Physics-based protein structure prediction and design using the confinement method

Arijit Roy, Justin L. MacCallum, Alberto Perez, and Ken A. Dill

Laufer Center for Physical and Quantitative Biology
Stony Brook University
Stony Brook, NY 11794-5252.

January 11, 2013

Abstract

Calculation of free energy differences is of central importance in the simulation of biochemical systems. It is particularly difficult to calculate between pairs of macromolecular conformations as well as a computationally expensive task with existing methods. In this work, confinement approach is used to calculate absolute free energies of biomolecular systems. This method provides two main advantages: it does not require a reaction coordinate or transition path and it is fast to compute. Free energy calculated can be decomposed into a per residue contribution in an approximate way. Per residue free energy allows us to identify the reason behind conformational preferences in biomolecules. Through out the article we show its use in different challenging modeling problems. In particular, we show its use in predicting the conformational preference of chamaleon sequences (sequences with high sequence identity and different folds). This sequence dependent conformational preferences and per residue free energy decomposition set the stage for the use of this method in protein design.

1 Introduction

In computational structural biology there are numerous cases where the free energy difference between two well defined states plays an important role. Examples range from small to large conformational changes of proteins due to ligand binding, change of pH or other conditions[1,2,3]. Free energy also plays an important role in the case of protein folding. Ground breaking work of Christian B. Anfinsen and coworkers showed that the native structures of small globular proteins have a unique, thermodynamically stable native structure with their conformation at the global free energy minimum[9]. Regardless of the starting point or even after unfolding it by changing different condition most proteins will go back to the native state after reinstituting native like conditions. Often, the changes in free energy between different conformations is small, but nonetheless crucial in favoring the native state since defects in protein folding may be the molecular cause of a range of human genetic disorders. For example a misfolded protein known as a prion can misfold correctly folded proteins when entering a healthy organism. Thus it is important for protein to have its native structure. As protein native structure stays in the global free energy minimum, free energy can be used as a scale for accurate identification of misfolded state from the protein native state. I'm not sure I understand this here These facts further give free energy a special importance in structural biology.

However, the theoretical calculation of conformational free energy change becomes difficult if the states of interest are very different[1]. In such scenario timescales involved in such conformational transitions may be beyond the sampling ability of classical molecular dynamics simulation. Generally, in order to calculate the free energy a pathway or reaction coordinate is needed with several intermediate structures. Methods like umbrella sampling[5], thermodynamic integration[6] along with classical MD can suffer from overlap problem and requires a larger computational effort. The idea of a reaction coordinate became even more complicated in the case of protein folding. As free energy landscape for protein folding is rough[7,8], the molecule can get trapped in an energy well during conformational sampling. Even with advanced methods like replica exchange molecular dynamics, the three dimensional structure may be trapped in a local minima for a considerable time, giving rise to nonconverged distributions where

misfolded states might be selected as native like

The calculation of free energy have been successfully attempted by a number of groups[1],[11] -[13]. In recent years methods like the Reference system method[11], Decativated Morphing[12], Orthogonal Space Random Work[13] or the Confinement Method[14],[15] to name a few have been used to calculate free energy differences. In this work we have applied the confinement method which was originally devoled by Tyka et. al.[14] and Cecchini et. al.[15]. This method relies on the fact that the free energy difference is a state function and thus it is independent of the sampling path. This approach uses a thermodynamic cycle where first, non-harmonic degrees of freedom are removed by applying a series of restraints to the system. Then the free energy of the remaining harmonic system is obtained using a normal mode calculation and combined with the non-harmonic part to give the total absolute free energy. Previously, this method was applied to calculate the conformational free energy of a 16 amino acid residue peptide, known as BHP. We have first reproduced that known result. Next we have applied this method to larger proteins for the first time. The free energy difference of pairs of conformations in proteins with similar sequence but completely different fold was calculated. i made this part too long, have to summarize here and define better inside text?–¿ Encouraged by the result, we applied the method to help tackle the problem of protein conformation ranking. We did this by selecting different targets from the Critical Assessment of Structure Prediction (CASP) event. In this event, different research groups try to find the 3D structure of a protein given a sequence in a blind way using their methodologies. Upon comparing with the real structure it is often seen that several groups are able to generate better structures than the ones they submit, but are not able to identify them. This is known as the ranking problem, and this method is able to correctly rank structures in the majority of cases we have tried. Finally, we have been exploring the ability of working on a per residue basis to identify which residues are the major players in the stabilization/destabilization of a pair of conformations. Our initial experience makes us thinkg that this method might be helpful in identifying key residues for protein design.

2 Results and Discussion

2.1 The confinement method produces correct results in control experiments

This is benchmarking, good internal, but not relevant. Should go to SI,out, or summarized in one sentence: "After reproducing the results of Cecchini and coworkers we were interested in the ability to apply this method to larger systems"

As a first step, we performed several control experiments to verify that our implementation of the confinement method produces results compatible with previous calculations reported in the literature.

The method has previously been applied to a 16 amino acid residue β-hairpin from protein G, known as BHP[15]. We calculated the free energy difference between two different conformations of the peptide: (1) the native conformation, called bhp1, with a two-stranded β-sheet; and (2) a conformation, called bhp3, which has a three-stranded β-sheet. Analysis of long (4 μs) equilibrium simulations[15,17] shows that bhp1 is the more favorable configuration by 1.8 kcal/mol. Using the confinement method, we obtain a value 1.7 kcal/mol, which is in good agreement with the equilibrium simulations and with previous calculations using the confinement method[15] (see Supporting Information for further details).

Previous applications of the confiment method have focused on relatively short peptides, up to 17 residues in length. In this work we apply the confinement method to larger proteins. On that direction, we first test the confinement method for a chamelon sequence and found that it can correctly pick out the preferential structure of the chamelon sequence. Other applications we explore in the present work is the re-scoring, or metaprediction, of structures submitted during the Critical Assessment of Structure Prediction (CASP) experiment (described in detail later). We computed the relative free energies of predictions submitted during CASP9, with the expectation that the most native-like prediction will have the lowest free energy. As a control, for several targets we also calculated the relative free energy of the experimentally determined structure. should not start with results yet, first we are describing what we are going to show them As Table 1 shows, in most cases, the confinement method correctly

assigns a lower free energy to the experimentally determined structure than to any of the decoys. Most interestingly, we decomposed the free energy into its per residue component, which help us to identify the residues which stabilize or destabilize a particular conformation of the protein.

[Table 1 about here.]

2.2 The confinement method correctly predicts the structural preferences of chameleon sequences

In general, proteins with similar sequences tend to have similar structures. This idea is the basis of comparative modeling and fold recognition in protein structure prediction. There are, however, examples—often referred to as chameleon sequences—of proteins with similar sequences that have remarkably different structures. Orban and co-workers have designed a sequence of 56-residues that is marginally stable in one of two possible folds. By mutating key residues in this sequence they are able to stabilize one fold or the other (see Figure 4). Sequences that adopt a mixed alpha/beta structure similar to Protein G are denoted as "GB", while sequences that form a three-helix bundle are denoted as "GA". One pair of sequences (GA88/GB88) are 88 percent identical in sequence and differ in seven positions. Another pair (GA95/GB95) are 95 percent identical and differ in three positions. The last pair (GA98/GB98) differ only in a single tyrosine to alanine mutation.

The fact such small changes in sequence can lead to such dramatic changes in structure is rather remarkable. Accurately predicting the structural preferences of these structures presents a serious challenge for computational methods[18–21].

We initially approached this problem by making a model of each sequence with the same backbone structure as its partner chameleon sequence. For example, we took the sequence of GA88 and built a model with the same overall structure as GB88. We then used the confinement method to assess the free energy difference between the experimentally determined structure of GA88 and the model (with the GA88 sequence and the GB88 structure). The confinement method was able to predict the conformational preferences correctly for all six sequences (data not shown). This is however not

surprising. It is well known [CITE] that it is easy to distinguish computational models from native structures. Despite the fact that this models are built by having a huge amount of structural information, it might be possible that we were able to make correct predictions simply because artefacts of our modeling procedure always lead to the model having a higher free energy than the experimental structure.

To avoid this potential problem, we instead computed relative free energy of two different structural models for each sequence. One model is based on the GA structure and the other on the GB structure (see Supporting Information for details on the modeling procedure). This is a much more realistic test of the confinement method's ability to accurately calculate relative free energies.

The results of these calculations are presented in Figure 4. Confinement method identifies the correct structure for all five sequences. One of the hypothesis for fold switching of chamelon sequence is that the structural transitions require states with diminished stability. It is widely believed that if the free energy of the native state and the alternative state is within a range of $5Kcal/Mol$ then it can quickly change fold when the stability of the native state decreases. The stability of the native state can decrease for a number reason ranging from chemical modification, breaking of disulphide bonds or mutations as in this case. The free energy differences, that came out from our calculation ranges from around 3.5 to 5.0 kcal/mol, which is consistent with the above hypothesis[19],[20].

2.3 The per residue free energy calculation can identify the mechanistic detail behind conformational pereference of a particular residue

We decomposed the calculated free energy into per residue decomposition as described in the method section. This helps us to identify important residues that stabilize a particular conformation. In this direction, we first calculate the per residue free energy for the GA95 and GB95 sequences. The per residue free energy, $\Delta\Delta G((4\beta + \alpha)$ - $(3\alpha))$ is plotted in Figure 4(F). In this plot, the negative peak indicate the residue will stabilize the $4\beta + \alpha$ conformation of a particular residue and the positive peak will stabilize the

3α structure. As, the differences of amino acid residues are only at 3 position, there are some common features of residues in both GA95 and GB95 sequences in both 3α and $4\beta + \alpha$ conformation. Role of some important residues in stabilizing either $4\beta + \alpha$ or 3α conformation and possible reason for such conformational preferences are discussed in details in Figure 4 and Table 2. It is important to note that experimental observations classified the protein into two parts: Amino acids 9-51 are fully structured in both the folds, where as residues 1-8 and 52-56 are unstructured in 3α, but form β strand in $4\beta + \alpha$ structure. Most of the amino acid residues in the region 1-9 favor $4\beta + \alpha$ structure as can be seen from the plot Figure 4(F). Interestingly, we can explain most of the big peak in this plot. They are presented in Figure 4 and Table 2. Here, we explain only one case. The first big peak appears at position 7. Here, the hydrophobic Leu-7 stabilizes the $4\beta + \alpha$ conformation as its sidechain is oriented towards the protein hydrophobic core, whereas in 3α structure it exposed to the solvent (Figure 4(A)). Thus, if someone want to better design the 3α conformation of this protein, we propose to mutate Leu at this position with some hydrophilic residue. Similarly, we found that the free energy differences arise in two conformation for various reasons ranging from hydrogen/salt bridge formation to relative solvent exposure of hydrophobic residues etc. They are listed in Figure 4 and Table 2.

This is not Reader Friendly at all. we do not have to describe every little detail. I would rather have a table where we have the pairing and the reason for favoring one structure or the other. In the main text I would say that the main reason for preferences of one fold or another has to do with the relative solvent exposure of the residue in one conformation or another, or the possibility of stablishing hydrogen/salt bridges in one conformation. This would be my preference, ditching everything below until the next section. What do you think?

I have rewrite this part.There was a table. I made it better. I want to keep some discussion though if you agree.

Now, let us discuss the differences of 3α and $4\beta + \alpha$ fold due to specific difference of mutated amino acid residues in GA and GB sequence. The differences are in positions 20, 30 and 45. While GA95 has Leu-20-Ile-30-Leu-45, GB95 sequence has Ala-20-Phe-30-Tyr-45 at those position. One can find that the peak at position 49 stabilizes the GA sequence more

than that of GB sequence. In 3α fold, Ile-49 has hydrophobic interactions with Leu-20 and Ile-30 with GA sequence, whereas with GB sequence it has such interactions with smaller Ala-20 (Figure 4(I). In 3α, with GB sequence, Phe-30 being a larger group can not be completely accomodated in the hydrophobic core(Figure 4(I)) and therefore, partially exposed to the protein surface. There is a very interesting difference for residue at position 45 as well. In GB95 sequence and $4\beta + \alpha$ fold there is H-bonding between Tyr-45 and Asp-47 (Figure 4(K)), which is absent in the 3α conformation. On the other hand with GA conformation, this H-bond is absent for both the conformation as there is a Leu at position 45 (Figure 4(J)). Although, one can not find much difference between $4\beta + \alpha$ and 3α conformation for position 20 for both GA95 and GB95 sequence. There are significant difference for the residues at position 30 and 45. These residues have some long term effects as well. As already explained, residue Ile-49 has much more hydrophic interactions with residues at position 20 and 30 in GA sequence, which is less in GB sequence. There are difference for Asp-47 as well. With 3α conformation and GA95 sequence, Asp-47 has H-bonding with Lys-50, which is absent in the other conformation.

[Table 2 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

2.4 The confinement method correctly identifies the most native-like predictions from a subset of CASP predictions

This method is particularly well suited to pick out the native/native like structure from misfolded or decoys. We have tested this using different models from CASP (critical assessment of structure prediction). CASP is a blind test in which different groups world wide apply their methods to predict the 3D structure of proteins given their sequence. This is done with a strict 3 week limit on each target (3 day for servers) and each group is allowed to submit 5 possible structures (ranked from better to worst). In our

role as assessors during the CASP8 and CASP9[23] competition we observed no correlation at all between the ordering of structures submited by the groups and the real ranking compared to an experimental model[24]. The consequences of this go beyond those five structures; the deeper meaning is that groups producing ensembles of structures are generating structures that are better than the ones they submit, but they do not know about it. In fact, when querrying different groups after the results were known, most groups agree that this is the case. Beyond the CASP problem, this reflects on the modelers ability to correctly rank order models in many different environments, from structures for drug design leads to designing more stable proteins or peptide mimetics such as peptoids. One of the main culprits of this lack of accuracy is the fact that ranking is often done via a potential energy function, which in many cases lacks an entropy component. Other initiatives including knowledge based potentials do use some sort of free energy to rank order, but its accuracy is not enough. Our method provides a physics based solution to this problem. We have tried two experiments centered on CASP. First we tried to rank order some structures from the previous CASP9 experiment. We have tested different options for that purpose. First, we rank order submition from same group. Then we rank submition from different groups. Our second test was to check whether this method can do better quality assessment/metaprediction than the best group in that catagory in CASP9. In both methods, our ranking is determined as a free energy measure, and it is compared with the ranking given by a geometrical comparison between native and submitted models called GDT_TS (Global distance test score[25]). GDT_TS represents the average percentage of residues that are in close proximity in two structures optimally superimposed using four different distance cutoffs (1, 2, 4 and 8 Å).

We have chosen the inital models on which to test the methods based on server groups that have traditionally done good in past CASP events. There is a high correlation between the structures that are chosen for this method and its predictive accuracy. In particular, when GDT_TS scores are below 50, it is difficult for us to say anything about the models. Our approach is simple, with every subset of structures we choose, we use the confinement method in a pairwise procedure, and then rank the structures. In all cases, we have calculated the free energy with respect to the native structure either obtained from NMR or x-ray crystallography but we do not need to

make comparisons to native in order to do rank ordering.

2.4.1 confinement method can rank different models submitted by the same group and distinguish the native structure

The question that we want to answer here is whether we can rank models submitted by the same group.In all cases, we have calculated the free energy difference between the native structure and the submitted models as well. First we choose a protein BVU3908 from Bacteroides vulgatus whose PDB id and CASP target code are 2L01 and T0559 respectively. The native structure of this 69 residue protein was solved using NMR. The best predictor group for this target was "BAKER-ROSETTASERVER". We initlally checked similarity between the models submitted by "BAKER-ROSETTASERVER" and discarded two of them from the analysis on the basis of being too similar to some of the other models. The GDT_TS score and rmsd values as shown in Figure 4 indicate that the model 1 was predicted correctly, whereas the order of model 3 and 5 was wrong. Main difference between the model 3 and the rest of the model is that the orientation of first alpha helix of model 3 is opposite. On the other hand, as shown in Figure 4 confine and release method not only can differentiate the native structure from the submitted models , it can also correctly rank all the models like their The GDT_TS score.

To further test the method we have taken the example of protein BT2368 from Bacteroides thetaiotaomicron. The PDB id of this 74 amino acid residue protein is 2L02 and CASP target code is T0560. We have compared the free energy difference between the native structure and the two of the five submitted models from the group "Splicer". The remaining three models are discarded as again they are similar to the rest of the models. The comparison of GDT_TS score as shown in Figure 7 and published in final CASP result is only for models with residues 3 to 66. In order to keep consistency, we have also done our analysis with models and crystal structure consisting residues 3 to 66. As expected, the native state is identified correctly, and the two other models are ranked in the correct order in accordance with the GDT_TS score.

2.4.2 Confinement method can rank models submitted by different groups

Our next test was between models from the best submissions of different groups. For this purpose we have chosen the x-ray crystallographic structure of fas apoptosis inhibitory protein molecule whose pdb id and CASP target codes are 3MX7 and T0540 respectively. This protein contain 90 amino acid residues with 8 beta strands. We have chosen best models from groups ”LTB” and ”Mufold” for our analysis, which are labelled as Model 1 and Model 2 in Figure 8. As presented in Figure 8 we found that this method is able to correctly order the models which matches with the GDT_TS score.

2.4.3 Per residue free energy calculation identify the residues that are responsible for difference between two conformation

In order to understand the per residue contribution towards the free energy, we have first calculated the total free energy of NMR structure of a domain of adhesion exoprotein from Pediococcus pentosaceus and its submitted CASP9 model. The pdb id and CASP code of this 79 residue protein are 2KWY and T0569 respectively. We have compared the native structure with the best model structure submitted for this target in CASP9. This model structure was submitted by the group ” Mufold”. All the five model structures submitted by this group was found to be close, within a GDT_TS score of 72 to 78. Thus we proceed by comparing only the native structure and the best predicted model of this target. Visualization upon superimposition of two structures reveals that the difference between the two structure is at the region consist of residues 48 to 65, where model1 is much more disordered compare to the native structure(see Figure 10). Other regions look similar for both the cases with of model1 is within 2.6 Åbackbone rmsd value of the native structure. We found that, confinement method can still differentiate between the native structure and the generated model. We have also calculated the per residue free energy with a aim that this can identify the residue which are responsible for such free enerfy difference. The per residue free energy decomposition is shown in Figure 11(A). We indeed found that two hydrophobic residues Val-59 and Ile-61 are oriented towards the solvent in the protein surface in the generated model (Figure 11(B). This the region, where there is a beta sheet

in the original crystal structure, which is absent in the model. There are also some region, which stabilizes the generated model compare to that of the crystal. A close inspection reveal that the hydrophilic Glu in position 67 is completely exposed to the protein surface and solvent in the best model, whereas it is partially exposed in the crystal. This is presented in Figure 11(C). The peak at position 76 is due to the H-bond between Lys-76 and Asp-11, which is absent in original crystalographic structure (Figure 11(D)).

2.4.4 Confinement method fails ?

There are however some instances where the method can fail, specially when GDT scores and all the structures are very close. We have found this for an engineered protein from Asr4154 protein (PDB ID: 2L09 and CASP Target T0538). This protein contains 54 amino acid residues with three alpha helixes and two very short beta strand. The model that is closest to the native structure was generated by the group PconsR with a GDT_TS value of 96.23. We have also considered one model from group "Shell" (GDT_TS = 90.09) and "FOLDIT" (GDT_TS = 86.32) for our analysis along with the native crystal struture. The models from PconsR, Shell and FOLDIT are labelled as model 1, model 2 and model 3 respectively in Figure 5. As seen in Figure 5, one can find that the model from PconsR has lowest free energy value, making it a more stable structure. Although ranking of the other models remain consistent, it may happen that the theoretical model produced by the group "PconsR" much more stable than the crystal structure. This is particularly difficult target for rank ordering as all the models those are considered for this calculation are within RMSD values of 1.6 - 2.0 Å. In order to understand the reason behind the mismatch of GDT_TS values of crystal and the model1, we have decomposed the total free energy into the per residue component. A close analysis of the per residue free energy value i(data not shown) reveal that there are number of sidechains oriented in a very different way in crystal and different model. This may account for differences in free energy values, inspite of close RMSD and GST_TS values. As shown in Figure 6(A), there is a salt bridge between Arg-32 and Glu-35 present in the crystallographic structure, which is absent in the model 1. In model 1, Arg-32 has another salt bridge with Glu-28. Glu-28 in crystal has the salt bridge with Lys-

24. In model 1, Lys-24 is exposed towards the protein surface. This is shown in Figure 6(A). There are more such interesting and may be crucial sidechain differences between the model and the crystallographic structure, which indicate the limitation of the GDT_TS based ranking of the protein structure. As, shown in Figure 6(B), Arg-26 and Glu-50 has a salt bridge in the crystallographic structure. Also, Phe-51 is oriented towards the protein core in the crystallographic structure. In model 1, the salt bridge between Arg-26 and Glu-50 is absent and the Phe-51 is oriented towards the protein surface.

2.4.5 What can we say about low resolution models?

We mentioned earlier that our predictive abilities are greatly decreased with low model qualities. It is better to have at least one good model to pick out the native/native like structures from the rest of the decoy. In order to test this hypothesis we performed another test with target T0531 of CASP9, in which we compared the crystal structure to five models from the MUFOLD server. Pdb id of this extracellular domain of the jumping translocation breakpoint protein is 2KJX and CASP target code is T0531. Looking at Figure 9 shows: 1.) native is correctly identified as expected and 2.) Surprisingly there is a high level of correlation between the GDT ranking and the free energy ranking for model 1 and model 3, the rest three structures with GDT_TS score less than 35 are ordered incorrectly. It is worth to note that models 2 and 3 have the same GDT and very different free energies, meaning that the actual ordering could change a lot[26]. It is encouraging that at least the method can pick out the best model with low GDT score of 44.

2.4.6 Can confinement method perform better quality assessment in protein structure prediction ?

A part of CASP experiment is dedicated to the assessment of the quality of predicted model[24]. Most of the top performing group in this catagory use consensus approaches for quality assessment[27]. A review of CASP9 quality assessment also pointed out that the methods which are based on clustering techniques perform much better compared to those methos

which are based on analysis of individual models[24]. In this background we try to answer, whether the confinement method can do better quality assessment compare to the best performing group (MUFOLD-WQA) in this catagory. In CASP9, the overall reliability of models in quality assessment (QA) was accepted in a mode known as QMODE 1. Here, predictors were asked to score each model on a scale from 0 to 1, with higher values corresponding to better models[24]. We have again picked up the CASP taget T0538 to answer this question. We have chosen the Model 3 submitted by PconsR (GDT_TS = 96, QMODE 1 = 0.5434) and model 5 from the MULTICOM-NOVEL (GDT_TS = 83, QMODE 1 = 0.5865). Both are server predicted models and they were further used for quality assessment by the group MUFOLD-WQA[27]. It is important to note that the model from PconsR was the most accurate predicted model for this target and the model from MULTICOM-NOVEL was predicted as best by the quality assessment experiments of MUFOLD-WQA. Also, the QMODE 1 value presented here is from the group MUFOLD-WQA. Our free energy analysis predict that the model from Pconsr is indeed better than the model from MULTICOM-NOVEL ($\Delta\Delta G(Model\_Pconsr - Model\_MULTICOM - NOVEL = -3.9Kcal/Mol)$, which matches well with the GDT_TS score. On the other hand the MUFOLD-WQA predicted the ranking in the reverse order.

We have also carried out the similar type of analysis for the case of Target 560 of CASP9 experiment. For that we have chosen the model 1 from the group PconsR (GDT_TS = 94, QMODE 1 = 0.5178), BilabEnable (GDT_TS = 87, QMODE 1 = 0.5344) and Multicom_Refine (GDT_TS = 74, QMODE 1 = 0.4770). If, one want to rank order with respect to the GDT_TS values, the model from PconsR is the best predicted model, but according to the QA assessment by MUFOLD-WQA, the model from Bilab-Enable was found to be best predicted model. All the QMODE 1 values are from MUFOLDWQA. In CASP 9, the sequence of 74 residues were given to predict the structure and for quality assessment, whereas during GDT_TS analysis only residues 3-66 were considered. When we considered the full 74 amino acid residues, the free energy calculated were ($\Delta\Delta G(Model\_Pconsr - Model\_BilabEnable = 3.3Kcal/Mol)$ and ($\Delta\Delta G(Model\_Pconsr - Model\_Multicom\_Refine = -30.0Kcal/Mol)$ which support the QA assessment by MUFOLD-WQA. On the other hand, if we calculate the same free energy with the trimmed version of the protein (with residues 3-66), the values of the free energy difference become -1.7

and -31.1 Kcal/Mol respectively, which support the GDT_TS score. The summary of this whole observation is presented in Table 4. There is no doubt that the consensus approach can predict the model quality in a faster manner. But we expect that confinement method can predict the model quality in a relatively expensive but much accurate way.

[Table 3 about here.]

3 Conclusion

In recent past there are number of studies directed towards understanding protein structure using theoretical tools. Experimentally, it is relatively easy to identify the native structure from the data obtained from x-ray crystallography or NMR. On the other hand using theoretical tools it is often difficult to identify the native structure using both physics based methods and knowledge based method. In order to rank structure from the available data, some groups use clustering algorithms and the structure of the most populated cluster is picked out as the most probable structure. However, it may happen that the structure may be trapped and spend most of its time in a kinetic trap, which may give a wrong interpretation of the clustering result. On the other hand, some groups also depend on potential energy functions to pick out their best structure, lacking an entropic component. In this work we attempted to identify the native/native like protein strcuture from the decoy using free energy as a scale. We have used previously developed confinement method to larger systems. We found that in majority of the case, the native structure is always the most stable and we can rank the protein structure upto 100 amino acid residues long. This method works very well if atleast one of the strcuture in the decoy is good. Main advantage of this method is that it does not require any reaction coordinate. Thus one can calculate the free energy difference between two very different conformations. An interesting application, that we have developed using this method is per residue free energy calculation. Although, we have calculated that in an approximate way, but it give important information regarding conformational preference of a particular residue. In future, this may help for protein design. First, we have tested this method for a series of chameleon sequences, which have

88, 95 and 98 percent sequence identity but one of them has a 3α structure and another one has $4\beta + \alpha$ structure. We found that the method can pick the the conformational preferance of a particular sequence, which matches with the experimental results. Per residue free energy calculation identify the residues that stabilize/destabilize a particular conformation and can help for better protein design. Encouraged by that result we have tested this method for different targets of CASP experiments. We found in most cases, it can pick out the native/native like structure in terms of lowest free enrgy value and rank the protein structure with terms of free energy. The method can be also give wrong result if all the models in the decoy are very poor. It may also be difficult to rank structures if all the structures are very very close to the native structure. But that does not matter at all as our main aim is to differentiate between the good and the bad structure. In this article we have extensively tested this method using examples from CASP experiment. This method can be computationally expensive in actual CASP experiment if one want rank hundreds of submitted models in CASP experiments. However, in real life experiment with a particular target, we expect this method can give accurate result. The important fact about confinement method is that all the part of the calculation can be done independently. There is another big bottlenecks in this calculation. This has to do with the amount of simulations that have to be performed to confine the ensemble of conformations close to a microstate into a single microstate. We have tackled this issue by using GPU (Graphical Processing Unit) technology as opposed to the classical CPUs, which gives us a two order of magnitude boost in computational efficiency. For example, we have mostly carried out our calculation with Amber program[28] in GPU computer. With a avarage 56 residue we found that it take only 4 hour to complete 20 ns of the confinement step. So, with available computer power this method will be easy to use for real problems.

This method can be further used for calculating free energy due to very large conformational change. Another application may be for the case of ligand binding. It can also be used to distinguish the most stable structure during protein design. Decompition of total free energy into the per residue component help identifying residues that give stability to a particular conformation. Another, important application may be to study the protein folding kinetics, specially to study the phi value analysis.

4 Method

The confinement method has been described in details in ref. by Tyka et.al.[14] and Cecchini et. al.[15]. The basic approach of confinement approach is same in both these paper. However, there are some technical differences. Here we briefly describe what we follow to compute free energy ΔG_{AB} between A and B conformation of the protein.

1. We first minimized both the conformation, A and B. These minimized conformation(A* and B*) are the reference conformation of that state

2. Backbone dihedral angle of both the full protein is restrained by a harmonic potential. This is required to keep the configurational state of the protein. This approach is particularly important, so that the any misfolded part of the protein does not chage its configuration.

3. The free energy is calculated using a thermodynamic cycle. The first part of this approach is the confinement, where the state A and B are confined to the reference state A* and B*. This is done gradually by applying larger and larger harmonic restraint on all the atoms of the biomolecule. This is done by running 21 molecular dynamics simulation each of 20 ns long, where the harmonic restraint constant was varied from 0.00005 to 81.92 untill the free and the restrained state overlap well. The final restrained state was chosen so high so that the rotational contributional of the protein frozen out and only remaining contribution remain that of vibrational free enrgy. The fluctuation from the reference structure was recorded and the free energy is calculated using a numerical approach developed by Tyka et. al.[14]. The confinement free energy calculated in this way recorder as $\Delta G_{A,A*}$ and $\Delta G_{B,B*}$ as shown in Figure 3.

4. Finally the thermodynamic cycle is closed by calculating the free energy between the final restrained state A* and B* using normal mode analysis or quasiharmonic analysis. The free energy calculated in this way is shown as $\Delta G_{A*,B*}$ in Figure 3.

5. The full free energy, $\Delta G_{A,B}$ between the two state A and B is calculated using the equation $\Delta G_{A,B} = \Delta G_{A,A*} - \Delta G_{B,B*} + \Delta G_{A*,B*}$

In all the calculation amber 11 program is used with ff99SB with GB/SA implicit solvent. Interestingly, we extend the method for calculation of per residue free energy in an approximate way. For this purpose, the confinement energy, $\Delta G_{A,A*}$ and $\Delta G_{B,B*}$ of each residue is calculated in the usual numerical way as described above. The internal energy of each residue is calculated using the decomp module of amber from the final restrained trajectory. We call this method approximate as we ignore the contribution from the normal mode or quasiharmonic analysis. However, this contrbution to the total free energy is very less (less than $1.0Kcal/Mol$), which allow us to study the mechanistic details of conformational preference of each residue.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

References

[1] Meirovitch, H. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. Current Opinion in Structural Biology, 2007, 17, 181-186.

[2] Chipot, C.; Shell, M.S.; Pohorille, A. Introduction, in Chipot, C., Pohorille, A., editors. Free Energy Calculations: Theory and Applications in Chemistry and Biology. Springer Series in Chemical Physics, vol. 86. Berlin and Heidelberg: Springer; 2007, p. 132.

[3] Jorgensen, W.L. The many roles of computation in drug discovery, Science 2004, 303, 18138.

[4] Gilson, M.K.; Zhou, H.X. Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct. (2007) 36, 21-42.

[5] Torrie, G. M.; Valleau, J. P. iNonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling (1977) J. Comput. Phys. 23, 187

[6] Tironi, I.G.; van Gunsteren, W.F. A molecular-dynamics simulation study of chloroform. Mol. Phys. (1994) 83, 381-403.

[7] Dill, K.A.; H.S. Chan. From Levinthal to Pathways to Funnels: The "New View" of Protein Folding Kinetics. Nature Structural Biology 4, 10-19 (1997)

[8] Dill, K.A.; Ozkan, S.B.; Shell, M.S.; Weikl, T.R. The protein folding problem. Annual Review of Biophysics (2008), 37, 289-316.

[9] Anfinsen. C.B. Principles that Govern the Folding of Protein Chains. Science (1973) 181, 223-230.

[10] Christ, C.D.; van Gunsteren, W.F. Enveloping distribution sampling: A method to calculate free energy differences from a single simulation, J. Chem. Phys. (2007), 126, 184110.

[11] Ytreberg, F.; Zuckerman, D. Simple estimation of absolute free energies for biomolecules. J. Chem. Phys. 2006, 124, 104105.

[12] Park, S.; Lau, A.; Roux, B. Computing conformational free energy by deactivated morphing. J. Chem. Phys. 2008, 129, 134102

[13] Zheng, L.; Chen, M.; Yang, W. Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems, Proc. Natl. Acad. Sci. 2008, 105 (51), 20227.

[14] Tyka, M.; Clarke, A.; Sessions, R. An Efficient, Path-Independent Method for Free-Energy Calculations. J.Phys.Chem. B 2006, 110, 17212-17220.

[15] Cecchini, M., Krivov, S.V., Spichty, M., Karplus, M. Calculation of free-energy differences by confinement simulations. Application to peptide conformers. J. Phys. Chem. B 113, p. 9728-9740 (2009).

[16] Strajbl, M.; Sham, Y.Y.; Vill, J.; Chu, Z.-T.; Warshel, A. Calculations of Activation Entropies of Chemical Reactions in Solution. (2000) 104, 4578-4584.

[17] Krivov, S.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding Proc. Natl. Acad. Sci. U.S.A. 2004, 101, (41), 14766.

[18] Alexander, P.A.; He, Y.; Chen, Y.; Orban, J. Bryan, P. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc. Natl. Acad. Sci. 2007, 104 (29), 11963-11968.

[19] He, Y.; Chen, Y.; Alexander, P.A.; Orban, J. NMR structures of two designed proteins with high sequence identity but different fold and function. Proc. Natl. Acad. Sci. 2008, 105 (38), 14412-14417.

[20] Alexander, P.A.; He, Y.; Chen, Y.; Orban, J. Bryan, P. A minimal sequence code for switching protein structure and function. Proc. Natl. Acad. Sci. 2009, 106(50), 21149-21154.

[21] Shortle, D. One sequence plus one mutation equals two folds. Proc. Natl. Acad. Sci. 2009, 106(50), 21011-21012.

[22] Sheffler, W.; Baker, D. RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Science. 2009, 18(1), 229-239.

[23] MacCallum, J.; Perez, A.; Schnieders, MJ.; Hua, L.; Jacobson, M.P.; Dill, K.A. Assessment of protein structure refinement in CASP9. Proteins, 2011, 79, 74-90.

[24] Kryshtafovych, A.; Fidelis, K; and Tramontano, A. Evaluation of model quality predictions in CASP9. Proteins, 2011, 79, 91106

[25] Zemla, A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003, 31, 33703374.

[26] Perez, A.; Yang, Z.; Bahar, I.; Dill, K.A.; MacCallum, J.L.; FlexE: Using Elastic Network Models to Compare Models of Protein Structure. J. Chem. Theory Comput., 2012, 8, 3985-3991.

[27] Wang, Q.; Vantasin, K.; Xu, D.; Shang, Y. MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction. Proteins, 2011, 79: 185195.

[28] Case, D.A.; Cheatham, III, T.E.; Darden, T.; Gohlke, Luo, H.R.; Merz, Jr., K.M.; Onufriev, A; Simmerling, C.; Wang, B.; R. Woods, R. The Amber biomolecular simulation programs. J. Computat. Chem. (20005) 26, 1668-1688.

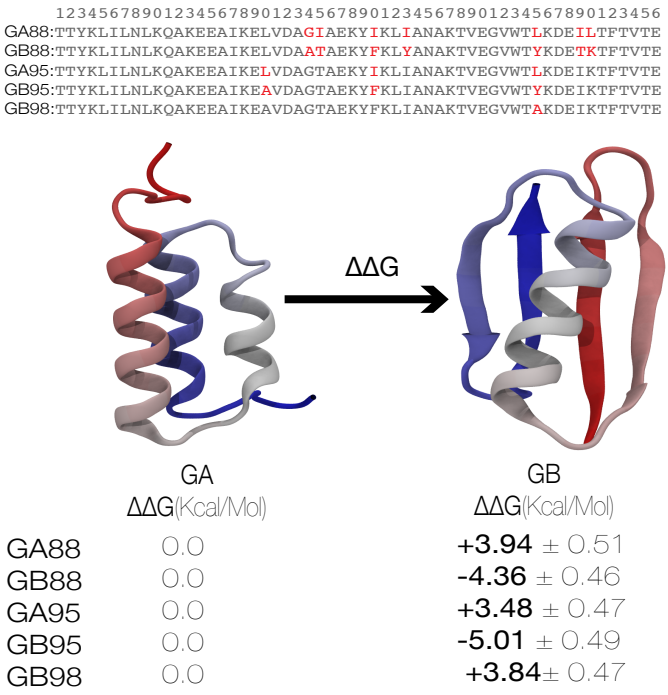

Figure 1: Confinement method correctly predicts the structural preferences of six chameleon sequences. (A) The six sequences used in this study. (B) Each sequence adopts either a Protein G-like fold (denoted GB) or a three-helix bundle fold (denoted GA). The relative free energies of the two folds are reported for each sequence.

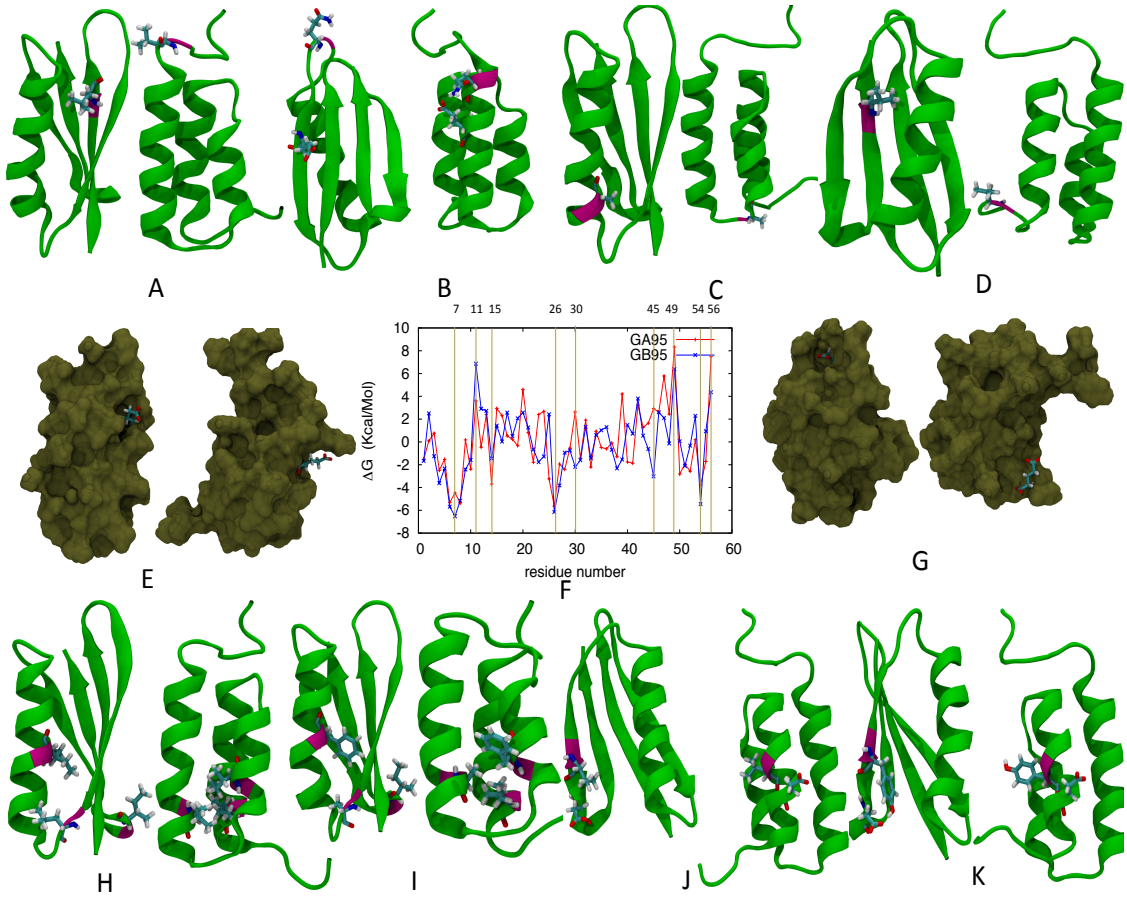

Figure 2: Sidechain orientation of some important residues which stabilize/destabilize either of the $4\beta + \alpha$ or 3α conformation. The details are in Table 2 and the main text. A. Leu-7. B. Gln-11 has H-bond with Glu-15 in 3α, which is absent in $4\beta + \alpha$ C. Ala-26 D. Val-54. E. Glu-14. F. plot of free energy difference for GA95 and GB95 sequence. Positive peak stabilize 3α form and negative peak stabilize $4\beta + \alpha$. G. Glu-56 . H. sidechain orientation of Hydrophic Ile-49, along with Leu-20 and Ile-30 with GA95 sequence I. sidechain orientation of Hydrophic Ile-49, along with Ala-20 and Phe-30 with GB95 sequence J. Tyr-45 forms H-bond with Asp-47 with GB95 sequence and $4\beta + \alpha$ conformation K. No H-bond possible between Leu-45 and Asp-47 with GB95 sequence and $4\beta + \alpha$ conformation .

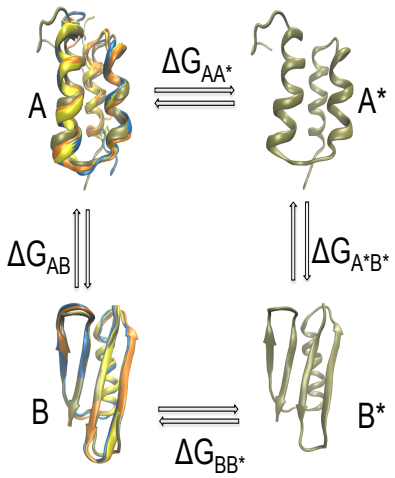

Figure 3: Graphical representation of the thermodynamic cycle involving confinement method.

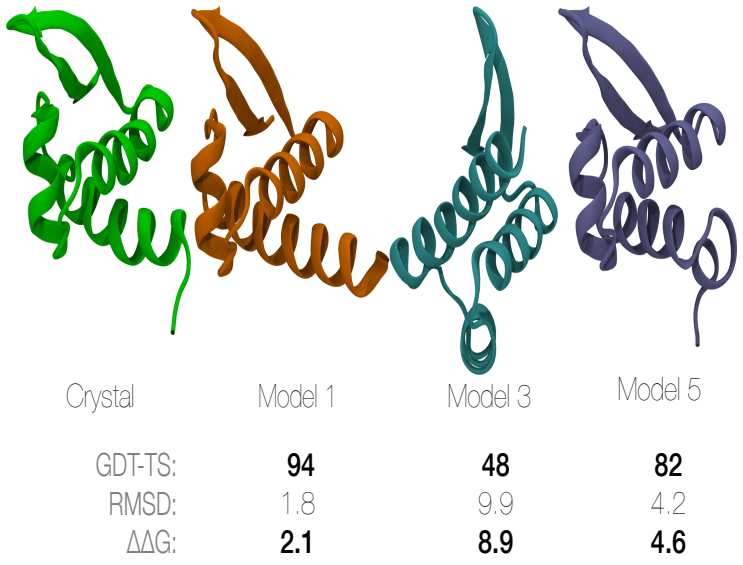

Figure 4: The native and three submitted model structure, along with their GDT_TS, RMSD and relative Free energy values of protein BVU3908 from Bacteroides vulgatus (PDB id: 2L01 and CASP code: T0559).

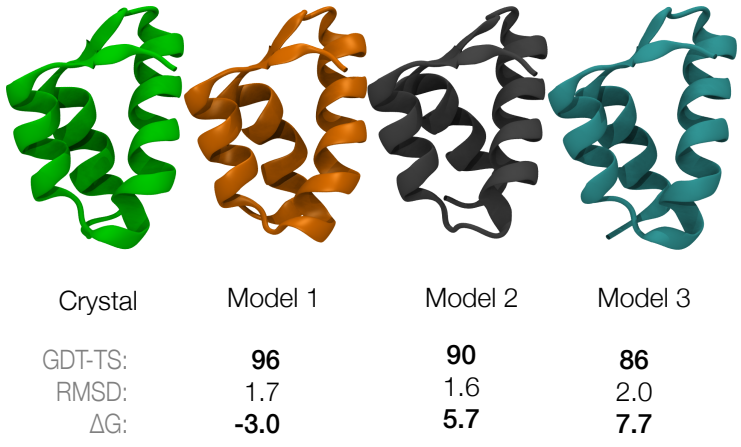

Figure 5: The native and three model structure of engineered protein from Asr4154 protein (PDB ID: 2L09 and CASP code:T0538). The model 1,2 and 3 are from the group PconsR, Shell and FOLDIT respectively.

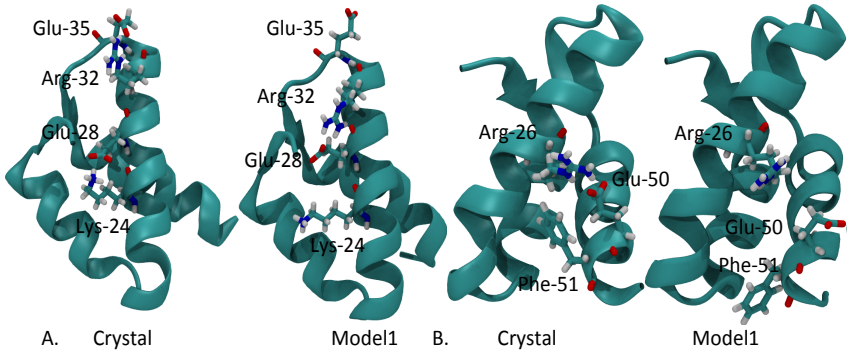

Figure 6: Difference of the sidechain orientation between model 1 and crystallographic structure in Target-538. A. The salt bridge between Glu-35 and Arg-32 in crystal structure. In model 1, this is compensated by H-bonding between Arg-32 and Glu-28. There is another salt bridge between Glu-28 and Lys-24. B. Salt bridge between Arg-26 and Glu-50 in crystal structure. Orientation of Phe-51 is different in crystal than in model1.

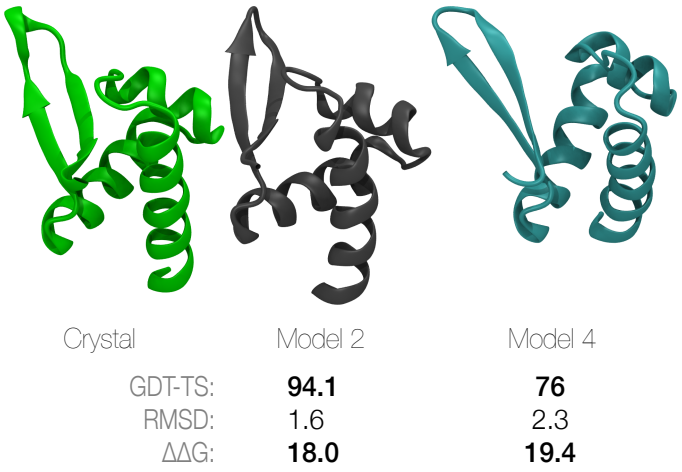

Figure 7: Native and two model structure of protein BT2368 from Bacteroides thetaiotaomicron (pdb id: 2L02 and CASP code: T0560). The two models were from the group ”Splicer”.

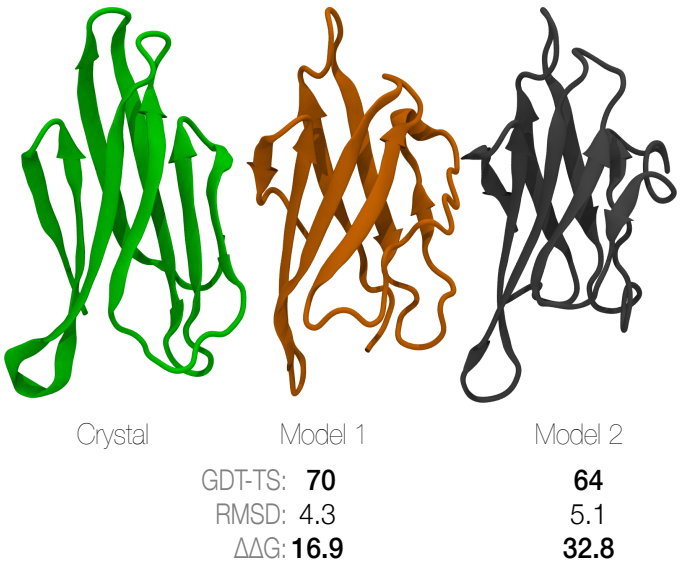

Figure 8: X-ray crystallographic structure and two submitted models of fas apoptosis inhibitory protein (pdb id: 3MX7 and CASP code: T0540). Model 1 and Model 2 in this analysis were submitted by the group LTB and MUFOLD respectively.

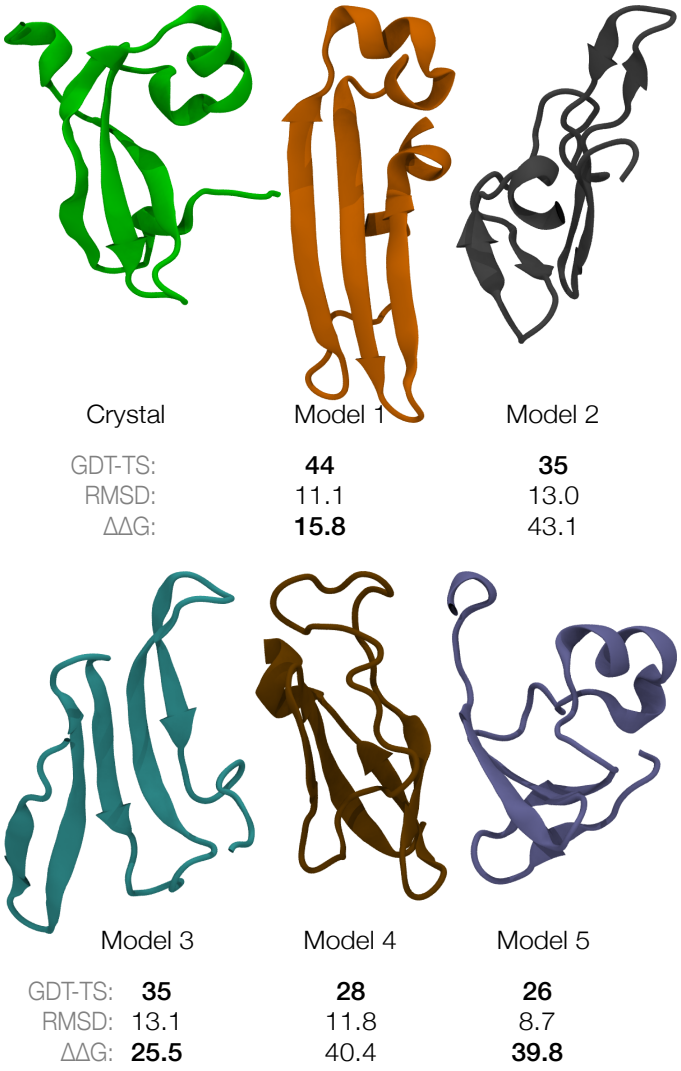

Figure 9: The native structure and 5 models of extracellular domain of the jumping translocation breakpoint protein (pdb id: 2KJX and the CASP code: T0531).

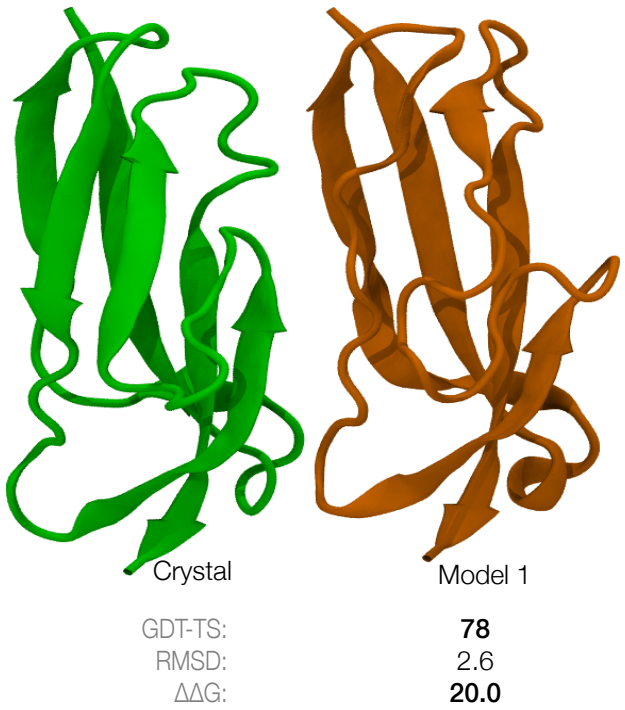

Figure 10: Native and best model structure of a domain of adhesion exoprotein from Pediococcus pentosaceus (pdb id: 2KWY and CASP code: T0569). The best model was from the group ″Mufold″.

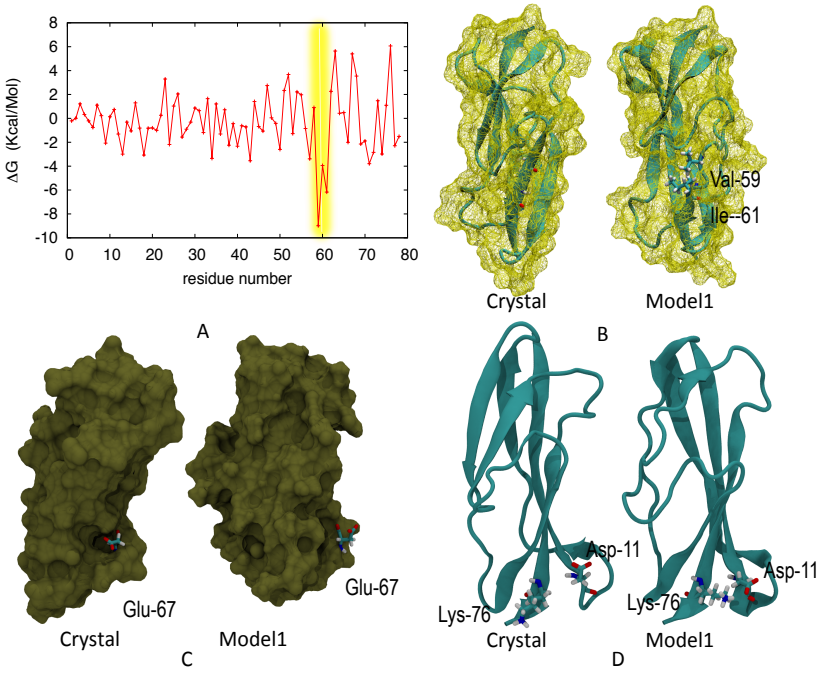

Figure 11: A. Plot of per residue free energy difference between the model and the crystal structure of Traget 569 (pdb id: 2KWY) of CASP9. B. In model 1, the beta sheet containing Val-59 and Ile-61 is disordered. These hydrophobic residues are exposed to the solvent. C. Glu-67 is partially exposed to the solvent in crystal and it is fully exposed in model 1. D. There is a H-bond between Lys-76 and Asp-11 in model 1, which is absent in the crystallaographic structure.

Table 1: The confinement method assigns a more favorable free energy to the experimentally determined structure than to computer-generated predictions. For each target, we examined as many as five predictions submitted by CASP participants. Positive $\Delta\Delta G$ values indicate that the experimental structure is predicted to be more favorable than any of the decoys.

CASP Target	PDB Identifier	$\Delta\Delta G_{native \to bestdecoy}$ (kcal/mol)
T0531	2KJX	11.15 ± 0.70
T0538	2L09	-3.00 ± 0.47
T0540	3MX7	16.94 ± 0.49
T0559	2L01	2.10 ± 0.24
T0560	2L02	18.00 ± 0.49
T0569	2KWY	20.01 ± 0.69

Table 2: Analysis of per residue free energy decomposition of GA95 and GB95 sequence reveal that certain residues prefer either 3α or $4\beta + \alpha$ structure. These observations and the reason behind such conformational preferences are listed in this table.

Information from per residue plot	What is happening in 3α conformation	What is happening in $4\beta + \alpha$ conformation
Leu-7 favors $4\beta + \alpha$ (Figure 4(A))	hydrophobic residue is exposed to the solvent	oriented towards the hydrophobic core
Gln-11 stabilizes 3α (Figure 4(B))	Gln-11 has H-bond with Glu-15	no such H-bond
Glu-14 favors $4\beta + \alpha$ (Figure 4(E))	hydrophilic residue partially exposed to the solvent	completely exposed to the solvent
Ala-26 favors $4\beta + \alpha$ (Figure 4(C))	hydrophobic residue oriented towards surface	oriented towards the hydrophobic core
Ile-49 favors 3α (Figure 4(H) and (I))	hydrophobic interaction inside the protein	sidechain exposed to the solvent
Ile-49 favors GA more than GB (Figure 4(H) and (I))	more hydrophobic interactions with residues at 20, 30 in GA	unchanged
Val-54 favors $4\beta + \alpha$ Figure 4(D))	hydrophobic residue exposed to the solvent	hydrophobic interaction inside the protein
Glu-56 favors 3α (Figure 4(G))	hydrophilic residue partially exposed to the solvent	completely exposed to the solvent
Phe-30 in GB sequence stabilizes $4\beta + \alpha$	exposed to the solvent in protein surface	oriented towards the hydrophobic core
Tyr-45 in GB favors $4\beta + \alpha$ (Figure 4(K))	No H-bonding	H-bond with Asp-47
Asp-47 in GA sequence favors 3α	H-bond with Lys-50	No H-bond, No Tyr-45 in GA sequence

Model	$\Delta\Delta G$ (1-74)	$\Delta\Delta G$ (3-66)	GDT_TS	QMODE 1
PconsR	0.0	0.0	94	0.5178
BilabEnable	-3.3	1.7	87	0.5344
Multicom_Refine	30.0	31.1	74	0.4770

Table 3: Comparison of quality assessment by the group MUFOLDWQA and the confinement method of the CASP9 target 560 (pdb id: 2L02). The GDT_TS score obtained from CASP9 website, was based on the trimmed model containing residue 3-66, whereas, the QMODE 1 values are from quality assessment by the group MUFOLDWQA and based on residues 1-74.