Is Side-Chain Entropy an Important Factor in Protein Folding?

Joshua Ballanco

November 13, 2008

Abstract

Protein folding is a process which involves thermodynamic contributions from forces which favor the folded state and other forces which favor the unfolded state. When the balance of these forces tips in favor of the folded state, a protein will fold spontaneously into a unique conformation representing a thermodynamic minimum. The difficulty in predicting just how, and when, a protein will fold stems from a fundamental lack of knowledge about the forces involved. This is especially true in the case of the conformational entropy of amino acid side-chains (side-chain entropy).

Because of the difficulty associated with either directly measuring side-chain entropy or calculating its value from molecular simulations, there is uncertainty as to the effect it has on protein folding. The side-chains of amino acids range in size and flexibility, and some of the longer side-chains contain polar or even charged groups. It's been proposed that side-chains might counteract a loss of entropy by a gaining interactions with other amino acids. Additionally, the hydrophobic residues commonly found in the core of a folded protein are not the most flexible, and so they will not be affected as significantly by the corresponding space constraints.

On the other hand, buried residues will loose some degree of side-chain entropy. Even though the cores of proteins rarely represent a closest-packed configuration, the side chains of long, charged residues found in internal salt bridges will undoubtedly be restricted in their motion. Furthermore, it is now understood that what is commonly refered to as the "native protein structure" is actually an ensemble of closely related structures. It is possible that side-chain entropy may not be important in determining the global fold of a protein, but may be vital in understanding the dynamics associated with function.

In this paper I will look at a variety of experiments which approach the question of side-chain entropy from differing perspectives. The first series of approaches involves investigating the structure and sequence of natural proteins, specifically focusing on side-chain entropy. The alternative approach is to perform *de novo* protein design either using or neglecting side-chain entropy in the calculations. Both of these approaches are, necessarily, indirect, and each leads to a different conclusion. I will look at the results in context, and discuss whether or not a real controversy exists. Finally, I will propose a course of investigation which might shed more light on the role of side-chain entropy in protein folding.

Topics

Thermodynamics

Protein Folding

Statistical Mechanics

Molecular Dynamics

X-ray Crystallography

de novo Protein Design

Introduction

Proteins are the work-horses of biology. They are responsible for the vast majority of the biochemical processes that take place in living organisms, and consequently there is keen interest in understanding how they function. Normally, this would be a question of brute-force experimentation, approaching and investigating proteins one at a time. Two aspects of proteins, however, make this a problem that is tantalizingly close, and yet still frustratingly far, from a complete solution.

The first useful aspect of proteins is the relationship between their structure and their function. Knowing a protein's structure goes a long way to understanding how it might function. This is particularly important when developing pharmaceutical agents to treat disease. Second, a protein's structure is completely determined by its amino acid sequence, which can in turn be determined from available genetic data. This was originally shown in 1961 by Anfinsen [10] in his work with ribonuclease. This notion, of 3-dimensional structure data encoded in a 1-dimensional array of elements, has been verified for proteins in all but a few rare cases.

Thus, since knowing a protein's structure was already more than half the battle, and because it should be possible to determine that structure with readily available information, work began almost immediately on elucidating the mechanism drives proteins to form their unique structures. This work has been ongoing for the 47-years since.

Shortly after Anfinsen's discovery, Brandts proposed a two-state model for protein folding where a protein exists in an equilibrium, D $\rightleftarrows$ N, between the denatured and native state [4]. This two-state model of protein folding is still the prevailing accepted model for folding, and it has some interesting implications. One of the first implications was recognized by Levinthal in 1968, and has since become known as Levinthal's paradox.

Essentially, what Levinthal realized was that Brandts' two-state model implied an absence of long-lived or marginally stable conformations. In other words, the denatured protein, which is essentially a random coil, must find the single lowest energy conformation from all of the possible conformations in one step. Levinthal's rough calculations implied that a robust search of all possible conformations would require more than the age of the universe to complete yet, paradoxically, proteins fold on a millisecond to second time scale. Obviously, proteins do not randomly sample conformation space. Instead, there must be a driving force that guides a denatured protein, no matter its conformation, toward the native state.

This understanding of Levinthal's Paradox lead Dill to propose the concept of the folding energy landscape as a funnel [5], which has become an iconic symbol of the protein folding problem. The funnel captures the essence of the relationship between energy and conformational space. At high enough energies, proteins are free to explore all conformations, and are therefore "denatured" (in fact, the "denatured" state allows for some fraction of protein molecules to explore a native or native-like conformation, but this fraction is small enough to be inconsequential). As the energy of the protein is lowered, molecules with conformations near the center of the funnel will simply decrease in energy. However, those along the sides of the funnel will be steered toward more and more native-like conformations. In this way, Levinthal's paradox is resolved since all the molecules of a protein, regardless of their starting conformation, will converge on a point in conformation space as the energy is lowered.

Dill's funnel also captures another important aspect of protein folding: entropy. At any given energy, the volume of conformation space which remains inside the funnel corresponds to the conformational entropy of the protein. As a protein works its way down the ever narrower funnel, its entropy is decreasing. In order to offset this effect, the folded state of a protein must contain a host of favorable interactions. These will be the same sorts of interactions which define the funnel in the first place, and thus the problems of determining protein folding pathways, protein unfolding pathways, and native structures are all linked.

Ultimately, the question of entropy's effect on protein folding is a rather complicated one. Entropy is, unlike enthalpy or potential energy, an ensemble property. That is, it cannot be determined by looking at a single folded molecule, but rather depends on how many different folds are possible for a given molecule at a given energy [8]. This poses a significant problem in the calculation of entropy from molecular dynamics experiments. An exact calculation of the entropy of folding ($\mathbf{\Delta S_{fold}}$) would require not only enumerating all of the possible native-like conformations accessible to a folded protein, but also to enumerating over all of the possible denatured protein conformations. While the former is merely an extremely computationally intensive problem, the later is intractable. A number of techniques have been developed to provide good estimates for this value, but even some of these techniques are prohibitively difficult.

Entropy is also not very easy to separate into component contributions in the same manner as enthalpy or potential energy [2]. Where two hydrogen bonds in a protein would be expected to have twice the stability of a single hydrogen bond, the same cannot be said for the entropy of two side-chains. This makes it difficult to attack the problem of side-chain entropy with a piecemeal approach without making certain assumptions. This also means that directly measuring the contribution of one component of a protein to the protein's entropy via experimental methods is practically impossible.

In particular, free energy, a term which captures both the potential energy of a protein and its entropy, depends on quadratic and higher order terms of the potential energy [3]. If we can integrate the system from absolute zero to a given temperature, a technique known as Thermodynamic Integration, then these higher order terms become temperature derivatives of singular potential energy contributions. In this way, it is possible to separate free energy into contributory components, but only along certain informative paths.

To fully understand the contribution of entropy, specifically side-chain entropy, to the process of protein folding creative approaches are needed. Below, I will look at two different, yet complementary, approaches to this question. As noted above, the forces that govern the process of protein folding are the same forces that determine the native structure of a given sequence of amino acids. If we look at the amount of side-chain entropy in native protein structures, we should be able extrapolate the relative importance of side-chain entropy in the folding process. This can be accomplished either by calculation of the side-chain entropy for known structures, or by analyzing proteomic sequences with an eye toward an amino acid bias toward side-chains with more or less degrees of freedom.

Another means of assessing how much side-chain entropy participates in protein folding is by recreating, in a sense, protein evolution. Evolution is restricted by the physics of protein folding, and so it will favor certain amino acid sequences over others because they contain the right combination of physical properties to fold to the needed structure. By attempting to design proteins from scratch, we can investigate which aspects of amino acid physics are important by either including or excluding them in the selection process. If the design algorithms come to a solution which is close to the one resulting from natural evolution, we can assume that the algorithm contains the physical properties important for folding.

Side-Chain Entropy in Naturally Occurring Proteins

In holding with a classic scientific tradition, we can listen to nature and learn what she has to tell us. In this case, that means gathering information about naturally occurring proteins and subjecting this information to analytical techniques that might reveal something about the role of side-chain entropy. This can be an especially fruitful line of investigation when it comes to aspects of proteins simply because nature has a vastly larger laboratory than any man could create and experiments have been on-going for literally billions of years. All we have to do is frame our question in such a way that evolution will have already provided an answer.

One way we can do that is to ask a question about protein stability. Our interest is in the forces that give shape to a protein structure, but it has been well established that protein structures are marginally stable. In order to increase the stability of these structures, the forces favoring a native protein structure will need to be strengthened. To find which forces are important, we need only look at which forces have increased in magnitude in an environment where protein stability is key. Luckily, hyperthermophilic bacteria thrive in exactly that sort of an environment.

This is the technique adopted by Berezovsky, *et al.* [1] in comparing protein sequences from mesophilic and thermophilic bacteria and looking at their amino acid content and side-chain entropy. They begin with an interesting simplification of a protein model commonly known as the Gō model. This model simplifies the calculation of potential energy by only considering amino acid interactions present in the native state. While this might seem like a rather drastic simplification, it allows the experimenter to focus on features of the amino acid backbone and side-chains. In other words, they sacrifice detail in the folding pathway, but are able to retain atomic detail of the structure.

Using the number of accessible rotameric states as a measure of side-chain entropy, the first observation this group makes is that, in folded state of hydrolase H from both *Escherichia coli* and *Thermus thermophilus*, lysine retains more of its side-chain entropy than arginine (an average of 20.1 vs 3.5 accessible rotamers per residue, respectively, in the folded state). This is mildly surprising as both lysine and arginine have similar chemical properties, they are both positively charged, and the both have the same total number of rotamers. The difference between the two stems from the guanadinium group of arginine, which is bulky and restricts its motion in the core of the protein structure.

Next, they expanded their investigation to 18 pairs of proteins from *E. coli* and *T. thermophilus.* This time, instead of simply looking at absolute numbers of rotamers, they carried out Monte Carlo unfolding simulations and compared the number of observed rotamers in the folded and unfolded states for lysine and arginine (Figure 1). Consistent with their initial findings, the lysines have a much smaller discrepancy between the observed rotamers in the folded and unfolded states than the arginines.

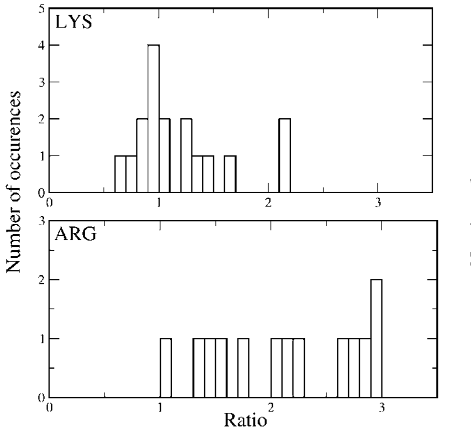

Figure 1: (Borrowed from [1]) Distribution of the Ratios of the Number of Rotamers in Unfolded and Folded States in a Representative Set of Proteins.

From these observations, they form a hypothesis that replacing arginines with lysines in the hydrolase H from *E. coli* will increase its temperature stability. There are a few difficulties with this sort of experiment. First, switching out residues like this without making corresponding changes to other interacting residues will decrease the number of favorable interactions in the native state. Second, even using the Gō model approximations, it is still difficult to collect enough Monte Carlo samples to definitively say if switching lysine for arginine has a stabilizing effect. Their results seem to indicate that this is the case but are far from convincing.

In addition to attempting to model their hypothesis, Berezovsky, *et al.* look to the available proteomes of 38 mesophilic and 12 hyperthermophilic bacteria. Here, they find that the arginine content of the hyperthermophiles is less than that of the mesopiles, and correspondingly, the hyperthermophilic lysine content is greater (Figure 2). Looking at the underlying genomic sequences for each proteome, they were able to rule out the possibility of this being an artifact of DNA-base bias in all but two cases.

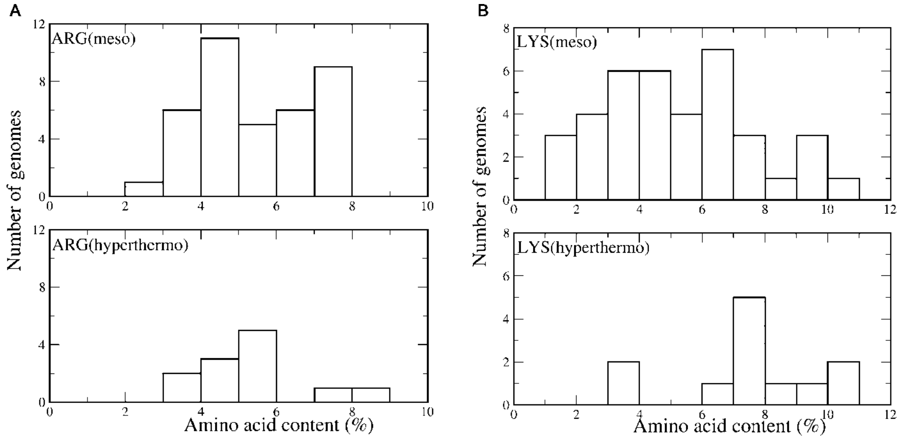

Figure 2: (Borrowed from [1]) Histograms of the Content of Charged Amino Acid Residues in Hyperthermophilic Genomes Compared with Mesophilic Genomes. (A) Arginine content of mesophilic genomes, top, and hyperthermophile genomes, bottom. (B) Lysine content of mesophilic genomes, top, and hyperthermophile genomes, bottom.

Instead of looking for clues in protein sequences, Zhang and Liu [11] decided to look to protein structures. They use a variation on a Sequential Monte Carlo method to approximate, from a set of experimentally determined protein structures, the entropy resulting from side-chain motions. Their initial results using this technique show that side-chain entropy increases linearly with chain length, and that as chains grow beyond a threshold length, the buried residues in a structure contribute about half of the total side-chain entropy for a protein. Both of these results are consistent with what is already known of side-chain entropy.

As a next step, they calculated the side-chain entropy of 24 distinct proteins and a corresponding set of decoy protein structures. The logic here is that, if side-chain entropy is important, then it should be useful as a metric to discriminate between the real and the decoy proteins. What they found was very convincing. If they plotted the side-chain entropy versus the radius of gyration, a measure of compactness, then in half of the cases studied the real structure had a clearly greater side-chain entropy than for any of the structures of similar compactness. In the half where this was not the case, there were extra constraints which led to the native structure having a lower entropy. Even for dimeric proteins, they were able to plot side-chain entropy versus the number of interfacial contacts and found that the native structures were convincing outliers.

Finally, they looked at the side-chain entropy of 23 protein X-ray crystal structures and their corresponding NMR structures. Because NMR structures represent solution structures, whereas X-ray crystal structures represent closely packed molecules, they were curious to see the difference in side-chain entropy between the two classes of structure. What they found was that, while the X-ray and NMR structures had similar radii of gyration in general, the X-ray structures had greater side-chain entropy in nearly all of the cases (Figure 3). What this implies is that even though the compactness of the X-ray and NMR structures is not different, the X-ray structures are able to pack side-chains more "intelligently", such that they have greater side-chain entropy.

From all of this evidence, Zhang and Liu hypothesized that a free-energy calculation which includes side-chain entropy should be better at discriminating real from decoy structures better than an enthalpy calculation based only on potential energy. To test this hypothesis, they calculated both the backbone conformational enthalpy and Gibbs free energy for the 24 proteins and decoy sets used above. The result

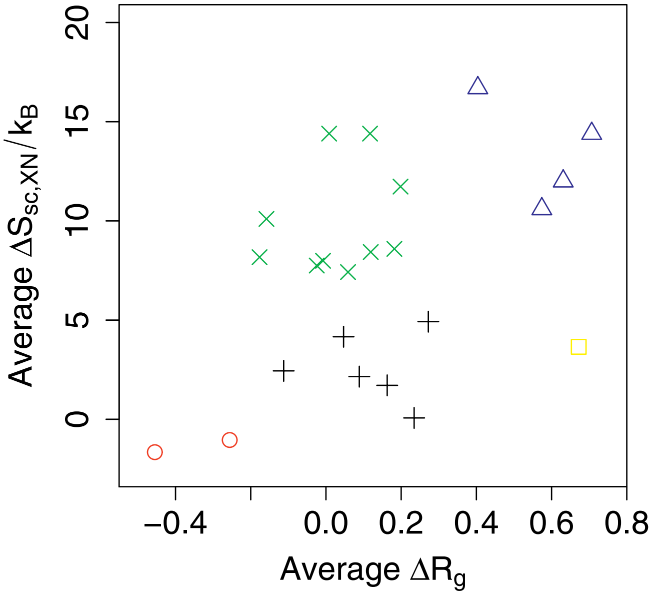

Figure 3: (Borrowed from [11]) Side-Chain Entropy of NMR and X-Ray Structures vs R_g. 'X's are proteins whose X-ray structures have much higher side-chain entropy than but similar R_g to the corresponding NMR structures. Triangles are proteins whose X-ray structures gain considerable side-chain entropy by packing a little looser. Circles are proteins whose X-ray structures pack tighter than NMR structures but with comparable side-chain entropy. '+'s are small proteins with both small R_g and ΔS_{XN}.

was that the Gibbs calculation was better at picking the real protein 73% of the time, was on par with the enthalpy calculation in 20% of the cases, and was slightly worse in only one instance.

What we can learn from the work of both of these groups is that the evolutionary process seems to account for side-chain entropy when picking a protein structure. In situations where protein stability is key, such as the high temperature environments occupied by hyperthermophiles, evolution will limit the number of arginines in favor of more motile lysines. In the case of X-ray crystal structures as compared to NMR or energy-minimized decoy structures, evolution favors configurations which allow more room for packed side-chains to explore conformational space, and therefore have increased side-chain entropy. While it is possible that the effects observed here are artifacts of the evolutionary process, it is not likely. Evolution is a random process, and should have mixed out any bias toward increased side-chain entropy if it was not important for protein structure.

Side-Chain Entropy in Protein Design

Another approach to the question of side-chain entropy, and one that has only become feasible with recent advances, is to look at its role in the *de novo* design of proteins. The prospect of explicitly calculating all of the thermodynamic parameters involved in folding a protein is already a daunting challenge. Performing these calculations while iterating through all of sequence space to find a specific protein fold is flat out impossible. For that reason, *de novo* design algorithms make simplifying assumptions about what sorts of thermodynamic parameters are important in determining a protein's shape [7]. Most of these algorithms will randomly iterate through sequence space, stopping at regular intervals to minimize the protein structure (for example, by simulated annealing) followed by calculation of scoring function which will determine if the structure should be kept or rejected.

RosettaDesign is just such an algorithm, and Hu and Kuhlman have used it to investigate the role of side-chain entropy in the *de novo* design process [6]. The advantage of this approach is that it requires only a small modification to the design algorithm. The portion of the algorithm which handles iteration through sequence space and minimization of the mutated structures does not need to be modified. Instead, Hu and Kuhlman include a step which involves Monte Carlo sampling to generate an ensemble of structures used to calculate side-chain entropy and free energy. They then substitute free energy as the Metropolis acceptance criterion in place of minimized energy, which was used as the criterion in the original algorithm.

To calculate side-chain entropy, they look at the side-chains in the Monte Carlo ensemble and, for each residue, they iterate through the available rotamers, summing a probability based entropy. In equation form, this is:

$$S = -R \sum_{i=1}^{nres} \sum_{r=1}^{nrot} p(r,i) ln[p(r,i)]$$

where $nres$ is the number of residues in the protein and $nrot$ is the number of rotamers for the residue. What's notable about this method of calculating entropy is that it explicitly neglects pair-wise terms. The entropy for each residue is calculated independently of the other residues in the chain. This significantly simplifies the task of determining side-chain entropy, but is it a realistic assumption?

To answer that question, they first calculated side-chain entropy in three different ways. The first was simply the enumeration, given above, that they planed to use in the design algorithm. The second method involved treating clusters of 6 neighboring side-chains as a unit and enumerating over the various combinations of rotamers for each cluster. The final method involved treating each dihedral angle in each side-chain independently. Comparing these three techniques, they find that the difference between an independent treatment of the side-chains and treating clusters of side-chains is on the order of $10^{-2} kcal/mol$, which the authors judge to be insignificant. Treating each dihedral angle independently yields a more significant overestimation of the side-chain entropy, indicating that there is covariant motion within a side-chain but not between side-chains.

With these results in hand, they proceeded to carry out repacking simulations (essentially, the design algorithm without mutating any residues) using both the original scoring metric and the modified metric including side-chain entropy. Dividing the residues between those which are near the surface and those buried in the protein's interior, they were able to measure a change in side-chain entropy upon burial (Figure 4). As expected, they find that the longer, more flexible side chains suffer a greater entropy penalty when buried, but when accounting for the difference between the average energy of the side-chains and the energy of the most stable side-chain conformation, they find the difference is less noticeable. Only four amino acids (Met, Arg, Gln, and Glu) have a greater than $0.3 kcal/mol$ advantage when positioned near the surface.

From this result, the authors expected that inclusion of side-chain entropy in their design algorithm would not have a significant impact on the results. To verify this, they carried design experiments for 110 naturally occurring proteins in an attempt to recover the native sequence. Comparing the designed sequences from both the original algorithm and the modified algorithm, the success rate is not materially affected by the inclusion of side-chain entropy. In both cases, approximately one-third of the native sequence residues were selected by *de novo* design. There was a noticeable change in the frequency with which the longer amino acids were buried when including side-chain entropy, but not as large an effect as when other parameters of the energy calculation, such as solvation energy, were omitted.

While these results would appear to indicate that side-chain entropy is not vital for determining protein structure, the technique used depends on the pairwise entropy interactions being ignored. Even though a proof of concept experiment seemed to indicate that this was a valid simplification, entropy is a tricky quantity to characterize. Because of the fact that it is an ensemble property, what might appear to be a

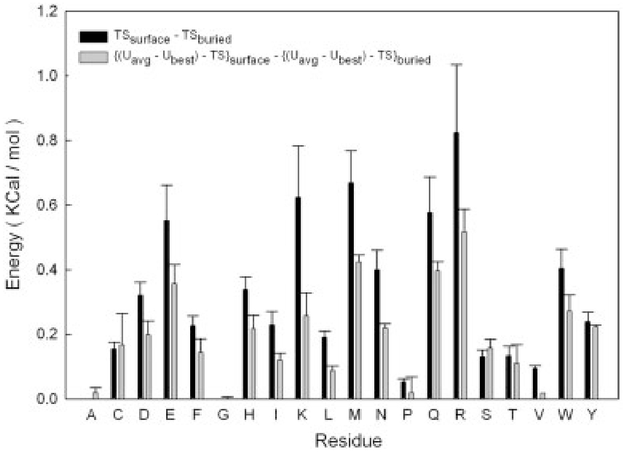

Figure 4: (Borrowed from [6]) Changes in side-chain conformational entropy and free energy between surface and buried positions. The black bars show the change in entropy (TS) when a residue is buried whereas the gray bars compare average free energies (U_{avg} - TS) obtained with the explicit side-chain entropy model to the energies that are calculated with the standard RosettaDesign model (U_{best}). The gray bars indicate the net effect that the explicit side-chain entropy model has on the environmental preferences of the amino acids.

small discrepancy in a small experiment, might become a large error in a full scale simulation. So, what happens if pairwise contributions are accounted for?

This is precisely the question Sciretti, *et al.* attempted to answer [9]. Using a pairwise approximation applied via the Belief Propagation (BP) technique, they are able to calculate free energy and use it as a minimization criteria. Before undertaking a full protein design experiment, however, they begin by evaluating the conformational side-chain entropy for a three residue stretch in an SH3 domain. Considering all possible rotamer combinations for all possible amino acid combinations for this stretch leads to a few million calculations. This yields an exact calculation of the side-chain entropy for the stretch, which they compare to the same calculations performed using the BP technique. The results validate the use of BP for side-chain entropy calculations, as the results using BP are essentially identical.

As further validation, they allow the side-chains of the remaining residues (beside the three being mutated) to relax normally. The problem in this case is intractable for an exact calculation, but they were able to verify that the BP results from these simulations were essentially the same as those from the above case where all but the three mutated residues had their positions fixed. As a final validation of the BP method, they measure free energy as the temperature nears absolute zero and find that free energy converges with potential energy (the entropy contribution disappears at absolute zero). Armed with this evidence that BP is a powerful technique for evaluating the free energy of protein structures, they then undertook a series of *de novo* design experiments.

In the first set of design runs, they carried out calculations at absolute zero using potential energy as the evaluation criteria. In the second set of runs, they carried out calculations at a temperature of $T = 0.6$ (in RT units), and using the BP technique, evaluated structures by free energy. They then held the structures fixed, and calculated free energy at a range of temperatures for both sets (Figure 5). The structures evaluated by potential energy at absolute zero were, unsurprisingly, also the lowest in terms of free energy at absolute zero. What was unexpected was that, while the structures selected using BP at $T = 0.6$ had significantly higher free energy at the low temperatures, they had significantly lower free energies at the more physiologically relevant temperature than the first set of candidates.

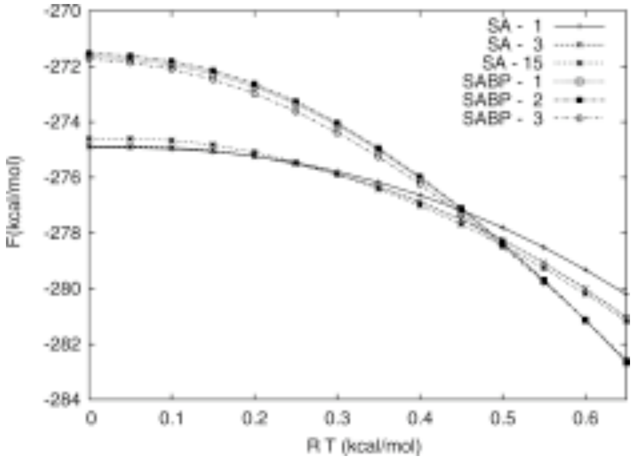

Figure 5: (Borrowed from [9]) Free energies vs temperature for some of the best ranking sequences as found by Simulated Annealing alone (SA) and Simulated Annealing with Belief Propagation (SABP). The authors have checked that BP free energies converge to SA energies at low temperature, suggesting that for these sequences the pair-wise approximation is reliable also at low temperatures.

In fact, the authors point out that the difference between these two sets at both the high and low temperatures is great enough that there would be little chance of finding the structures in one set using the selection criteria from the other for the design phase, and then switching scoring methods when picking candidates. In other words, a *de novo* protein design methodology which accounts for free energy will give very different results than one which does not. Perhaps the most surprising outcome of these experiments, however, is that the designed proteins in both sets recover only 35-55% of the native structure, a result not entirely inconsistent with the work of Hu and Kuhlman.

Is Side-Chain Entropy Important?

What can we conclude about the importance of side-chain entropy in determining the folded structure of a protein?

- Design vs Structure
- Not yes or no, but how much?
- Both important and unimportant simultaneously

How Important Is Side-Chain Entropy?

With the possibility that side-chain entropy may be important for some aspects of protein folding and not for others, it is vital that we understand better the roles that side-chain entropy plays.

- Mechanical experiment for validation
- New techniques for simulating proteins

References

[1] Igor N Berezovsky, William W Chen, Paul J Choi, and Eugene I Shakhnovich. Entropic stabilization of proteins and its proteomic consequences. *PLoS Comput Biol*, 1(4):e47, Sep 2005.

[2] G P Brady and K A Sharp. Entropy in protein folding and in protein-protein interactions. *Curr Opin Struct Biol*, 7(2):215–21, Apr 1997.

[3] G P Brady, A Szabo, and K A Sharp. On the decomposition of free energies. *J Mol Biol*, 263(2):123–5, Oct 1996.

[4] A Clay Clark. Protein folding: are we there yet? *Arch Biochem Biophys*, 469(1):1–3, Jan 2008.

[5] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. The protein folding problem. *Annual review of biophysics*, 37:289–316, Jan 2008.

[6] Xiaozhen Hu and Brian Kuhlman. Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins*, 62(3):739–48, Mar 2006.

[7] Shaun M Lippow and Bruce Tidor. Progress in computational protein design. *Current Opinion in Biotechnology*, 18(4):305–11, Aug 2007.

[8] Hagai Meirovitch. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Curr Opin Struct Biol*, 17(2):181–6, Apr 2007.

[9] D Sciretti, P Bruscolini, A Pelizzola, M Pretti, and A Jaramillo. Computational protein design with side-chain conformational entropy. *Proteins*, Jul 2008.

[10] Robert F Service. Problem solved* (*sort of). *Science*, 321(5890):784–6, Aug 2008.

[11] Jinfeng Zhang and Jun S Liu. On side-chain conformational entropy of proteins. *PLoS Comput Biol*, 2(12):e168, Dec 2006.