Entropic Landscape: the methods to predict folding patterns and regional stability of proteins

Kentaro ONIZUKA
PsiPhiFoldings Co. Ltd
onizuka.kentaro@jp.panasonic.com

1 Introduction

The entropy of a system indicates the diversity of the system. The more micro-states the system has, the higher the entropy of the system is. Since in general complex systems have a large number of micro-states, their entropy should be high. However, some system with a large number of *possible* states may have low entropy when the number of *highly probable* state is limited, because the diversity is statistically small, no matter how many *possible* states they have.

This very popular principle of statistical mechanics seems to explain why most proteins quickly fold into a particular conformation. Their *highly probable* conformations to form must be just a few, no matter how many *possible* conformations they could form. If we consider a protein as a system in canonical ensemble, each possible conformation of the protein is considered a micro-state of the system. Then, such a protein that quickly folds into a particular fold should have very low entropy because the polypeptide stay in the state of the optimally stable conformation for a long time, once they complete folding. The speed of folding should also be correlated to the entropy. Relatively high entropy systems have many suboptimal states, in each of which they could stay for a bit of time before they reach the optimal state while low entropy systems have just a few suboptimal states to stay in. They could reach the optimal state before long. Thus, they fold quickly.

This analogy could be applicable to the folding speed of each sequential region of a protein. Considering the sequential region of the fixed length, say five-residue long, some regions have low entropy, and some high. Very low entropy regions should have just a few favourable conformations to form, while high entropy ones have a large number of even probable conformations. We can, therefore, assume that low entropy regions set out to fold early on, and high entropy regions linger in folding until the neighbouring lower entropy regions urge or induce them to fold. Inevitably, the conformations into which high entropy regions fold are moulded by the neighbouring lower entropy regions already having formed their own favoured conformations that affect the folding of high entropy region sequentially neighbouring or sequentially separated but happening to approach spatially.

Hence, by comparing the entropies of protein-sequence fragments over all possible positions and lengths, we could predict the folding pathway, where the lower the entropy is the earlier the region should be optimized, and the shorter the region, the earlier. The pathway is virtually *written* in the sequence as the form of entropic landscape, which we could decipher by calculating the entropies of sequence fragments.

The entropy of a sequence fragment is difficult to compute and even to define, particularly because of the difficulties associated with the folding in water solution which weakens electrostatic force fields, and gives rise to hydrophobic interaction.

Actually total entropy of a given protein sequence fragment is very difficult to compute, However, if we separate the whole entropy of a fragment into two parts, 1) conformational (sequence-independent) entropy S^C which is the average entropy of a sequence-fragment consisting of average residue of any time, and 2) sequence-dependent entropy S^S, then conformational entropy S^C is constant for all fragments of the same length. Then ΔS, the difference between sequence-dependent and sequence-independent entropy is easily calculated as the sum of the difference of sequence-dependent and independent pair-wise entropy over all possible combinations of main-chain atoms.

1.1 Conformational Entropy of Polypeptides

Given a system that has N micro-states, and suppose that we are able to observe the probability of the system in each micro-state along with the fluctuation of the system, entropy S of the system is calculated as follows,

$$S = -K_B \sum_{i=1}^{N} P_i \ln P_i, \tag{1}$$

where P_i is the probability of the system in micro-state i, and K_B is Boltzmann's constant.

When the system's micro-state is continuous with respect to indexing quantity $\mathbf{x}$, entropy S of the system is calculated from the probability density $\rho(\mathbf{x})$ of the continuous micro-state with respect to $\mathbf{x}$.

$$S = -K_B \int_{\mathbf{x}} \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x} \tag{2}$$

The entropy of a polypeptide fragment could be associated with the probability density of the fragment forming each main-chain conformation indexed by $\mathbf{x}$. For a short peptide fragment, a set of main-chain dihedral angles, ϕ, ψ, ω could be quantity $\mathbf{x}$ which indexes the conformation. For a long peptide fragment, a set of the distances between main-chain atoms could be $\mathbf{x}$, the indexing quantity. In any case for polypeptide conformations, the indexing quantity $\mathbf{x}$ is multi-dimensional. Note that the side-chain conformations are ignored at all in this study. If we could observe the probability density $\rho(\mathbf{x})$, the entropy of a polypeptide conformation is computable.

1.1.1 Conformational Entropy of Ideal Free Polypeptides

Before we figure out the method for calculating the entropy of true polypeptides, let's start with the entropy of ideal *free* polypeptides. Ideal free polypeptides have no interaction

between constituent atoms except for those between tightly bonding atoms. Indeed, ideal free polypeptides are completely sequence dependent. The bond lengths and bond angles between main-chain atoms are fixed. For the main-chain dihedral angles ϕ, ψ, ω, the chance of choosing any possible angle is completely even for ϕ and ψ, and for ω, it is either 0 or π. Any combination of straight ϕ, ψ, ω along the main-chain is allowed and the probability density of that combination is completely even. The entropy for each dihedral angle is, thereby, given independently as below,

$$\begin{aligned} S_\phi &= -K_B \int_{-\pi}^{\pi} \rho(\phi) \ln \rho(\phi) d\phi = K_B \ln 2\pi, \\ S_\psi &= -K_B \int_{-\pi}^{\pi} \rho(\psi) \ln \rho(\phi) d\phi = K_B \ln 2\pi, \\ S_\omega &= -K_B \sum_{i=0}^{1} P(i \times \pi) \ln P(i \times \pi) = K_B \ln 2, \end{aligned} \tag{3}$$

where,

$$\begin{aligned} \int_{-\pi}^{\pi} \rho(\phi) d\phi &= 1 \cdots \phi \\ \int_{-\pi}^{\pi} \rho(\psi) d\psi &= 1 \cdots \psi \\ P(0) + P(\pi) &= 1 \cdots \omega \end{aligned} \tag{4}$$

Hence the entropy of free polypeptides per residue is $K_B\{\ln(2\pi) + \ln(2) + \ln(2\pi)\} = K_B \ln(8\pi^2)$. Because the choice of each dihedral angle is completely independent, the total entropy of the M-residue long ideal free peptide is given as follows,

$$\begin{aligned} S &= (M-1)(S_\psi + S_\omega + S_\phi) + S_\psi \\ &= K_B(M-1)\ln(8\pi^2) + K_B \ln(2\pi), \end{aligned} \tag{5}$$

where the last term $K_B \ln(2\pi)$ is for C-terminal ψ angle, and the degree of freedom for N-terminal ϕ angle is ignored. Note that ideal free polypeptides allow clash between non-bonded atoms because there is *no* repulsiv interaction between them that prevents clash.

Now we look at a pair of main-chain atoms that are non-bonded, and the distribution density of the distance between these atoms. Here, we observe distance r between two C^α atoms that are separated by k residues along the main-chain. The distribution density $\rho^f_{C^\alpha C^\alpha}(k, r)$ with respect to distance r between the atoms is not even, although all dihedral angles along the main-chain segment between the two atoms are chosen completely at random. The number of ϕ angles between these two C^α atom that affect distance r between these two are $k-1$, the number of ψ is $k-1$, and the number of ω is k. Entropy $S^f_{C^\alpha C^\alpha}(k)$ of the main-chain segment between these two C^α's in this case is $K_B\{(k-1)\ln(8\pi^2) + \ln(2)\}$. This entropy must be equivalent to the entropy of distance distribution density $\rho^f_{C^\alpha C^\alpha}(k, r)$ even though $\rho^f_{C^\alpha C^\alpha}(k, r)$ is not flat. Thus partition function $Z^f_{C^\alpha C^\alpha}(k)$ of $\rho^f_{C^\alpha C^\alpha}(k, r)$ fulfills the equation below,

$$\ln Z^f_{C^\alpha C^\alpha}(k) = \frac{S^f_{C^\alpha C^\alpha}(k)}{K_B} = (k-1)\ln(8\pi^2) + \ln(2), \tag{6}$$

where,

$$Z^f_{C^\alpha C^\alpha}(k) = \int_0^\infty \rho^f_{C^\alpha C^\alpha}(k,r)dr. \tag{7}$$

Since ideal free peptides we are considering are completely free from the interactions between non-bonded atoms, and the probability of main-chain dihedral angles is completely even, the entropy of ideal free peptides have the highest diversity and thus highest entropy among any conceivable polypeptides of the same length. This must be also the case for the entropy of the distance between a pair of atoms along the main-chain.

1.1.2 Sequence-independent Entropy of Polypeptides

Now we are quite ready to think about the entropy of true polypeptides. First, let us consider sequence-independent entropy of the true protein structure segment between particular pair of main-chain atoms, where all amino-residue-types of the protein sequence are set to *any* or *average* type in terms of statistics. It is tedious but not difficult to observe the distance between atoms along protein main-chain in the protein structure database, such as PDB. We can, thereby, obtain the distribution density of the distance between a particular pair of main-chain atoms over a lot of protein structures in PDB. As we did in case of ideal free polypeptides before, we also consider the distance distribution density between two C^α atoms that are separated by k residues along the main-chain. The distribution density with respect to distance r between the two C^α is denoted by $\rho^{xx}_{C^\alpha C^\alpha}(k,r)$, where xx denotes that both of C^α atoms are of residues of *any* type because amino-acid-sequence-independent distribution is being discussed here. If distribution density $\rho^{xx}_{C^\alpha C^\alpha}(k,r)$ were very similar to $\rho^f_{C^\alpha C^\alpha}(k,r)$, the distribution density for ideal free polypeptide, we could assume that entropy $S^{xx}_{C^\alpha C^\alpha}(k)$ of the k-residue long main-chain segment between these C^α atoms should be nearly equal to $S^f_{C^\alpha C^\alpha}(k)$, the entropy for the equivalent segment of ideal free polypeptides. Then ratio $\rho^{xx}_{C^\alpha C^\alpha}(k,r)/\rho^f_{C^\alpha C^\alpha}(k,r)$ is nearly constant. However, most of the cases distribution density $\rho^{xx}_{C^\alpha C^\alpha}(k,r)$ is very different from $\rho^f_{C^\alpha C^\alpha}(k,r)$ due to the interactions exerted to the atoms in the segment. For some particular r, distribution density $\rho^{xx}_{C^\alpha C^\alpha}(k,r)$ could be very high, while the equivalent distribution density for ideal free polypeptides at that r would be flat. And also $\rho^{xx}_{C^\alpha C^\alpha}(k,r)$ could be 0 mainly due to the atomic clash, which the free polypeptides allow. Assuming the Boltzmann distribution for protein structures, high frequency of a particular conformation is assumed to be caused by the conformation's having low energy, and low frequency by high energy. However, we should note that there must be plural different conformations that have the equal distance between the atoms we are looking at in the conformations . For example, the distance between C^α atoms of adjacent residues is always 3.8 Åfor any combination of ψ, ϕ angles between the residues as long as ω between the two residues is π. And the distances between main-chain atoms in L-helix are almost identical to the equivalent distances in R-helix. When there are more than two conformations whose distance between that pair of atoms is equal, we have no way to know which conformation among them is frequent just by observing the distribution density at that distance. Thus, as the way for approximation, let us assume that when the distribution density at a particular distance is high the frequency of all conformation that has that distance for the atom pair is equally high. Then ratio $\rho^{xx}_{C^\alpha C^\alpha}(k,r)/\rho^f_{C^\alpha C^\alpha}(k,r)$, (distance distribution density of true

protein conformation per that of ideal free polypeptide,) would indicate the energy of the main-chain segment between two C^α atoms. As we are assuming Boltzmann distribution for protein structures, the energy of the conformation segment between the pair of C^α atoms we are looking at is estimated as below[1],

$$E^{xx}_{C^\alpha C^\alpha}(k,r) = -K_B T \ln \frac{\rho^{xx}_{C^\alpha C^\alpha}(k,r)}{\rho^{f}_{C^\alpha C^\alpha}(k,r)} + E^{xx}_{0,C^\alpha C^\alpha}(k). \tag{8}$$

Here $E^{xx}_{0,C^\alpha C^\alpha}(k)$ is the constant for all r, and would be determined when partition function $Z^{xx}_{0,C^\alpha C^\alpha}(k)$ of this partial system is calculated.

Now we are ready to estimate the entropy of the main-chain segment between the pair of C^α atoms we are looking at, because energy $E^{xx}_{C^\alpha C^\alpha}(k,r)$ with respect to r for this partial system is approximately estimated. At $r = r_0$, the distance distribution density for free polypeptides is $\rho^{f}_{C^\alpha C^\alpha}(k,r_0)$, and that for true protein conformation segment is $\rho^{xx}_{C^\alpha C^\alpha}(k,r_0)$. The probability density $\rho(r_0)$ for the calculation of entropy is proportional to the ratio of two distribution density, thus $-K_B T \ln \rho^{xx}_{C^\alpha C^\alpha}(k,r_0)/\rho^{f}_{C^\alpha C^\alpha}(k,r_0)$, the natural logarithm of this ratio multiplied by $-K_B T$ is the energy difference. Therefore at $r = r_0$, probability density $\rho(r)$ should be multiplied by the distance distribution density for free polypeptides $\rho^{f}_{C^\alpha C^\alpha}(k,r_0)$. Hence the entropy of this segment is given as below.

$$\begin{aligned} S^{xx}_{C^\alpha C^\alpha}(k) &= -K_B \int_0^\infty \rho^{f}_{C^\alpha C^\alpha}(k,r)\rho(r)\ln\rho(r)dr \\ &= -K_B \int_0^\infty \rho^{xx}_{C^\alpha C^\alpha}(k,r) \ln \frac{\rho^{xx}_{C^\alpha C^\alpha}(k,r)}{\rho^{f}_{C^\alpha C^\alpha}(k,r)} dr \\ &= -K_B \int_0^\infty \rho^{xx}_{C^\alpha C^\alpha}(k,r)\ln\rho^{xx}_{C^\alpha C^\alpha}(k,r)dr + K_B \int_0^\infty \rho^{xx}_{C^\alpha C^\alpha}(k,r)\ln\rho^{f}_{C^\alpha C^\alpha}(k,r)dr \end{aligned} \tag{9}$$

This is the estimated sequence-independent entropy of the main-chain segment between two C^α atom that are separated by k residues in the structure. Now we can estimate the entropy of any main-chain segment such as the segment between H atom and O atom that are separated by k residues in the structure, which would be given as follows,

$$S^{xx}_{HO}(k) = -K_B \int_0^\infty \rho^{xx}_{HO}(k,r)\ln\rho^{xx}_{HO}(k,r)dr + K_B \int_0^\infty \rho^{xx}_{HO}(k,r)\ln\rho^{f}_{HO}(k,r)dr \tag{10}$$

For the convenience, we denote uv for a particular pair of main-chain atoms. u and v could be H,N,H^α,C^α,C,O, and uv could be any combination of these atoms. Hence the sequence-independent entropy of main-chain segment between u and v that are separated by k residues in the structure is given as follows.

$$S^{xx}_{uv}(k) = -K_B \int_0^\infty \rho^{xx}_{uv}(k,r)\ln\rho^{xx}_{uv}(k,r)dr + K_B \int_0^\infty \rho^{xx}_{uv}(k,r)\ln\rho^{f}_{uv}(k,r)dr. \tag{11}$$

1.1.3 Sequence-dependent Entropy of Polypeptides

Now we begin to think about the sequence-dependent entropy of the segment. Because we ignore side-chain conformations, we have to distinguish main-chain atoms by the type of

residues to which they belong. For example, C^α atom of Alanine-residue should be distinguished from C^α atom of, say, Leucine-residue. Actually H,N, H^α, C^α atoms are not common to all residue-types because Proline does not have H, Proline's N is chemically different from that of other residue-types, and C^α of Glycine is different in partial charge from that other residue-types. However, C and O atoms are all common to all residue-types in terms of chemical properties. To avoid the redundancy we consider that H, N, H^α, and C^α of residue-type a is different from those of other residue-type b, while C and O are common to all residue-types. The distribution density with respect to distance r between C^α of Glycine and C^α of Proline must be very different from that between C^α of Aspartic residue and Lyzine residue and are also different from that between two C^α of average (or any or sequence-independent) residue. We need, however more discussions about which atoms are considered residue-type-dependent, and which independent. We may set that only C^α is type-dependent and others are independent, as a simplest way.

The sequence-dependent entropy of the segment between atom u of residue-type a and v of residue-type b that are separated by k residues in the amino acid sequence should be naturally calculated as below,

$$S_{uv}^{ab}(k) = -K_B \int_0^\infty \rho_{uv}^{ab}(k,r) \ln \rho_{uv}^{ab}(k,r) dr + K_B \int_0^\infty \rho_{uv}^{ab}(k,r) \ln \rho_{uv}^{f}(k,r) dr. \qquad (12)$$

where uv is one of the combination of H, N, H^α, C^α and thus residue-type a and b are specified. If u is one of H, N, H^α, C^α and v is either C or O, residue-type b is not specified, thus,

$$S_{uv}^{ax}(k) = -K_B \int_0^\infty \rho_{uv}^{ax}(k,r) \ln \rho_{uv}^{ax}(k,r) dr + K_B \int_0^\infty \rho_{uv}^{ax}(k,r) \ln \rho_{uv}^{f}(k,r) dr. \qquad (13)$$

And if u is either C or O and v is one of H, N, H^α, C^α, residue-type a is not specified, thus,

$$S_{uv}^{xb}(k) = -K_B \int_0^\infty \rho_{uv}^{xb}(k,r) \ln \rho_{uv}^{xb}(k,r) dr + K_B \int_0^\infty \rho_{uv}^{xb}(k,r) \ln \rho_{uv}^{f}(k,r) dr \qquad (14)$$

1.2 Entropy of Sequence Fragments

Since our final goal is to calculate the entropy of a whole protein structure, or a fragment of the protein structure, we need to find the way to estimate such entropies from the sequence-dependent and sequence-independent entropy of the main-chain segments building the whole structure. We denote the sequence-dependent entropy of a k-residue long fragment by $S^s(k)$ and that of sequence-independent (conformation) by $S^c(k)$.

The entropy of a fragment $S^s(k)$ is not calculated just by the summation of segment entropies over all component segments in the fragment. However, $\Delta S(k) = S^s(k) - S^c(k)$, the difference between sequence-dependent and sequence-independent entropy of a k residue long structure fragment turned out to be computable, which is just the sum of the difference between sequence-dependent and sequence-independent entropy of the segments over all possible segments within the fragment.

$$\begin{aligned}\Delta S(k) &= S^s(k) - S^c(k) \\ &= \sum_{1\le i\le j\le k}\sum_{uv}\begin{pmatrix} S^{ax}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u = H, N, H^\alpha, C^\alpha, v = C, O \\ S^{ab}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u, v = H, N, H^\alpha, C^\alpha \\ S^{xb}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u = C, O, v = H, N, H^\alpha, C^\alpha \end{pmatrix}. \end{aligned} \tag{15}$$

$\Delta S(k)$ is no less valuable than $S^s(k)$ for the purpose of analyzing the physical stability of a sequence fragment in protein structures. And sequence independent entropy $S^c(k)$ is by definition independent of the amino-residue sequence of the fragment but dependent only on length k of the fragment, and could be roughly estimated as entropy $S^{xx}_{uv}(k-1)$, though this estimation would not be accurate enough, particularly when the fragment length is long.

1.3 Scrutinizing The Entropy Calculation Methods

Here we need to scrutinize the quality of $\Delta S(k)$ and see what it actually mean. When we estimated $S^{xx}_{uv}(k)$ above, we assumed that the frequency of a particular distance relative to the that of free polypeptides is caused by the low or high energy of the conformation including the segment. We assumed that if the energy is low the frequency is high, and if high energy then low frequency, assuming Boltzmann distribution for conformations. We then did not check whether the low energy is caused by the interaction between atoms within the fragment or otherwise. For example, α helix is very frequent structure in which the distance between two O atoms of straight two residues is short and is around 3.5 Å, thus the distribution density of the distance between those two O atoms has a crest of frequency at the distance being about 3.5 Å. Considering that both O atoms are negatively charged, the interaction between these O atoms must be repulsive, thus the energy is not low at that distance, 3.5 Å. When we are just looking at two-residue long fragments, the high frequency at the distance 3.5 Åbetween two O atoms in the fragment would seem mysterious. Actually the high frequency at 3.5 Åturns out to be caused by the hydrogen bonding between H and O atoms at least one of which is located outside of that two residue long fragment we are considering. Hence, very frequent conformations of fragments do not always have low energy and are not necessarily stable when the fragments are cut out and isolated alone. It would probably be the case for strand conformations forming β sheet structures. Strands alone should be stable due to the alternating CONH dipole groups along the main-chain, but could be further stabilized by coupling with other strands to form β sheets in which low energy hydrogen bonds would be abundant. Also one-turn helices would be further stabilized by more hydrogen bonds just by extending themselves to multi-turn helices.

1.3.1 Absolute Entropy for The Final Stages of Foldings

Simply put, the entropy calculation method above is actually trying to calculate the entropy of the fragments in the final fold, where fragments are stabilized both by intra-fragment and inter-fragment interactions. $\Delta S(k)$ of the sequence fragments would, thereby, suggest which part of the sequence is expected to be stable or unstable in the final fold of the protein. We call this method of entropy calculation "absolute method" which is based on the assumption that the high or low frequency of a particular distance between a particular atomic pair is caused by low or high energy, respectively, of all interactions exerted to the structure segment between the two atoms. The absolute method is calculated by equations 13, 12, 14, 11. They are summerized into the following equations.

$$\begin{aligned}
S_{uv}^{ax}(k) &= -K_B \int_0^\infty \rho_{uv}^{ax}(k,r) \ln \rho_{uv}^{ax}(k,r) dr \\
&\quad +K_B \int_0^\infty \rho_{uv}^{ax}(k,r) \ln \rho_{uv}^{f}(k,r) dr, \cdots u = H, N, H^\alpha, C^\alpha, v = C, O \\
S_{uv}^{ab}(k) &= -K_B \int_0^\infty \rho_{uv}^{ab}(k,r) \ln \rho_{uv}^{ab}(k,r) dr \\
&\quad +K_B \int_0^\infty \rho_{uv}^{ab}(k,r) \ln \rho_{uv}^{f}(k,r) dr, \cdots u, v = H, N, H^\alpha, C^\alpha \\
S_{uv}^{xb}(k) &= -K_B \int_0^\infty \rho_{uv}^{xb}(k,r) \ln \rho_{uv}^{xb}(k,r) dr \\
&\quad +K_B \int_0^\infty \rho_{uv}^{xb}(k,r) \ln \rho_{uv}^{f}(k,r) dr, \cdots u = C, O, v = H, N, H^\alpha, C^\alpha \\
S_{uv}^{xx}(k) &= -K_B \int_0^\infty \rho_{uv}^{xx}(k,r) \ln \rho_{uv}^{xx}(k,r) dr \\
&\quad +K_B \int_0^\infty \rho_{uv}^{xx}(k,r) \ln \rho_{uv}^{f}(k,r) dr \cdots \text{always.}
\end{aligned} \tag{16}$$

1.3.2 Net Entropy for Early Stages of Foldings

When a protein sets out to fold, the initial main-chain conformation of the protein is assumed to be rather stretched and then the non-bonded main-chain atoms should be spatially separated from each other. Considering that the electrostatic force fields (and resultant hydrogen-bonds) which are assumed to be the dominant force fields stabilizing protein structures (particularly secondary structures) at the final folds would be very weak due to water's large relative permittivity, the electrostatic interaction between atoms that are sequentially separated would be very weak. Not only electrostatic interaction, but hydrophobic interactions between non-bonded parts of a molecule are also week when two hydrophobic parts are spatially separated far away from each other. Then, the entropy of structure fragments calculated in the above-mentioned way does not represent the stability of the fragment at the initial stage of folding nor the strength of the fragment's tendency (or could be termed as "desire") to fold into particular conformations. We need to, thus, seek for another way to calculate the entropy for the initial stage of folding. In order to calculate the proper entropy of the peptide fragment, we need to distinguish intra-fragment interactions (those interac-

tion within the fragment we are looking at) from all interaction exerted to that fragment (including inter-fragment interactions). The exact method of calculating the fragment proper entropy should be, thereby, energy calculation of the fragment over all possible conformations of that fragment in order to determine energy E_i of each state i. This method sounds infeasible, considering the enormous amount of required calculation particularly for long fragments whose number of possible conformations is uncountably large. One of the simple and feasible but challenging methods for approximation is to ignore whatever main-chain interactions altogether for sequence-independent entropy. Then the sequence-dependent net energy for atomic interactions is the subject for calculating the entropy. The sequence-dependent net energy is calculated as below,

$$\begin{aligned} E^{ax}_{uv}(k,r) &= E^{ax}_{0,uv}(k) - K_B T \ln \frac{\rho^{ax}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \cdots u = H, N, H^{\alpha}, C^{\alpha}, v = C, O \\ E^{ab}_{uv}(k,r) &= E^{ab}_{0,uv}(k) - K_B T \ln \frac{\rho^{ab}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \cdots u, v = H, N, H^{\alpha}, C^{\alpha} \\ E^{xb}_{uv}(k,r) &= E^{xb}_{0,uv}(k) - K_B T \ln \frac{\rho^{xb}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \cdots u = C, O, v = H, N, H^{\alpha}, C^{\alpha} \end{aligned} \tag{17}$$

Here $E^{ax}_{0,uv}(k)$, $E^{ab}_{0,uv}(k)$, $and E^{xb}_{0,uv}(k)$, the energy constants with respect to distance r are determined when partition functions are calculated. Therefore, they can be ignored in this stage of calculation. Thus, entropies are given as below,

$$\begin{aligned} S^{ax}_{uv}(k) &= -K_B \int_0^{\infty} \frac{\rho^{ax}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \ln \frac{\rho^{ax}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} dr \cdots u, v = H, N, H^{\alpha}, C^{\alpha} \\ S^{ab}_{uv}(k) &= -K_B \int_0^{\infty} \frac{\rho^{ab}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \ln \frac{\rho^{ab}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} dr \cdots u = H, N, H^{\alpha}, C^{\alpha}, v = C, O \\ S^{xb}_{uv}(k) &= -K_B \int_0^{\infty} \frac{\rho^{xb}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \ln \frac{\rho^{xb}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} dr \cdots u = C, O, v = H, N, H^{\alpha}, C^{\alpha} \\ S^{xx}_{uv}(k) &= -K_B \int_0^{\infty} \frac{\rho^{xx}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} \ln \frac{\rho^{xx}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} dr \cdots \text{always.} \end{aligned} \tag{18}$$

Here sequence-independent entropy $S^{xx}_{uv}(k)$ is not zero nor constant because the ratio $\rho^{xx}_{uv}(k,r)/\rho^{xx}_{uv}(k,r)$ could be 0 when $\rho^{xx}_{uv}(k,r) = 0$.

We call these entropies from net energy, "net entropy."

1.3.3 Cross Entropy for Intermediate Stages of Foldings

We can also conceive one more way to calculate the entropy, which we hope to represent the entropy at the intermediate stage of folding between initial and final. That could be called "cross entropy" which is calculated as follows.

$$S^{ax}_{uv}(k) = -K_B \int_0^{\infty} \rho^{ax}_{uv}(k,r) \ln \frac{\rho^{ax}_{uv}(k,r)}{\rho^{xx}_{uv}(k,r)} dr \cdots u, v = H, N, H^{\alpha}, C^{\alpha}$$

$$
\begin{aligned}
S_{uv}^{ab}(k) &= -K_B \int_0^\infty \rho_{uv}^{ab}(k,r) \ln \frac{\rho_{uv}^{ab}(k,r)}{\rho_{uv}^{xx}(k,r)} dr \cdots u = H, N, H^\alpha, C^\alpha, v = C, O \\
S_{uv}^{xb}(k) &= -K_B \int_0^\infty \rho_{uv}^{xb}(k,r) \ln \frac{\rho_{uv}^{xb}(k,r)}{\rho_{uv}^{xx}(k,r)} dr \cdots u = C, O, v = H, N, H^\alpha, C^\alpha \\
S_{uv}^{xx}(k) &= -K_B \int_0^\infty \rho_{uv}^{xx}(k,r) \ln \frac{\rho_{uv}^{xx}(k,r)}{\rho_{uv}^{xx}(k,r)} dr = 0 \cdots \text{always.}
\end{aligned} \tag{19}
$$

In this case, the sequence-independent entropy is always 0, thus the $\Delta S(k)$ is the equal to sequence-dependent entropy $\Delta S^s(k)$. This "cross" method calculates entropy considering the energy of forming structure to be the "net" energy while distribution density to be "absolute" for $\rho_{uv}^{ax}(k,r)$, $\rho_{uv}^{ab}(k,r)$, and $\rho_{uv}^{xb}(k,r)$.

1.4 Entropic Landscape

We now have thee methods for calculating entropies of given fragments of amino sequence. What we want to do next is to see which region (or fragment) of the whole protein sequence (or structure) has low entropy and which has high. Then we need to cut the whole sequence into many fragments. In some cases, a single residue could play a special role in folding or binding to other molecules, and then the residue which could be considered a single residue fragment has low or high entropy associated with its role. Or in other cases, a very long fragment plays a special role such as forming a hydrophobic core that induce the folding of other sequential regions, then the long fragment could have low entropy. Calculating a given sequence fragment does not suffices. We need to calculate the entropy of all possible fragments of given sequences. The number of possible fragments of an M residue-long sequence is $M(M+1)/2$, which is equivalent to the number of entropies to calculate. It is not difficult to calculate this number of entropy given an entropy table whose components are $S_{uv}^{ab}(k), S_{uv}^{ax}(k), S_{uv}^{xb}(k), S_{uv}^{xx}(k)$ for all possible combination of uv and k, because the target entropies of all fragments are just the summations over component entropies.

We denote the difference of the sequence-dependent and sequence independent entropy of the fragment as $\Delta S_{i_0 \sim j_0}$, where i_0 is the position of the fragment's first residue in the whole sequence, and j_0, the last. $\Delta S_{i_0 \sim j_0}$ is the sum of entropy difference $S_{ij}^s - S_{ij}^c$ over all combination of ij where $(i_0 \le i \le j \le j_0)$. The component sequence-dependent entropy S_{ij}^s and S_{ij}^c are given by the sum of $S_{uv}^{ab}(k = j - i), S_{uv}^{ax}(k = j - i), S_{uv}^{xb}(k = j - i), S_{uv}^{xx}(k = j - i)$ over all combination of atoms uv, where a is the residue type of i-th residue, and b the type of j-th. These component entropies needed to calculate the fragment entropies are given as in the entropy table before the sequence is given.

The entropic landscape of the protein sequence is the set of $\Delta S_{i_0 \sim j_0}$, which demonstrates what fragment of and what position $(i_0 \sim j_0)$ of what length $(j_0 - i_0)$ has lower or higher entropy than other fragments of the same length within the sequence. And three different methods, absolute, cross, and net of entropy calculation show what fragment is stable or unstable in what stage of folding. Low entropy fragments in net entropic landscape could

be considered to fold into conformations that are usually unstable in the final stage of folding because low entropy fragments in net entropic landscape usually have high entropy in absolute entropic landscape. And low entropy fragments in absolute entropic landscape normally correspond to regular secondary structures like α helices or β strands which are stabilized by abundant hydrogen bonds.

2 Results

Fig 1 and Fig 2 show the entropic landscape of protein L (1K50A). The high entropy regions by the absolute method and low entropy regions by the cross and net methods correspond to turn region or the transition between Strand to Helix or vice-versa particularly when k= 4 (the fragments are 5 residue long).

Fig 3 shows the entropic landscape of small protein sh3 (1ABQ in PDB). The folding pathway of this protein has been analyzed by experimentalists, demonstrating that the anti-parallel sheet consisting of straight three strands is the dominant structure at the early stage of folding [2]. The low entropy bottoms by the net and cross methods seem to be corresponding the strands region near the hairpin-like turns, which seem to have a significant role in the formation of that anti-parallel sheet.

Fig 4 shows the entropic landscape of protein DCTD (1QKKA), which consists of Greek-key like fold with a long helix around the c-terminal region. The turns and transitional regions from helix to strand or vice versa corresponds to the low entropy regions computed by the net and cross methods.

Fig 5 shows the entropic landscape of a small protein *Pleurotus ostreatus* proteinase A inhibitor 1 (POIA1, 1ITP in PDB). The low entropy region is located around the hairpin turn. The formation of the hairpin sheet the packing of the sheet with helix should take place at an early stage of folding. This assumption of pathway suggested by entropic landscape almost completely agrees with the folding process proposed by experiments[3].

Another story of this protein, POIA1, is even more intriguing. The propeptide of a serine protease subtilisin BPN (the P chain of 1SPB in PDB) have a structure almost identical to that of POIA1 in terms of topology when it is bound to the other unit of the complex. Although POIA1 folds into a stable structure by itself, the propeptide does not when it is alone[4]. The sequence identity between POIA1 and the propeptide is roughly 20%. The entropic landscape of the propeptide yields an insight into why the propeptide does not fold by itself. The region for the hairpin turn between strand 2 and 3 does not have low entropy compared to the equivalent region of POIA1. As is mentioned above, this hairpin region of POIA1 is considered to be the fold starter. Naturally, the propeptide devoid of the fold-starter region would not fold. This hairpin turn region turns out to be the very site which binds to the two-helix bundle of subtilisin BPN. The folding of the propeptide must be triggered by the binding to the other subunit, and then the hairpin turn is formed. The overall entropic landscape of the propeptide is, in short, similar to that of POIA1 except for the hairpin region, as shown in Fig6. The true propeptide's folding is slightly different. The propeptide is a part of the whole sequence of subtilisin BPN when synthesized. The whole sequence folds into a structure, in which the hairpin region of the propeptide binds to the helix bundle of subtilisin BPN.

Finally, Fig7 shows the entropic landscapes of very popular globins, that are human hemoglobin A-chain and human myoglobin. They share almost identical conformations, but the entropic landscapes differ from each other. This might be relevant to the fact that hemoglobin normally does not fold in the absence of hem while myoglobin folds.

Figure 1: The Entropic Landscape of Protein L (1K50A) for k=0,1,2.

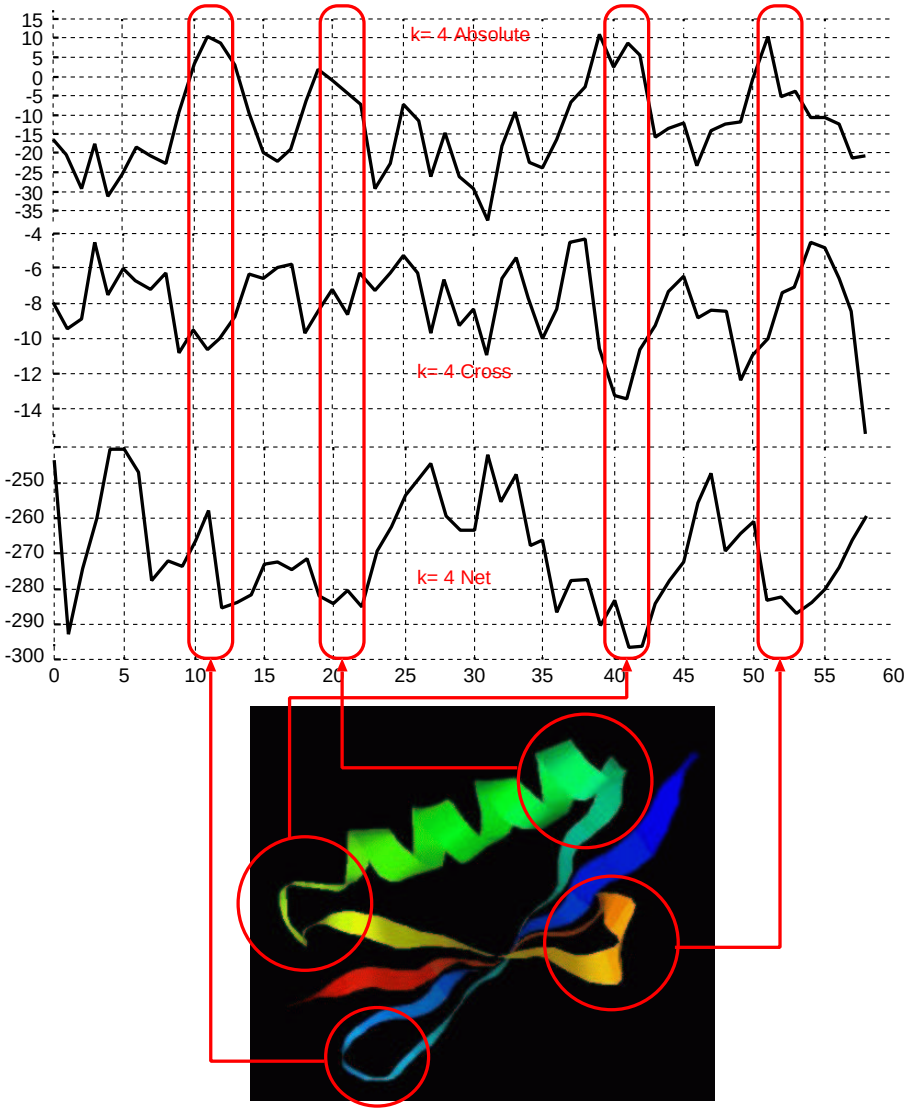

Figure 2: The Entropic Landscape of Protein L (1K50A) for k=4.

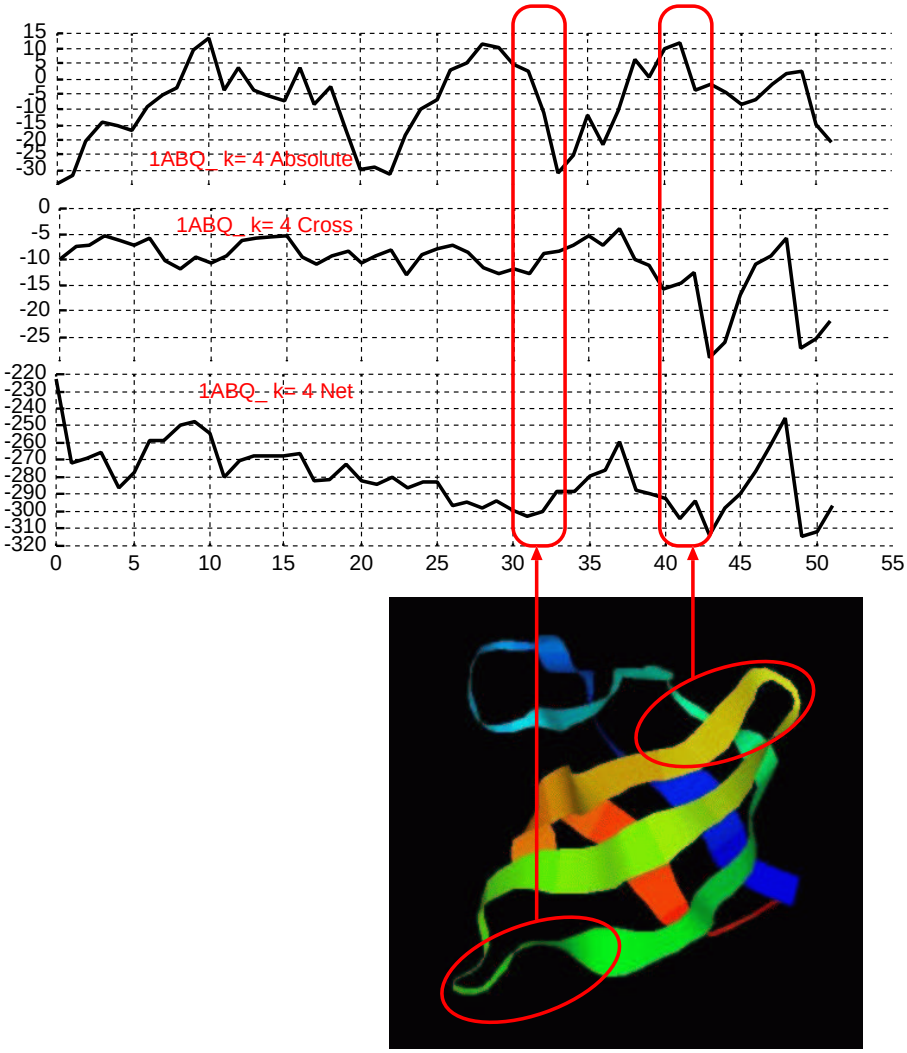

Figure 3: The Entropic Landscape of sh3 (1ABQ) for k=4.

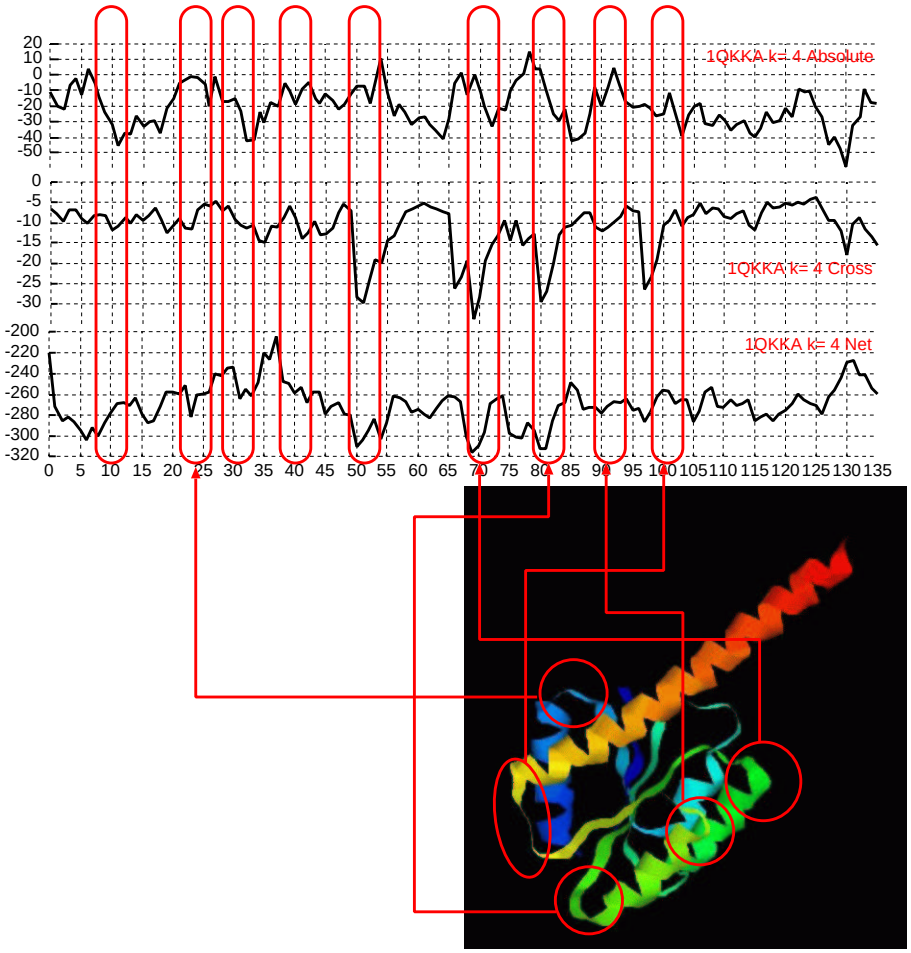

Figure 4: The Entropic Landscape of DCTD (1QKKA) for k=4.

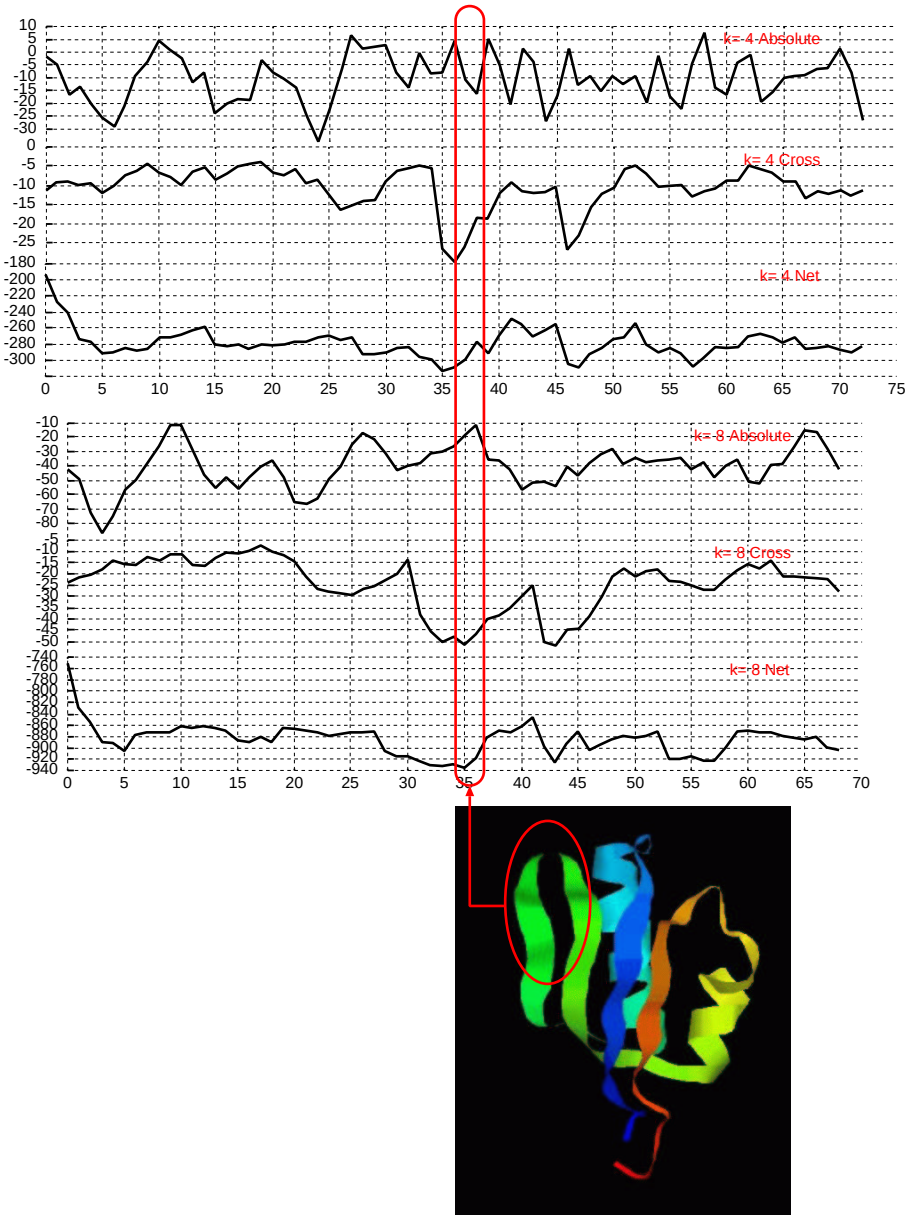

Figure 5: The Entropic Landscape of POIA1 (1ITPA) for k=4.

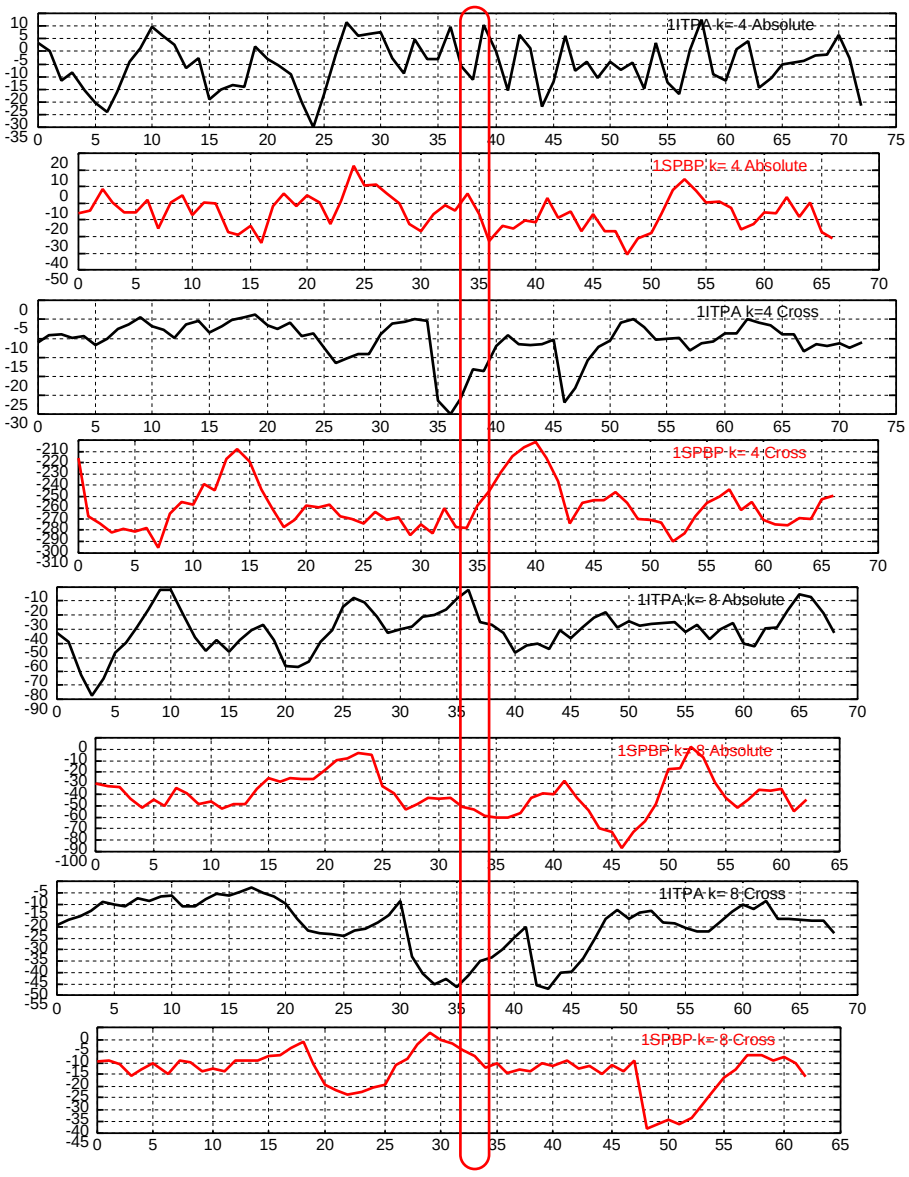

Figure 6: The Entropic Landscape of POIA1 and Subtilisin Propeptide (1ITPA and 1SPBP) for k=4.

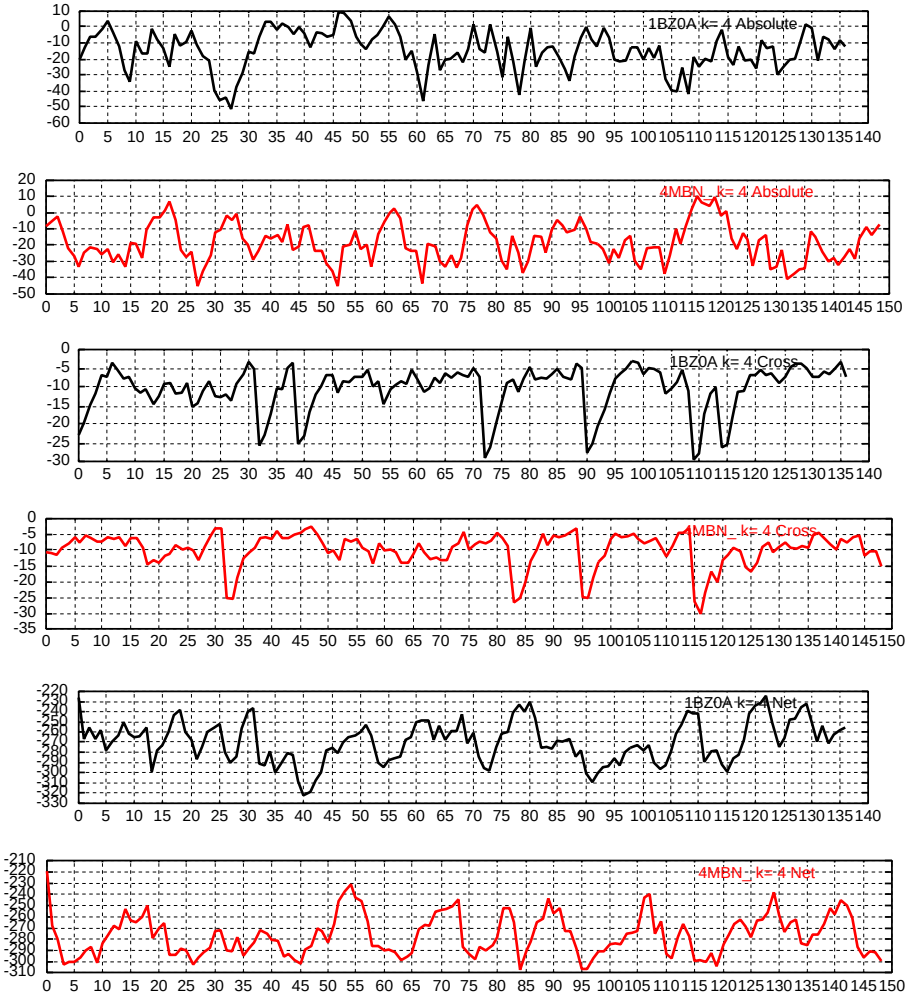

Figure 7: The Entropic Landscape of Hemoglobin and Myoglobin for k=4.

3 Discussion

Here we have three different methods to calculate the entropy of sequential regions (or fragments). The entropic landscape, which is the set of fragment entropies over all possible fragments of a given protein sequence, shows which region is statistically stable. The landscape computed by the absolute method seems to show the each fragment's stability at the final stage of folding where each fragment is stabilized not only by the interactions within the fragment but by other neighbouring fragments in the fold. Hence low entropy regions usually correspond to α helices or β strands forming sheets. On the other hand, The landscape by the cross or net method shows which region has a strong desire not to folding into regular α or β structures, where the low entropy regions usually correspond to turns or the transitional regions between α or β structures. Because turns and transitional regions are not stable at the final stage of the fold, they normally have high entropy computed by the absolute method.

The computer programs for the calculation of entropic landscape is commercially available. The software package contains programs for reading PDB entry files and compiling them into the statistical data required to calculate entropy tables. Given a protein sequence, the entropic landscape calculation program calculate the entropic landscape of the sequence.

The choice of sequence-dependent and sequence-independent atoms out of main-chain atoms is one thing we need to decide. In the present study sequence-dependent atoms are H, N, C^{α}, H^{α}, and C, O are considered sequence-independent. But when only Cα is set to sequence-dependent and other interactions are ignored, the entropic landscape becomes more reasonable in terms of the lowentropy regions' correspondence to secondary structures. There must be wiser way to calculate entropic landscape then the three methods proposed in this paper.

4 Method

4.1 Statistical Data Required for Entropy Calculation

In order to calculate the entropy of a sequence fragment, we need the statistical data below obtained from analysis of protein structures selected from PDB.

4.1.1 The distribution density with respect to sequence separation k and distance r

Distribution density $\rho(k,r)$ with respect to sequential separation k and distance r varies according to the pair of atoms and the types of residues to which each atom belongs. And to calculate entropy by the absolute method, we have to generate random conformations for distribution density of random conformation.

- $\rho^{f}_{uv}(k,r)$

  The distribution density with respect to distance r between atom u and v sequentially separated by k residues, of *ideal free* polypeptides. This is obtained from artificially generated random peptide conformation by setting random dihedral angles along the main-chain. The range of ϕ is $(-\pi < \phi < \pi)$, that of ψ is $(-\pi < \psi < \pi)$, and ω is either 0 or π. However, to generate more realistic or natural random structure, choosing a combination of ω, ϕ, ψ at random for each residue out of all possible combination of ω, ϕ, ψ in the statistically observed data, where observed ϕ and ψ angles are rounded to $n\pi/12$, and ω is rounded to either 0 or π, in this study.

- $\rho^{xx}_{uv}(k,r)$
  The distribution density with respect to distance r between atom u of any residue-type x and atom v of any residue-type x sequentially separated by k residues, analyzed from the selected protein structures from PDB.

- $\rho^{ax}_{uv}(k,r)$,$\rho^{ab}_{uv}(k,r)$,$\rho^{xb}_{uv}(k,r)$
  The distribution density with respect to distance r between atom u of residue-type a (or any x) and atom v of residue-type b (or any x) sequentially separated by k residues, analyzed from the selected protein structures from PDB.

In any case, when k is larger than a certain threshold such as 15, the distribution density $\rho_{uv}(k,r)$ is considered *almost* identical and those are combined altogether.

The distribution density of distance with respect to the distance cannot directly obtained just by observing the data set. First we need to make a histogram of the distribution. The bin size of the histogram with respect to distance is variable. When the distance is short, the bin size should be small enough (such as 0.2 Å) to detect the small difference in distance, while long, the size can be large ($\geq$ 1.0 Å). The distribution density with respect to the distance is approximately calculated by interpolation using second order Bezier curve.

Here we need to consider how to deal with N-terminal -NH_3^+ group and C-terminal -COO^- group. Group -NH_3^+ is positively charged as a whole, and the orientation of the group is not supposed to affect the conformation strongly. In this study, group -NH_3^+ is considered a single atom N^T and independent of the residue-type. On the other hand, -COO^- is considered to consist of three atoms C^T,O^T, and O^{XT} which are all independent of residue-types.

Consequently, the atoms dependent on residue-type are, H,N,H^α,C^α, and the atoms independent of residue-type are N^T,C,O,C^T,O^T,O^{XT}.

- The combinations of atoms uv for the residue-type dependency being ax are those 20 below
  - H-C, H-O, H-C^T, H-O^T, H-O^{XT},
  - N-C, N-O, N-C^T, N-O^T, N-O^{XT},
  - H^α-C, H^α-O, H^α-C^T, H^α-O^T, H^α-O^{XT},
  - C^α-C, C^α-O, C^α-C^T, C^α-O^T, C^α-O^{XT}.
- The combinations of atoms uv for the residue-type dependency being ab are those 16 below,
  - H-H, H-N, H-H^α, H-C^α,
  - N-H, N-N, N-H^α, N-C^α,
  - N-H, N-N, N-H^α, N-C^α,
  - H^α-H, H^α-N, H^α-H^α, H^α-C^α,
  - C^α-H, C^α-N, C^α-H^α, C^α-C^α.
- The combinations of atoms uv for the residue-type dependency being xb are those 12 below,
  - N^T-H, N^T-N, N^T-H^α, N^T-C^α,
  - C-H, C-N, C-H^α, C-C^α,
  - O-H, O-N, O-H^α, O-C^α.
- The combinations of atoms uv for the residue-type dependency being xx are all those whose u is one of N^T, H, N, H^α, C^α, C, O, and whose v is one of H, N, H^α, C^α, C, O, C^T, O^T, O^{XT}. In total there are 63 compbinations.

4.1.2 The distribution density with respect to dihedral angles ϕ, ψ, ω

The distance dependent distribution density is not accurate in terms of determining the relative configuration of residues. For an internal degree of freedom for a residue, and the relative configuration of straight two residues along residues, the combination of two or three straight dihedral angles is the best indexing quantity to represent the conformation.

- $\rho^x(\phi, \psi)$
  The distribution density with respect to the combination of dihedral angles ϕ, ψ of any residue-type x, analyzed from the selected protein structures from PDB.

- $\rho^a(\phi, \psi)$
  The distribution density with respect to the combination of dihedral angles ϕ, ψ of residue-type a, analyzed from the selected protein structures from PDB.

- $\rho^x(\phi)$

  The distribution density with respect to ϕ of any residue-type x, analyzed from the selected protein structures from PDB.

- $\rho^a(\phi)$
  The distribution density with respect to ϕ of residue-type a, analyzed from the selected protein structures from PDB.

- $\rho^x(\psi)$

  The distribution density with respect to ψ of any residue-type x, analyzed from the selected protein structures from PDB.

- $\rho^a(\psi)$
  The distribution density with respect to ψ of residue-type a, analyzed from the selected protein structures from PDB.

- $\rho^{xx}(\psi, \omega, \phi)$
  The distribution density with respect to the combination of dihedral angles ψ, ω, ϕ between straight two residues of any types, analyzed from the selected protein structures from PDB.

- $\rho^{ab}(\psi, \omega, \phi)$
  The distribution density with respect to the combination of dihedral angles ψ, ω, ϕ between straight two residues whose type is a and b respectively, analyzed from the selected protein structures from PDB.

The distribution density with respect to angle combination is obtained from the histogram of the distribution with respect to angle combination, where the bin size is $\pi/12$ for ϕ and ψ, and for ω, π. The distribution density is approximately calculated by smoothing the histogram, using two dimensional second-order Bezier interpolation surface for the combination of ϕ and ψ, and for ω, cis and trans are considered two discrete states.

4.2 Statistical Data Compilation from Selected PDB structures

The PDB structures used in this study are listed below. The total number of chains is 1538. The PDB entries are from PDB Release 102. The sequence identity between any two chains is less than 30 %. The shortest chain is 40-residue long.

The list of 1538 protein chains is given in Appendix.

4.3 Entropy Table

The entropy table consists of element entropies below,

- $S^x_{\phi\psi}$,$S^a_{\phi\psi}$
  The entropies of $\phi\psi$ angle combination for a residue. $S^x_{\phi\psi}$ is for any residue type, and $S^a_{\phi\psi}$ for a specific residue type a. These are given as follows,
  - Absolute method

$$\begin{aligned} S^x_{\phi\psi} &= -K_B \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \rho^x(\phi,\psi)\ln\rho^x(\phi,\psi)d\psi d\phi \\ S^a_{\phi\psi} &= -K_B \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \rho^a(\phi,\psi)\ln\rho^a(\phi,\psi)d\psi d\phi \end{aligned} \quad (20)$$

  - Cross method

$$\begin{aligned} S^x_{\phi\psi} &= 0 \\ S^a_{\phi\psi} &= -K_B \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \rho^a(\phi,\psi)\ln\frac{\rho^a(\phi,\psi)}{\rho^x(\phi,\psi)}d\psi d\phi \end{aligned} \quad (21)$$

  - Net method

$$\begin{aligned} S^x_{\phi\psi} &= -K_B \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\rho^x(\phi,\psi)}{\rho^x(\phi,\psi)}\ln\frac{\rho^x(\phi,\psi)}{\rho^x(\phi,\psi)}d\psi d\phi \\ S^a_{\phi\psi} &= -K_B \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\rho^a(\phi,\psi)}{\rho^x(\phi,\psi)}\ln\frac{\rho^a(\phi,\psi)}{\rho^x(\phi,\psi)}d\psi d\phi \end{aligned} \quad (22)$$

- S^x_ϕ,S^a_ϕ,S^x_ψ,S^a_ψ
  The entropies of ϕ or ψ angle for a residue. S^x_ϕ,S^x_ψ are for any residue type, and S^a_ϕ,S^a_ψ are for a specific residue type a. These are given as follows,
  - Absolute method

$$\begin{aligned} S^a_\phi &= -K_B \int_{-\pi}^{\pi} \rho^a(\phi)\ln\rho^a(\phi)d\phi \\ S^x_\phi &= -K_B \int_{-\pi}^{\pi} \rho^x(\phi)\ln\rho^x(\phi)d\phi \\ S^a_\psi &= -K_B \int_{-\pi}^{\pi} \rho^a(\psi)\ln\rho^a(\psi)d\psi \\ S^x_\psi &= -K_B \int_{-\pi}^{\pi} \rho^x(\psi)\ln\rho^x(\psi)d\psi \end{aligned} \quad (23)$$

– Cross method

$$\begin{aligned} S^a_\phi &= -K_B \int_{-\pi}^{\pi} \rho^a(\phi) \ln \frac{\rho^a(\phi)}{\rho^x(\phi)} d\phi \\ S^x_\phi &= 0 \\ S^a_\psi &= -K_B \int_{-\pi}^{\pi} \rho^a(\psi) \ln \frac{\rho^a(\psi)}{\rho^x(\psi)} d\psi \\ S^x_\psi &= 0 \end{aligned} \tag{24}$$

– Net method

$$\begin{aligned} S^a_\phi &= -K_B \int_{-\pi}^{\pi} \frac{\rho^a(\phi)}{\rho^x(\phi)} \ln \frac{\rho^a(\phi)}{\rho^x(\phi)} d\phi \\ S^x_\phi &= -K_B \int_{-\pi}^{\pi} \frac{\rho^x(\phi)}{\rho^x(\phi)} \ln \frac{\rho^x(\phi)}{\rho^x(\phi)} d\phi \\ S^a_\psi &= -K_B \int_{-\pi}^{\pi} \frac{\rho^a(\psi)}{\rho^x(\psi)} \ln \frac{\rho^a(\psi)}{\rho^x(\psi)} d\psi \\ S^x_\psi &= -K_B \int_{-\pi}^{\pi} \frac{\rho^x(\psi)}{\rho^x(\psi)} \ln \frac{\rho^x(\psi)}{\rho^x(\psi)} d\psi \end{aligned} \tag{25}$$

- $S^{xx}_{\psi\omega\phi}$,$S^{ab}_{\psi\omega\phi}$
The entropies of the combination of angles $\psi\omega\phi$ between two straight residues. $S^{xx}_{\psi\omega\phi}$ is for any two straight residues, and $S^{ab}_{\psi\omega\phi}$ is for the two straight residues of residue-type a and b. These are given as follows,

– Absolute method

$$\begin{aligned} S^{xx}_{\psi\omega\phi} &= -K_B \sum_{i=0}^{1} \int_{-\pi}^{\pi} \int_{\pi}^{\pi} \rho^{xx}(\psi, \omega = i \times \pi, \phi) \ln \rho^{xx}(\psi, \omega = i \times \pi, \phi) d\phi d\psi \\ S^{ab}_{\psi\omega\phi} &= -K_B \sum_{i=0}^{1} \int_{-\pi}^{\pi} \int_{\pi}^{\pi} \rho^{ab}(\psi, \omega = i \times \pi, \phi) \ln \rho^{ab}(\psi, \omega = i \times \pi, \phi) d\phi d\psi \end{aligned} \tag{26}$$

– Cross method

$$\begin{aligned} S^{xx}_{\psi\omega\phi} &= 0 \\ S^{ab}_{\psi\omega\phi} &= -K_B \sum_{i=0}^{1} \int_{-\pi}^{\pi} \int_{\pi}^{\pi} \rho^{ab}_{(}\psi, \omega = i \times \pi, \phi) \ln \frac{\rho^{ab}(\psi, \omega = i \times \pi, \phi)}{\rho^{xx}(\psi, \omega = i \times \pi, \phi)} d\phi d\psi \end{aligned} \tag{27}$$

– Net method

$$S^{xx}_{\psi\omega\phi} = -K_B \sum_{i=0}^{1} \int_{-\pi}^{\pi} \int_{\pi}^{\pi} \frac{\rho^{xx}(\psi, \omega = i \times \pi, \phi)}{\rho^{xx}(\psi, \omega = i \times \pi, \phi)} \ln \frac{\rho^{xx}(\psi, \omega = i \times \pi, \phi)}{\rho^{xx}(\psi, \omega = i \times \pi, \phi)} d\phi d\psi$$

$$S^{ab}_{\psi\omega\phi} = -K_B \sum_{i=0}^{1} \int_{-\pi}^{\pi} \int_{\pi}^{\pi} \frac{\rho^{ab}(\psi, \omega = i \times \pi, \phi)}{\rho^{xx}(\psi, \omega = i \times \pi, \phi)} \ln \frac{\rho^{ab}(\psi, \omega = i \times \pi, \phi)}{\rho^{xx}(\psi, \omega = i \times \pi, \phi)} d\phi d\psi \tag{28}$$

- $S^{ax}_{uv}(k)$,$S^{ab}_{uv}(k)$,$S^{xb}_{uv}(k)$, $S^{xx}_{uv}(k)$
  The entropies calculated from distribution densities $\rho^{ax}_{uv}(k,r)$, $\rho^{ab}_{uv}(k,r)$, $\rho^{xb}_{uv}(k,r)$, $\rho^{xx}_{uv}(k,r)$ respectively, and in case of absolute method, $\rho^{f}_{uv}(k,r)$. They are calculated by equations 16,19,18, as mentioned before.

4.4 Calculation of The Entropic Landscape

Finally we are now ready to calculate the entropic landscape of given sequence. The entropic landscape is a set of $\Delta S_{i_0 \sim j_0} = S^s_{i_0 \sim j_0} - S^c_{i_0 \sim j_0}$, the difference between sequence-dependent and sequence-independent entropy of the fragment whose first residue's position is i_0 and last residue's j_0 in the given sequence. Since $i_0 \leq j_0$, the entropic landscape is a triangular matrix of which each component is $\Delta S_{i_0 \sim j_0}$. a fragment's Sequence-dependent entropy $S^s_{i_0 \sim j_0}$ is the sum of residue-pair-wise entropy S^s_{ij} over all possible combinations of $i_0 j_0$ but $i_0 \leq j_0$ in the sequence. So too is sequence-independent entropy. Thus, the entropic landscape of given protein sequence is calculated as follows,

$$\begin{aligned}
\Delta S_{i_0 \sim j_0} &= S^s_{i_0 \sim j_0} - S^c_{i_0 \sim j_0} \\
&= \sum_{i_0 \leq i \leq j \leq j_0} \{S^s_{ij} - S^c_{ij}\} \\
&= \sum_{i_0 \leq i \leq j \leq j_0} \sum_{uv} \begin{pmatrix} S^{ax}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u = H, N, H^\alpha, C^\alpha, v = C, O \\ S^{ab}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u, v = H, N, H^\alpha, C^\alpha \\ S^{xb}_{uv}(j-i) - S^{xx}_{uv}(j-i) & \cdots u = C, O, v = H, N, H^\alpha, C^\alpha \end{pmatrix},
\end{aligned} \tag{29}$$

where a is the residue type of i-th, and b is the that of j-th residue in the given protein sequence.

In the previous subsection, entropies for dihedral angles ϕ, ψ, ω were introduced, and to make use of them and for the sake of accuracy, the calculation of $S^s_{ii}, S^c_{ii}, S^s_{i,i+1}, S^c_{i,i+1}$ should be slightly modified as below,

$$S^c_{ii} = S^x_{\phi\psi} - S^x_{\phi} - S^x_{\psi}$$

$$S^s_{ii} = S^a_{\phi\psi} - S^a_{\phi} - S^a_{\psi} \tag{30}$$

$$S^c_{i,i+1} = S^{ab}_{\psi\omega\phi} \tag{31}$$

$$
\begin{aligned}
& + \; S^{xx}_{HH}(1) + S^{xx}_{HN}(1) + S^{xx}_{HH^{\alpha}}(1) + S^{xx}_{HC^{\alpha}}(1) \\
& + \; S^{xx}_{HC}(1) + S^{xx}_{HO}(1) \\
& + \; S^{xx}_{NO}(1) + S^{xx}_{C^{\alpha}O}(1) + S^{xx}_{H^{\alpha}O}(1)
\end{aligned}
$$

$$
\begin{aligned}
S^{s}_{i,i+1} \; = \; & S^{ab}_{\psi\omega\phi} \\
& + \; S^{ab}_{HH}(1) + S^{ab}_{HN}(1) + S^{ab}_{HH^{\alpha}}(1) + S^{ab}_{HC^{\alpha}}(1) \\
& + \; S^{ax}_{HC}(1) + S^{ax}_{HO}(1) \\
& + \; S^{ax}_{NO}(1) + S^{ax}_{C^{\alpha}O} + S^{ax}_{H^{\alpha}O}(1)
\end{aligned}
\tag{32}
$$

5 Appendix

The list of protein chains used in the statistical analysis is below,

1aym1, 1bev1, 2bpa1, 1bvp1, 1cov1, 1d4m1, 1ev11, 1g291, 1g6q1, 2mev1,
1pvc1, 1rvv1, 1sva1, 1tme1, 1tph1, 1bmv2, 1ilr2, 1ryp2, 1qqp3, 1cd34,
1lmb4, 12asA, 256bA, 1a0aA, 1a0hA, 1a0lA, 1a12A, 1a25A, 1a2oA, 1a2pA,
1a2zA, 1a34A, 1a3aA, 2a3dA, 7a3hA, 1a49A, 1a4mA, 1a4yA, 1a5dA, 1a5iA,
1a65A, 1a6cA, 1a6dA, 1a6yA, 1a73A, 1a79A, 1a81A, 1a88A, 1a8rA, 1a8uA,
2a8vA, 1a99A, 1a9nA, 1aa7A, 1aapA, 1aayA, 1abrA, 1adeA, 1adjA, 2ae2A,
1afrA, 1afsA, 1agjA, 1ahjA, 1ahsA, 1ahuA, 1aihA, 1ajsA, 1akhA, 1alvA,
2alcA, 1an4A, 9antA, 1aocA, 1aorA, 1aoxA, 1aozA, 1apxA, 1aq0A, 1atlA,
1au1A, 1auoA, 1auuA, 1auwA, 1avaA, 1avpA, 1avqA, 1awcA, 1awqA, 1ay9A,
1ayaA, 1ayfA, 1ayoA, 1ayrA, 2ay1A, 1azpA, 1azsA, 1azzA, 1b00A, 1b07A,
1b08A, 1b09A, 1b0pA, 1b0xA, 1b16A, 1b1bA, 1b22A, 1b2pA, 1b34A, 1b3aA,
1b3tA, 1b3uA, 1b43A, 1b4aA, 1b4kA, 1b5eA, 1b5tA, 1b63A, 1b66A, 1b6tA,
1b6vA, 1b71A, 1b74A, 1b77A, 1b78A, 1b7fA, 1b7yA, 1b87A, 1b8aA, 1b8dA,
1b8mA, 1b8oA, 1b8tA, 1b8wA, 1b93A, 1b94A, 1b9hA, 1b9lA, 1b9oA, 1b9rA,
1b9wA, 1b9zA, 1baiA, 1bazA, 1bbpA, 2bbvA, 1bd0A, 1bd3A, 1be3A, 1be9A,
1bebA, 1befA, 1behA, 1bfrA, 1bftA, 1bgvA, 1bh9A, 1bhdA, 1bhgA, 1bhtA,
1bihA, 1bjaA, 1bjfA, 1bjyA, 1bk7A, 1bkcA, 1bkrA, 1bl0A, 1blxA, 1bm9A,
1bmtA, 3bmpA, 1bn7A, 1bndA, 1bouA, 2bopA, 1bpyA, 1bqsA, 1bqyA, 1bs0A,
1bs2A, 1bs4A, 1bsmA, 1bsrA, 2btvA, 1bu1A, 1bu7A, 1bucA, 1bueA, 1bunA,
1buoA, 1bupA, 1bvoA, 1bvzA, 1bwdA, 1bx4A, 1bxdA, 1bxoA, 1bxrA, 1bxyA,
1by1A, 1byqA, 1byrA, 1bz0A, 1bz4A, 1bzkA, 1bzsA, 1bzyA, 1c01A, 1c02A,
1c0nA, 1c0pA, 1c16A, 1c1dA, 1c1kA, 1c1lA, 1c1yA, 1c1zA, 1c20A, 1c24A,
1c2aA, 1c39A, 1c3gA, 1c3mA, 1c3pA, 1c3yA, 1c44A, 1c4qA, 1c4xA, 1c4zA,
1c55A, 1c5eA, 1c5kA, 1c75A, 1c7jA, 1c7nA, 1c7qA, 1c7sA, 1c7uA, 1c7vA,
1c8dA, 1c8kA, 1c8uA, 1c8zA, 1c96A, 1c9hA, 1c9oA, 1cb8A, 1cb9A, 2cblA,
1cc8A, 1ccvA, 1ccwA, 1cczA, 1cd1A, 1cd9A, 1cdcA, 1cdtA, 1cdzA, 1ceuA,
1cf5A, 1cf9A, 1cfzA, 1cg2A, 1cg7A, 1cghA, 1ch4A, 1chmA, 1cipA, 1citA,
1cizA, 1cjcA, 1cjwA, 1ck2A, 1cklA, 1ckmA, 1cktA, 1cl2A, 1cl8A, 1cleA,
1cliA, 1clpA, 1cm0A, 1cm5A, 1cmbA, 1cmiA, 1cmlA, 1cmoA, 1cmzA, 2cmkA,
1cnzA, 1co4A, 1co6A, 1couA, 1cozA, 1cp2A, 1cpzA, 1cq3A, 1cqdA, 1cqmA,
1cqqA, 1cqxA, 1cqyA, 1crxA, 1cs6A, 1ctqA, 1cuoA, 1cvrA, 1cvuA, 1cw5A,
1cwvA, 1cwxA, 1cx1A, 1cxwA, 1cxyA, 1cy5A, 1cydA, 1cz1A, 1cz4A, 1cz9A,
1czfA, 1czpA, 1cztA, 1czyA, 1d0bA, 1d0dA, 1d0qA, 1d0vA, 1d1dA, 1d1qA,
1d1rA, 1d2dA, 1d2fA, 1d2kA, 1d2nA, 1d2oA, 1d2rA, 1d3vA, 1d4aA, 1d4bA,
1d4fA, 1d4oA, 1d4tA, 1d4vA, 1d5tA, 1d5vA, 1d66A, 1d6gA, 1d7cA, 1d7qA,
1d7uA, 1d7yA, 1d8bA, 1d8dA, 1d8jA, 1d8lA, 1d8uA, 1d8zA, 1d9cA, 1d9nA,
1d9sA, 1d9tA, 1dabA, 3daaA, 1dbfA, 1dbgA, 1dc7A, 1dciA, 1dcqA, 1dctA,
1dczA, 1dd3A, 1dd5A, 1ddgA, 1ddjA, 1ddlA, 1ddvA, 1de3A, 1de6A, 1delA,
1deoA, 1derA, 1deuA, 1devA, 1dgfA, 1dgnA, 1dgsA, 1dguA, 1dgwA, 1dhpA,
1di2A, 1diiA, 1dixA, 1dj0A, 1dj7A, 1dj8A, 1djnA, 1dk0A, 1dk5A, 1dk8A,

1dkqA, 1dkzA, 1dl6A, 1dljA, 1dluA, 1dlwA, 2dldA, 1dm5A, 1dm9A, 1dmhA,
1dmnA, 1dmtA, 1dmuA, 1dnyA, 1dosA, 1dovA, 1dozA, 2dorA, 1dp0A, 1dp4A,
1dpfA, 1dpgA, 1dptA, 1dpuA, 1dq3A, 1dqaA, 1dqeA, 1dqgA, 1dqiA, 1dqpA,
1dqtA, 1dqzA, 2drpA, 1ds1A, 1ds9A, 1dsyA, 1dszA, 1dt4A, 1dtjA, 1du2A,
1du5A, 1dujA, 1duzA, 1dv0A, 1dv2A, 1dv8A, 1dvjA, 1dvnA, 1dvoA, 1dw0A,
1dwmA, 1dwnA, 1dxeA, 1dxsA, 1dy2A, 1dykA, 1dymA, 1dynA, 1dysA, 1dytA,
1dz3A, 1dzfA, 1dzkA, 1dzlA, 1dzoA, 1e01A, 1e0bA, 1e15A, 1e17A, 1e19A,
1e1dA, 1e2aA, 1e2tA, 1e2wA, 1e30A, 1e39A, 1e3jA, 1e3yA, 1e53A, 1e5dA,
1e5kA, 1e5mA, 1e5qA, 1e68A, 1e6iA, 1e6uA, 1e6vA, 1e79A, 1e7kA, 1e85A,
1e87A, 1e88A, 1e8mA, 1e8oA, 1e8rA, 1e91A, 1e9tA, 1e9xA, 1ea5A, 1ebfA,
1ec8A, 1eceA, 1ecrA, 3ecaA, 1ed0A, 1ed1A, 1ed5A, 1ed8A, 1edhA, 1edoA,
1edqA, 1edxA, 1edyA, 1ee4A, 1ee8A, 1eejA, 1eeoA, 1eerA, 1eexA, 1ef4A,
1ef5A, 1ef9A, 1efvA, 1efyA, 1eg3A, 1eg7A, 1eg9A, 1egjA, 1egxA, 1eh9A,
1ehdA, 1ehiA, 1ehjA, 1ehxA, 1ei1A, 1ei5A, 1ei7A, 1ei9A, 1eigA, 1eijA,
1eikA, 1eiwA, 1eiyA, 2eifA, 3eipA, 1ej0A, 1ej3A, 1ej5A, 1ej8A, 1ejfA,
1ejnA, 1ek6A, 1ekgA, 1ektA, 1ekxA, 1ekzA, 1el0A, 1el5A, 1el6A, 1elkA,
1eluA, 1elwA, 1em9A, 1emvA, 1en7A, 1enfA, 1enwA, 1eo0A, 1eo1A, 1eokA,
1ep0A, 1epxA, 1eq6A, 1eq9A, 1eqfA, 1eqkA, 1eqoA, 1erxA, 3eraA, 3ertA,
1es5A, 1esjA, 1esrA, 1eteA, 1etpA, 1eu3A, 1eu4A, 1euaA, 1eucA, 1euhA,
1euoA, 1eupA, 1euvA, 4eugA, 1ev0A, 1evhA, 1ew0A, 1ew4A, 1ew6A, 1ewfA,
1ewwA, 1ex1A, 1ex2A, 1exbA, 1exjA, 1exkA, 1exmA, 1exrA, 1extA, 1eyhA,
1eylA, 1eyqA, 1ez3A, 1ezgA, 1ezrA, 1ezwA, 1f0kA, 1f0nA, 1f0yA, 1f0zA,
1f1zA, 1f21A, 1f24A, 1f2aA, 1f2dA, 1f2fA, 1f39A, 1f3hA, 1f3mA, 1f3uA,
1f3vA, 1f41A, 1f52A, 1f53A, 1f5vA, 1f5wA, 1f5xA, 1f5yA, 1f60A, 1f62A,
1f6aA, 1f6wA, 1f74A, 1f7dA, 1f7sA, 1f81A, 1f83A, 1f8mA, 1f8yA, 1f94A,
1faeA, 1fafA, 1faoA, 1fbvA, 1fc3A, 1fc9A, 1fcdA, 1fczA, 2fcbA, 1fd3A,
1fehA, 1fezA, 1ffyA, 1fgjA, 1fgxA, 1fh6A, 1fhgA, 1fhoA, 1fipA, 1fiqA,
1fiuA, 2fibA, 1fj2A, 1fj8A, 1fjkA, 1fjmA, 1fjsA, 1fkcA, 1fknA, 1fl0A,
1flcA, 1flgA, 1fliA, 1flkA, 1flmA, 1fm9A, 1fmcA, 2fmtA, 1fnsA, 1fnuA,
1fonA, 1fp0A, 1fp3A, 1fpoA, 1fpwA, 1fq3A, 1fqvA, 1fr3A, 1fr7A, 1frsA,
3fruA, 1fs1A, 1fs7A, 1ft5A, 1ftpA, 1ftrA, 1fugA, 1fuiA, 1fujA, 1furA,
1fv9A, 1fvkA, 1fwkA, 1fx4A, 1fxjA, 1fxlA, 1fxrA, 1fxyA, 1fy7A, 1fybA,
1fzqA, 1g17A, 1g25A, 1g2bA, 1g31A, 1g40A, 1g47A, 1g51A, 1g5tA, 1g61A,
1g66A, 1g6gA, 1g6sA, 1g71A, 1g72A, 1g73A, 1g79A, 1g7dA, 1g7eA, 1g8kA,
1g8qA, 1g8xA, 1g99A, 1ga6A, 1gakA, 1gaxA, 2gatA, 1gc7A, 1gcoA, 1gcuA,
1gd0A, 1gdhA, 1gdtA, 1gduA, 1gg3A, 1gh8A, 1gh9A, 1ghhA, 2gliA, 2gmfA,
1gp8A, 1gpeA, 2gsaA, 6gsvA, 1gtxA, 1guqA, 1gyfA, 1h8cA, 1hcnA, 1hcrA,
1hd0A, 1hdcA, 2hdcA, 2hddA, 1he1A, 1hf8A, 1hfgA, 1hg7A, 2hgsA, 2hhmA,
1hi7A, 1hjrA, 1hloA, 2hlcA, 1hm6A, 1hmdA, 1hnjA, 2hpaA, 5hpgA, 1hq3A,
1hqcA, 1hqmA, 1hryA, 2hrvA, 1hslA, 1hstA, 1hulA, 1huqA, 1huuA, 1hw1A,
1hw5A, 1hwmA, 1hx2A, 1hykA, 1hyuA, 1i02A, 1i06A, 1i0hA, 1i0vA, 1i11A,
1idaA, 1ihbA, 1ihfA, 1iibA, 1iieA, 1io7A, 1irsA, 1iscA, 1ithA, 1ivhA,
1ivyA, 1jbaA, 1jetA, 1jfrA, 1jhgA, 2jhbA, 1jlyA, 1jmcA, 1joyA, 1kbaA,

4kbpA, 1kdxA, 1kevA, 1khmA, 1klaA, 1knyA, 1kobA, 1kp6A, 1kptA, 1ksaA,
1kwaA, 1kxiA, 3ladA, 1lbeA, 1lckA, 1lehA, 2lefA, 1lfaA, 1lfdA, 2lisA,
1lklA, 1lktA, 1lopA, 1lpbA, 3lriA, 2masA, 1mdyA, 3mddA, 1meeA, 1memA,
1mfmA, 1mgsA, 1mgtA, 1mhdA, 6mhtA, 1mkaA, 1mkcA, 1mknA, 2mllA, 1mmsA,
1mnmA, 1mntA, 1molA, 1mpyA, 1mspA, 2msbA, 2mssA, 1mugA, 1mwpA, 1n45A,
1n5wA, 1n72A, 2nacA, 2napA, 1nbaA, 1nbcA, 1nciA, 1nfkA, 2ngrA, 1nmtA,
2nmbA, 1noyA, 1np1A, 1nscA, 2nsyA, 1ntcA, 1nubA, 1nzyA, 2oatA, 1ofgA,
1olgA, 1om2A, 1oneA, 1onrA, 1opbA, 1opmA, 1ordA, 1otcA, 1otfA, 1otgA,
1ounA, 1pa2A, 6paxA, 1pbwA, 1pcfA, 2pcfA, 3pcgA, 1pdkA, 1pdxA, 3pdzA,
1pfkA, 4pgaA, 1phnA, 1picA, 1pjcA, 2pldA, 1pmpA, 3pmgA, 1poiA, 1poxA,
2polA, 2pooA, 1psdA, 1pszA, 2pspA, 2pvaA, 3pviA, 1pyiA, 1qamA, 1qauA,
1qavA, 1qazA, 1qb0A, 1qb7A, 1qbhA, 1qc5A, 1qc7A, 1qciA, 1qckA, 1qcxA,
1qd9A, 1qddA, 1qdvA, 1qexA, 1qf6A, 1qf9A, 1qfeA, 1qfpA, 1qftA, 1qgiA,
1qgoA, 1qgxA, 1qh4A, 1qh5A, 1qhfA, 1qhkA, 1qhoA, 1qhvA, 1qhxA, 1qi7A,
1qi9A, 1qibA, 1qiuA, 1qj4A, 1qjtA, 1qjvA, 1qk8A, 1qk9A, 1qkjA, 1qkkA,
1qklA, 1qksA, 1ql0A, 1qlaA, 1qlmA, 1qlpA, 1qlwA, 1qm5A, 1qm9A, 1qmgA,
1qmhA, 1qmqA, 1qmuA, 1qmyA, 1qniA, 1qnjA, 1qnkA, 1qnrA, 1qnxA, 1qorA,
1qp2A, 1qpoA, 1qpzA, 1qq2A, 1qq4A, 1qq5A, 1qqfA, 1qqhA, 1qqiA, 1qqlA,
1qqsA, 1qqvA, 1qr0A, 1qr2A, 1qreA, 1qrrA, 1qryA, 1qs1A, 1qsaA, 1qsdA,
1qstA, 1qtfA, 1qtoA, 1qtrA, 1qtsA, 1qtwA, 1qu6A, 1quuA, 1qvaA, 1qvbA,
1r1rA, 1r2aA, 1ravA, 2ramA, 3rabA, 1rblA, 1rdzA, 1reqA, 1rfnA, 1rgeA,
1rl6A, 1rmvA, 1rnfA, 1rodA, 1rpjA, 1rpxA, 3rpbA, 1rthA, 1sacA, 2scpA,
3sdhA, 1seiA, 1shaA, 1shcA, 1shfA, 1shsA, 1smlA, 1sppA, 2spcA, 2sqcA,
1srrA, 1srsA, 1sryA, 1stmA, 3stdA, 1svpA, 1swuA, 1t1dA, 1tbaA, 1tcoA,
1tf4A, 1tfaA, 1tgoA, 1tgxA, 3thiA, 1tkiA, 1tl2A, 1tlfA, 4tmkA, 2tnfA,
1toaA, 1tpkA, 1trkA, 1tsrA, 1tx4A, 1tyfA, 1u2fA, 4ubpA, 1ugiA, 1unkA,
2up1A, 1urnA, 1uroA, 1uteA, 1vcaA, 1vcbA, 1vfrA, 1vfyA, 1vhrA, 2viuA,
1vrkA, 1vsrA, 1wapA, 1wdnA, 1wjbA, 1wtuA, 1wwcA, 1xbrA, 1xgsA, 1xnaA,
1xvaA, 1xyzA, 1yacA, 1yagA, 1ybvA, 1ycqA, 1yrgA, 1ytbA, 1ytiA, 1yuiA,
1zeiA, 1zpdA, 830cB, 1a28B, 1a4iB, 1a6bB, 1a6jB, 1a95B, 1a9nB, 1abrB,
1afwB, 2ahjB, 1aisB, 1allB, 1am9B, 1aoeB, 1aohB, 1aojB, 2arcB, 1atzB,
1auiB, 1auyB, 1avbB, 1avsB, 1awcB, 1ay7B, 2azoB, 1b4fB, 1b5qB, 1b67B,
1b6cB, 1b6sB, 1b72B, 1b79B, 1b7yB, 1b8iB, 1babB, 3bamB, 1bcmB, 1bcpB,
1bdmB, 1bdyB, 1be3B, 1bf6B, 1bh9B, 1bi2B, 1bi7B, 1bkpB, 1bouB, 1bpoB,
1bquB, 1bteB, 1btkB, 1buhB, 1bunB, 1byfB, 1c1yB, 1c30B, 1c9kB, 1c9pB,
2cb5B, 1ccwB, 7ceiB, 1cf7B, 1cjxB, 1cksB, 1cpcB, 2cpgB, 1cruB, 1cwpB,
1cxzB, 1d02B, 1d09B, 1d4vB, 1d8dB, 1dbwB, 1dceB, 1devB, 1dfoB, 1dgtB,
1dj7B, 1dl5B, 1dokB, 1ds6B, 1dszB, 1dtdB, 1dtwB, 1e0jB, 1e3uB, 1e42B,
1e7aB, 1ecfB, 1ecmB, 1ee2B, 1eerB, 1eesB, 1eexB, 1eg9B, 1egaB, 1egiB,
1ej1B, 1ek1B, 1ekbB, 1emvB, 1eo6B, 1eo9B, 1ep3B, 1epfB, 1euvB, 1evlB,
1exzB, 1eyvB, 1f0jB, 1f37B, 1f4qB, 1f51B, 1f5mB, 1f5qB, 1f60B, 1f8rB,
3fapB, 1fbqB, 1fcjB, 1fecB, 1fhwB, 1fiqB, 1fjeB, 1fjgB, 1fo1B, 1fr1B,
1fvaB, 1fx3B, 1fxoB, 1fzrB, 1g1eB, 1g6uB, 1g72B, 1g8kB, 1gdoB, 1gg1B,

1gnkB, 1gotB, 1guxB, 2gwxB, 1hcqB, 1hq3B, 1hxrB, 1hyoB, 1i4gB, 1iakB,
1ibrB, 1if1B, 1itbB, 1jckB, 1jkmB, 1jswB, 1latB, 1lucB, 1mpgB, 1mroB,
1nddB, 2ngrB, 2nllB, 2occB, 1otcB, 1p35B, 1pdkB, 1plfB, 1poiB, 8prkB,
1psrB, 1qb2B, 1qbeB, 1qd1B, 1qdlB, 1qfhB, 1qg3B, 1qg7B, 1qgtB, 1qipB,
1qj5B, 1qjbB, 1qlaB, 1qn2B, 1qopB, 1qrjB, 1qrvB, 1qs8B, 1qvcB, 3ranB,
1rfnB, 1rrpB, 1rypB, 1scjB, 2scuB, 1semB, 1sltB, 1smtB, 1smvB, 1sppB,
1tc1B, 1theB, 2tpsB, 1tr1B, 1tvxB, 4ubpB, 1ueaB, 1vpsB, 1wdcB, 1xikB,
1ycsB, 1ytfB, 1zagB, 1zymB, 2afgC, 1ak4C, 1avaC, 1b35C, 1bbzC, 1bc8C,
1c04C, 1c0mC, 1c8nC, 1cvsC, 1d2zC, 1dazC, 1dceC, 1dcoC, 1de4C, 1diiC,
1duxC, 1dxrC, 1dxxC, 1eaiC, 1ebdC, 1efaC, 1f15C, 1f2nC, 1f2rC, 1fcdC,
1fiqC, 1fjgC, 1fjlC, 1fsgC, 1fvuC, 1fxkC, 1g3kC, 2hapC, 1hx6C, 1jsuC,
2kauC, 1meyC, 1mroC, 1mseC, 1n5wC, 1npoC, 1pdnC, 1qb3C, 1qbjC, 1qh8C,
1qjzC, 1qqrC, 2rslC, 4sbvC, 1tc3C, 1tfxC, 1ubdC, 1vcbC, 1wdcC, 1xxaC,
1ytfC, 1zmeC, 1agqD, 1bcpD, 1be3D, 1d2zD, 1dpsD, 1dubD, 1eayD, 1ej6D,
1f2lD, 1fjgD, 1fkaD, 1fm0D, 1fm9D, 1g6wD, 1g73D, 1h7wD, 4lipD, 1mnmD,
1mtyD, 1qo3D, 1qsmD, 1quqD, 1tf6D, 1thfD, 1agrE, 1apmE, 1cseE, 1cvjE,
1cziE, 1d3bE, 1e08E, 1epmE, 1exbE, 1fqjE, 1gecE, 1gegE, 1hagE, 2occE,
1pekE, 1qo0E, 1sgpE, 6tmnE, 1zfpE, 1aotF, 1bcpF, 1bvyF, 3chbF, 1f9aF,
2occF, 2pjrF, 1pueF, 1c4rG, 1deeG, 1eexG, 1ekjG, 1fjgG, 1g9nG, 1gotG,
1hq3G, 1mtyG, 1qsgG, 1yagG, 2bbkH, 1bi6H, 1c5cH, 1d0iH, 1danH, 1djrH,
1dlfH, 1duvH, 1dxrH, 1hq3H, 2irfH, 2occH, 1an1I, 1avgI, 1cewI, 1cseI,
1dx5I, 1e0fI, 1e79I, 1ejlI, 1f00I, 1f2rI, 1fjgI, 1fleI, 1hrtI, 1icfI,
1jrhI, 8rucI, 1sgpI, 4sgbI, 3sicI, 1smpI, 1stfI, 1tgsI, 1ypcI, 1fjgJ,
1fjgK, 2bbkL, 1d3bL, 1fakL, 1fjgL, 1fjsL, 1h2rL, 1hfeL, 1bqqM, 1d7pM,
1e6qM, 1fjgM, 1pprM, 1a02N, 1b33N, 1czaN, 1efdN, 1fjgN, 1aonO, 1gd1O,
1ospO, 1dfuP, 1dp7P, 1e4cP, 1fjgP, 1i1iP, 1kapP, 1sknP, 3ygsP, 1fjgQ,
1fjgR, 1hiwR, 1tbrR, 1tocR, 1xdtR, 1fjgS, 1h2rS, 1hfeS, 1f02T, 1fjgT,
1fltW, 1hr0W, 1fltX, 1regX, 1wwbX, 1wwwX, 1bryY, 1bu6Z,

References

[1] Sippl, M.J. Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structure in Globular Proteins. *J. Mol. Biol.,* **213**,859-883.(1990)

[2] Riddle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I., & Baker, D. Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.,* **6**, 1016-1024.(1999)

[3] Tanaka, S., Kojima, S., & Tamura, A. Mapping the position of the transition state in the folding of small α/β protein, POIA1. *Chem. Phys.,***307**, 233-242.(2004)

[4] Morimoto, S., & Tamura, A. Key Elements for Protein Foldability Revealed by a Combinatorial Approach among Similarly Folded Distantly Related Proteins *Biochem.,* **43**, 6596-6605.(2004)