Predicting the conformational preferences of proteins using a physics-based free energy method

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

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

April 9, 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

Protein structure and conformational changes are of central importance in biochemistry. The populations of different conformational states are determined by their free energies. The ability to compute these free energies would allow us to understand and make predictions about a wide variety of biochemical processes—including the mechanism of enzymes, the function of structural proteins, binding of proteins to other proteins or DNA, and protein folding.

Computer simulations have proven to be a powerful tool for studying biomolecular systems at atomic detail [ref]. However, calculating the free energies for conformational changes remains a challenging problem[1], primarily because the timescales for such conformational transitions may be too long to simulate directly. For example, in protein folding the free energy landscape is very rough[7;8], which limits conformational sampling and leads to large uncertainties in the relative populations of different states.

Even with advanced methods like replica exchange molecular dynamics [ref], the molecule may become trapped in a local minima. Sometimes this problem may be circumvented by defining a pathway or reaction coordinate connecting the states of interest. The free energy along this reaction coordinate can then be determined using methods such as umbrella sampling[5], targeted MD, or thermodynamic integration[6]. However, accurately sampling this reaction coordinate may still be difficult and require a large computational effort. In the case of protein folding, it is often difficult to even devise a meaningful reaction coordinate that could connect the unfolded and folded states efficiently.

The calculation of protein conformational free energy has been successfully attempted by a number of groups[1;11;13]. In recent years, methods such as the reference system method[11], deactivated morphing[12], orthogonal space random walk[13], and the confinement method[14;15] have been used to calculate free energy differences. Include accelated MD, metadynamics

In this work we explore the confinement method[14;15], which relies on the fact that the free energy difference is a state function and thus is independent of the sampling path. This allows us to use the alchemical thermodynamic cycle shown in Figure 3. First, the two end states are slowly "frozen" into a very tight, harmonically restrained ensemble and the free energies

for these two processes are calculated. Next, the free energy difference between the two restrained ensembles is calculated using normal mode analysis. This allows us to calculate the free energy difference between the two end states without needing to define a physical path or reaction coordinate connecting the two states.

The confinement method has previously been applied to several small proteins [refs]. Here, we use the power of GPU computing (using graphic cards to run calculations [cite amber gpu]), which allows us to extend the method to larger systems while maintaining reasonable computation times. We first demonstrate that our implementation reproduces the previously reported results for small peptides. Next, we have applied this method to larger proteins in two different case studies. In the first case study, we examine the conformational preferences of a series of chameleon proteins designed by Orban and co-workers that can switch between two completely different folds with only small changes in amino acid sequence. In the second, we attempt to pick the most native-like structures from a set of models produced during the Critical Assessment of Structure Prediction (CASP) experiment.

2 Results and Discussion

2.1 Control tests of the confinement method produce the expected outcomes

As a first step, we performed several control experiments to verify that our implementation of the confinement method produces the expected results.

The method has previously been applied to a 16-residue β-hairpin from Protein G, known as BHP[15]. We calculated the free energy difference between two different conformations of the peptide: bhp1, the native conformation with a two-stranded β-sheet; and bhp3, a non-native conformation with 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].

Previous applications of the confinement method have focused on relatively short peptides, up to 17 residues in length. In this work we apply the method to larger proteins. As a first test, we compared the free energies for several predictions submitted during the CASP9 experiment with the corresponding experimentally determined structures from the Protein Data Bank (PDB). As Table 1 shows, in 5 out of 6 cases, the confinement method assigns a lower free energy to the experimentally determined structure—as would be expected. For target T0538, one of the submitted models has a more favorable free energy than the crystal structure, however, the submitted model is very close to the crystal structure (GDT-TS=96) and the free energy difference is small. Target T0559 also has a small free energy difference and the predicted structure is also very close to the NMR structure (GDT-TS=94).

Arijit: please add GDT-TS to the best decoy to the table below

[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 family of 56-residue sequences that are stable in one of two possible folds. By mutating key residues they are able to stabilize one fold or the other (see Figure 1).

[Figure 1 about here.]

The two structures will be named according to the dominant secondary structure. One of the structures is a simple three-helix bundle and will be called "α". The other structure is similar to Protein G and will be called "β". The sequences will be named according to the structure that they adopt. Sequences that adopt the α fold will be called "GA", while sequences

that adopt the β fold will be called "GB". There are five sequences, that can be arranged in three pairs: GA88/GB88 are 88 percent identical and differ at seven positions, GA95/GB95 are 95 percent identical and differ at three positions, GA98/GB95 are 98 percent identical and differ only by a single tyrosine to alanine mutation at position 45. 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 sequences presents a serious challenge for computational methods[18;21].

We initially approached this problem by making a model for 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 β structure. We then used the confinement method to compute the free energy difference between the experimentally determined GA88/α structure and the GA88/β model. The confinement method was able to predict the conformational preferences correctly for all six sequences (data not shown). This is, however, not a particularly stringent test. It is well known that it is easy to distinguish computational models from native structures [ref]. To avoid this potential problem, we instead computed the relative free energy difference between two different computer generated models for each sequence—one α model and one β model (see methods for details). This is a much more realistic test of the confinement method's ability to accurately calculate free energy differences between two computer generated models.

The confinement method identifies the correct structures for all five sequences (Figure 1). One hypothesis for the mechanism of fold switching by chameleon sequences is that the structural transitions require states with diminished stability [ref]. It is believed that if the free energy of the native state and the alternative state are within about 5 kcal/mol, then it is possible to switch structural preference with only small changes in sequence. The calculated free energy differences range from around 3.5 to 5.0 kcal/mol, consistent with the above hypothesis[19;20].

Thic calculations have the potential to be used for protein design. We can use the confinement method to screen for different mutants to determine sequences compatible with a target structure. In this way we can concentrate the mutational efforts in a few regions of interest.

2.3 Per-residue free energy calculations can identify the details behind conformational preferences

To understand the details behind the conformational preferences of the different sequences, we decomposed the calculated free energy into per-residue contributions as described in the methods section. This allows us to identify important interactions that stabilize each fold. The per residue free energy, $\Delta G(\beta - \alpha)$ is shown in Figure **??** for both the GA95 and GB95 sequences with both the α and β structures. In this plot, residues colored in blue favor the α structure, residues in red favor the β structure, and white residues have no preference.

[Figure 2 about here.]

Although the overall free energy difference between the two structures is small (of the order of XXX), the individual residues can show marked preferences for being in either α or β conformation. Such differences can be undestood by looking in detail at the local environemnt for those resiudes. For example, the region around residue 7 forms a random coil in the α structure, whereas it is forming a beta sheet structure in the β structure. These residues strongly favour the locally well packed and hydrogen bonded environment found in the β sheet. Overall, favoring α or β structure is a delicate balance, were the relative global free energy difference is small and the contributions of different per residue tendencies balance out. It is therefore very likely that by changing key resiude preferences the global fold preferences can be changed.

Three residues are mutated between GA95 and GB95, at positions 20, 30, and 45. In GA95, these residues (L20, I30 and L45) stabilize the α structure. In GB95, two of these residues (F30 and Y45) favor the β structure, because they have large solvent exposed surface areas in the α structure, but are more buried in the β structure. Additionally Y45 forms a hydrogen bond with D47 in the β structure. On the other hand, residue A20 from GB95 still favors the α structure, although not as strongly as L20 in GA95.

The GA and GB sequences are nearly identical, so there are some common features observed for all sequences. Roles of some important residues in stabilizing either the α or β conformation and possible reasons for such

conformational preferences are discussed in details in Figure **??** and Table **??**. The experimental observations classified the protein into two parts: Amino acids 9-51 are fully structured in both folds, whereas residues 1-8 and 52-56 are unstructured in the α fold, but form a β-strand in the β fold. Most of the amino acid residues in the region 1-9 have negative per-residue free energies, which means that these residues favor the β structure.

In addition to the direct effects of the mutation, there are also indirect effects due to small perturbations due to the mutations. For example, the L20A mutation causes a slight repacking around residue 20. This causes large changes in the per-residue free energies of nearby residues T25 and A26. However, these differences largely cancel and the overall effect of the L20A mutation is weaker than either the I30F or L45Y mutations.

Such per-residue free energy decompositions allow us to understand the driving forces behind protein folding and conformational change and may be useful for designing proteins with specific structures and functions.

2.4 We have applied the confinement method to structure prediction

We have applied the confinement method as a "meta-predictor" for structure prediction. Here, the task is to correctly identify the most correct models out of a set of "decoys" generated by different methods during the Critical Assessment of Structure Prediction (CASP) experiment. CASP is a blind test in which different groups apply methods to predict the 3-dimensional structure of proteins from their sequences. This is done with a either a three week (for humans) or three day (for fully automated servers) time limit on each target. Each group is allowed to submit five possible structures, which they are supposed to rank from best to worst. We have performed two experiments centered on CASP.

In the first experiment, we examined several targets and tried to rank-order the predictions produced by other groups. In some cases the structures all came from the same group, in others the structures were drawn from several different groups. The goal of this experiment is to determine if the physics-based confinement method can correctly identify more native-like structures as having low free energies.

Our second experiment was to see if the confinement method can identify structures that are missed by other meta-prediction servers. Most successful meta-prediction servers are based on the idea of consensus: if many different prediction methods produce similar results, then that is probably a correct prediction. This is often a powerful heuristic, but it can miss cases where there is a very good result that is only predicted by one method. We chose several cases where such structures were missed and assessed if the confinement method can correctly identify these accurate models.

As is common in the CASP experiment, we assess our results in terms of Global Distance Test Total Score (GDT-TS)[25], which is a Cα based measure of structural accuracy. It can be understood roughly as the percentage of residues that are correctly positioned in the model (higher is better, range 0 to 100).

We have chosen the initial models on which to test the methods based on server groups that have traditionally done well in past CASP events.

2.4.1 The confinement method can correctly rank different models generated by the same methodology and distinguish the native structure

The question that we want to answer here is whether we can rank models that have been generated with the same methodology. 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 initially 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 model 1 was predicted correctly, whereas the order of model 3 and 5 was wrong. The main difference between model 3 and the rest of the models is that the orientation of the first alpha helix of model 3 is opposite??. On the other hand, as shown in Figure 4 the confine and release method not only can differentiate the native structure from the submitted models,

the ranking also correlates perfectly with 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 the 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 The confinement method can rank models generated by different methods

Our next test was between models that are produced with different methodologies. We do this by looking at some of the best CASP submissions from different research 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 identifies the residues that are responsible for differences between two conformations

We used one of the CASP targets (pdb id=2KWY, casp id=T0569) to study the per residue contribution between this NMR structure and the best CASP9 predicted model (from the 'Mufold' group, GDT_TS=78). 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, the 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. Interestinlgy, some regions are more stable in the generated model. A close inspection reveals 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 (Figure 11(C)). The peak at position 76 is due to the H-bond between Lys-76 and Asp-11, which is absent in the original crystalographic structure (Figure 11(D)).

2.4.4 Failures in the confinement method

Despite the great success in most of the studied systems, we found a few failures, specially when the GDT scores of the compared 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 strands. 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 (RMSD values between 1.6 and 2.0Åfrom native, right?all atom? CA?). 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 remains consistent, model 1 is predicted to be more stable than the crystal structure. To rationalize this, we have decomposed the total free energy into the per residue components. The results show us that despite the small variation at the backbone level (as

shown by high GDT_TS scores and low RMSDs), the sidechains are oriented in very different ways, giving raise to large differences in the stabilization of certain residues. In particular, some of the differences arise from different salt bridge patterns (Arg-32 with Glu-35 and Glu-28 with Lys-24 in crystal versus Arg-32 with Glu-28 in model 1) and certain flexible polar residues exposed to the surface (Lys-24 in model 1, which has an entropic gain from not forming a salt bridge and is stabilized by interactions with the solvent). This is summarized in Figure 6(A) Similar patterns arise in other places of the protein (see Figure 6(B)), where the salt bridge between Arg-26 and Glu-50 is absent in model 1 and is further stabilized? by Phe-51 having a very different conformation that in the crystal structure. All in all, this unexpected result just shows us that this method is sensitive to local interactions such as those happening from side chain reorientation. At the same time, the results explained in past sections showed that for larger changes the accuracy of the method is very high.

2.4.5 What can we say about low resolution models?

So far we have seen that the method is good at predicting preferences when the structures are not very close to each other, how far from native can we go? In this section we want to explore what happens when we use the method to describe low quality models. We use one of the CASP targets (pdb id = 2KJX, CASP id = T0531) . We mentioned earlier that our predictive abilities are greatly decreased with low model qualities. Comparing low qualitive models between themselves is not very informative, however, comparing all of them to a native/native like structure, allows to rank order this structures. 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 even

though it is got a low GDT score: 44.

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

A part of the CASP experiment is dedicated to the assessment of the quality of predicted models[24]. Most of the top performing groups in this category 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 methods 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 compared 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 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 a similar type of analysis for the case of Target 560 from the 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 wants to rank order with respect to the GDT_TS values, the model from PconsR is the best predicted model, but according to the QA as-

sessment 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 $(\Delta\Delta G(Model\_Pconsr - Model\_Multicom\_Refine = -30.0Kcal/Mol)$ -30 sounds huge! 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 the confinement method can predict the model quality in a relatively expensive but much more accurate way.

[Table 2 about here.]

3 Conclusion

The last 50 years represent our effort to understand what the native state of a protein looks like by using theoretical methods. Experimental evidence coming from NMR and x-ray cristallography remain so far the only real ways to solve the structure of the native state. However, this methods are expensive and slow. Theoretical methods on the other hand are cheaper, but despite improvement still have a long way to go. Often this computational pipelines produce ensembles of diverse structures and identifying which structure is more native like is still a problem. Here we have shown how the application of the confinement method can help identify native like states. The high success rate balances with the computational expense. However, with GPU technology and the rate at which computers improve, this computational expense is becoming

less of a bottle neck.

I don't agree with some of the points here, but I think I covered what you

meant in the previous paragraph

In the recent past there have been a 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 ***kinetic traps in MD,not other methods, local minima***, 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 the previously developed confinement method, applying it to larger systems.

We found that in majority of cases, the native structure is always the most stable and we can rank protein structures upto 100 amino acid residues long. This method works very well if at least one of the strcuture in the decoy set has high quality. The 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 from this method that we have exploited is the partition of free energies on a per residue basis. Although, it is an approximate partition, it gives important insight regarding the conformational preferences of a particular residue. The use of this decomposition in understanding the effect of a small number of aminoacids in our test chamaleon sequences allows us to think that in the future, this method may be used for protein design. 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 we want to ignore this fact 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 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 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.

less of a bottle neck. We foresee great potential for this methodology for two reasons: independence from a reaction coordinate (just need start and end state) and avility to decompose the free energies on a per residues basis. This last point can be very useful in protein design. Finally, we have presented data for structure comparison of proteins, but we are currently exploring the possibility of using this methodology for getting free energies of ligand-protein binding.

Do we mention this in the text? too technical for conclusions: There is another issue, which is normal mode analysis calculations are Oh(N**3) and require large amounts of memory, effectively limiting the size of the systems under study. In most cases quasiharmonic 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 the confinement

approach is the same in both these papers. However, there are some technical differences. Here we briefly describe the procedure to compute free energy ΔG_{AB} between two conformations A and B of a protein. The basic idea to compute the free energy between conformations A and B is to perform a thermodynamic cycle:

1. Minimization of A and B. These minimized conformations (A* and B*) are the reference conformation of that state. During this minimization the backbone is kept restrained to the original position to avoid large conformational changes.

2. The free energy of confining the ensemble (A or B) to a microstate (A* or B*) is calculated. This is done by gradually applying larger and larger harmonic restraints on all the atoms of the biomolecule. This is done by running 21 molecular dynamics simulation (20 ns long each), where the harmonic restraint force constant was scaled from 0.00005 units?? (mostly free) to 81.92 (frozen in one microstate). In this final restrained state, the rotational contribution to the free energy is frozen out and the only remaining contribution is the vibrational part. The free energy for this step is estimated from the fluctuations around the reference structure using a numerical approach developed by Tyka et. al.[14]. The confinement free energy calculated in this way is recorded as $\Delta G_{A,A*}$ and $\Delta G_{B,B*}$ as shown in Figure 3.

3. 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.

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

All calculations where performed with the amber 11 suit of programs in combination with ff99SB and 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 much smaller (less than $1.0Kcal/Mol$that sounds huge, i would say what percent of error we stimate, I think it was lower than 1kcal/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

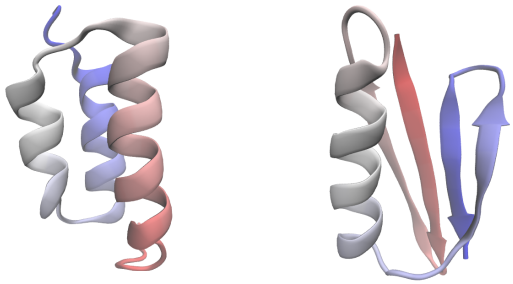
[25] Zemla, A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003, 31, 33703374.

[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.

```
2jws: TTYKLILNLKQAKEEAIKELVDAGIAEKYIKLIANAKTVEGVWTLKDEILTFTVTE GA88
2jwu: TTYKLILNLKQAKEEAIKELVDATAEKYFKLYANAKTVEGVWTYKDETKTFTVTE GB88
2kdl: TTYKLILNLKQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE GA95
2kdm: TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE GB95
Y45A: TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTAKDEIKTFTVTE GA98
```



α β

Sequence	Experimental Fold		Calculated Fold		Calculated $\Delta G_\beta - \Delta G_\alpha$ (kcal/mol)
	α	β	α	β	
GA88	✔		✔		3.9 ± .5
GB88		✔		✔	-4.4 ± .5
GA95	✔		✔		3.5 ± .5
GB95		✔		✔	-5.0 ± .5
GA98	✔		✔		3.8 ± .5

Figure 1: Chameleon sequences and computed free energies.

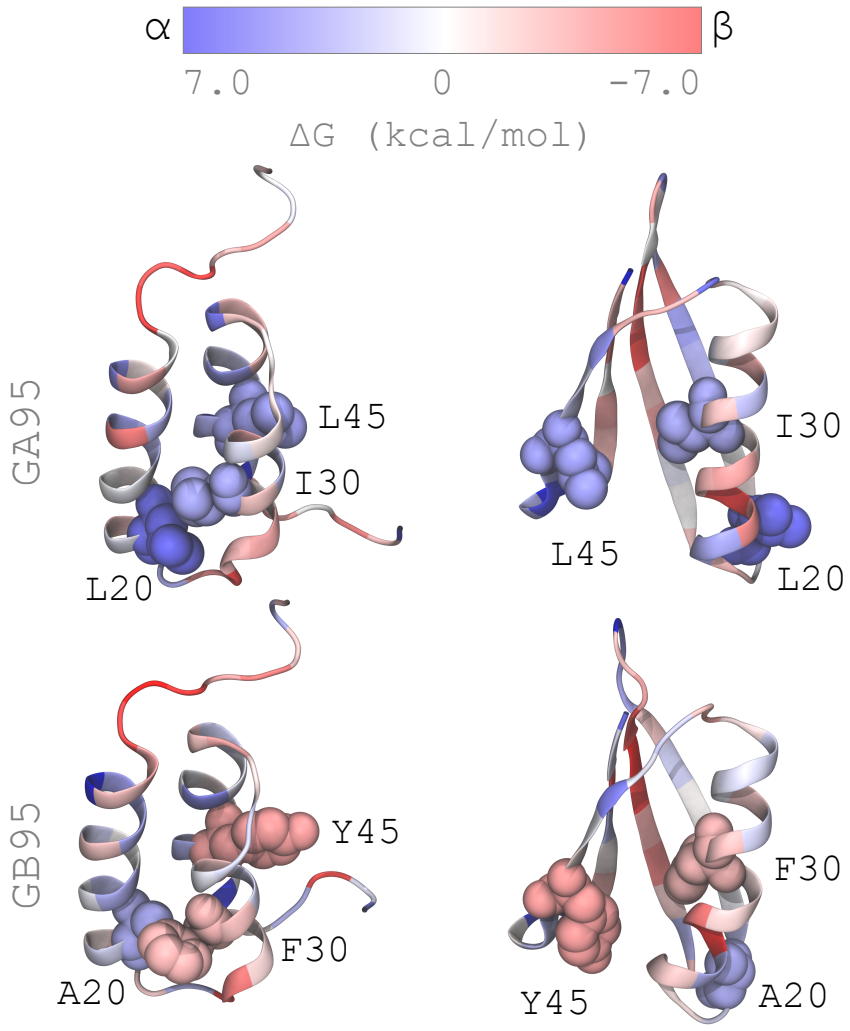
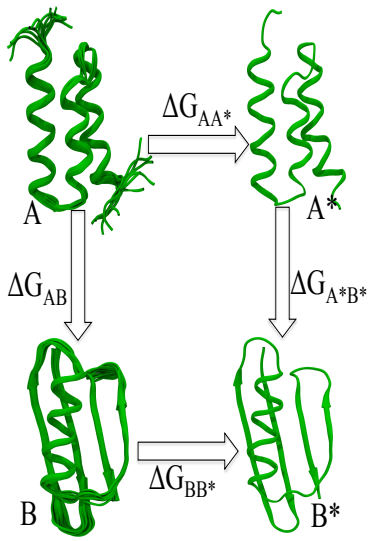


Figure 2: Per-residue ΔG.



$$\Delta G_{AB} = \Delta G_{AA^*} - \Delta G_{BB^*} + \Delta G_{A^*B^*}$$

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

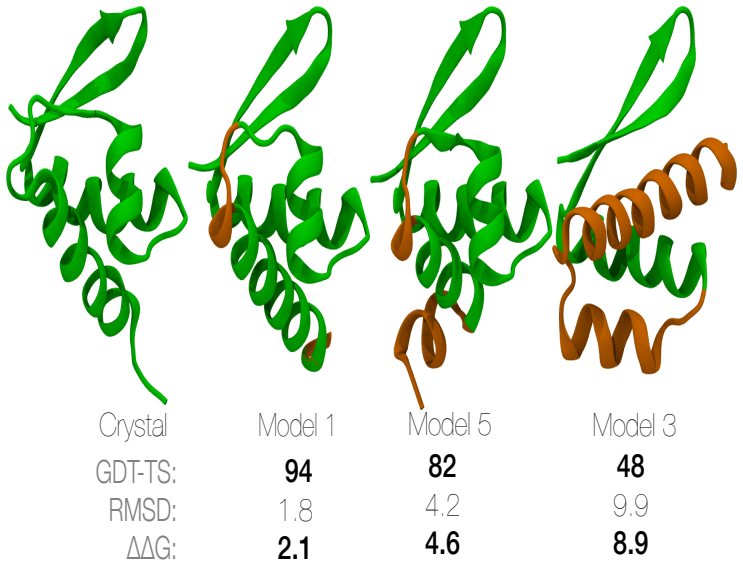


Figure 4: The native and three submitted model structure, along with their GDT_TS, RMSD and relative Free energy values of protein BVU3908 from Bacteroides vulgatus (PDB id: 2L01 and CASP code: T0559).

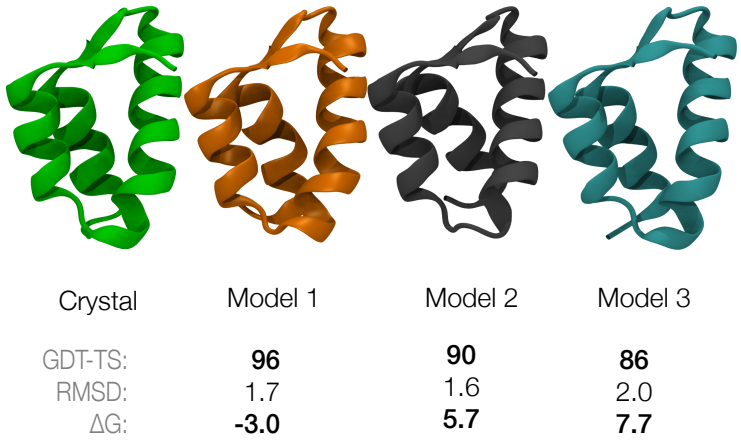


Figure 5: The native and three model structure of engineered protein from Asr4154 protein (PDB ID: 2L09 and CASP code:T0538). The model 1,2 and 3 are from the group PconsR, Shell and FOLDIT respectively.

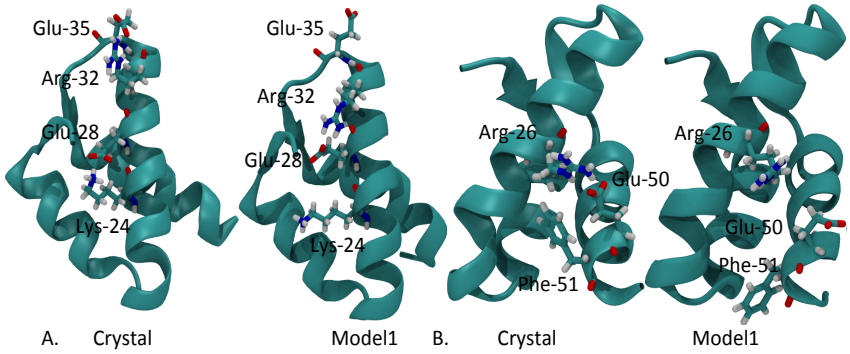


Figure 6: Difference of the sidechain orientation between model 1 and crystallographic structure in Target-538. A. The salt bridge between Glu-35 and Arg-32 in crystal structure. In model 1, this is compensated by H-bonding between Arg-32 and Glu-28. There is another salt bridge between Glu-28 and Lys-24.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

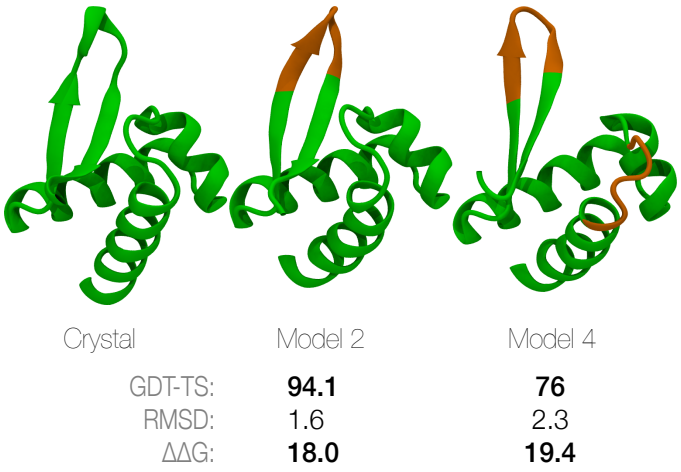


Figure 7: Native and two model structure of protein BT2368 from Bacteroides thetaiotaomicron (pdb id: 2L02 and CASP code: T0560). The two models were from the group ”Splicer”.

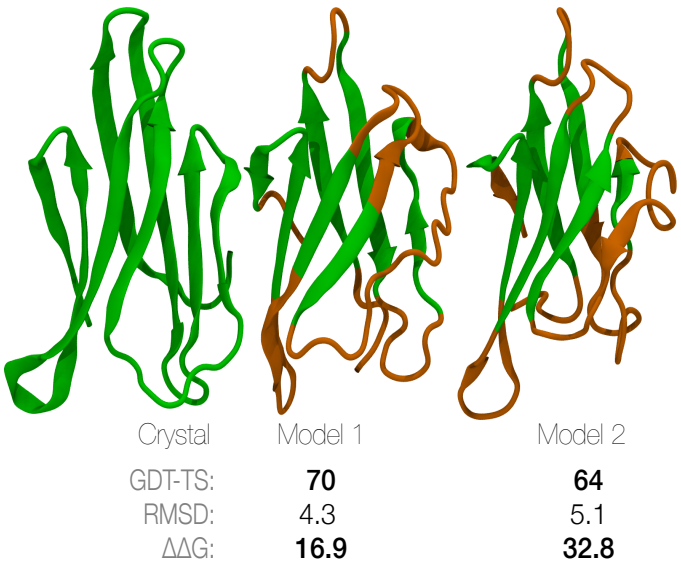


Figure 8: X-ray crystallographic structure and two submitted models of fas apoptosis inhibitory protein (pdb id: 3MX7 and CASP code: T0540). Model 1 and Model 2 in this analysis were submitted by the group LTB and MUFOLD respectively.

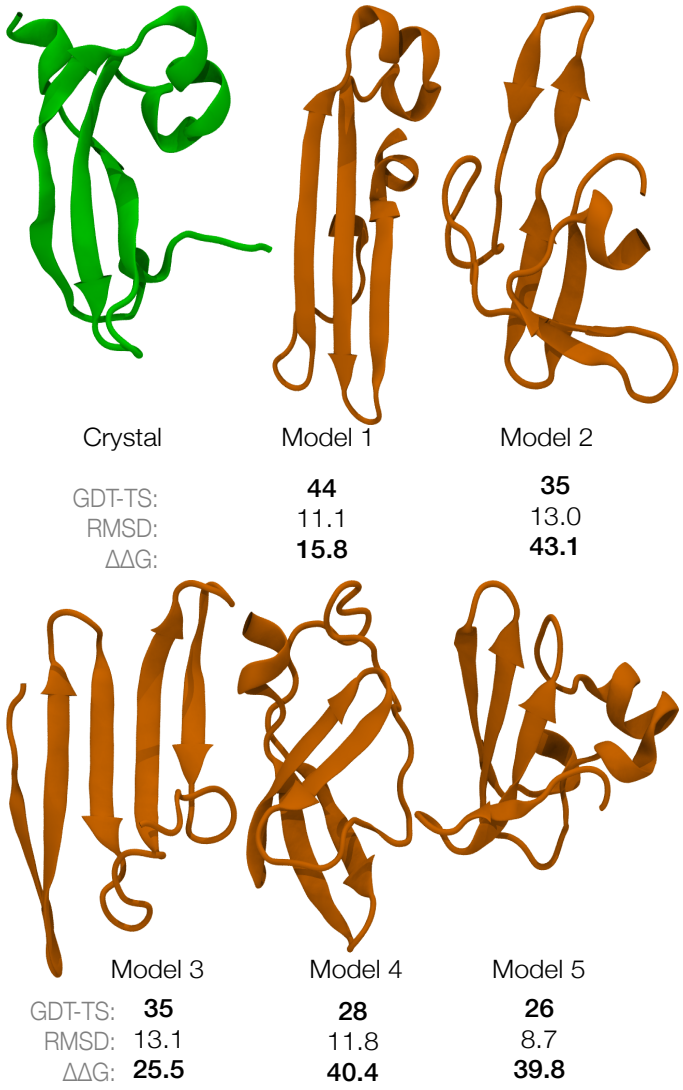


Figure 9: The native structure and 5 models of extracellular domain of the jumping translocation breakpoint protein (pdb id: 2KJX and the CASP code: T0531).

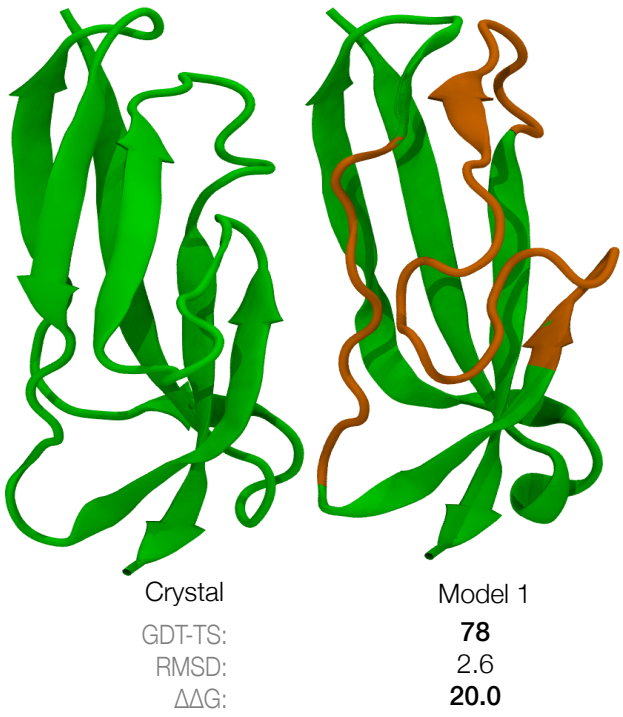


Figure 10: Native and best model structure of a domain of adhesion exoprotein from Pediococcus pentosaceus (pdb id: 2KWY and CASP code: T0569). The best model was from the group ″Mufold″.

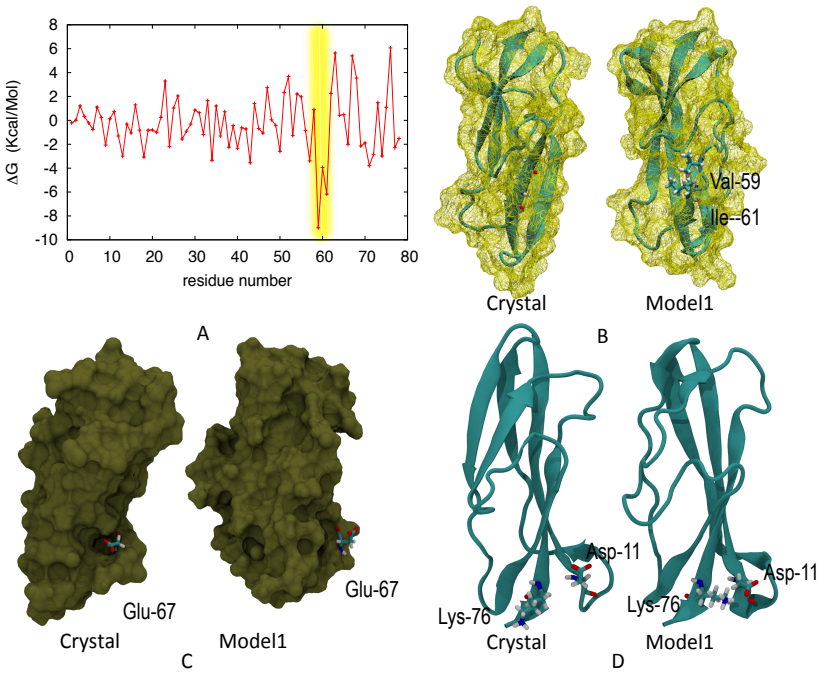


Figure 11: A. Plot of per residue free energy difference between the model and the crystal structure of Traget 569 (pdb id: 2KWY) of CASP9. B. In model 1, the beta sheet containing Val-59 and Ile-61 is disordered. These hydrophobic residues are exposed to the solvent. C. Glu-67 is partially exposed to 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 gives lower free energies for native states than for computer-generated predictions. For each target, we examined as many as five predictions submitted by CASP participants. We report the free energy difference between the most favorable decoy and the experimentally determined structure. Positive ΔG values indicate that the experimental structure is predicted to be more favorable than any of the decoys.

CASP Target	PDB Identifier	$\Delta G_{native \rightarrow bestdecoy}$ (kcal/mol)
T0531	2KJX	11.2 ± 0.7
T0538	2L09	-3.0 ± 0.5
T0540	3MX7	16.9 ± 0.5
T0559	2L01	2.1 ± 0.2
T0560	2L02	18.0 ± 0.5
T0569	2KWY	20.0 ± 0.7

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 2: 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.