $N(E)$ and its derivatives: $\Omega(E)$ and $d\Omega(E)/dE$ for 3660 proteins. This calculation makes it possible to estimate $-\left.\frac{dG(E)/dE}{G(E)}\right|_{E=E_n}$ that we denote by β_n .

The above definition of β_n is similar to the usual estimate of the Boltzmann factor in statistical mechanics in configuration space (Feynmann 1982). One may adapt the language of statistical mechanics and identify β_n with an inverse temperature of selection -- $\beta_n = 1/T_n$.

There are two points worth emphasizing here: (a) $G(E)$ depends on more than the energy (we already mentioned above function and flexibility and it is likely that more variables are relevant). The coupling between the different variables is not quantified and has the potential of complicating our analysis. The relatively simple picture described below for the temperature of selection across the whole protein databank (Berman, J. et al. 2000) is therefore even more striking. (b) The inverse temperature β_n is a property of a single protein structure, which is obtained by computing the selection variable for one protein shape (and many fitting sequences). There is no obvious reason why β_n is the same or similar for different proteins, and we investigate the properties of β_n in order to answer this question.

Analyzing sequence space using statistical mechanics tools was pioneered by Gutin and Shakhnovich (Shakhnovich and Gutin 1993). In particular they have shown the analogy

of the protein design problem to the ferromagnetic Ising model and demonstrated sequence optimization for four proteins and lattice models. Further analysis of sequence space was provided by a calculation of the entropy (Shakhnovich 1994) using an alternative protocol to the approach that we use below. A few groups had followed with further theoretical analysis of sequence space and its relation to structure. Betancourt and Thirumalai (Betancourt and Thirumalai 2002) had suggested a new measure for protein foldability, and Saven and Wolynes (Saven and Wolynes 1997) considered an alternative analysis of sequence space and protein designability that is based on counting with constraints (using Lagrange multipliers).

Significant advances were made in the understanding of general principles. However, the detailed simulations were limited to model systems or to only a few protein shapes. It is of interest to have a more global view on the currently known space of protein folds and to examine its relationship to sequence space. Providing the desired global view is the purpose of the present manuscript.

Numerical studies of sequence to structure fitness were done by Koehl and Levitt (Koehl and Levitt 2002) and by Larson et al (Larson, England et al. 2002). A number of sequences and shapes (from the protein databank) were considered and examined; atomically detailed models were employed to generate realistic structural models. Nevertheless, the atomically detailed models are expensive to compute, and make it difficult to efficiently sample sequence space and to compute statistical mechanic functions. In the present manuscript we use simple protein energies that can be computed

rapidly. These simplified representations were used extensively in protein annotation, and identification of correct folds. While not as accurate as atomic potentials, they do capture structural fingerprints of many families, and we therefore use them here.

At the core of the statistical mechanic computation is a counting procedure. We count the number of sequences, $N(E)$, that fit into a structure as a function of energy, E. And, as discussed above, we compute from $N(E)$ the functions $\Omega(E)$ and T_n for 3660 proteins.

To summarize the major finding of this manuscript: We have introduced an efficient algorithm to count sequences that fit a given shape, and apply it to obtain estimates of the absolute number of sequences that fit 3660 representative structures of the protein databank.

The first observation is that it is easy to find sequences that are significantly more stable than the native sequence, and this is for a number of fitness measures. It means that structural stability and energy considerations are clearly insufficient to account for the current (natural) distribution of proteins. The current set of proteins is not optimized for structural stability. While we are able to account for other factors of selection by the "temperature", the origin of the selection temperature is not clear and further studies are warranted. For example, since it is so easy to generate more stable proteins, what are the mechanisms that lead to the relatively high natural selection temperature(s)?

Perhaps the most striking (second) observation of the present manuscript is the calculation of a singly peaked distribution of selection temperatures, suggesting a similar selection mechanism throughout the currently known shapes of proteins. A comparison of

the selection temperatures to an estimate of the physical folding temperatures shows weak correlation between the two.

II. The database and energy functions

II.1 The database of protein structures and native sequences

Our goal is to compute the statistical properties of sequence space for all known folds of proteins, one structure at a time. The first objective is to determine which structures will be studied. Using our experience in protein annotation that requires similar databases (Meller and Elber 2002), we adopted a reduced and non-redundant set of protein shapes that is used for fold recognition. This set includes 3660 structures, which represents well the known folds of the protein databank. It can be found at http://www.cs.cornell.edu/~leonidm/jm_list.txt. We removed from the original set (with 3904 structures) the longest proteins (longer than 500 amino acids) for which the convergence of the counting procedure was poor. The structures were selected according to (yet another) structural alignment algorithm of our own design (Meller and Elber, unpublished – see below). The complete protein databank was scanned to avoid shape redundancy, and only shapes with RMS distances of at least 6Å to the rest of the structures in the set were added to the database. The comparison of any two structures was done in two steps. In the first step the structures were aligned, and in the second step the RMS distance between the two aligned structures was computed.

To compute an alignment, contact maps were used. The geometric centers of the amino acid side chains (called here (structural) sites) provide the Cartesian representation of the

structure. From the Cartesian representation all the contacts between the sites are computed and stored (by definition, a pair of structural sites is in contact if it is separated by a distance smaller than 6.5Å).

In our structural alignment algorithm two layers of contacts are used providing more detailed description of structural sites. Consider a primary site i that is in contact with J sites. We record the number of contacts, m, to each of the J sites and use it as a second identifier. A contact to site i has two indices (j,m) $j=1,...,J$. Comparing two structures, sites are examined for the identity of their contacts, and are aligned using dynamic programming. Structural sites that are not aligned against gaps are used in the second step of an RMS calculation. In the RMS calculations the aligned residues are overlapped by overall translation and rotation to minimize the distance between the two Cartesian representation of the structures (Kabsch 1978). The prime advantage of this heuristic algorithm is of efficiency; however, other structural comparison algorithms could have been used. Indeed, the set of 3904 structures was later tested using the CE structural alignment algorithm (Shindyalov and Bourne 1998). We obtained very similar results, indicating that our database provides a non-redundant representation of the protein databank.

II.2 The energy functions

Two energy functions are used to count sequences with energies lower than the native energy: THOM2 (Meller and Elber 2001) and TE-13 (Tobi and Elber 2000). We determined the potential parameters by solving linear inequalities. We required that

energies of native sequences embedded into native structures are lower than the energies of native sequences embedded into decoy structures. Both functions are based on the geometric centers of the side chains and were discussed and tested extensively in the past. We briefly describe the two energy functions below.

In computing the THOM2 energy, we use the same definition of contacts as the one described in the structural comparison section above. The geometric centers are computed for the native structures and are used also for the probe sequences even if the amino acid side chain was changed. This procedure is computationally efficient and was found superior to the alternative use of C_α-based potentials (Tobi and Elber, unpublished). Potentials based on the geometric centers recognize more native folds than potentials based on C_α positions (with the same number of parameters).

The energy is given in a table with 3 indices -- $E_{ij}\left(\alpha_i, j_i, m_j\right)$ where E_{ij} is the energy associated with the contact between amino acids in sites i and j, α_i is the type of the amino acid in site i, j_i the index of a contact to site i, and m_j is the number of contacts to site j. The total energy in THOM2 is given by a sum over all pairs (i, j). Table I provides the THOM2 energies

*** PLACE TABLE I HERE ***

Since THOM2 is independent of the identity of the amino acid at site j, it is possible to use THOM2 to score efficiently an optimal alignment of a sequence to a structure (with dynamic programming). This is the primary advantage of the THOM2 energy function. The other energy function that we use (TE13) cannot be scored optimally with such

techniques. However, TE13 is more accurate than THOM2 in the sense that more native folds are recognized in an extensive test (Tobi and Elber 2000).

The TE13 model is a distance-dependent pair-energy. The pairwise interaction between structural sites i and j is extracted from a table, $E_{ij}\left(\alpha_i,\alpha_j,r_{ij}\right)$. The table entries depend on the two types of the amino acids, α_i,α_j, and on the distance between the sites r_{ij}. The distance is binned into 13 steps for detailed parameterization. The total number of TE13 parameters is therefore $210\times13=2730$ that can be found at http://www.cs.cornell.edu/~leonidm/te13/te13.htm . To compute the total energy all pairwise interactions are summed.

Note that in the counting described below we do not align the sequences and structures since they are of the same length, L. The two energy functions are used to test the "sensitivity" of our results to the details of the potentials.

III. The algorithm

For each protein in the set we calculate the number of sequences $N(E)$, which is exponentially large in the protein length. Therefore, it is necessary to avoid working with the full search space explicitly. The difficulties are similar to those encountered in the calculations of partition functions in computational statistical physics. In fact, $dN(E)/dE$ plays the role of the partition function for the microcanonical ensemble. The above observation suggests that approaches to estimate the partition function are likely to be useful here as well. An alternative algorithm for the calculation of the sequence

entropy as a function of temperature was put forward by Shakhnovich (Shakhnovich 1994).

The algorithm we employ is closely related to umbrella sampling (Torrie and Valleau 1977), and to randomized algorithms for approximate enumeration (Jerrum and Sinclair 1996). The basic step is the estimation of the ratio $N(E+\Delta)/N(E)$, for given E and Δ. This ratio can be estimated efficiently with a Markov chain in sequence space. Let A_i^0 be the current sequence. From the current sequence we generate at random a new sequence A_i^1 in which (at most) two amino acids are modified. Two types of random changes to the sequence (mutations) are considered: (i) a site is picked at random and the amino acid in that site (a_k) is modified (at random again) to $a_{k'}$; and (ii) two sites are picked at random and their corresponding amino acids exchange positions. Option (i) is more complete but may over-count sequences with low sequence diversity (e.g. homopolymers) that can be artificially stable in our models; option (ii) is more restricted (the composition of the amino acids is fixed), but it avoids the problem with sequences of low complexity (e.g. the creation of poly-cysteine as the ultimate most stable sequence). There were more studies with a fixed composition (Shakhnovich and Gutin 1993; Koehl and Levitt 2002) but studies with variable composition are also available (Saven and Wolynes 1997; Saven 2002). The energy of the new sequence in the fixed structure is computed according to the current model (either THOM2 or TE-13). To further explore measures of sequence-to-structure fitness we also consider the effect of reverse sequence in which the energy (in this case THOM2) is computed for the current sequence, A, and the inverse of that sequence, A_r. The score energy in the last case is $E_{THOM2}(A)-E_{THOM2}(A_r)$. The idea is

to mimic a random sequence by the inverse sequence and to compute a "cheap" measure of significance. This measure was used effectively in Hidden Markov models (Karchin, Cline et al. 2003). Finally, we also tested a random sequence (with the same amino acid composition) instead of an inverse sequence in the formula above. The results were remarkably similar to the reverse sequence and are therefore not shown in the manuscript.

If the energy is lower than $E+\Delta$, the sequence is used in the counting described below, otherwise the sequence is rejected and another random selection is attempted. This process is run for $l(E+\Delta)$ accepted steps, where we define $l(E)$ to be the number of these steps in which the sequence had energy below E. The ratio $N(E+\Delta)/N(E)$ is estimated as $l(E+\Delta)/l(E)$. For sufficiently small Δ, the ratio is close to 1 and the counting converges rapidly. Multiple calculations of ratios close to 1 (called telescoping) make it possible to estimate ratios significantly different from 1. For example, (k is a positive integer)

$$N(E+k\cdot\Delta)/N(E)=\prod\left[\frac{N(E+k\cdot\Delta)}{N(E+(k-1)\Delta)}\right]\left[\frac{N(E+(k-1)\cdot\Delta)}{N(E+(k-2)\Delta)}\right]\left[\begin{matrix}\ldots\\ \ldots\end{matrix}\right]\left[\frac{N(E+\Delta)}{N(E)}\right]$$

where each term in the product is computed by a separate Markov chain.

Using the "telescoping" procedure, we are able to compute the function $N(E)$ up to an unknown normalizing constant c; in other words, we are computing $\bar{N}(E)=N(E)/c$.

The constant c affects the calculation of the number density (and the entropy defined below) by the addition of a constant number and does not affect the temperature, since we have

$$S = \log\left(\Omega(E)\right) = \log\left(c\, d\bar{N}/dE\right) = \log\left(d\bar{N}/dE\right) + \log(c)$$

$$T = \left(dS/dE\right)^{-1} = \left[\frac{d\log\left(dN/dE\right)}{dE}\right]^{-1} = \left[\frac{d\log\left(c\cdot d\bar{N}/dE\right)}{dE}\right]^{-1} = \left[\frac{d\log\left(d\bar{N}/dE\right)}{dE}\right]^{-1}$$

Estimating c (in addition to $\bar{N}(E)$) cannot be done with the telescoping procedure described above, and an alternative is required. The problem is similar to the determination of absolute free energies, which is known to be computationally harder than the determination of free energy differences.

Nevertheless, in the present case shortcuts can be found. Let $N(\infty)$ be the total (known) number of sequences. It is 20^L for option (i) of sequence sampling, and for option (ii) it is $\frac{L!}{\prod_{\alpha=1}^{m} l(\alpha)!}$ where L is the length of the sequence, α is the type of the amino acid, and $l(\alpha)$ is the number of amino acids of type α (Saven and Wolynes 1997). Let $\langle E\rangle$ be the energy averaged over all possible sequences that can be computed efficiently by a direct sampling procedure as the average energy of randomly sampled sequences: $\langle E\rangle = \frac{1}{K}\sum_{k=1} E(A_k, X)$ (where A_k is a random sequence). We then sample an additional set of K random sequences. Let $K_{\downarrow}/K$ be the fraction of sequences that are below $\langle E\rangle$;

we therefore have $N(\langle E\rangle) = N(\infty)\cdot K_{\downarrow}/K$. The ratio $N(E)/N(\langle E\rangle)$ is computed by telescoping and the known value of $N(\langle E\rangle)$ is used to determine the absolute value of $N(E)$.

Finally we wanted to test the hypothesis of a sharply peaked distribution of native energies; an assumption that we used in the beginning of the discussion on the selection function. For a sharply peaked distribution we expect the energies of the homologous sequences embedded in the native structure to be narrowly distributed. To find homologous proteins each of the sequences in our list of proteins was aligned against the NR database of protein sequences (Benson, Karsch-Mizrachi et al. 2000). Using the BLAST algorithm with an E value of 0.001 we have found about 360,000 matches. For each of the matches we computed the energy per-amino-acid of the aligned segments since the homologous sequences are not necessarily of the same length. A sample from the 360,000 data points is shown in figure 1.a demonstrating that the energies per amino acid of the homologous proteins are indeed peaked. For comparison we also show a histogram of the variances of the energies for homolog and random sequences in figure 1.b.

Results

The number of sequences up to the native energy, the number density, and the native temperature were computed for 3660 proteins using a Markov chain of sequences that were fitted into a single structure at a time. Typical counting results for individual proteins of the same length (150 amino acids) are shown in figure 2. All of the plots approach the same value since the maximum number of sequences is the same 20^{150} .

However, the rate at which they approach this maximal value depends on the protein shape. We have computed the function of $N(E)$, the derivative of this function $\Omega(E)$ was computed by Shakhnovich and Gutin (1993). When we compute the derivative of the number of sequences the results are qualitatively similar. We note however that the simple Gaussian model of sequence entropy is missing a phase transition in sequence space that we observed frequently in our set (Meyerguz, Elber, and Kleinberg, unpublished)

Despite the significant variation in the number of sequences (of the same lengths) with energy below the native sequence, we found the length of the protein to be the dominant factor in determining the sequence number as is demonstrated in figure 3. The figure includes counting with four different protocols: (i) counting with TE13, (ii) counting with THOM2 (iii) counting with THOM2 and fixed composition of amino acids, and (iv) THOM2 counting using reverse sequences. The counting is performed for the representative set of the protein databank described above. Note that (i), (ii) and (iv) are very similar and only the counting with fixed composition shows different quantitative behavior (but similar qualitative behavior).

The results for the temperature calculations are presented as a function of the protein lengths for the different counting protocols in figure 4. Each protein is presented as a dot. In figure 4 we also show a histogram plot for the individual temperatures computed with TE13 for proteins longer than 200 amino acids.

The linear relationship of figure 3 is observed on a logarithmic scale, and the straight line is rather thick. The explicit calculations of the temperature, using the derivative of the entropy, show deviations from a constant value. The temperature is roughly a constant only for proteins longer than 200 amino acids and even then there is a spread around the average. This pattern repeats for the two energy functions that we considered, suggesting it to be a general phenomenon. Note that this is at variance with the results for homologous sequences (figure 1) for which the normalized protein energy is essentially a constant for all protein lengths.

It is of interest to compare the selection temperature to the physical temperature of folding (i.e. the temperature in which the protein chain melts) or (alternatively) to the stability energy of the folded state and to check if the two are correlated. We have computed the stability energy for a selected set of 235 proteins. The unfolded state was set to the zero energy using the procedure described in (Betancourt and Thirumalai 1999) and the difference from the native energy was computed. Hence, we estimate the stability energy by changing the reference state. Instead of a reference state of a random sequence, the reference state is of contacts with only water molecules. These stability energies are correlated with the folding temperature. However, when we examine the correlation of the stability energies with the selection temperatures the correlation was weak (correlation coefficient of 0.6).

Discussion

Before continuing to the main conclusion of the present study it is worth comparing the results using different counting procedures and energy functions. We first comment on the calculation of the function, $N(E)$, the number of sequences up to a given energy E. There are two striking observations in the log plots of figure 3. The first is the remarkable similarity in the counting results using two very different energy functions TE13 and THOM2. The TE13 is a distance dependent pairwise potential that is the most accurate function we have at our disposal. The THOM2 energy is a one body potential that we can compute and align efficiently. On the log plot of figure 3.a both energies lead to $\log(N(E))$ that are similar, suggesting that conclusions based on this function are likely to hold for alternative forms of energy functions.

On the other hand significant variations in $\log(N(E))$ are observed for different counting protocols. Alternative sampling procedures were introduced in the past in an attempt to overcome the problem of "positive" and "negative" design (see below). We consider the matches of alternative sequences into a single structure (we repeat the calculations of the matches to each of the 3660 structures, one structure at a time). Our design is therefore "positive" in the sense that we check if the present sequence fits the probe structure. We do not check if alternative structures do not match the same sequence ("negative" design). Consider a poly-cysteine. This amino acid is highly hydrophobic and adds significantly to the stability of the protein (negative energy). For all compact structures and energy functions (we know), poly-cysteine is a (very) low energy sequence. However, in reality, homopolymers do not adopt a unique globular shape since all compact structures have about the same energy, making it impossible to identify a

single lowest energy minimum. The positive design of sequences does not consider alternative structures that may match the probe sequence. We therefore conclude that counting all sequences introduce an artifact of over counting, since homopolymers are added to the counting incorrectly.

How significant is the extra counting of homopolymers?

In principle this problem can be addressed by testing the fitness of a new sequence to all alternative structures and not only to a single shape (Meyerguz, Kleinberg, and Elber, work in progress). However, exact enumeration in this case is considerably more computationally demanding (if at all possible), and it is desirable to have a cheaper solution.

One suggestion is to consider only sequences with the native composition of amino acids (Shakhnovich and Gutin 1993; Koehl and Levitt 2002). This constraint does not allow for homopolymers to develop; however, it reduces significantly the space of all available sequences. Figure 3.b shows that not only the log of the number of sequences, $\log\left[N(E)\right]$, is significantly smaller (seen even on a log scale) but also the fluctuations in $\log\left[N(E)\right]$ are much larger. Is this change "real" in the sense that fixing the composition shifts the distribution in the correct direction?

A gentler fix to the homopolymers problem was proposed in a different context (Karchin, Cline et al. 2003) (see the **Results** section). The score energy is defined as the difference $E_{THOM2}(A) - E_{THOM2}(A_r)$ where A_r is the reverse sequence. The energy of a homopolymer is now zero regardless of the structural template. The zero energy is usually quite high, which helps eliminating the contributions of homopolymers to the counting. Moreover, there is no constraint on the amino acid composition, so we

anticipate more complete statistics. The results for counting with the above energy difference (figure 3.b) are essentially identical (on a log scale) to the full counting in figure 3.a, using THOM2 or TE13. Hence, the contribution of homopolymers to the statistic of sequences is quite small and either direct counting in full sequence space or counting with the reverse sequences as a reference can be used. The fixed composition, however, behaves differently. In retrospect, this is not surprising. The number of homopolymers is exponentially smaller than the number of heteropolymers. At the native energies (which are reasonably high), their impact is small.

The main conclusion of our study is the universality of the selection function throughout the protein data bank for proteins longer than 200 amino acids. The function G appears to have the same dependence on energy in the neighborhood of the native state over an extensive set of proteins. A plausible guess for a functional form for G will therefore be: $G(E,\Gamma;\beta) = f(\Gamma)\exp(-\beta E)$ where β is the inverse selection temperature, E is the stability energy of the protein, and Γ represents the rest of the variables (as protein function, flexibility, and other relevant features that we did not include in our studies). The (unknown) function $f(\Gamma)$ assigns weights to all these features.

Our study suggests that to a good approximation β is a universal parameter across different families of proteins. It is not at all obvious (at least to the authors) that the "sequence-energy excitation" with respect to the lowest energy sequence will be comparable in all protein families. This suggests that the mutation mechanism (at least as measured by stability energy for proteins longer than 200 amino acids) is universal. Such universality is consistent with two models: (1) Models of connectivity within different

clusters in protein space (Maynard 1970). (2) A single mutation mechanism that produces similar statistical distribution of sequences in isolated (structural) islands. While evidence for extensive connectivity has been found in simplified models of protein structure (Lau and Dill 1990; Lipman and Wilbur 1991; Huynen, Stadler et al. 1996; Kleinberg 1999) it has been difficult to demonstrate this connectivity in sequence space for realistic structural models of proteins. At present we are unable to differentiate between (1) and (2).

Our method thus has the effect of treating sequence space, which is far too large to analyze directly, by analogy with a physical system that can be "probed" at various points to test whether it is well-mixed. Within this framework, the uniformity of the temperature at these probe points offers evidence for a background mixing process on distinct protein families. This background mixing can be achieved by a uniform selection mechanism or by direct mixing (Migration) of one protein family to the other.

Is the "selection" temperature a consequence of the "physics" of folding and therefore it does not contain information on evolutionary processes?

We comment that in our hands the selection temperature is only weakly correlated with the stability energy. Suggesting that there are other contributions to the selection temperature not accounted for by simple physical arguments (e.g. number of contacts). Another evidence that not only "physics of folding" is playing a role here is the observation that the native energies as a function of sequence are high on the energy scale and are not fully optimized as far as structural stability is concerned. The selection temperature is a measure of how other factors influence the observed distribution of proteins and the amusing observation that is it peaked across many families (at relatively

high energies) suggests a uniform selection mechanism unrelated to stability. Another dissimilarity between selection and folding temperature can be extracted from a paper by (Kumar, Tsai et al. 2003) in which experimental *folding temperatures* correlated negatively with the protein length. This correlation is in disagreement with the studies here in which positive (and linear) correlation with length was found for the *selection temperature*.

Acknowledgements

This work was supported by an NIH grant GM67823 to RE. We thank David Shalloway for useful discussions. Correspondence should be addressed to Ron Elber, ron@cs.cornell.edu

Table I

The table of the THOM2 energy as a function of the contact type and the amino acid type (i is the primary site, j the secondary site). Note that the number of neighbors to a site is "coarse-grained" and means the following actual number of neighbors

$1 \to 1,2 \quad 3 \to 3,4 \quad 5 \to 5,6 \quad 7 \to 7,8 \quad 9 \to \geq 9$

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
(1,1)	0.225	-0.029	-0.033	-0.082	-0.822	-0.259	0.091	0.286	0.072	-0.117
(1,5)	-0.207	-0.257	-0.103	0.196	-1.109	-0.005	-0.075	0.002	0.029	-0.306
(1,9)	-6.011	-4.086	-5.419	-6.137	-7.266	-5.878	-5.801	-5.808	-4.753	-5.455
(3,1)	-0.006	-0.096	-0.172	0.023	-0.496	-0.091	0.108	0.307	0.043	-0.104
(3,5)	-0.078	0.177	0.153	0.129	-0.693	0.115	0.236	0.037	-0.029	-0.287
(3,9)	-0.295	0.056	-0.327	0.082	-0.780	0.182	0.018	-0.128	-0.469	-0.597
(5,1)	0.134	-0.206	0.045	0.222	-0.147	-0.113	0.076	0.480	0.191	-0.148
(5,3)	0.064	0.165	0.202	0.169	-0.596	0.040	0.127	0.183	-0.038	-0.245
(5,5)	-0.654	0.681	-0.264	-0.195	-0.821	-0.092	0.427	-0.365	-0.194	-0.469
(7,1)	6.291	5.499	5.558	6.020	5.090	5.547	5.681	6.102	5.697	5.591
(7,5)	0.172	0.289	0.363	0.386	-0.276	0.285	0.450	0.327	0.277	-0.080
(7,9)	0.082	0.409	-0.003	-0.154	-0.297	0.038	-0.275	0.052	0.685	0.039
(9,1)	10.000	4.497	6.050	5.215	3.999	5.936	10.000	10.000	10.000	10.000
(9,5)	0.259	0.305	0.261	0.712	0.412	-0.017	0.323	0.828	-0.091	1.256
(9,9)	0.195	0.042	-0.367	-1.340	-1.186	0.469	1.374	-1.358	1.055	-1.991
(0,0)										

	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	GAP
(1,1)	-0.159	-0.016	0.213	-0.204	0.029	0.047	-0.065	-0.502	-0.637	-0.280	8.900
(1,5)	-0.230	-0.132	-0.147	-0.292	-0.231	0.067	-0.093	-0.605	-0.398	-0.358	5.700
(1,9)	-5.855	-4.905	-4.967	-5.826	-6.169	-5.887	-5.886	-5.254	-6.791	-6.989	10.000
(3,1)	-0.099	0.106	-0.196	-0.170	-0.015	0.405	0.061	-0.311	-0.295	-0.053	10.000
(3,5)	-0.213	0.141	0.080	-0.315	-0.054	0.058	0.079	-0.364	-0.278	-0.168	10.000
(3,9)	-0.487	0.086	-0.851	-0.065	0.195	0.234	0.150	-0.151	0.034	-0.272	10.000
(5,1)	-0.319	-0.056	-0.152	-0.271	0.169	0.190	0.342	-0.068	0.016	0.190	10.000
(5,3)	-0.187	0.258	-0.259	-0.283	0.089	0.114	0.017	-0.365	-0.297	-0.270	10.000
(5,5)	-0.423	0.336	0.319	0.074	0.549	0.218	0.005	0.038	-0.459	-0.584	10.000
(7,1)	5.262	6.082	5.642	5.797	5.819	5.226	5.477	6.419	5.170	5.530	10.000
(7,5)	-0.008	0.497	0.243	-0.158	0.421	0.126	0.337	0.042	-0.083	-0.029	10.000
(7,9)	-0.175	0.668	0.061	0.032	-0.706	0.825	0.242	-0.362	0.142	-0.246	10.000
(9,1)	6.222	5.593	4.915	6.021	9.614	10.000	10.000	5.885	10.000	10.000	10.000
(9,5)	-0.150	0.525	-0.194	0.431	3.066	0.426	0.524	-0.080	0.081	0.206	10.000
(9,9)	-0.248	-0.293	1.411	-1.330	6.939	3.223	-0.538	0.815	-0.533	-0.515	10.000
(0,0)											1.000

Figure Legends

1. **The energies of homologous proteins embedded in native shapes.** For the 3660 different proteins with known shapes (see text for more details) we compute the THOM2 energies for all the homologous sequences that we found in the NR (non redundant proteins) database. The energies per site are represented in figure 1.a We also computed the variance in energy per site and show the binned distribution in figure 1 (blue) for homologous sequences. For comparison we also show the distribution of variances of energies of random sequences in the same figure (1.b red).
2. **A: The number of sequences as a function of energy.** Five proteins, lengths of 150 amino acids, from the set of 3660 proteins that we analyze are shown in detail. The proteins are (from left to right): 1f3g, 1nul, 1ash, 1br1, 1bbr. We also show the structures of 1ash (right) and 1f3g (left) to demonstrate the significant variation in folds. The energy function was THOM2 and no constraints were used on sequence composition. **B: Cartoon plots of two of the proteins for which counting was performed:** The cartoon plots were copied from the protein databank site.
3. **The natural logarithm of the number of sequences below native energy, $N\left(E_{native}\right)$, as function protein length**. The results for 3660 proteins are shown. For every protein structure, this number is found by first computing the ratio $R = N\left(E_{native}\right) / N\left(\langle E \rangle\right)$ via the stochastic counting procedure outlined in the article, using the TE13 energy function with sequence mutations. The expected

energy $\langle E \rangle$ is of randomly-sampled sequences. The resulting graph demonstrates an exponential dependence of the number of sequences more stable than the native sequence and the protein length. Figure 3.a: Blue dots are the results with the THOM2 energy, the red dots are from TE13. Figure 3.b: Blue dots the counting with fixed composition and the red dots with reverse sequences (both with the THOM2 energy). Note that the least similar dots are with the fixed composition that has the largest effect on the counting.

4. **The native temperature, T_{native}, as a function of protein length**. The graph shows the temperature computed for each protein using (4.a) the THOM2 and (4.b) TE13 energy functions. The resulting graph shows that the overall behavior of the native temperature is similar for the two different energy functions. In figure 4.c we show a histogram plot for the temperatures computed with reverse energy THOM2. This time for proteins longer than 200 amino acids. Note the long tail for proteins with high selection temperature.

Figure 1.a

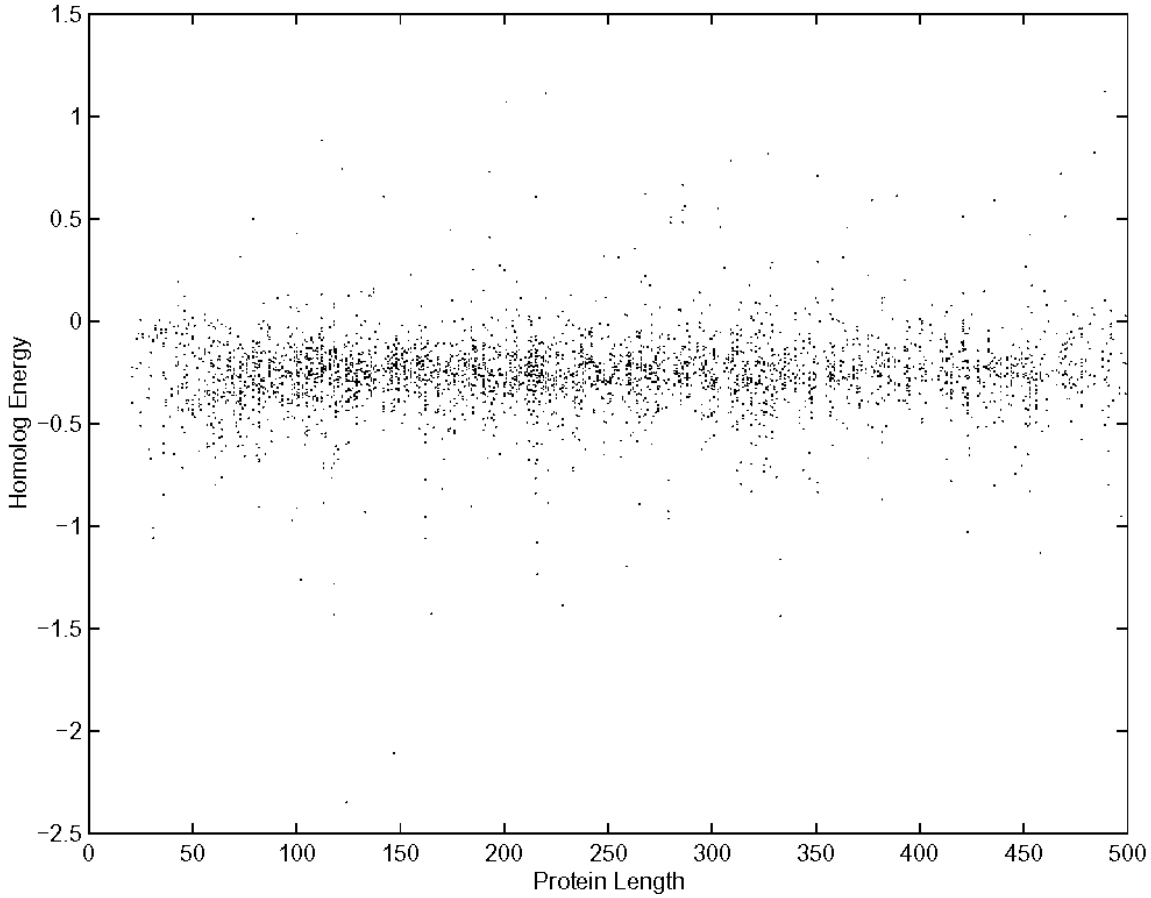

Homolog Energy
Protein Length
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
50
100
150
200
250
300
350
400
450
500

Figure 1.b

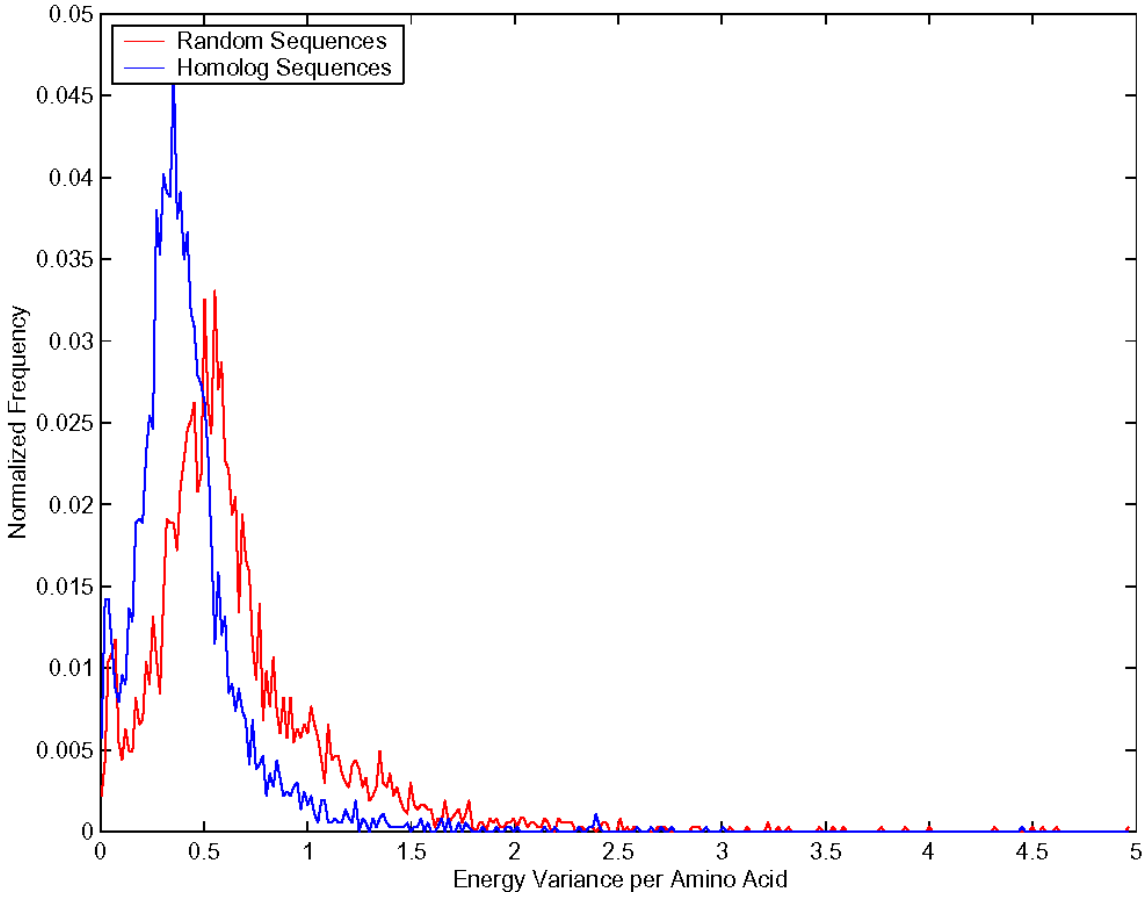
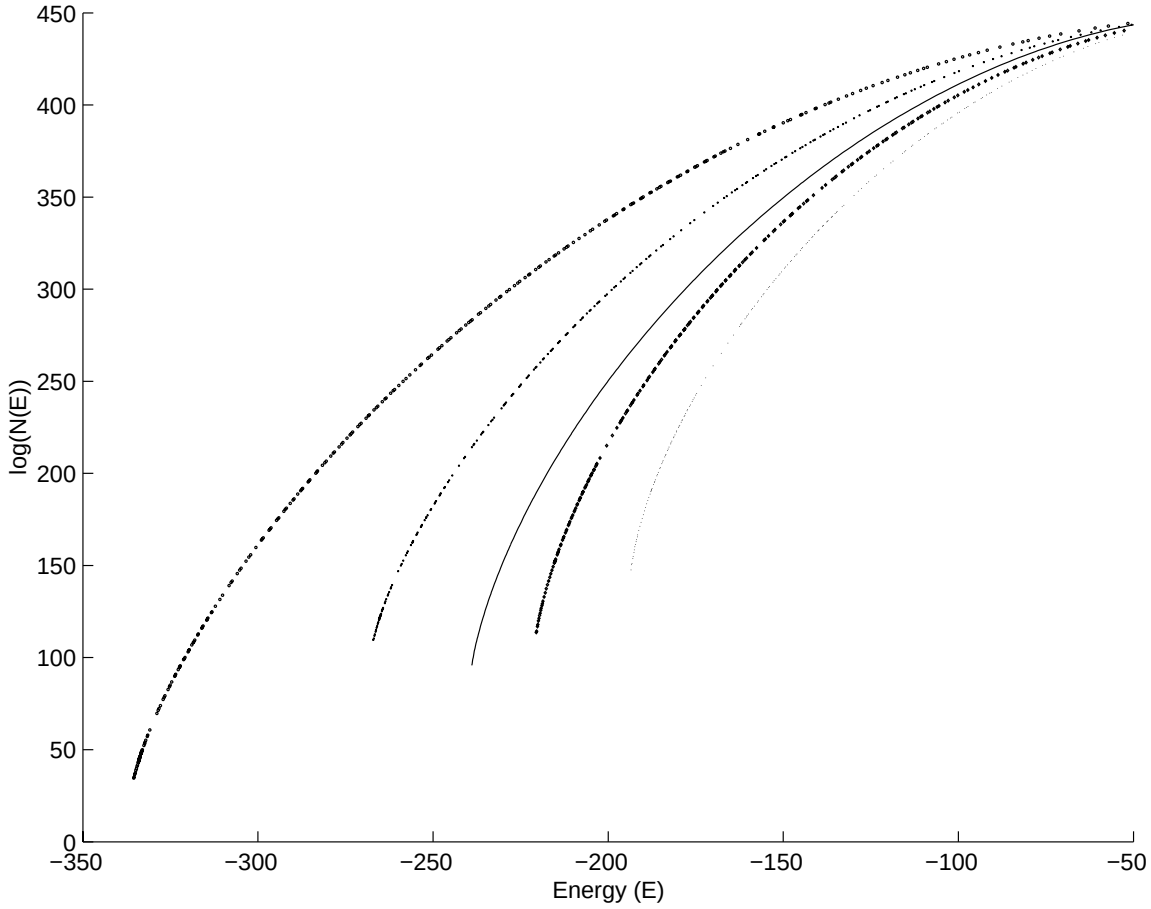

Random Sequences
Homolog Sequences
Normalized Frequency
Energy Variance per Amino Acid
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Figure 2


450
400
350
300
250
200
150
100
50
0
log(N(E))
−350
−300
−250
−200
−150
−100
−50
Energy (E)

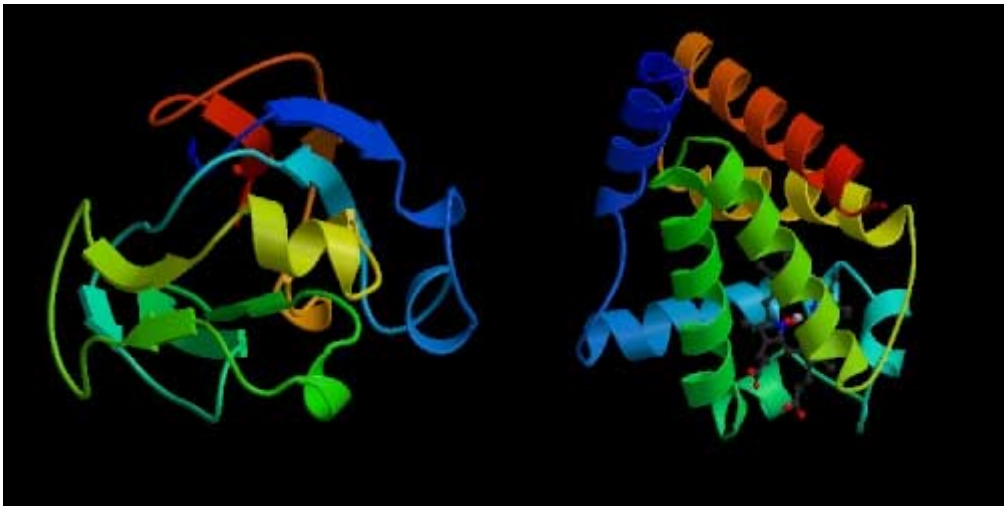


Figure 3.a

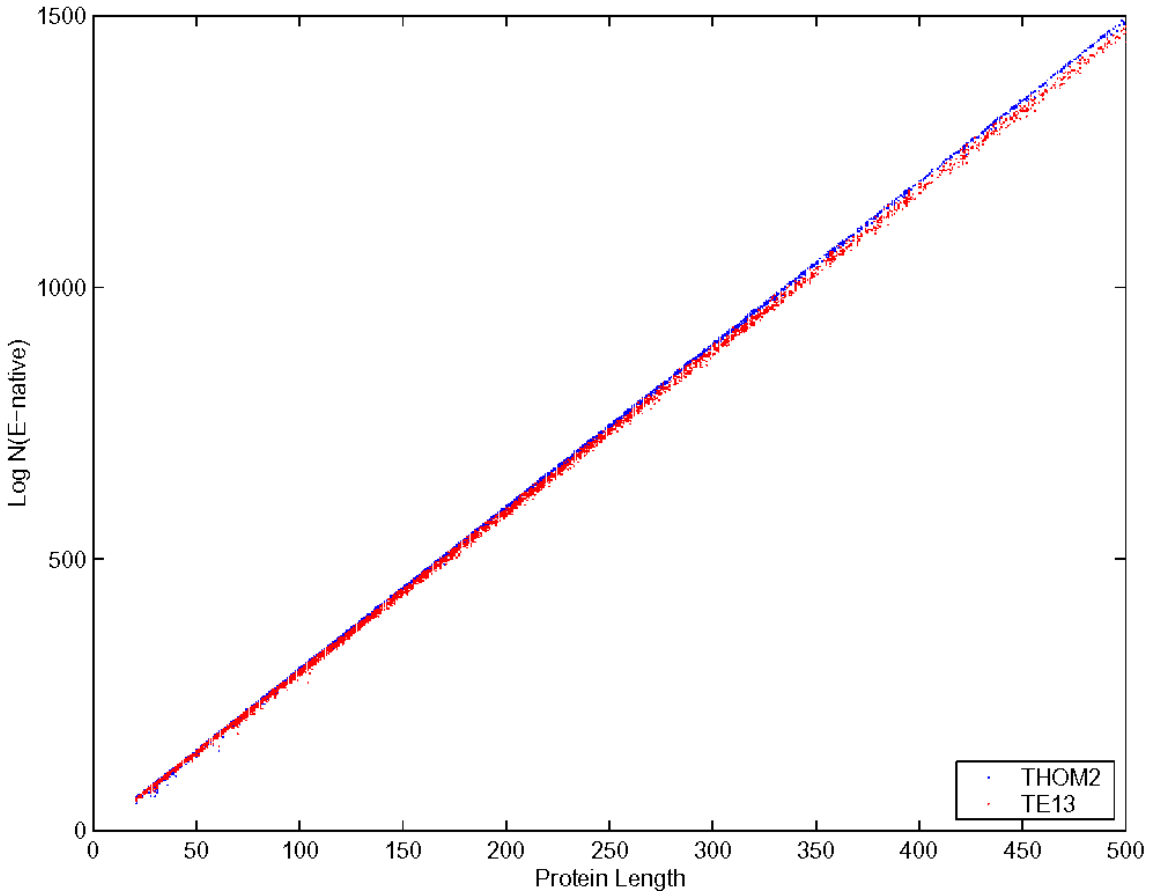

1500
1000
500
0
Log N(E−native)
0
50
100
150
200
250
300
350
400
450
500
Protein Length
THOM2
TE13

figure 3.b

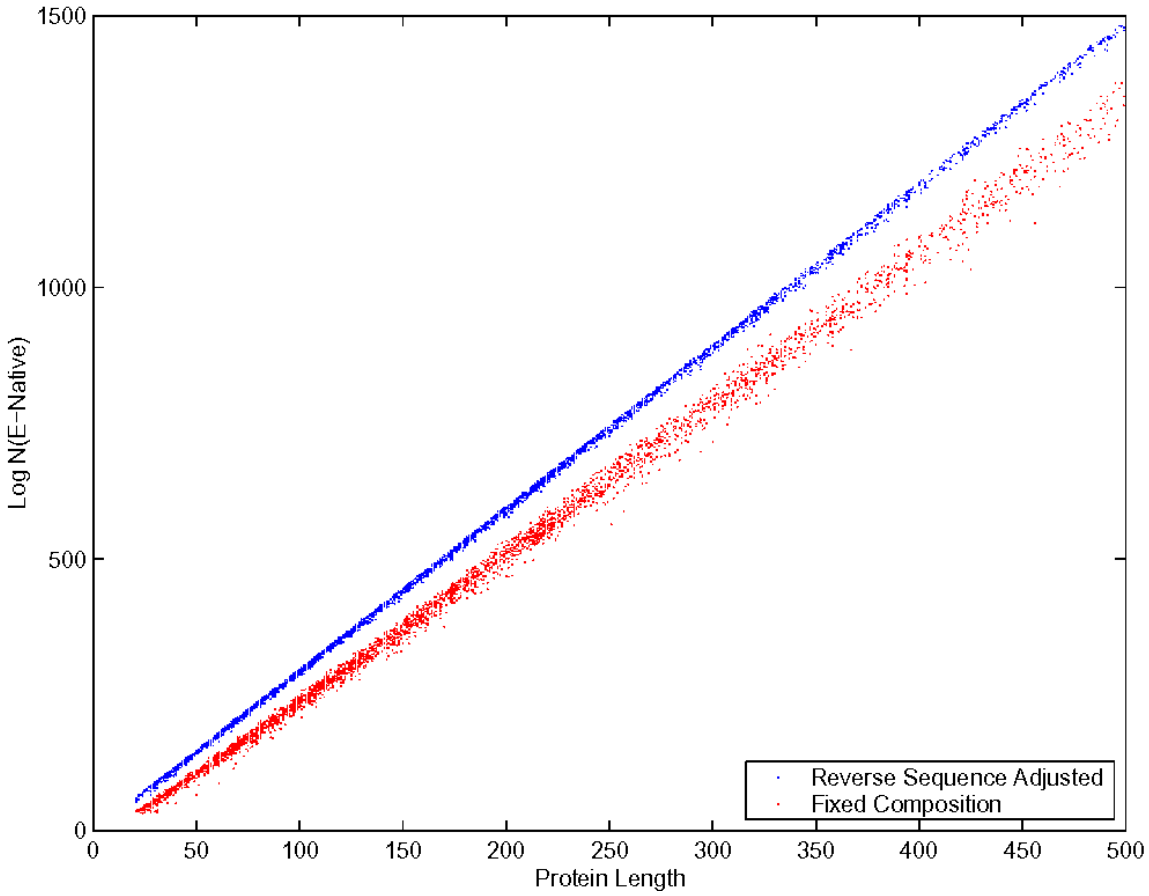

1500
1000
500
0
Log N(E−Native)
0
50
100
150
200
250
300
350
400
450
500
Protein Length
Reverse Sequence Adjusted
Fixed Composition

Figure 4.a

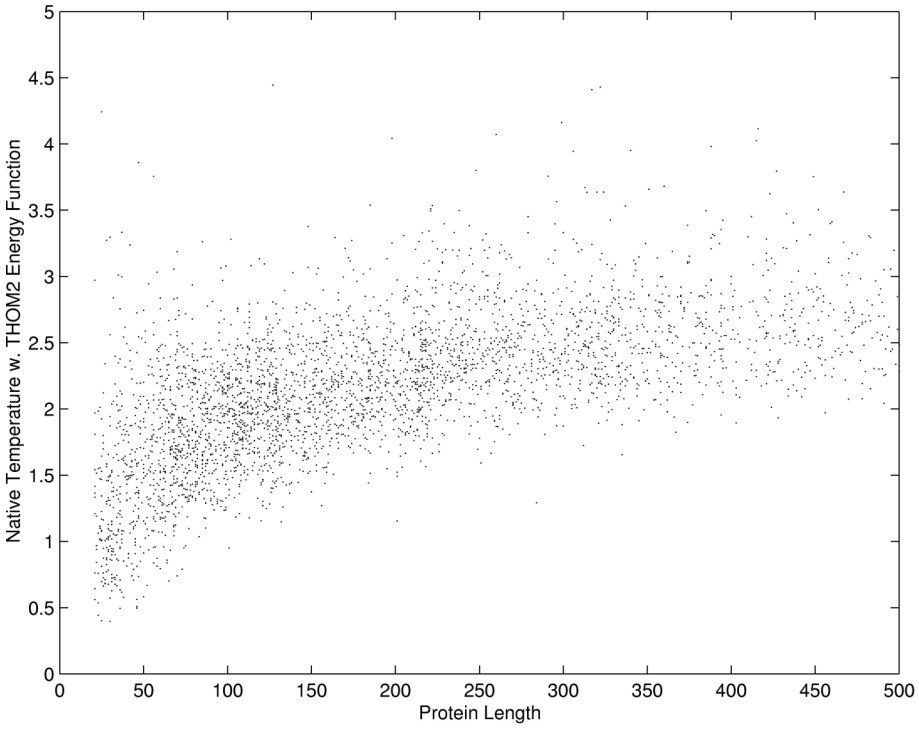

Native Temperature w. THOM2 Energy Function
Protein Length
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
50
100
150
200
250
300
350
400
450
500

Figure 4.b

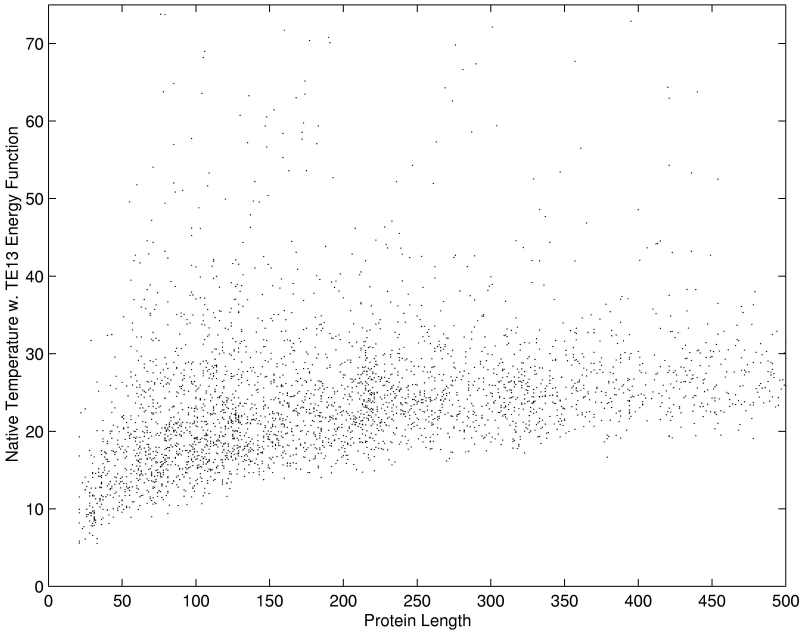
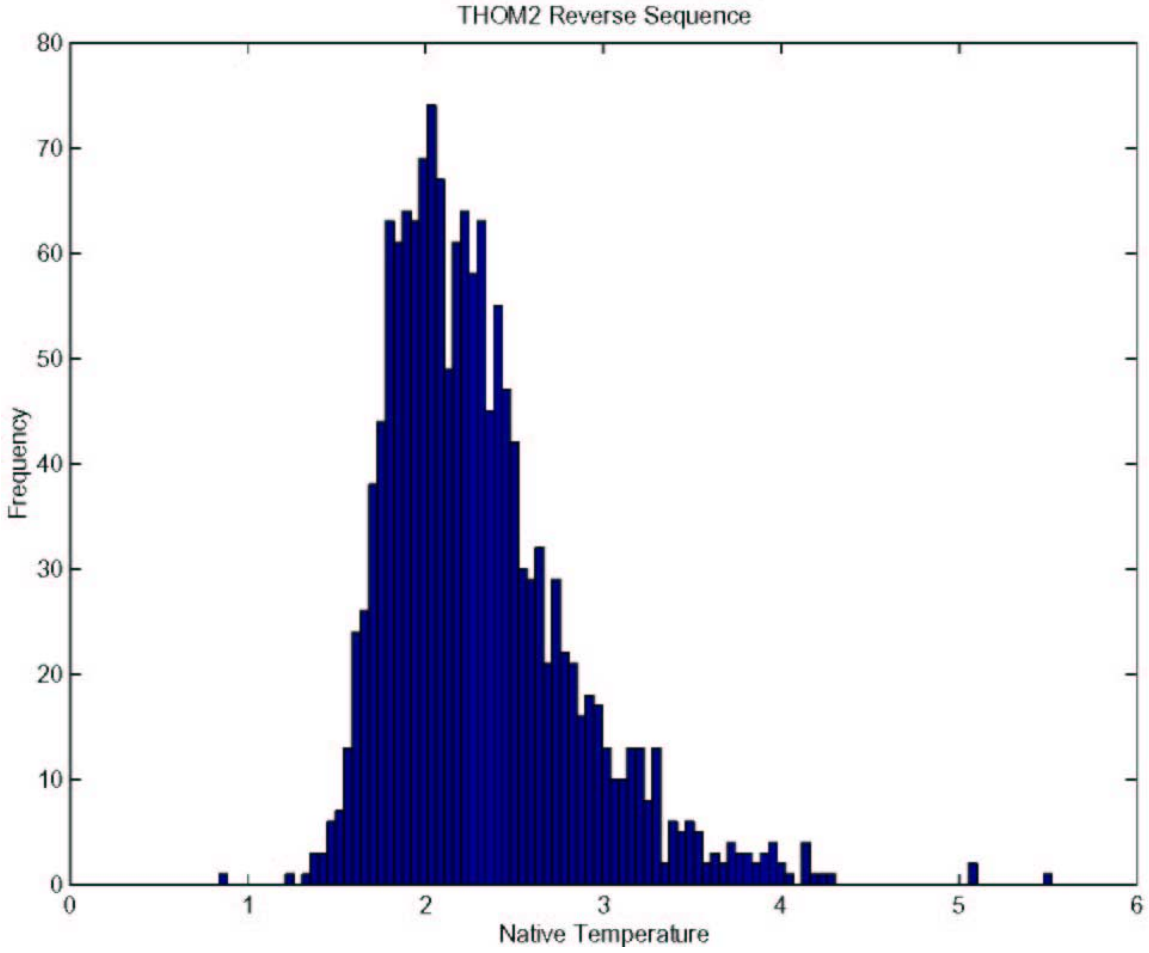

Native Temperature w. TE13 Energy Function
Protein Length
0
10
20
30
40
50
60
70
0
50
100
150
200
250
300
350
400
450
500

Figure 4.c


THOM2 Reverse Sequence
Frequency
Native Temperature
0
10
20
30
40
50
60
70
80
0
1
2
3
4
5
6

References

Benson, D. A., I. Karsch-Mizrachi, et al. (2000). "GeneBank." Nucleic Acid Research **28**(1): 15-18.

Berman, H. M., W. J., et al. (2000). "The protein data bank." Nucleic acids research **28**: 235.

Betancourt, M. R. and D. Thirumalai (1999). "Pair potentials for protein folding: choice of reference states and sensitivity of predicted states to variations in the interaction schemes." Protein science **2**: 361-369.

Betancourt, M. R. and D. Thirumalai (2002). "Protein sequence design by energy landscaping." Journal of Physical Chemistry **106**: 599-609.

Feynmann, R. P. (1982). Statistical Mechanics: a set of lectures. Reading, Massachusetts, Benjamin.

Huynen, M., P. Stadler, et al. (1996). "Smoothness within ruggedness: The role of neutrality in adaptation." Proceeding of the National Academy of Science USA. **93**: 397.

Jerrum, M. and A. Sinclair (1996). The Markov chain Monte Carlo method: an approach to approximate counting and integration. Approximation Algorithms for NP-hard Problems. D. S. Hochbaum. Boston, PWS.

Kabsch, W. (1978). "Discussion of solution for the best rotation to relate 2 sets of vectors." Acta Crytsllographica Section A **34**: 827-828.

Karchin, R., M. Cline, et al. (2003). "Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry." Proteins, Structure, Function and Genetics **51**(4): 504-514.

Kleinberg, J. (1999). Efficient algorithms for protein sequence design and the analysis of certain evolutionary fitness landscapes. ACM RECOMB.

Koehl, P. and M. Levitt (2002). "Protein topology and stability define the space of allowed sequences." Proceeding of the National Academy of Science USA. **99**: 1280.

Kumar, S., C. J. Tsai, et al. (2003). "Temperature range of thermodynamic stability for the native state of reversible two-state proteins." Biochemistry **42**: 4864-4873.

Larson, S. M., J. L. England, et al. (2002). "Thoroughly sampling sequence space: Large-scale protein design of structural ensembles." Protein science **11**(12): 2804-2813.

Lau, K. F. and K. Dill (1990). "Theory for protein mutability and biogenesis." Proceeding of the National Academy of Science USA. **87**: 638.

Lipman, D. and W. Wilbur (1991). "Modeling the neutral and selective evolution of protein folding." Proceeding of the Royal Society London B **245**: 7.

Maynard, S. J. (1970). "Natural Selection and the concept of protein space." Nature **225**: 563.

Meller, J. and R. Elber (2001). "Linear Optimization and a double statistical filter for protein threading protocols." Proteins, Structure, Function and Genetics **45**: 241.

Meller, J. and R. Elber (2002). Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models. Advances in chemical physics. F. Richard, John Wiley & Sons. **120:** 77-130.

Saven, J. G. (2002). "Combinatorial protein design." Current Opinion in Structural Biology **12**: 453.

Saven, J. G. and P. G. Wolynes (1997). "Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules." Journal of Physical Chemistry B **101**: 8375-8389.

Shakhnovich, E. I. (1994). "Proteins with selected sequences fold into a unique native conformations." Physical Review Letters **72**(24): 3907-3911.

Shakhnovich, E. I. and A. M. Gutin (1993). "A new approach to the design of stable proteins." Protein Engineering **6**(8): 793-800.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Engineering **11**: 739-747.

Tobi, D. and R. Elber (2000). "Distance dependent, pair potential for protein folding: Results from linear optimization." Proteins, Structure, Function and Genetics **41**: 40.

Torrie, G. M. and J. P. Valleau (1977). "Non-physical sampling distributions in Monte-Carlo free energy estimation - umbrella sampling"." Journal of Computational Physics **23**: 187.