Probing sequence-structure relationships in proteins: Application of simple energy functions to the inverse folding problem

NATO school in soft matter physics

Ron Elber, Jian Qiu, Leonid Meyerguz, and Jon Kleinberg
Department of Computer Science
Upson Hall 4130
Cornell University
Ithaca, NY 14853

Abstract
A brief description of the protein-folding and inverse-folding problems is provided. Design of energy functions for protein recognition based on machine learning approaches is discussed. The energy functions are applied to estimate the sequence capacity of all known protein folds, and to compute the evolutionary temperature(s) of embedding sequences in known protein structures.

This manuscript is divided into three sections. We start with a brief introduction to proteins, continue with the design of energies for recognition of protein folds, and conclude with an application to protein evolution, studying the sequence capacity of different structures.

I. Introduction

Proteins are linear polymers that are sometimes cross-linked (via sulfur bonds) but are never branched. They serve diverse and numerous functions in the cell as facilitators of many biochemical reactions, signaling processes, and providers of essential skeletal structures. These linear polymers consist of 20 different types of monomers that share the same backbone atoms (exceptions are proline and glycine) and have different (short) side chains. The chemical composition of a protein molecule is determined by the linear sequence of amino acids, called the primary structure. The sequence starts at the so-called N terminal and ends at the C terminal; the two ends are not equivalent, i.e. running the sequence backward does not produce the same protein. Typical lengths of protein sequences are a few hundred amino acids. The extremes are a few tens to a few thousands of amino acids.

One of the remarkable features of proteins is their ability to fold into a well-defined three-dimensional structure in aqueous solutions, the so-called protein-folding problem. This manuscript is concerned with a few indirect aspects of this problem. The working hypothesis is that the three-dimensional shape is determined uniquely by the sequence of the amino acids, and is the thermodynamically stable state. Anfinsen [1] put forward this extraordinary hypothesis that protein molecules are stable in isolation with no support from other components of the living system. From a computational and theoretical viewpoint, the Anfinsen hypothesis makes it possible to define and use in predictions an (free) energy function of an isolated protein molecule (in an aqueous solution). This function leads to significant simplification and saving of computational resources compared to studying a complete cellular environment. The free energy has a global minimum that coincides with the three-dimensional structure observed experimentally.

Structures of proteins are classified in terms of secondary structure elements, domains and individual chains (tertiary structure), and packing of tertiary structural elements (quaternary structure). Secondary structure is determined according to the hydrogen bond patterns of the backbone atoms that form small structural elements (ten to twenty amino acids). These elements are assembled to form the stable three-dimensional compact structure of the protein. Typical elements of secondary structure are the helices and sheets, where helices provide local chain structure and beta sheets connect pieces of the chain that can be far apart along the sequence (but close in space). The formation of local structure restricts the number of allowed conformations of the peptide chain and facilitates more accurate and rapid folding compared to a comprehensive search through all self-avoiding “walks” of the polymer chain.

Domains are fragments of a protein chain. Each domain includes several secondary structure elements and is “self-sustained.” It is expected that the average number of

contacts between amino acids that belong to the same domain (the total number of contacts in the domains divided by the number of amino acids) is much larger than the average number of contacts between amino acids situated at different domains. Domains have evolutionary implications. Empirically, domains were shown to swap between genes and proteins suggesting an evolutionary mechanism in which a significant segment of one protein (a domain) is inserted in, or exchanged with, another protein. This is in contrast to the alternative evolutionary mechanism of a point mutation, a process that modifies one amino acid at a time. Identifying relevant domains is likely to assist us in the characterization of basic building blocks of evolutionary processes and the mechanisms that guide them.

A complete (single) protein chain defines the tertiary structure. The quaternary structure is an aggregate of a few protein chains that work cooperatively on a biological task. The discussion in the present paper considers only isolated chains, and we therefore stop at the tertiary structure. This is clearly an approximation since some of the relevant interactions arise from nearby chains. Nevertheless, as in to the domain picture, we anticipate that at least some of the individual chains are stable and can be studied in isolation.

In statistical mechanics the folded conformation of a protein can be found by minimizing the potential of mean force that we loosely call the free energy, F . The free energy is defined by the following integration:

$$F(X) = -kT \log\left[\int \exp\left[-\frac{U(X,R)}{kT}\right] dR\right] \tag{1}$$

The probability of finding the system in equilibrium specified by a temperature T at X is proportional to $\exp\left[-F(X)/kT\right]$. The microscopic potential is $U(X,R)$, k is the Boltzmann constant, and T is the absolute temperature. In equation (1) the free energy is a function of a subset of the total number of coordinates, X , which includes (for example) bond rotations. The vector R includes the remaining coordinates that we eliminate by the integration on the right hand side equation. Examples of typical coordinates of the R vector are the positions of the solvent (water) molecules, and bond vibrations within a protein. The free energy is defined in terms of protein coordinates (e.g. torsions) that remain quite large in number. The number of torsions more than doubles the number of amino acids in the protein and is therefore between a few hundreds to a few thousands for a single protein chain. Since each of the torsions has about three rotamer states, a significant entropic contribution to the reduced free energy remains and a minimum alone cannot be the whole story. However, in the discussion below we do not consider the question of stability or chain entropy (i.e. if the minimum of the potential of mean force is sufficiently deep to overcome the entropy of the misfolded state). At present we are happy to identify the minimum with the correct structure, even if the stability energy is not available.

It is clear from equation (1) that for a reasonable microscopic potential $U(X,R)$ (so that the integral is well defined) the free energy $F(X)$ is computable. However, we cannot determine for the general case a simple and transferable functional form for $F(X)$, even if the microscopic potential is known (and this is not guaranteed, either). By transferable potentials we mean a single set of parameters for a given type of an amino acid regardless of the position of the amino acid along the sequence or the specific protein chain the amino acid is embedded in. The transferable formulation is similar in spirit to that of the microscopic potential and leads to more general, and simpler parameterization.

Therefore to ensure transferable potentials and ease of computations many applications to protein folding assume the functional form of $F(X)$, and do not compute it as outlined in equation (1). A set of potential parameters is optimized within the preset functional form. Assuming an empirical functional form for $F(X)$ is a natural extension of the approach used for the atomically detailed potential, $U(X,R)$. The last is also set empirically in most applications to proteins since the full exact calculations (including explicitly the electrons in the system) are just too expensive. Only a limited number of calculations (that are severely restricted in time) employ the full electronic structure model. For example, the free energy functional below is assumed to be a sum of pair interactions between all amino acids, a convenient but an ad-hoc proposition.

$$F = \sum_{i>j} F_{ij}\left(\alpha_i, \beta_j, r_{ij}\right) \qquad \forall i,j \ \ (i,j \text{ amino acids}) \tag{2}$$

The distance between the geometric centers of the amino acid side chains is r_{ij} [2]. The free energy of each interacting pair depends on the distance between the pair, r_{ij}, and their type, α_i, β_j (but not their position along the sequence).

The impact of the averaging formulated in equation (1) is subtle. For example, averaging of the solvent interactions yields repulsive potential of mean force between charged amino acids even if they have opposite electric charges that attract in vacuum. The preferred state of charged amino acids is to be surrounded by water molecules, far from the low dielectric medium typical of the interior of proteins. The tendency to be well solvated is observed only indirectly (since the solvent is not present explicitly in the model), and results in effective repulsion between well solvated (charged) amino acids. The hydrophobic (apolar) residues "attract" each other since they disrupt the hydrogen bond structure of the water molecules and their aggregation minimizes this effect. These interactions are weak, require the cancellation of many large terms, and are difficult to reproduce by direct averaging for proteins. The solvent-induced interactions are small in magnitude and the integrals in equation (1), which are performed stochastically, may not be accurate to the level required to fold proteins.

An alternative approach that avoids the integration in equation (1), and which we consider in the present manuscript (section II), is to "learn" the free energy surface from a set of experimentally determined protein structures.

We conclude the discussion on energy in the Introduction with another comment from the school of skeptics. The hypothesis that the native structures of proteins are global (free) energy minima is not always true. Some post-folding modifications (for example, cutting a leading peptide, the start of the protein chain) make the global (free) energy minimum of the original chain different from the native (modified) structure. A classic example is of the protein insulin [3]. Other examples are proteins that do not fold spontaneously and require external help of other macromolecules (chaperones) to adopt their correct three-dimensional shape. Nevertheless, despite the considerable complexity of the biological machinery that folds proteins (which suggests that some proteins cannot be studied in isolation), we do find numerous proteins that follow the Anfinsen hypothesis. Therefore the discussion below, seeking a functional form for the free energy and its global minimum, is a valid approach to determine structures of many proteins.

While the path from sequence to structure is considered to be the protein-folding problem, the present manuscript focuses on another intriguing question: the inverse folding problem. The Anfinsen's hypothesis argues that every sequence corresponds to one unique structure of the protein. Is the reverse true, i.e., can any structure of a protein be linked to a unique sequence of amino acids? This question, the reverse of the protein-folding problem (from structure to sequence), is answered by a definite "no." There are many sequences that are known (experimentally) to fold to the same or similar shapes. Consider the Protein Data Bank [4] (PDB http://www.rcsb.org) which is the digital repository of protein shapes. The PDB includes 25,960 protein structures as of June 15, 2004. These structures include many redundant shapes and can be reduced to a few hundred distinct protein families. The structural families are defined by shape similarities regardless of the amino acid sequences of the compared proteins. Hence, on the average, there are hundreds of sequences in the protein databank that fold into the same protein shape. The "seed" shape defines a fold family.

Consider another important database of proteins, nr (non-redundant) [x] that includes sequences only. A significant fraction of the millions of sequences in the nr database can be associated with a known fold family. The observed redundancy in mapping from sequences to structures in nr is even larger than the redundancy implied by the PDB. It is a mapping from the many (sequences) to the relatively few (structures). Evolutionary processes that modify and generate new protein sequences by changing one amino acid at a time are "stuck" in the neighborhood of individual structures (islands in sequence space) and produce new proteins that have essentially the same shape (note that the evolutionary process we consider here is not the domain-swap mentioned earlier). The seed shape is used over and over again for alternative sequences. The variations in sequences in the neighborhood of a given fold may adjust the function of the protein while maintaining the same overall structure. For example, a small change in activity would create a modified enzyme with enhanced (or reduced) affinity to the same ligand. A large change will use the same structural template for enzymes with different ligands, or chemical reactions. Since there are numerous examples of the second kind (large change in function) it is difficult to predict protein function based on structure similarity only.

An intriguing follow-up research direction is of sequence capacity of a structure. Given a shape of a protein X and energy E (which is a function of the sequence and the structure) what is the number of sequences $N(E,X)$ that fit this shape with energy lower than E? We will demonstrate that the number of sequences is so large that statistical mechanic analysis is suggestive. Following the usual notion of entropy in statistical thermodynamics we define a "selection temperature" for the ensemble of sequences that fit a particular structural family. We finally speculate on evolutionary implications of our work.

II. Energy functions for fold recognition

For meaningful calculations of macromolecular properties we must have a free energy function that weighs the importance of different structures. The lower the free energy, the more probable the structure. The design, choice of functional form, and optimization of parameters for the free energy function are the focus of the present section. Traditionally, energy functions for simulations of condensed phases and macromolecules (and proteins are macromolecules) were built according to chemical principles, starting with small molecular models, and interpolating to the large macromolecules, such as proteins. A typical atomically detailed energy function is of the following form

$$U(X,R) = U_c(X,R) + U_n(X,R) \tag{2}$$

The energy $U_c(X,R)$ includes the covalent terms: bonds, angles, and torsions. These are two, three and four body terms respectively.

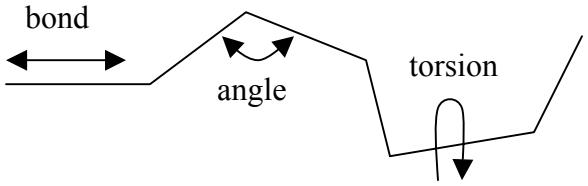

Figure 1. A schematic drawing of a polymer chain with covalent degrees of freedom denoted by arrows. A "stick" connects two atoms at the edges. A bond term describes the distance between the two atoms. An angle (term) is between two connected bonds, and a torsion measures the angle between the planes defined by three sequential bonds (the first and the second bond define the first plane, the second and the third bonds define the second plane).

Most of the time, the bonds and angles of the protein chain remain near their equilibrium values. It therefore makes sense to model the bond and the angles with (stiff) harmonic energy terms. An alternative is to use holonomic constraints, and to fix them at their ideal values.

In the present paper we do not directly consider the contribution of the covalent part U_c but focus instead on $U_n(X,R)$. The covalent term is considered (and enforced) indirectly by using a subset of structures that satisfy the holonomic constraints on bonds and angles.

The energy of the above covalent coordinates at their ideal value is zero, which makes it unnecessary to add the contributions of the bond and angle energies.

The case of torsions is different from the bonds and the angles. Torsions are allowed to change significantly, and are not constrained. However, the torsion energy contribution is small and is set to zero in some potential functions, which is the approach we take here. An exception to the "rule" of torsions with small energy contribution is rotations around double bonds (e.g. amide planes). The rotations around double bonds are fixed at their ideal values similarly to the bonds and the angles.

The considerations above leave us with only the non-bonded energy term $U_n(X,r)$. The lower case r denotes coordinates to be integrated out (e.g. water coordinates) to obtain the free energy. The set r no longer includes bonds and angles that were removed using holonomic constraints. The set r is therefore smaller than R.

As we argued in the introduction, the function $F(X)$ may be computed from the detailed potential $U(X,R)$ or $U_n(X,r)$ following the integration outlined in equation (1). However, this is computationally intractable, and so far no one has done it comprehensively and accurately for proteins (even if we are willing to forget about the transferability issue). A more pragmatic approach is to accept that a function $F(X)$ exists, and to search for a functional form and parameters that make general physical sense, and are successful in identifying the correct folds of proteins.

II.1 Statistical potentials

The idea of statistical potentials was pioneered by Scheraga [5] and popularized by Miyazawa and Jernigan [6]. It is probably the most widely functional form of an energy function in the protein-folding field. At this point it is useful to introduce some probabilistic arguments to motivate the computational approach below. As argued above the free energy is directly related to the probability of finding the system at a particular state. Consider the following question: what is the probability that in the set of known protein folds we will find an amino acid of type a_k in a (given) structural site with exposed surface area A_i and secondary structure s_j? Here is the formula

$$P(a_k \mid A_i, s_j) = \frac{P(A_i, s_j, a_k)}{P(A_i, s_j)} \tag{4}$$

The probability of the event z is $P(z)$. The amino acid type is a_k. We will define the secondary structure by a discrete variable s_j (0=α helix, 1=beta sheet structure, 2=3/10 helix, 3=bend and turn, 4= π helix, and the rest), and the exposed surface area A_i is binned into eight discrete states. Note that the variables A_i and s_j take the role of the X

coordinate vector that we discussed abstractly in the previous section. Identifying the relevant reduced variables is of crucial importance and here we are using our intuition on protein structure and energy. A surface term motivates polar residues to be on the surface of the protein, and such events should be observed with high frequency. Similarly, hydrophobic residues are buried in the protein matrix, a frequent observation.

The above conditional probability is related to the inverse protein-folding problem that was mentioned in the introduction. Alternatively, and more related to the protein folding problem, we may consider the probability that an amino acid a_k will be found in a structural site characterized by $\left(A_i, s_j\right)$. We will use the exposed surface area and secondary structure as non-bonded variables to describe the state of the protein.

$$P\left(A_i, s_j \mid a_k\right) = \frac{P\left(A_i, s_j, a_k\right)}{P\left(a_k\right)} \tag{5}$$

For a protein chain with a sequence $a_1 a_2 ... a_L$ we write the probability of having a sequence of structural sites characterized by $\left(A_1, s_1\right)\left(A_2, s_2\right)...\left(A_L, s_L\right)$ as a product. This is clearly an approximation in which we assume no correlation between the sites. Nevertheless, let us push a little further in that direction.

$$P\left(\left(A_1, s_1\right)...\left(A_L, s_L\right) \mid a_1 ... a_L\right) = \prod_{l=1} \frac{P\left(A_l, s_l, a_l\right)}{P\left(a_l\right)} \tag{6}$$

Since the free energy F, of a state X, is related to the probability of observing that state, $P\left(X\right) \propto \exp\left[-F\left(X\right)/kT\right]$, we can use reverse engineering and write the free energy of folding as

$$F\left(\left(A_1, s_1\right)...\left(A_L, s_L\right) \mid a_1 ... a_L\right) = -kT \sum_l \log\left[P\left(A_l, s_l, a_l\right)\right] + kT \sum_l \log\left[P\left(a_l\right)\right] \tag{7}$$

Note that approximating the probability as a product results in a free energy that is a sum. The free energy components depend on the properties of site i only. Of the two functions at the right hand side of equation (7), $P\left(a_i\right)$ is trivial to estimate (and probably irrelevant if our focus is on a fixed sequence with only the protein coordinates as variables). The more challenging function to estimate is $-kT \log\left[P\left(A_l, s_l, a_l\right)\right]$. It is based on averaging over all possible conformations of the protein chain and solvent coordinates that are consistent with the values of the predetermined secondary structure and surface area (equation (1), here we go again).

Besides the technical difficulties, it is important to note that the absolute free energy as given in equation (7) is not necessarily what we need. It is more useful to consider the free energy difference between the folded and unfolded states, since a protein is always in

one of these states and we are attempting to estimate which of the two states is more probable.

$$\Delta F_{FU} = F_F - F_U = -kT\sum_l \log\left[\frac{P\left(A_{lk}^F, s_{lk}^F, a_{lk}\right)}{\sum P\left(A_{lk}^U, s_{lk}^U, a_{lk}\right)}\right] \tag{8}$$

The index l runs over the sequence, and k is used to denote the type of the site characterized by surface area, secondary structure, and the amino acid embedded in it. The summation in the denominator includes all structures that we assigned to the unfolded state.

The expression in equation (8) is very general, so more details on computability are required. To make the formula meatier we need to come up with a feasible computational scheme of the free energy per structural site. The first step is to construct a model for the unfolded state, since direct summation over all possible unfolded coordinates is impossible in practice. In the unfolded state we expect the structural characteristics (surface area and secondary structure) to be weakly dependent on the amino acid types. We also expect it to be independent of the specific misfolded structure under consideration. Hence, rapidly exchanging misfolded structures are expected to be similar on the average. Note that we differentiate above between misfolded and unfolded structures. Unfolded structures make a larger set than misfolded structures. The last set includes non-compact shapes that do not resemble true protein conformations. Misfolded structures are protein-like shapes that (nevertheless) are incorrect. This assumption makes it possible to estimate the direct sum in the denominator (right hand side of equation (8)) using a statistical argument.

$$\Delta F_{FU} = F_F - F_U = -kT\sum_l \log\left[\frac{P\left(A_{lk}^F, s_{lk}^F, a_{lk}\right)}{(N-1)\cdot \overline{P}\left(A_{lk}^U, s_{lk}^U\right)\overline{P}\left(a_{lk}\right)}\right] \tag{9}$$

The total number of structures at hand is N. One of the structures is correct and the rest of the structures $(N-1)$ represent a misfolded state. The symbol $\overline{P}$ denotes probability of a structural site averaged over the set of misfolded structures. Since $(N-1)$ is fixed it adds a constant value to the free energy difference. This constant affects the absolute stability of the current model, but not the ranking of the structures according to their probability. Accurate estimation of the free energy of stability is important but hard to obtain computationally since it requires comprehensive summation of all possible (unfolded) structures. The good news is that absolute stability is not required to detect which of the candidate structures in our set is more likely to be the correct fold. We approach the more moderate goal by considering two fold candidates i and j, and compare the free energy differences ΔF_i and ΔF_j. This is a good point to define and use the statistical potential, $V\left(A_{lk}, s_{lk}, a_{lk}\right)$

$$V\left(A_{lk},s_{lk},a_{lk}\right)=-kT\log\left[\frac{P\left(A_{lk},s_{lk},a_{lk}\right)}{\overline{P}\left(A_{lk}^{U},s_{lk}^{U}\right)\overline{P}\left(a_{lk}\right)}\right] \tag{10}$$

The statistical potential can be used to estimate which fold is preferred. We have

$$\Delta F_i-\Delta F_j=\sum_l V\left(A_{lk}^{i},s_{lk}^{i},a_{lk}\right)-\sum_l V\left(A_{lk}^{j},s_{lk}^{j},a_{lk}\right) \tag{11}$$

All that remains is to estimate the numerical value of the entries to the table $V\left(A_p,s_q,a_r\right)$ (the triplet of indices $\left(p,q,r\right)$ identifies the type of the structural site and the amino acid, and replaces the single index k used in equation (11)). Perhaps the most remarkable feature of the statistical-potential approach to fold recognition (identifying the correct fold) is the way in which the table is generated. The probabilities in equation (10) are estimated directly from the protein databank. Having a set of non-redundant protein structures defines the N candidate structures that we are using to generate the tables $P\left(A_p,s_q,a_r\right)$, and $P\left(A_p^U,s_q^U\right)$ (computing $P\left(a_r\right)$ is trivial). We first consider all correctly folded proteins. For each protein we have binned the number of occurrences of the triplet A_p,s_q,a_r. We have a non-redundant sample of about 6000 proteins with lengths between a few tens to a thousand of amino acids. The number of bins is $20\times5\times8=800$ which is significantly smaller than (roughly) 1,000,000 data points, allowing for sufficient sampling. The next task of estimating the probability of misfolded sites, $P\left(A_p^U,s_q^U\right)$, is done in a similar way by collecting the same structural data in 40 bins. By ignoring the correlation between structural sites and sequences we assume that the distribution of the structural sites represents misfolded (but compact) structures. As argued earlier, our prime interest is in ranking, proposing plausible folds. We avoid the more difficult calculation of stability, which must take into account truly unfolded non-compact structures in order to estimate the free energy of stability.
The set representing the misfolded structures should include $\left(N-1\right)$ shapes that are incorrect and exclude the correct fold. However, removing the native shape from the set of 6000 structures will have a small effect on the statistics and will make it necessary to generate separate $P\left(A_p^U,s_q^U\right)$ for every fold. It is much simpler to generate this function only once including all the structures. The difference in the probabilities is expected to be negligible anyway.

Note also that the set of structures that we considered above has nothing to do with the normal thermal energy (after all, these are folded structures picked from the protein databank and not from thermal denaturation experiments). The multiplying factor kT in formula (10) determines the energy scale and not the relative ordering of different structures. It can be chosen arbitrarily and in the calculations that follow we set it to 1.

Below we show statistical potentials parameterized by exposed surface area, secondary structure, and type of amino acid. We show three cases (different amino acids) of two-dimensional cross-sections of the computed statistical potential.

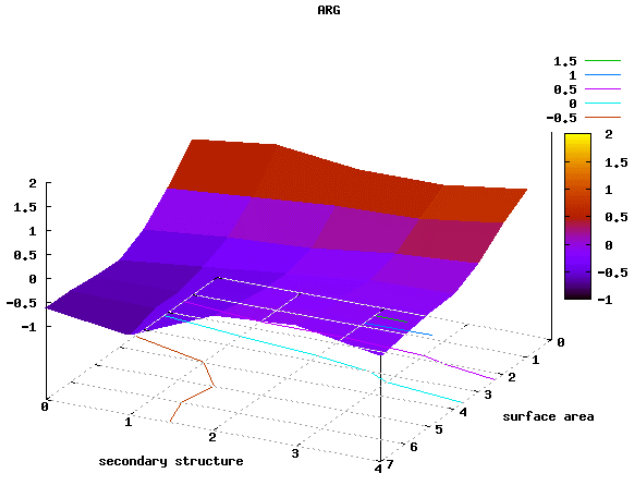

Figure 2a. The statistical potential of an arginine is plotted as a function of the secondary structure and of the fraction of solvent-exposed surface area. The secondary structure is parameterized as follows (0=α helix, 1=beta sheet structure, 2=3/10 helix, 3=bend and turn, 4=π helix, and the rest). The exposed surface area is normalized with respect to a maximum found in a tri-peptide Gly-X-Gly or in the proteins. Note that the two variables are sometimes correlated.

The first example is of arginine, a charged residue. It follows the usual expectation from polar residues. Like other charged residues it has a significant tendency to form a helix, though from the plot the weight of a beta sheet structure is similar.
The second example is of another charged residue (glutamic acid).

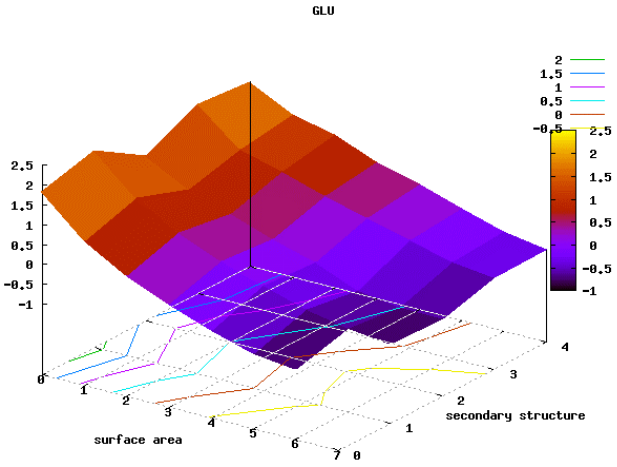

Figure 2.b Statistical potential for glutamic acid. See legend of figure 2.a for more details.

Glutamic acid also strongly prefers maximal exposure. It has a tendency to an alpha helical structure, with a 3/10 helix the second best.

Our third and last example of this kind is of hydrophobic residue (valine).

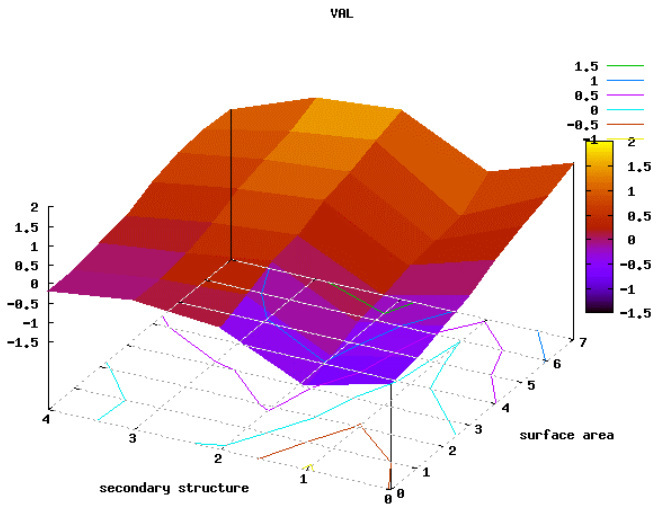

Figure 2.c Statistical potential for valine. Note the strong tendency of valine to be buried and to adopt a beta sheet conformation. See legend of figure 2.a for more details.

The potentials drawn above look reasonable and coincide with our physical and chemical intuition about proteins. However, they do not reflect the usual physical principles of free energies and their definition by statistical mechanics; the set of structures we consider is not thermal. These potentials were proven useful for "fishing" templates for structural modeling. However, they should not be used to estimate thermodynamic properties.

II.2 Potentials from mathematical programming

The attractive idea of statistical potentials is the use of experimentally determined protein structures to learn folding potentials, a radically different approach from the chemical physics bottom-up approach where parameters are derived from small molecules and the potential is scaled to large molecules (such as proteins). The difficulty in the chemical approach for proteins is that they are only marginally stable and slight accuracies in building up parameters for small molecules will be enhanced when applied to macromolecules (as proteins are).

The statistical potentials are easy to construct and to use, and were successful in identifying the correct folds in numerous cases. These advantages kept the statistical potentials in wide use. On the other hand, the derivation and design of the statistical potentials require numerous assumptions, putting into question our ability to use these entities in the calculation of physical and chemical properties (besides ranking the candidate structures for modeling). In this section we propose an alternative learning scheme that is considerably more flexible (in the choice of the functional form, and the parameter set) and makes it possible to pick a potential that is not inconsistent with known chemical and physical properties. The method we have in mind is that of mathematical programming. Just as in statistical potentials we learn the potential from protein structures, and not from data or calculations on small molecules. However, we learn it in a way that is consistent with the chemical physics principles of the system.

Consider the free energy, $F(X)$, which is a function of the reduced set of coordinates, X. A minimal requirement from this free energy, either from recognition or physical perspectives, is

$$F(X_i) - F(X_n) > 0 \qquad \forall i \tag{12}$$

Related inequalities were written and solved by Maiorov [7], and Vendruscolo [8]. We denote the coordinates of the correct (native) structure by X_n and the coordinates of a wrong (decoy) structure by X_i. The above condition, that the free energy of the correct structure is lower than the free energy of any alternative structure, is expected from the true energy function as well as from a successful measure of fold templates. How to use the flexible information in (12) to estimate functional form and parameters is a problem that can be addressed efficiently with mathematical programming tools. We first note that $F(X)$ (like any function) can be expanded by a (complete) basis set with linear coefficients. In the "learning formulation" below the decoy and the correct structures are known and the linear coefficients are the unknowns that we wish to determine.

$$F(X) = \sum_k a_k \phi_k(X) \tag{13}$$

Substituting the linear expansion in equation (12), we have

$$\sum_k a_k \left[\phi_k(X_i) - \phi_k(X_n) \right] = \sum_k a_k \cdot \Delta\phi_k(X_i, X_n) > 0 \qquad \forall i, n \tag{14}$$

Equation (14) defines a set of linear inequalities in a_k (for all decoy and native structures) that we wish to determine. We call $\Delta\phi$ the structural difference function. Linear inequalities can be solved efficiently using mathematical programming techniques. We may write equation (14) as a condition on a scalar product of two vectors $\hat{a}$ and $\Delta\hat{\phi}$. The two vectors must be parallel to satisfy the constraint (positive scalar product). Every inequality divides the space of parameters (linear coefficients) into two, a half that satisfies the inequality, and another half that does not. Gathering the constraints of all the inequalities can result in one of two outcomes: (a) there is a feasible volume of parameters, every point that belongs to that volume satisfy all the inequalities, or (b) there is no set of parameters for which all the inequalities are satisfied, i.e. the problem is infeasible.

A schematic drawing of the determination of two parameters with two inequalities is shown below (figure 3). Note that the actual number of inequalities that we solve in practice is much larger than the number of potential parameters that we wish to determine (the linear expansion coefficients). Typically, millions of constraints are solved with a few hundred parameters. In fact, we can use the number of constraints that have been solved as a test of the quality of the model. The more inequalities we are able to solve with the same number of parameters, the better the functional form is that we have

chosen for the energy function. Hence, in some cases increases in potential complexity (and number of parameters) are not justified since the number of inequalities that we solve after adding more parameters does not increase in a substantial way. In the field of "machine learning" in computer science, such an ineffective way of increasing model complexity, and over fitting parameters is major concern and called "over-learning."

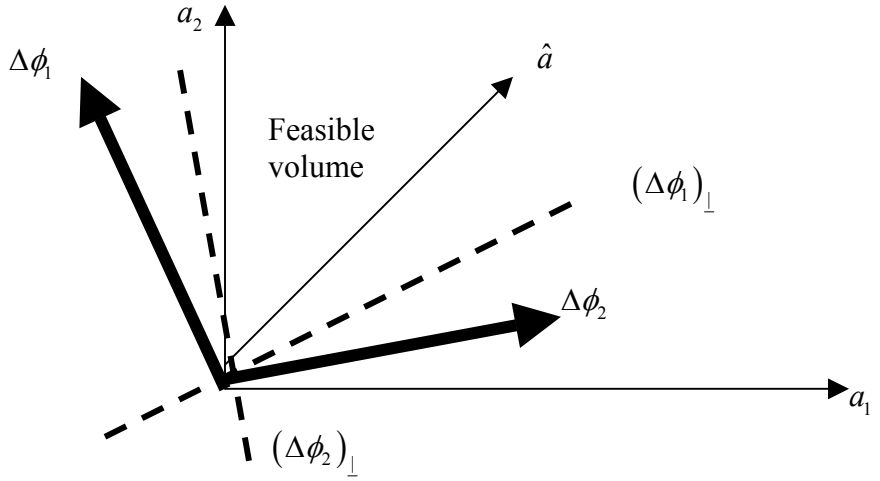

Figure 3. A schematic drawing of the parameter space and the inequalities we used to solve for the potential parameters. The drawing is for two parameters and two constraints. The dotted lines are hyperplanes (only lines in two dimensions) perpendicular to the corresponding $\Delta\phi_k$ - the structural difference vectors, and are called $(\Delta\phi_k)_\perp$. A solution, the vector $\hat{a}$, must be between the enclosing hyperplanes, and a thin arrow denotes a sample solution. Note that the norm of the solution, $|\hat{a}|$, is arbitrary.

For the set of proteins that follow the Anfinsen's hypothesis, we expect that the free energy function exists, and the set of inequalities is feasible (after all nature already solved that problem). However, since in practice our base functions are always incomplete, infeasibilities are at least as likely to indicate the failure of the current model as a failure of the Anfinsen's hypothesis. We therefore use the onset of infeasibility as a sign that the model is not good enough and seek a better basis set. This observation is in contrast to the statistical potential approach that does not provide such a self-test. A statistical potential that fails to recognize the correct fold of a protein does not offer an alternative path to further improve the potential.

A limitation of the mathematical programming approach, which is similar to the statistical potential calculations, is the energy scale. It is not possible using inequality (14) to determine the absolute value of the coefficients a_k. For any solution $\hat{a}$, the product $\lambda \cdot a_k$ where λ is a positive constant is also a solution. The norm of the parameter vector that solves the inequalities of (14) is unbounded. The scale can be determined using experimental information (when available). For example, measurements of the stability energy, or the free energy gap between the folded and unfolded states, can be used to determine the absolute scale. The connection is, however,

not trivial since the stability energy is a free energy difference that requires a model with all the uncertainties associated with it. This is the typical case, and most experimental observations that determine an energy scale require expensive computational averages to be accounted for by simulations.

An important advantage of the mathematical programming approach compared to the statistical potentials is the ability to learn from incorrect shapes. Statistical potentials "learn" only from correct structures. Misfolded structures are learned in an average way, and unfolded shapes are not considered at all. The inequalities make it possible to consider all alternatives shapes, misfolded and unfolded structures alike (provided that they are available). The limitation is technical and not conceptual (how many decoys we can effectively solve), and our collaborators are working to develop codes for parallel study of exceptionally large sets of inequalities [9].
Another limitation of the mathematical programming approach is the ability of this approach to optimize (efficiently) only convex quadratic functions. Clearly, general thermodynamic properties will have more complex dependence on the potential parameters, and this restriction affects our ability to make the best choice of a parameter set from a feasible volume. Nevertheless, there are a few guidelines that help us make an educated guess. These guidelines are statistical in nature and are based only on the information we have at hand (a limited set of protein structures). We therefore do not use thermodynamic information in the procedure described below.

Note also that the mathematical programming approach learns from the tail of the distribution of free energy difference (and not the average as is done in statistical potentials). The tail is of prime importance since we wish to put the native shape at the extreme left of the distribution. For the mathematical programming algorithm every new inequality can have a significant impact if it cuts through so-far feasible parameter space. The statistical potentials learn only average misfolded structures. Adding new information in the form of one or a few new conditions (after considerable data were already put in) is unlikely to change significantly statistical potentials. In contrast one or a few new data points can have a profound effect on potentials computed with mathematical programming. We emphasize that our data is without noise and we do expect to find a true potential that solves all the data exactly.

The training procedure that we described above will always be limited by the availability of data. The space of alternate protein conformations is tremendous in size and is unlikely to be explored in full for proteins of average size. For average number of conformations per amino acid, Z, and protein of length L, a rough estimate to the number of possible states of the chain is Z^L (for $Z=3$ and $L=100$ we have $5\cdot 10^{47}$). The largest set of constraints that we solved is of the order 10^7, much smaller than 10^{45}. Given the sparsity of the data the feasible volume of parameters will never be determined exactly, and significant deviations, especially near the boundaries of the feasible volume, are expected. We therefore do not wish to select a parameter set near the boundaries, since the uncertainties may result in false predictions. We anticipate however that deep in the feasible volume (assuming that some depth is available…) interpolations to new datasets are more likely to be accurate. In other words, we expect that the center of the feasible

volume will be sufficiently far from the boundaries, which are not well determined and are more prone to errors. If our learning is sound most new data points will fall in the neighborhood of the borders, the center of the feasible volume is expected to remain feasible.

How do we define and find the center of the feasible volume? We are working with two different approaches. In the first approach, we exploit the properties of the interior point algorithm [10], an optimization procedure to solve constrained problems of the type of equation (14).
In the interior point algorithm continuous logarithmic barriers replace the inequalities. A continuous minimization problem is solved that is guaranteed to converge in a polynomial number of steps. If the system is bound, the minimum of the target function will be the analytical center, a position in which all the forces from the logarithmic barriers balance each other. The analytical center is the sum of the forces of all inequalities that were used; some of the inequalities are redundant. An example for a trivial redundancy are the two constraints $\alpha_1 > 3$ and $\alpha_1 > 5$. Clearly the inequality $\alpha_1 > 5$ is sufficient. However the interior point algorithm uses both inequalities to generate forces towards the center. In that sense a direction with many redundant inequalities is more repulsive than a direction that was sampled sparsely. The result is therefore not the geometric center of the feasible volume that is defined by a minimal set of inequalities. Instead, the center of the interior point algorithm (with no function to minimize) is a weighted average of forces from all the constraints.
In practice the analytical centering procedure, which means using the interior point algorithm without a function to optimize, provided the best potentials measured by a maximal number of proteins recognized with a minimal number of parameters.

The second option is an intriguing subfield of machine learning in computer science, namely the SVM approach (Support Vector Machine [11]). In the language of the problem at hand, it is possible to use statistical learning theory to come up with a mathematical programming formulation similar to the inequalities (14) and to obtain meaningful results even if the set is not feasible. Here we consider only the feasible case (i.e., there are parameters such that all the inequalities are satisfied). The discussion about the infeasible set is beyond the scope of the present manuscript.
Returning to the task at hand, we cosmetically adjust the inequalities in (14) to read

$$\sum_k a_k \left[\phi_k \left(X_i \right) - \phi_k \left(X_n \right) \right] = \sum_k a_k \cdot \Delta \phi_k \left(X_i, X_n \right) > \delta \qquad \forall i, n \qquad (15)$$

The new variable δ defines an energy gap (the difference in energy between the folded and the misfolded/unfolded shapes). We wish to maximize this distance to increase the stability of predictions made by the energy function. This is only a cosmetic change since maximizing the gap directly is unproductive. The energy gap according to (15) and the norm of the vector of coefficients, $\hat{a}$, are not bound. The undetermined energy gap is a result of the missing energy scale that we mentioned earlier. To get around this problem we redefine the coefficient vector $\hat{a} \leftarrow \hat{a}/\delta$, which set the energy (and the energy gap) to

be dimensionless. Minimization of the newly defined vector of parameters will maximize the dimensionless energy gap. The problem we solve is

$$\sum_k a_k \cdot \Delta\phi_k\left(X_i, X_n\right) > 1 \qquad \min\left[\hat{a}^t \hat{a}\right] \qquad \forall i, n \tag{16}$$

If an energy scale is determined by other sources, we can always enforce the scale by replacing the "1" on the hand right side by the appropriate constant. The parameter vector so determined is maximizing the distance from the planes that are closing the feasible volume. This procedure is much closer in spirit to a geometric interpretation of the center of the feasible volume than the analytical center of the interior point algorithm mentioned above. Nevertheless, emphasizing the importance of constraints that are sampled very frequently even if they are redundant (as is done in the interior point algorithm) does have a merit. In practical applications, the potentials we derived from analytical centers tend to perform better than potentials derived from the SVM procedure. Below we describe a specific potential (THOM2 [12]) that was calculated with the centering of the interior point algorithm.

THOM2 is a specific realization of the structural function, $\phi_k\left(X_i\right)$, based on biochemical intuition, which was motivated by the lukewarm success of another potential -- THOM1 (see below). THOM1 and THOM2 are exploiting (in a different way) properties similar to the solvent-exposed surface area that we discussed earlier. Instead of surface area we count the number of contacts to a site as another measure of solvent accessibility. A contact is defined between the geometric centers of two side chains that are separated by no more than 6.5Å. A site with a large number of contacts (to other protein residues) is less likely to be exposed to the solvent. This type of site is likely to host apolar amino acids such as phenylalanine, or isoleucine. On the other hand, sites with a small number of contacts are appropriate for charged residues such as lysine that strongly prefer a water environment. THOM1 is an energy function that builds on the above intuition. We construct a table $T1(\alpha, n)$ that assigns an energy value to a site along the protein chain, according to the type of the amino acid (α) embedded in the site, and the number of contacts with the site (n). The total (free) energy of a protein is given by the sum of contributions from different sites.

$$F\left(X\right) = \sum_l T1\left(\alpha_l, n_l\right) \tag{17}$$

The summation index l is over the protein sequence (and structural sites). We have assumed separability of the free energy function to decoupled contributions from individual sites. This separation is similar to what we have done with the statistical potentials. It is convenient to write formula (17) with a sum over all the types of sites, K (K is the product of the number of amino acid types, times the number of neighbors a site may have).

$$F\left(X\right) = \sum_{k=1}^{K} m_k \cdot T1\left(\alpha_k, n_k\right) \tag{18}$$

The integer m_k is the number of times a site of a given type was sampled in a structure (for example, we may have in a specific protein five alanine residues embedded in sites with exactly four neighbors, in which case the corresponding *m* will be five). Using the last formulation, inequalities for THOM1 parameter training are written

$$\Delta F = \sum_{k=1}^{K} \left(m_k^i - m_k^{(n)} \right) T1\left(\alpha_k, n_k\right) > 0 \qquad \forall i, (n) \tag{19}$$

The table entries are the unknown coefficients to be determined (in this case with the interior point algorithm). The indices of the inequalities are for misfolded structures i, or native shape (n). The number of parameters for THOM1 is 200 (twenty amino acids and contact numbers vary from 0 to 9) which was determined using a few millions inequalities [2]. It turns out that THOM1 capacity to recognize native shapes is limited. We therefore were looking for a more elaborate model with a better recognition capacity, hence THOM2.

The THOM2 scoring scheme is also about contacts. In contrast to THOM1 which scores sites, THOM2 scores individual contacts. Different contacts score differently according to the number of contacts to that site and the amino acid embedded in the prime site. Consider a site with n_1 neighbors that we call the primary site. One of the contacts of the prime site is with a secondary site that has n_2 neighbors. THOM2 is an energy table that scores a contact between the two sites according to the type of amino acid in the primary site, and the number of contacts n_1 and n_2 - $T2\left(\alpha_1, n_1, n_2\right)$. The free energy of a protein in the THOM2 framework is therefore written as

$$\Delta F = \sum_{k=1}^{K} \left(m_k^i - m_k^{(n)} \right) \cdot T2\left(\alpha_k, n_{1k}, n_{2k}\right) > 0 \qquad \forall i, (n) \tag{20}$$

The sum in inequality (20) is over contact types (not sites). The counters for the unfolded structure m_k^i and the native shape $m_k^{(n)}$ are characteristics of the structure that are scored according to table $T2\left(\alpha, n_1, n_2\right)$ to be determined. The index k is equivalent to the triplet $\left(\alpha, n_1, n_2\right)$ and is used in formula (20) in addition to the triplet for clarity. The THOM2 energy was designed subject to about 30 million constraints [12]. The set that was found feasible with the 300 parameters, comprises the entries to the *T2* table of THOM2. It is remarkable that only 300 parameters capture the information contained in 30 million constraints, suggesting that this functional form is indeed useful.

It is also amusing that some of the entries to the table are undetermined (the entries with values of 10.00). Hence the number of parameters that we actually required to satisfy all the inequality constraints was even smaller than 300 (291 parameters). The combination of a site with the maximal number of neighbors, interacting with a site with the smallest

number of neighbors was exceptionally rare in our data and left many of these parameters (for different types of amino acids) undetermined.

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 of 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

The THOM2 potential derived as discussed above will be used in the study of evolutionary capacity of structures in the next section.

III. The evolutionary capacity of proteins [13]

One of the remarkable properties of proteins is the redundancy of sequence space with respect to structure space. There are numerous sequences that fold into the same shape. An obvious question is how large "numerous" is, and in this section we attempt to address this problem. More concretely, we compute the entity we name "structure capacity" – the number of sequences that a particular protein can accommodate up to an energy *E*. We consider protein sequences that improve on the stability of the native structure, i.e. sequences that are more stable than the native sequence of a particular (experimentally determined) protein structure. We find an exponentially large number of "better" (more stable) sequences.

The observation that one may improve stability (in a considerable way) compared to the natural sequence is perhaps not that surprising, since protein sequences are not optimized for structural stability only. True biological sequences are subject to constraints that are related to their function. Proteins need to be flexible, to have recognition sites, and other biological features that are at variance with the single criterion we use here, which is stability. Despite the limitations of studying stability only, there is still considerable interest in it, providing insight into the space in which further design and evolutionary refinement of sequences can be made. The stability constraint is an obvious one. It is always part of the equation and therefore studying it in isolation is likely to provide meaningful information, even if it is highly permissive (as we indeed find out).

So much for philosophy, to be concrete we compute the function $N(\Delta F)$. It is the number of sequences with free energy gaps larger than the free energy gap (ΔF). The problem now resembles the calculation of a microcanonical partition function, with a small (but important) difference. The microcanonical partition function is the number density - the number of sequences in the neighborhood of ΔF

$$\Omega(\Delta F) = \frac{dN(\Delta F)}{d\Delta F} \tag{21}$$

It is useful to reiterate the definition of the problem. The space in which we count events is of sequences and not of Cartesian coordinates. The sequence space is discrete and the maximum number of sequences that may fit to a protein of length L is 20^L (twenty types of amino acids). During the counting the structure is kept fixed while we generate sequences that may fit into this particular structure with a present stability criterion ΔF. The total number of sequences with free energy gap below ΔF is given by $N(\Delta F)$. The number density is a useful entity to build on a "thermodynamic" description of sequence space. The entropy, S, of sequence space (constrained to the neighborhood of one structure) is given by

$$S = \log\left[\Omega(\Delta F)\right] \tag{22}$$

To obtain a comprehensive view (as much as possible) on the structural templates of sequence evolution we repeat the calculations of sequence capacities for all distinct folds in the protein databank. To determine the distinct folds we employed a similarity measure of our design and compared all the structures in the protein databank against each other. Starting from a seed structure, new structures were added to the non-redundant set only if they were sufficiently different from all the structures already included in the set. This procedure gave 3660 non-redundant shapes [2]. Repeating the procedure with a different similarity measure (a measure produced by the CE structural alignment program [14]) provided comparable results.

The sequence space of each of these folds was counted separately. This counting is approximate since we ignore potential overlap of sequence space between different protein shapes. For example, the same sequence A may be found to have a low energy in two proteins P_1 and P_2. Obviously the sequence A can match with one structure only. Computing the sequence space for one structure at a time ignores this possibility and over-counting of sequences is a possibility. The extent of the over-counting is unclear and is a topic of future work.

A restricted counting is made in which no deletions or insertions of amino acids are allowed during our model of the evolutionary process. That is, the lengths of the template structure and the sequence are the same and are fixed. Related counting and evolutionary studies that did not probe the complete protein data bank were pursued by other groups [15-22].

III.1 The counting algorithm

We emphasize that the algorithm below is **not** Metropolis though it is still a randomized algorithm. The procedure below is based on the umbrella sampling of Torrie and Valleau [23] and of knapsack algorithm of Morris and Sinclair [24]. We consider a sequence A_0 embedded in a structure X with a free energy difference $\Delta F_0 \equiv \Delta F\left(A_0, X\right)$. We wish to determine the ratio $N\left(\Delta F^{(1)}\right) \Big/ N\left(\Delta F^{(2)}\right)$ where $N\left(\Delta F^{(i)}\right)$ is the number of sequences with energies up to $\Delta F^{(i)}$. The energy of the starting sequence ΔF_0 is set below $\Delta F^{(2)}$. The algorithm goes as follows:

1. Pick at random one of the amino acids, a_{ij}, in the current sequence A_i and change it at random to one of the twenty amino acids to generate a new intermediate sequence $\overline{A}_i$.
2. Check the energy of the intermediate sequence $\Delta F\left(\overline{A}_i, X\right)$. If it is larger than $\Delta F^{(2)}$ reject the step, change the sequence back to the original sequence A_i, and return to 1. Otherwise accept the step (set A_{i+1} to be equal to $\overline{A}_i$), and continue to 3.

3. Compare $\Delta F\left(\overline{A}_i, X\right)$ to $\Delta F^{(1)}$ and $\Delta F^{(2)}$. Updates the counters l_1 and l_2 (l_i is the number of sampled sequences with energy smaller than $\Delta F^{(i)}$).
4. Check stopping criteria (number of steps, convergence of the ratio $l_1/l_2 \cong N\left(\Delta F^{(1)}\right)\Big/ N\left(\Delta F^{(2)}\right)$ that approximate the function we are after). Go to 1 if criteria were not satisfied.

It is necessary for the energies $\Delta F^{(1)}$ and $\Delta F^{(2)}$ to be sufficiently close to each other, so the ratio will be close to one and converging rapidly. Calculation of ratios could be aggregated together (a collection of rapidly converging randomized counting). From the above equation it is clear that we can get a sequence of ratios similar to

$$\prod_{i=1,..,n} \frac{l_i}{l_{i+1}} = \frac{l_1}{l_N} \cong \frac{N\left(\Delta F_1\right)}{N\left(\Delta F_N\right)} \tag{23}$$

From equation (23) we can estimate the number of sequences for any energy ΔF_N provided that we know the density at anchor energy ΔF_1. Anchors are not hard to get if we know the sequence with the lowest possible energy, since the number of alternative sequences in the neighborhood of that sequence is small and countable directly. If the minimum energy is not known one could use energy that can be sampled easily. For example, it is not difficult to estimate the median energy and the number of sequences below the median (exactly half of the total number of sequences, 20^L).

III.2 computing temperatures for all protein folds

We have computed the number of sequences for all relevant free energy differences $N\left(\Delta F\right)$. This function has a strong (exponential) dependence on the protein length, which is easy to rationalize. The total number of possible sequences is exponential in length (20^L). The actual number of accepted sequences is expected to grow like $M^L \quad \left(M < 20\right)$ (still grows exponentially with the length). Every length extension of the protein molecule, and the addition of a new structural site will allow a few more amino acids (per site) to be accommodated increasing exponentially the number of accessible sequences. Counting for the complete set of 3660 proteins that differ significantly in length was performed. In figure 4 below we show $\log\left(N\left(\Delta F\right)\right)$ as a function of ΔF. The obvious linearity of the plot strongly supports the above assertion of exponential growth in $N\left(\Delta F\right)$ as a function of the protein length L.

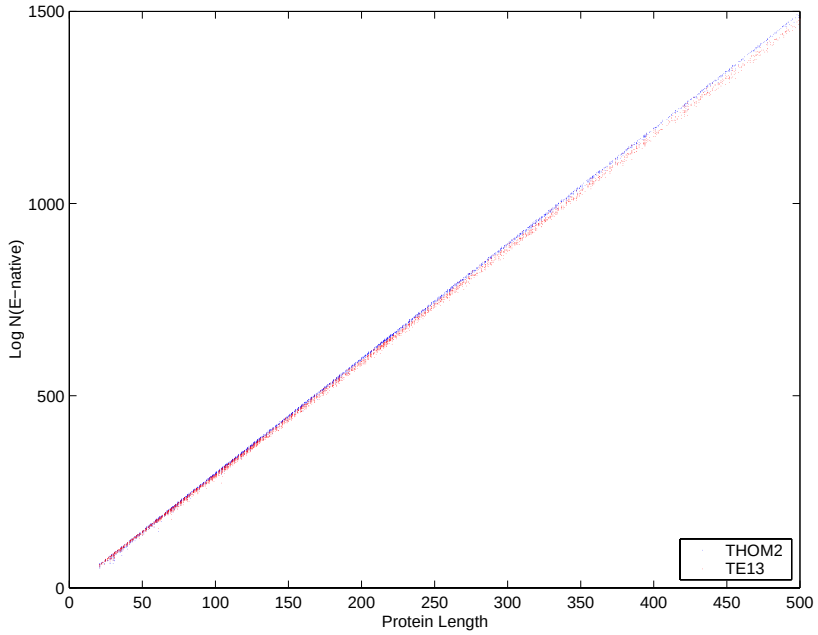

Figure 4. Counting the number of sequences that can be embedded in different folds (3660 total) with energies better than the energy of the native sequence. The log of the number of sequences is plotted as a function of protein length to emphasize the exponential dependence. While the most obvious feature is the linear dependence, we should note that the line has significant thickness which is significant since a log function was used. The plots include counting results from two potentials. One set is from THOM2, the potential that was discussed in this paper. The second potential (TE13 [25]) was discussed elsewhere.

In figure 5 we show a sample of a few complete curves of $\log\left[N\left(\Delta F\right)\right]$ versus the free energy difference ΔF .

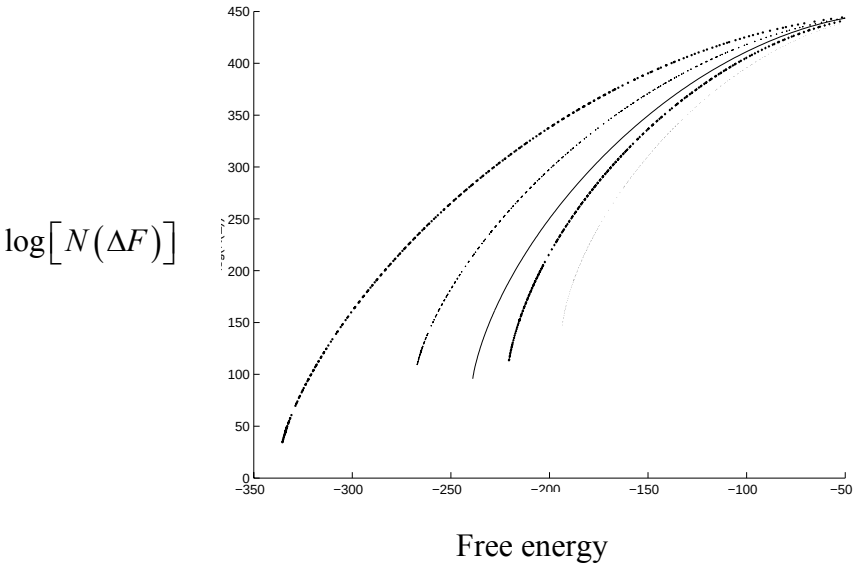

Figure 5. Computing sequence capacity for five proteins of the same length (150 amino acids), from the set of 3660 proteins that we analyze are shown in detail. The proteins are (from left to right): 1f3g, 1nul, 1ash, 1br1, 1bbr.

The energy that we used for the counting was THOM2, for which the determination of the lowest energy sequence is trivial, making it possible to identify the lowest energy sequence and its corresponding degeneracy. We finally compute the temperature associated with the energy of the native sequence using

$$\frac{1}{T} = \frac{d}{dE_n} \log\left(\Omega\left(E\right)\right) \tag{24}$$

In figure 6 we showed the distribution of temperatures computed for THOM2 energy and for the set representing the protein databank. The distribution of temperatures is highly peaked but still quite broad.

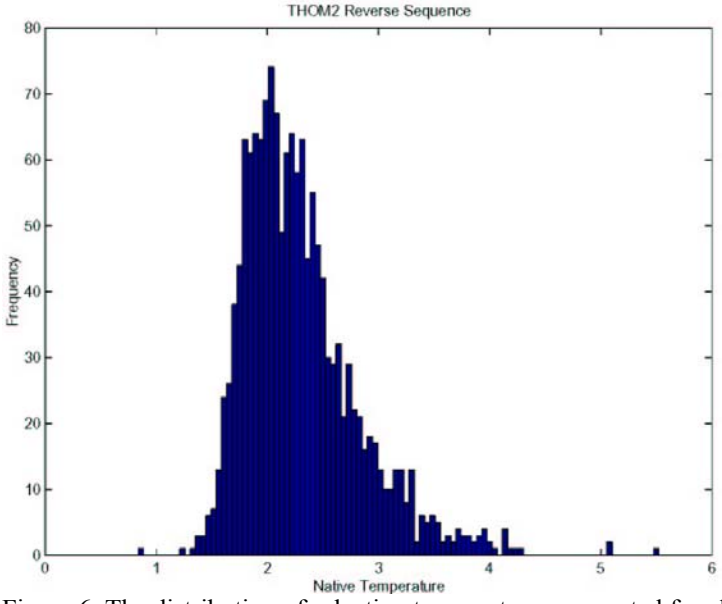

Figure 6. The distribution of selection temperatures computed for all 3660 folds at their native energies.

The calculations of the temperature employ a standard thermodynamic formula [26]. However, the meaning of this temperature is not obvious. What are the implications of the temperatures? Can we propose a mechanism that might lead to this set? Here is a possible model that can help us think about the data. We consider the selection of a particular native sequence with energy E_n and write the probability that it will be observed biologically, $P\left(S_n\right)$, as a product of two terms: The number of sequences at E_n -- $\Omega\left(E_n\right)$ and a selection function $G\left(E_n\right)$ ($P\left(S_n\left(E_n\right)\right) \equiv P\left(E_n\right) = \Omega\left(E_n\right) G\left(E_n\right)$). The selection function of nature depends on more than the energy. For example: flexibility, binding site, and electric field, are important to protein function and exert evolutionary pressures. The number of sequences at a particular energy E_n is a rapidly increasing function of the energy. To find an optimal (probable) sequence with a low energy, it is necessary that the selection function will be rapidly decreasing, leading to a maximum in $P\left(E_n\right)$. More conveniently we seek a (equivalent) maximum in $\log\left[P\left(E_n\right)\right]$. We have

$$\left.\frac{d\log[P]}{dE}\right|_{E=E_n} = \left.\frac{d\log(\Omega(E))}{dE}\right|_{E=E_n} + \left.\frac{d\log(G(E))}{dE}\right|_{E=E_n} = 0$$
$$\frac{1}{T_n} = -\left.\frac{d\log(G(E))}{dE}\right|_{E=E_n} \tag{25}$$

Hence, the selection temperature is telling us something about the selection function. Equation (25) makes it possible to compute a relationship between the number of sequences (that we can compute) and sequence selection. The selection functions computed for different protein shapes at their native energies are therefore quite similar (as the temperatures are).

How can a universal selection mechanism be realized? The simplest answer is the universality of the genetic code and mutation mechanisms (e.g. UV radiation on DNA basepairs). All genes coded on the DNA are likely to be mutated in roughly the same way, providing the same level of "sequence-thermal-excitation" (temperature) to all genes (proteins). The other option to explain the data, which is more intriguing (but not necessarily more correct), is to have all the folds connected via paths in sequence space, i.e., a sequence that belongs to one structural family can be mutated to a different structural family. In this case (regardless of the underlying mutation mechanism) the temperature should be the same. A way to prove or to disprove the above proposition is to try to identify plausible paths connecting sequence islands that are associated with a given structure. Simple models have been studied [27]. However studies of known protein folds and the interactions of their corresponding sequence spaces are still to be desired.

References

1. Anfinsen, C.B., *Principles that govern folding of protein chains.* Science, 1973. **181**(4096): p. 223-230.
2. Meller, J. and R. Elber, *Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models*, in *Advances in chemical physics*, F. Richard, Editor. 2002, John Wiley & Sons. p. 77-130.
3. Lehninger, A.L., *Principles of Biochemistry*. 1982, New York: Worth Publishers Inc. pp 723.
4. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**: p. 235.
5. Tanaka, S. and H.A. Scheraga, *Statistical Mechanical Treatment of Protein Conformation. 1. Conformational properties of amino acids in proteins.* Macromolecules, 1976. **9**(1): p. 142-159.
6. Miyazawa, S. and R.L. Jernigan, *Estimation of effective interresidue contact energies from protein crystal structures: Quasi chemical approximation.* Macromolecules, 1984. **18**(3): p. 534-552.
7. Maiorov, V.N. and G.M. Crippen, *Contac potential that recognizes the correct folding of globular proteins.* Journal of Molecular Biology, 1992. **227**: p. 876-888.
8. Vendruscolo, M., et al., *Comparison of two optimization methods to derive energy parameters for protein folding: Percepton and Z score.* Proteins, Structure, Function and Genetics, 2000. **41**(2): p. 192-201.
9. Wagner, M., J. Meller, and R. Elber, *Large-scale linear programming techniques for the design of protein folding potentials.* Mathematical Programming Ser. B, 2004.
10. Meller, J., M. Wagner, and E. R., *Maximum Feasibility Guidline in the Design and Analysis of Protein Folding Potentials.* J Comput Chem, 2002. **23**: p. 111-118.
11. Cristianini, N. and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. first ed. 2001, Cambridge: Cambridge University Press. 189.
12. Meller, J. and R. Elber, *Linear Optimization and a double statistical filter for protein threading protocols.* Proteins, Structure, Function and Genetics, 2001. **45**: p. 241.
13. Meyerguz, L., et al., *Computational analysis of sequence selection mechanisms.* Structure, 2004. **12**: p. 547-557.
14. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**: p. 739-747.
15. Shakhnovich, E.I. and A.M. Gutin, *A new approach to the design of stable proteins.* Protein Engineering, 1993. **6**(8): p. 793-800.
16. Saven, J.G. and P.G. Wolynes, *Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules.* Journal of Physical Chemistry B, 1997. **101**: p. 8375-8389.

17. Betancourt, M.R. and D. Thirumalai, *Protein sequence design by energy landscaping.* Journal of Physical Chemistry, 2002. **106**: p. 599-609.
18. Lau, K.F. and K. Dill, *Theory for protein mutability and biogenesis.* Proceeding of the National Academy of Science USA., 1990. **87**: p. 638.
19. Koehl P. and Levitt M., *Protein topology and stability define the space of allowed sequences.* Proceeding of the Natural Academy of Sciences USA, 2002. **99**: p. 1280.
20. Larson, S.M., et al., *Thoroughly sampling sequence space: Large-scale protein design of structural ensembles.* Protein science, 2002. **11**(12): p. 2804-2813.
21. Huynen, M., P. Stadler, and F. W., *Smoothness within ruggedness: The role of neutrality in adaptation.* Proceeding of the National Academy of Science USA., 1996. **93**: p. 397.
22. Lipman, D. and W. Wilbur, *Modeling the neutral and selective evolution of protein folding.* Proceeding of the Royal Society London B, 1991. **245**: p. 7.
23. Torrie, G.M. and J.P. Valleau, *Non-physical sampling distributions in Monte-Carlo free energy estimation - umbrella sampling".* Journal of Computational Physics, 1977. **23**: p. 187.
24. Morris, B. and A. Sinclair, *Random walks on truncated cubes and sampling 0-1 knapsack solutions*. Proc. IEEE Foundations of Computer Science. 1999. 130-240.
25. Tobi, D. and R. Elber, *Distance dependent, pair potential for protein folding: Results from linear optimization.* Proteins, Structure, Function and Genetics, 2000. **41**: p. 40.
26. Landau, L. and Lifshitz, *Statistical Physics I*. 1986, Oxford: Pergamon Press. 34-36.
27. Kleinberg, J. *Efficient algorithms for protein sequence design and the analysis of certain evolutionary fitness landscapes.* in *ACM RECOMB*. 1999.