The Entropic landscape of proteins revealing protein folding mechanism

Kentaro ONIZUKA

PsiPhiFoldings Co. Ltd.

It has long since been a mystery why most proteins fold within a flash of time into particular structures out of astronomically large numbers of possible conformations. Even more confusing is that protein folding in vivo is played out in rich solution containing various organic and non-organic, big and small molecules and ions which would potentially bind the protein molecules and prevent them folding. A possible answer to these mysteries might be, "Nature have favoured such proteins that quickly fold in rich solution through natural selection." Then what mechanism of folding has been favoured?

Here I show how to decipher protein sequences to reveal the folding mechanism. The entropic landscape of a protein sequence tells which region of the sequence sets out to fold first which next and last. Each step of the folding procedure is *programmed* in the sequence. This make it clear why proteins fold quickly and escape from surrounding molecules and ions. The folding pathways represented by the entropic landscape agree with the pathways experimentally proposed. Besides, the simulation of protein folding scheduled by the entropic landscape generates native-like conformations, where the lower the entropy of a sequential region is the earlier its conformation is optimized in terms of energy minimization. The attempt to simulate protein folding gives further insights into the folding mechanism.

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. In general complex systems have a large number of microstates thus 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 fold. 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 corresponds to 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 optimal stable conformation for good, 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 must 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.

This notion that the regional entropies of a protein determines the folding speed of the region would explain why proteins are able to fold in rich solution with various organic or non-organic molecules and ions. If a protein's productivity declined due to the new environment with new ions and molecules that the protein had never experienced previously through evolution, natural selection would favour such mutations that would accelerate or decelerate the folding speed of some particular regions to escape from disturbing new ions and molecules sacrificing the total stability of the protein in some cases, as long as the protein could work normally in the new environment. A folding polypeptide could even make use of surrounding molecules or ions in such a way that the binding or approaching of a particular molecule to a particular region of the polypeptide would trigger the folding of the regions as the folding starter. Whether the protein uses this kind of switching mechanism or not depends on one or two residues in the region, a mutation on which could drastically change the entropy of that region.

Hence, by comparing the entropies of fragments of a protein 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 or even to define, particularly when protein folding in rich solution is assumed. However, if we separate the whole entropy of a fragment into two parts, 1) conformational (sequence-independent) entropy S^C and 2) sequence-dependent entropy S^S, then conformational entropy S^C is constant for all fragments of the same length. As long as we compare the entropy of fragments of the same length, S^S can be ignored. Sequence-dependent entropy S^S is directly calculated from the knowledge-based potentials of mean force[1,6] which are compiled from the set of known protein structures, which are resultant conformations that polypeptides reached in rich solution.

Entropic Landscape

The entropic landscape of a protein is the matrix of fragment entropies S^S_{ij} which are the sequence-dependent entropy S^S of the fragment of length $j-i+1$ from i-th through j-th residue in the entire protein sequence. For the sake of easy interpretation of entropic landscape, we introduce $\bar{S}_{mk} = S^S_{ij} - < S^S_k >$, where $m = (i+j)/2$, $k = j-i$, and $< S^S_k >$ is the mean S^S of $k+1$-residue long fragments. By plotting the $\bar{S}_{mk}$ with respect to m, we can see which region of what size (or length $k+1$) around which position (m) in the entire sequence has lower or higher entropy than the mean entropy of the fragments of length $k+1$.

Fig 1 is the entropic landscape of protein L (1K50A). Each curve shows how $\bar{S}_{mk}$ changes with respect to m, the position of the region in the entire sequence (starting from 0-th residue

in the graph). Here six curves are shown, each of which represents $\bar{S_m}k$ for $k = 0, 1, 2, 4, 8, 16$ respectively The entropic landscape has more information, though the graph would be very loud and confusing when curves or all possible k are shown. The curve for $k = 16$ suggests that this protein sets out to fold from C-term side rather than N-term side, because $\bar{S}_{m16}$ is lower in C-term-side half than in N-term-side half. And $\bar{S}_{m16}$'s local minimum around $m = 46$ suggests that the formation of helix 1 and strand 3 proceeds the coupling of strand 3 and strand 4 to form a sheet. This is even clear if we look at the curve for $k = 8$. $\bar{S}_{m8}$ is the minimum around $m = 43$, the region between helix 1 and strand 3. The curve for $k = 4$ clearly shows where the regions between regular structures (α and β) are located. $\bar{S}_{m4}$ is locally minimum at $m = 14, 22, 42, 53$, which correspond to 1) the first hairpin between strand 1 and 2, 2) the region between strand 2 and helix 1, 3) between helix 1 and strand 3, and 4) the hairpin between strand 3 and 4, respectively.

$\bar{S}_{m0}$ represents the entropy of a single residue calculated from the residue's statistical preference to $\omega\phi\psi$ angles. Glycine has the lowest entropy among all common twenty residue types. This might sound strange because Glycine has the most flexible backbone due to the lack of side-chain, and in general flexibility strongly suggests diversity and high entropy. However it is also the case that Glycine quite often forms very unusual conformation because of the same reason. The nature of the sequence-dependent entropy is that the more frequently the residue (pair) forms unusual configuration (or conformation), the lower the entropy of the residue (pair) is. Hence, Glycine's low entropy is not surprising. Other low entropy residues are Aspartate, Proline, Isoleucine, and Valine. The highest entropy residue is Alanin, and other high entropy ones are Arginine, Phenylalanine, Serine, and Tyrocine.

The regions containing low entropy residues usually have low entropy for the length, and those regions quite often correspond to hairpin turn, loop, or irregular structures between regular structures like α helix or β strand particularly when Glycines and Prolines are involved. On the contrary, regular structures (α, β) usually occur around relatively high entropy regions. Possible interpretation for this is that those regular structures are stable not because of the sequence's strong preference to those regular structures but due to the hydrogen bonds between backbone atoms (i.e. CO-HN). In other words, any sequence could form regular structures like α and β unless the sequence has a strong preference to irregular structures. On the other hand, to be stable, irregular structures need supports from low entropy residues.

Folding Pathway

Let the entropic landscape of protein sequences talk about the folding mechanism of three proteins.

Fig 2 shows the entropic landscape of SH3 domain of ABL tyrosine kinase (1ABQ in PDB). The entropy at $k = 8$ is the lowest around the hairpin turn between strand 5 and 6. And the region around the other hairpin turn (between strand 4 and 5) has the second lowest entropy. Although the entropy around strand 5 is high, the formation of the two hairpin turns must urge the formation of the anti-parallel sheet of strand 4,5, and 6, among which coupling of strand 5 and 6 proceeds the coupling of strand 4 and 5. This assumed scenario is almost identical to that proposed by experiments[3], where sheet of strands 4,5, and 6 is the predominant region ordered in all three proteins with SH3 domain.

The second example is Fig 3 which shows the entropic landscape of a small protein *Pleurotus ostreatus* proteinase A inhibitor 1 (POIA1, 1ITP in PDB). The lowest entropy region is located around the hairpin turn between strand 2 and 3. The region around the end of helix 1 and the consecutive turn also has low entropy. Consequently the formation of the sheet consisting of the strand 2 and 3 and the packing of the sheet with helix 1 should take place at an early stage of folding. These early events should induce strand 1 to meet the sheet. The formation of helix 2 and strand 4 would take place independently. The final native structure is likely to form when helix 2 meets helix 1 and this urge coupling of strand 4 and 1. This assumption of pathway suggested by entropic landscape almost completely agrees with the folding process proposed by experiments[4].

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[5]. 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 unless some triggering events take place. This hairpin turn region turns out to be the very locale which binds to the two-helix bundle of subtilisin BPN. Thus the folding of the propeptide must be triggered by the binding to the other subunit, and then the hairpin turn is formed, which induces the folding of other regions. The overall entropic landscape of the propeptide is, in short, similar to that of POIA1 except for the hairpin

region, as shown in Fig4. 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. However, the assumed folding pathway does not contradict the true folding pathway.

Folding simulation

As a preliminary study, a crude but very essential folding simulation system was devised, whose scheduling of folding is perfectly faithful to the entropic landscape of given protein sequence, where the lower the entropy of the region is and the shorter the region is, the earlier the conformation of the region is optimized in terms of energy-minimization.

The folding simulation of the partial structure of POIA1 is one of the successful cases in this study as is shown in Fig 5. The 17-residue long fragment that forms the hairpin sheet coupling strand 2 and 3 in the native structure is cut out from the sequence of POIA1. The lowest entropy region at $k = 1$ is Pro-Gly at the middle of the cut-out fragment.

After the optimization at level 0 ($k = 0$) is complete, most regions are relatively stretched, except for Pro-Gly site, where Proline's ϕ is fixed around $-60°$. Quite early on, Pro-Gly site begins to form a tight hairpin turn, and gradually the neighbouring regions begin to make hydrogen bonds to form a sheet. This actually suggests that the formation of the hairpin turn which occurs at Pro-Gly site induces the hydrogen-bond coupling of neighbouring regions to form a sheet. Note that the residue pairs making hydrogen-bonds between strand 2 and 3 are not always favour strand formation and the coupling is statistically not always favoured. The formation of the sheet in

this case is, thus, not by the sequence's propensity to sheet, but by being induced by hairpin turn formation at Pro-Gly site.

Another case study is the folding simulation of the 31-residue long N-term region of 1QKKA. This region forms a strand-helix-strand conformation and strand 1 and 2 couple together to form a parallel sheet in the native structure of 1QKKA. The entropic landscape of this region for $k = 1$ shows that the entropy is the lowest globally around the region between helix and strand 2, and the lowest locally around the region between strand 1 and helix.

Through the simulation at low levels ($k < 6$), the helix formation is very slow, and a turn-like structure is generated around globally lowest entropy site which corresponds to the region between the helix and strand 2. The helix formation starts from N-term side around the locally lowest entropy site corresponding to the regions between strand 1 and the helix. When the helix extension reaches tight turn, the helix stops extending because the hydrogen-bonding donors and acceptors required for the helix formation is already consumed by hairpin-turn.

Insights into Protein Folding mechanism

The study of entropic landscape analysis and the folding simulations seem to have partially revealed the folding mechanism of proteins. First, the folding pathway of a protein interpreted from the entropic landscape agree with the pathway proposed by experiments. And also the folding simulation works well only when the scheduling of which region is optimized first, next, and last is perfectly reflecting the entropies of the regions, the entropic landscape. These strongly support the notion that a polypeptide in solution sets out to fold from the low entropy regions, and the high

entropy regions linger in folding until the neighbouring low entropy regions urge them to fold. Second, this study has suggested a lot of other insights into folding mechanisms. The followings are those.

- Low entropy regions tend to form irregular structures like loops, turns or transitional conformations between α helix and β strand. By turn, regular structures like α and β occur in relatively high entropy regions. Regular structures should, thus, be induced by the low entropy regions nearby forming irregular structures.

- By default, strand-like conformations are favoured because neighbouring dipole groups (CONH) try to be alternating, where all ψ angles of backbone are set to positive. Irregular local conformations with negative ψ's hardly exist ab initio unless the consisting residues or pairs strongly favour negative ψ. And those regions which have strong preference to negative ψ are particular kind of short low entropy regions.

- When short low entropy regions folds into a hairpin turn, the neighbouring regions remaining in forming strand-like conformation are induced to couple and finally a parallel sheet is formed.

- When short low entropy regions have their two or more dipoles (CONH groups) placed in parallel (thus some ψ's are negative), the electrostatic field exerted from those dipoles in parallel induces the neighbouring regions to form helix.

- Helices could ever extend unless they are stopped by a helix-stopper region, which has low entropy and forms some turn-like structure that have already consumed hydrogen-bonding

donors or accepters before the helix extension reaches there.

- High entropy regions forming strand-like conformations by default could meet other regions forming strand-like conformation if the coupling of those regions are arranged by the sequential regions between these strand-like regions. Then formed is either a parallel or anti-parallel sheet made up with those sequentially separated regions.

- A residue's preference to a particular conformation seems to be quite naturally caused by the interaction between side-chain and CONH dipole groups at both side of the residue. Glycine has the lowest entropy at $k = 0$ due to the lack of side-chain, which allows positive ϕ angle that would cause $O - C^{\beta}$ clash for other residue types. Aspargine's second lowest entropy must be caused by its side-chain's polar group that is pretty eager to form a hydrogen bond with a backbone polar group. Proline's low entropy is due to its restricted choice of ϕ. The size of side-chain seems to be more significant to the choice of $\phi\psi$ angles when it comes to the straight residue pairs, particularly the pair has Glycine at C-term side. Glycine's flexible choice of ϕ angle allows the N-term residue to choose their most favoured ψ angle particularly when its side-chain is big.

These insights seem to be responsible with the worse-than-expected accuracy of conventional secondary structure prediction, and with often betraying fold-recognition-based protein structure prediction. The secondary structure prediction must be difficult because the core of neither α helix or β strand could have low entropy and strong desire to fold into those regular structures. The fold-recognition could be unsuccessful because most of evaluation functions to detect the structural similarity focus on regular structures and coupling preferences of residues separated

in the sequence, paying much less attention to loop and irregular structure regions. And it is also the case that sequentially homologous and or structurally homologous two proteins could have completely different entropic landscape that suggest they have completely different folding pathways.

Methods

Calculation of Entropy

The entropy S of a system is given as follows,

$$S = -k_b \sum_l P_l \ln P_l, \tag{1}$$

where P_l is the probability of the system in state l, and k_b is the Boltzmann constant. If energy E_l of state l is observable or computable, assuming the Boltzmann distribution under the temperature T, P_l is calculated as follows,

$$P_l = \exp(-E_l/k_bT)/Z, \tag{2}$$

where Z is the partition function normalizing probability P_l.

$$Z = \sum_l \exp(-E_l/k_bT). \tag{3}$$

Introducing the average (or expected) energy of the system U, the entropy S is given as follows,

$$\begin{aligned} S &= -k_b \sum_l P_l(-E_l/k_bT - \ln Z) && (4) \\ &= U/T + k_b \ln Z && (5) \end{aligned}$$

where,

$$U = \sum_l P_l E_l \tag{6}$$

Note that $-k_bT \ln Z$ is the free energy.

The sequence fragment entropy S is separated into two terms, 1) conformational (sequence-independent) entropyS^C and 2) sequential (sequence-dependent) entropy S^S.

$$S = S^C + S^S. \tag{7}$$

Conformational entropy S^C is by definition constant for all fragments of the same length, thus ignored. Sequence dependent entropy S^S is a simple sum of pair-wise entropies S_k^{ab} over all residue pairs in the fragment.

$$S^S = \sum_{i \leq j} S_{k=j-i}^{ab}, \tag{8}$$

where i and j are positions of the two residues in the fragment. Pair-wise entropy S_k^{ab} is given as a sum of probability-weighted net pair-wise energy $P_k^{ab}(l)E_k^{ab}(r_l)$ of residue-types a, b at sequential separation k, over all spatial configurations r_l of two residues. Here r_l is the representative configuration of all those rs within l-th bin. Therefore the pair-wise entropy is given as follows,

$$S_k^{ab} = -k_b \sum_l P_k^{ab}(l)(E_k^{ab}(r_l)/k_bT - \ln Z_k^{ab}). \tag{9}$$

Here $P_k^{ab}(l)$ is the pair-wise probability of two residues a, b at sequential separation k being placed at configuration r within l-th bin. Z_k^{ab} is the pair-wise partition function. Net energy $E_k^{ab}(r)$ could be replaced with knowledge-based potential of mean force, calculated from distribution density $f_k^{ab}(r)$ of residues a, b, and $f_k^{xx}(r)$ of any residue pairs [1]. Both $f_k^{ab}(r)$ and $f_k^{xx}(r)$ are compiled from the set of selected protein structures from PDB. To avoid those probability densities being zero, we need $f_k^{rand}(r)$, the distribution density compiled from artificially generated random conformations, which is always > 0.

$$E_k^{ab}(r) = -k_bT \ln \frac{f_k^{ab}(r) + mf_k^{rand}(r)}{f_k^{xx}(r) + mf_k^{rand}(r)}, \tag{10}$$

where m is a small number $0 < m < 1$. In this study, $m = 0.1$. Note that pair-wise probability $P_k^{ab}(l)$ is directly calculated from $f_k^{ab}(r)$,$f_k^{xx}(r)$, and $f_k^{rand}(r)$.

$$P_k^{ab}(l) = \frac{(f_k^{ab}(r_l) + mf_k^{rand}(r_l))/(f_k^{xx}(r_l) + mf_k^{rand}(r_l))}{Z_k^{ab}}, \tag{11}$$

And pair-wise partition function Z_k^{ab} is given as follows,

$$Z_k^{ab} = \sum_l \frac{f_k^{ab}(r_l) + mf_k^{rand}(r_l)}{f_k^{xx}(r_l) + mf_k^{rand}(r_l)}. \tag{12}$$

The net potential of mean force $E_k^{ab}(r)$ given in equation (10) statistically includes the net interaction specifically between side-chain atoms of the residue a, b, and that between a's side-chain atoms and b's backbone atoms (NH-C$^\alpha$H-CO) and vice-versa, but excludes average interactions among side-chain atoms and backbone atoms. And $E_k^{ab}(r)$ is considered to be independent of the other neighbouring residues, thus entropy S_k^{ab} is also independent, and additive.

Denoting a_i, a_j as the residue types of i-th or j-th residue respectively, fragment entropy S^S is given as follows,

$$\begin{aligned} S^S &= -k_b \sum_{i \le j} S_{k=j-i}^{ab} && (13) \\ &= -k_b \sum_{i \le j} \sum_l \frac{P_{k=j-i}^{a_i a_j}(r_l)}{Z_{k=j-i}^{a_i a_j}} \ln \frac{P_{k=j-i}^{a_i a_j}(r_l)}{Z_{k=j-i}^{a_i a_j}} && (14) \\ &= -kb \sum_{i \le j} \left(\frac{1}{Z_{k=j-i}^{a_i a_j}} \sum_l P_{k=j-i}^{a_i a_j}(l) \ln P_{k=j-i}^{a_i a_j}(l) - \ln Z_{k=j-i}^{a_i a_j} \right). && (15) \end{aligned}$$

In this study, potentials of mean force are multi-dimensional with respect to 1) backbone dihedral angles $\omega\phi\psi$ of a single residue in case $k = 0$, 2) $\psi\omega\psi$ in case of straight two residue pairs, $k = 1$, or 3) quasi-five-dimensional $(r_{ij}, \theta_{ij}, \phi_{ij}) \times (\theta_{ji}, \phi_{ji})$ in case of pairs of sequentially separated residues, $k(= j - i) > 1$. The bin size of dihedral angles ($k = 0$ or $k = 1$), is 15°

for Φ, Ψ, and 180° for ω. And for polar coordinates ($k > 1$), the size is 1Å(or 100 pm in SI) for distance, 30° for angles θ, ϕ. These are revised version of statistical potentials in the literature[6].

Folding Simulation Scheduled by Entropic Landscape

The folding simulation whose scheduling is based on entropic landscape is very simple. The lower the fragment entropy is, the earlier the fragment's conformation is optimized, and the shorter the fragment, the earlier. At level k, when a fragment of length $k+1$ is optimized, the region for total energy calculation includes those neighbouring and overlapping fragments of the same length $k+1$ which have lower entropy and have been optimized before, as long as the region is continuous in the sequence. This region is called the "extended region" of the fragment to be optimized. At each level when the last (thus highest entropy) fragment is optimized, the whole structure is the subject of energy-minimization. Thus, when the optimization is complete at each level, the resultant conformation is sound without any residue-residue clash. The resultant conformation at level k is fed to the optimization at the next level $k+1$ as the initial conformation at the next level. The very initial conformation is determined at level 0, where each single residue's conformation is individually and independently optimized by setting the best combination of ω, ϕ, ψ angles of the residue giving the lowest energy. When k reaches the whole length N, the structure prediction is completed.

The extended region's both ends have CONH group in the local optimization. That is, the extended region has CO group at the end of the N-term-side and NH at the end of C-term-side, in order to take into account the effects from those polar groups, which seem to affect the choice of ψ angle near either end of the region. When the extended region is just a single residue, at $k=0$, the residue has CONH at both sides.

For each fragment, the conformation of the fragment is so optimized that the total energy

of the extended region is minimized. The total energy is the sum of 1) the electrostatic, 2) the Lenard-Jones potentials between all backbone atoms, 3) all the backbone torsion potentials, and 4) the multi-dimensional potentials of mean force between all residues within the extended region. The parameters for electrostatic, Lenard-Jones and backbone torsion potentials are of CHAMM22. The cut-off length for electrostatic and Lenard-Jones force field is 7Å. This very short cut-off distance is intended to take into account the effects of water molecules weakening those force fields. One thing to point is that all residues are set to Glycine, thus electrostatic, Lenard-Jones, and torsion potentials of poly-Glycine are calculated for extended regions of any sequence. The energy difference between true total energy and poly-Glycine structure is approximately adjusted by the difference between potentials of mean force of Gly-Gly interaction and that of corresponding residue-type pairs. When those potentials for molecluar dynamics are mixed with the potentials of mean force, they are too strong in intensity, thus certain adjustment is required. In this study, the potentials for molecular dynamics are reduced into a fifth.

For the optimization of fragment conformation, backbone dihedral angles are changed step by step, where $\omega\phi\psi$ of a residue or $\psi\omega\phi$ between two straight residues are changed simultaneously. Which set of dihedral angles within the fragment to optimize is determined by the entropy of the two straight residues (two-residue-long fragment at level $k = 1$) in the fragment. Among the six dihedral angles $\omega_i\phi_i\psi_i\omega_{i+1}\phi_{i+1}\psi_{i+1}$ along the straight two residues, angles $\omega\phi\psi$ of the lower entropy residue are optimized first, and then of the other one, and finally $\psi_i\omega_{i+1}\phi_{i+1}$. The optimization of $\omega\phi\psi$ or $\psi\omega\phi$ at each step is the selection of the best combination of these angles that minimizes the total energy of the extended region at the step, where in case of ϕ or ψ, the best angle is first chosen from $0, \pm 1/3\pi, \pm 2/3\pi, \pi$, and then refined by shifting $0, \pm 32°$,$0, \pm 16°$,$0, \pm 8°$,$0, \pm 4°$,$0, \pm 2°$,$\pm 1°$

sequentially.

The potentials of mean force for structure optimization are smooth functions with respect to the multi-dimensional relative configuration r, multi-dimensionally interpolated using the second-order Bezier functions.

Acknowledgements For the selection of the proteins for entropy analysis and the discussion I was much helped from Tanaka, Shuhei. Regarding the propeptide folding, the insights provided by Tamura, Atsuo at Kobe University was very informative. For the overall assessment of the entropy analysis, the advice from Takada, Shoji at Kyoto University was profound and helpful. The discussions at *CASP8* were very helpful, particularly the discussion with Harold A. Sheraga at Cornel University about the hydrogen-bonds affecting the fragment entropy lead me to focus on how those effects are included in this study. I am very grateful to Joel Sussman at Weizmann Institute who picked up my study in the poster session, and gave me the opportunity to have discussions with a lot of people there.

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. Galzitskaya, O.V., & Garbuzynskiy, S.O. Entropy Capacity Determines Protein Folding. *Proteins* **63**,144-154.(2006)

3. 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)

4. 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)

5. 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)

6. Onizuka, K., Noguchi, T., Akiyama, Y., & Matsuda, H., Using Data Compression for Multidimensional Distribution Analysis. *IEEE Intelligent Systems,* **17-3**, 48-54.(2002)

Figure 1 The entropic landscape of protein L (1K50A)

Figure 2 The entropic landscape of SH3 (1ABQ)

Figure 3 The entropic landscape of POIA1 (1ITP)

Figure 4 The entropic landscape of the propeptide and POIA1

Figure 5 The folding simulation of hairpin sheet formation of POIA1

Figure 6 The folding simulation of strand-helix-strand structure formation of 1QKKA.

The Entropic landscape of proteins revealing protein folding mechanism

Kentaro ONIZUKA

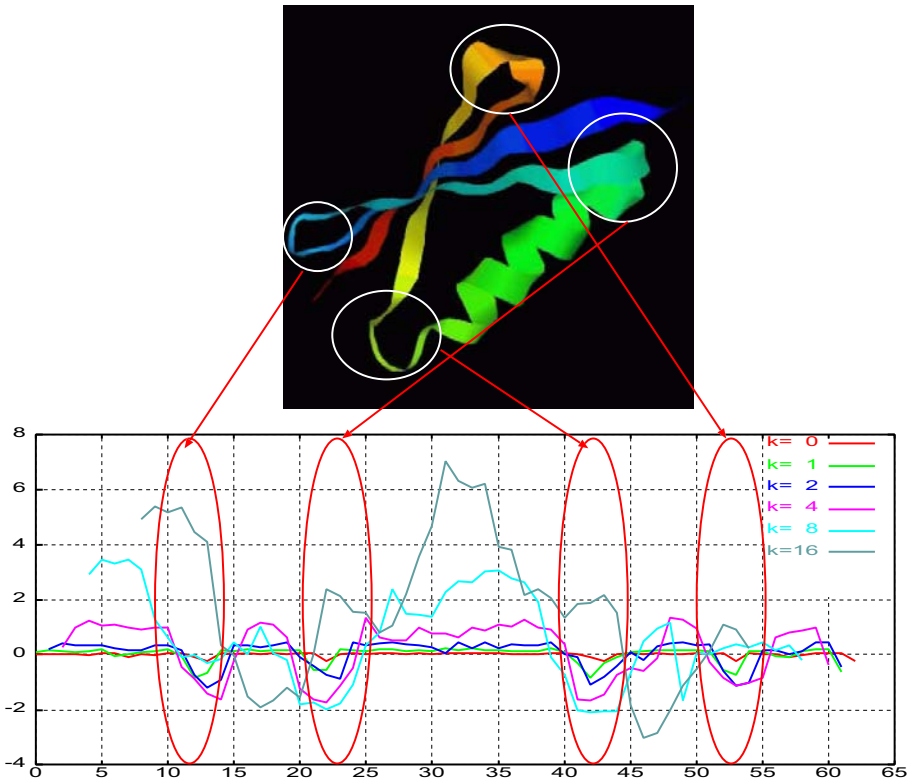

Figure 1: The entropic landscape of protein L (1K50A)

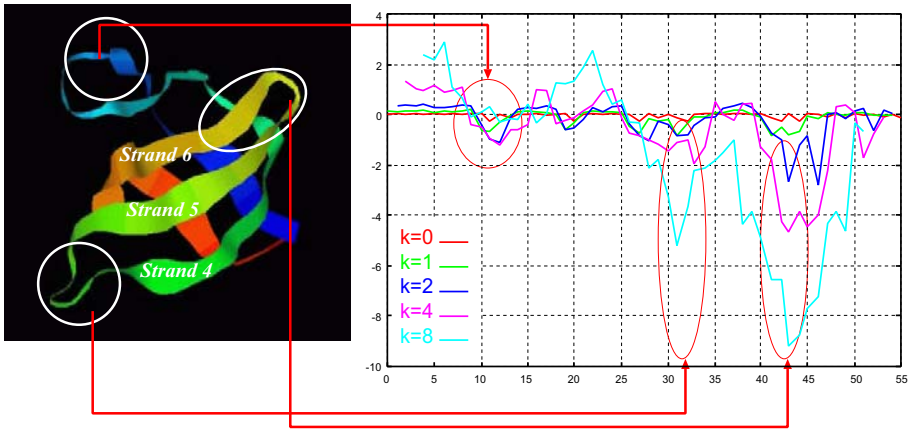

Figure 2: The entropic landscape of SH3 (1ABQ)

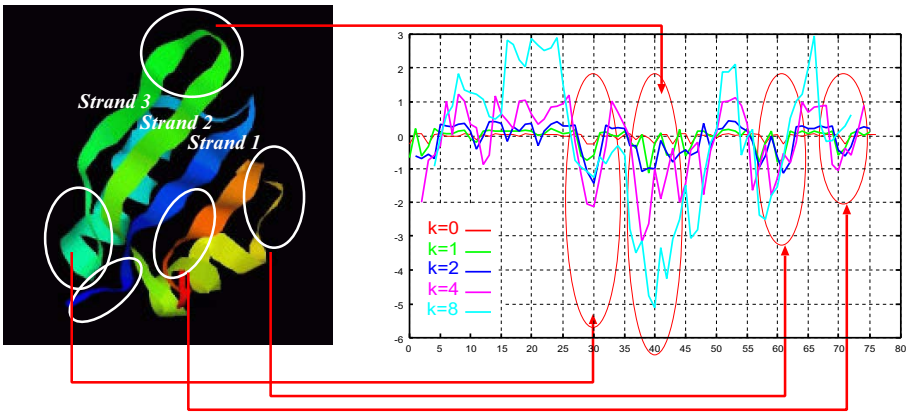

Figure 3: The entropic landscape of POIA1 (1ITP)

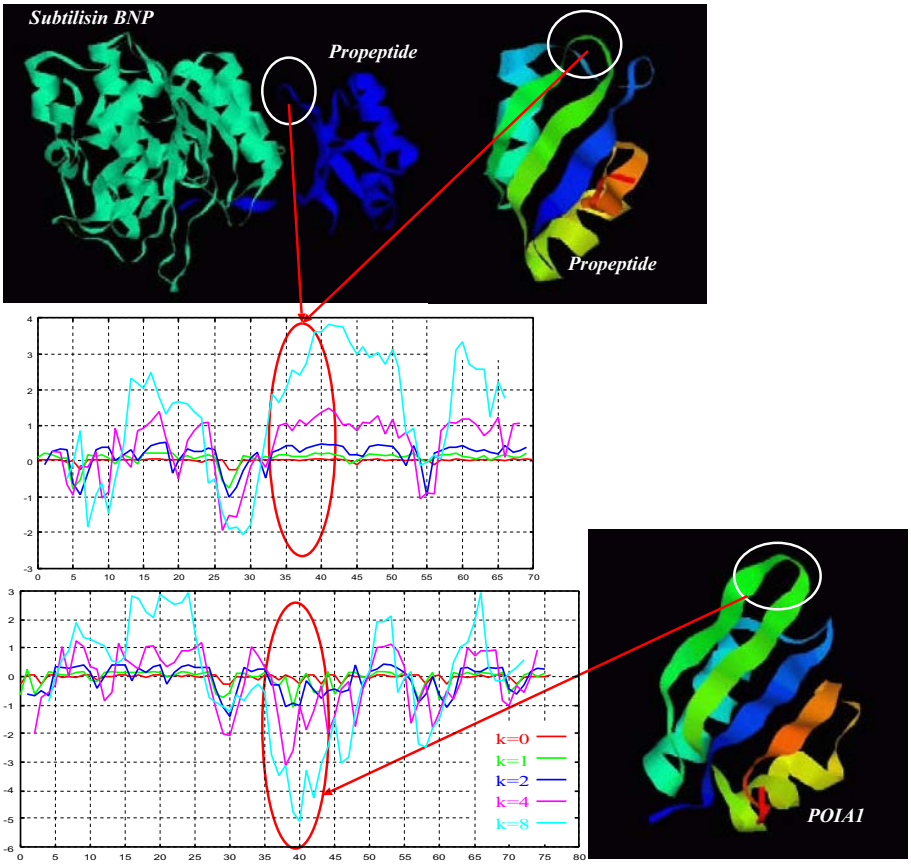

Figure 4: The entropic landscape of the propeptide and POIA1

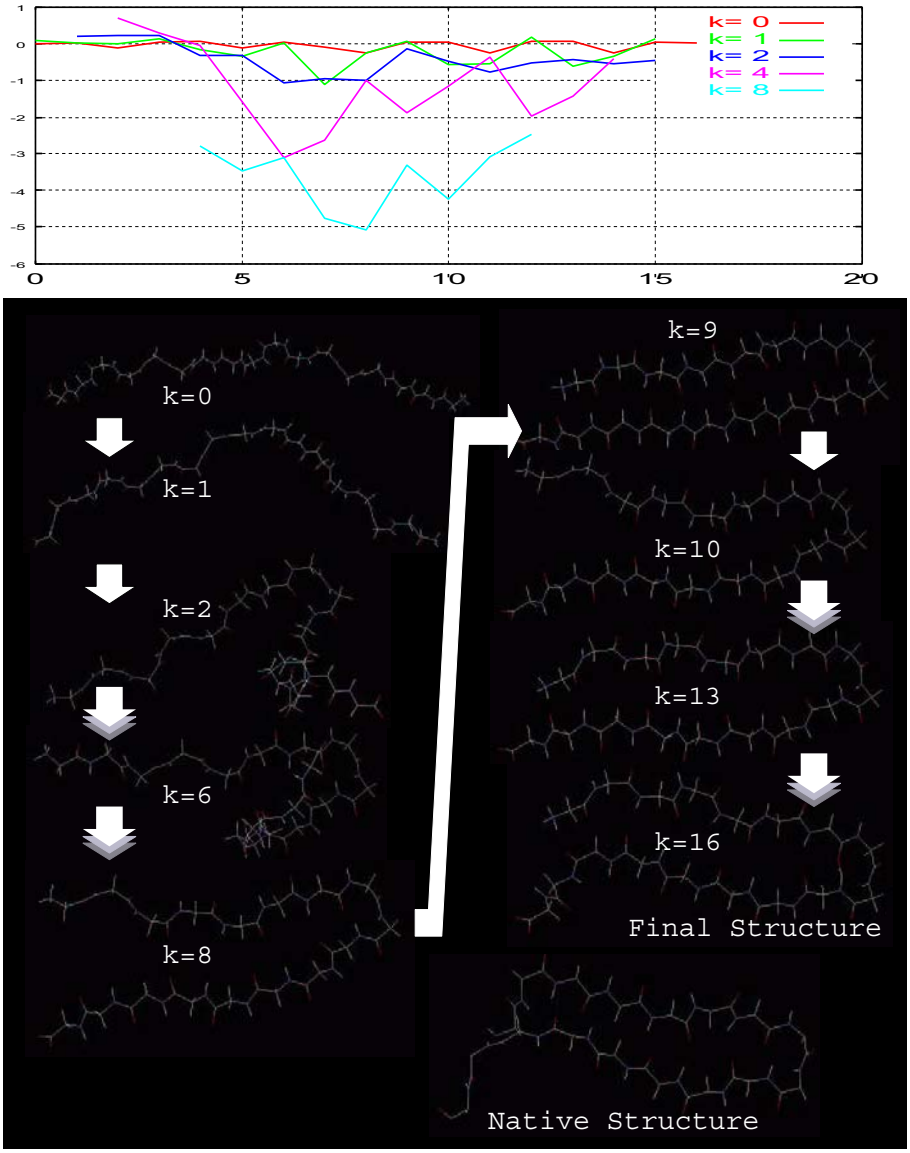

Figure 5: The folding simulation of a hairpin sheet formation of POIA1

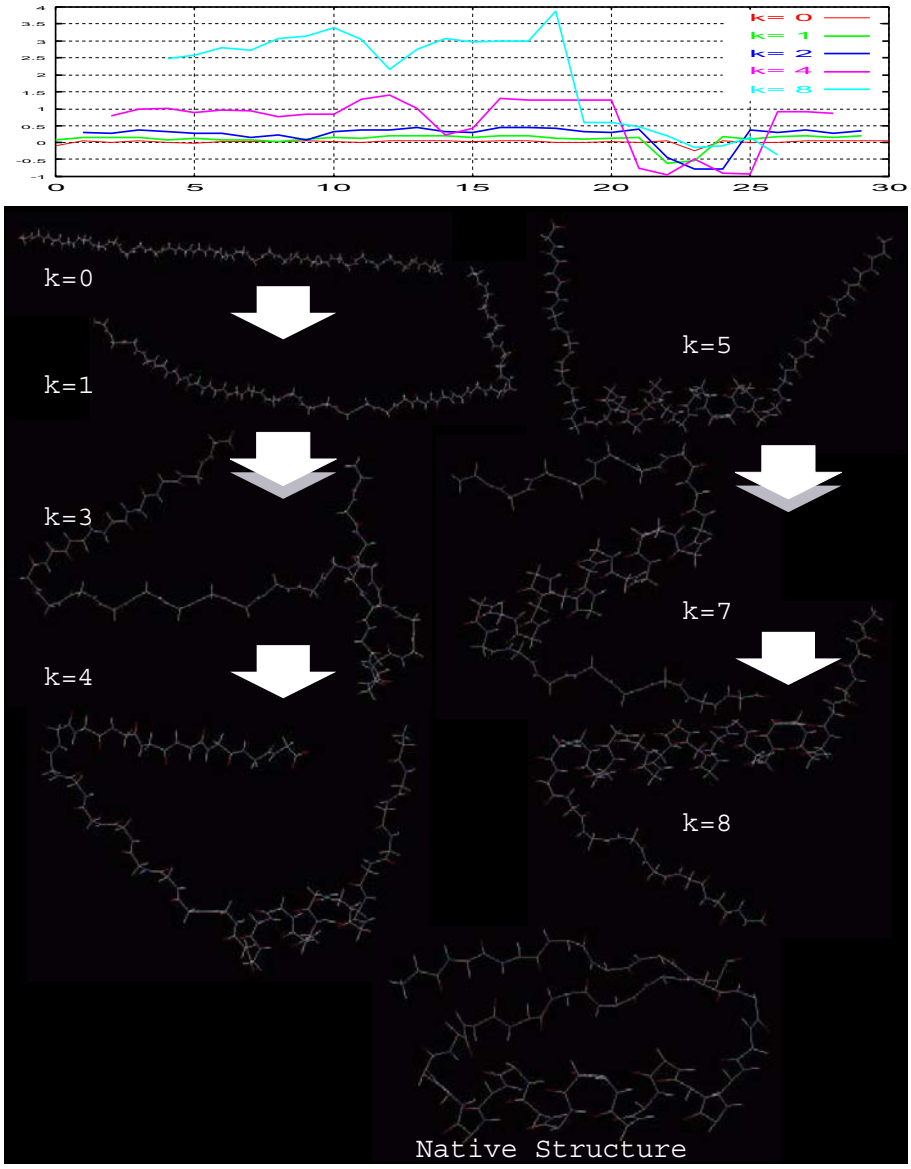

Figure 6: The folding simulation of a strand-helix-strand structure formation of 1QKKA.