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.

October 1, 2012

Abstract

The calculation of free energy differences is of central importance in the simulation of biochemical systems. The computation of the free energy difference between pairs of macromolecule with large conformational change is particularly a difficult task and computationally expensive with existing methods. In this work, an improved version of the confinement approach is used to calculate absolute free energies of bio molecular systems. The method does not require a reaction coordinate or transition path. It is fast to compute. We show that the method correctly picks out the state of lowest free energy of a pair of structures having similar sequence but different fold. We also show using models from CASP9 that the method picks out native-like structures from misfolded or decoys, provided at least one good structure is in the input set.

1 Introduction

In computational structural biology there are numerous cases where free energy between well defined states are necessary. Examples ranges from

small to large conformational change of protein due to ligand binding, change of pH etc. Free energy also play an important role in case of protein folding. The 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. Regardless of the starting point or unfolding it by changing different condition most proteins will finally assume the same structure. It is very important for proteins to achieve their native conformation since defects in protein folding may be the molecular cause of a range of human genetic disorder. For example a misfolded protein known as prion when enters any healthy organism it converts properly folded protein into misfolded one. Thus, Free energy can also act as a scale to identify between the misfolded and the native state of the protein. These observations further give free energy a special importance in structural biology.

However, the theoretical calculation of conformational free energy change become difficult if the states of interest are very different. In such scenario timescale involved in such tranformation may be beyond timescale of direct classical molecular dynamics simulation. Generally, in order to calculate free energy a pathway or reaction coordinate is created with the intermediate structures. Method like umbrella sampling, thermodynamic integration along with classical MD can suffer from overlap problem and require large computational effort. The idea of reaction coordinate became more complicated in case of protein folding. As free energy landscape of the protein folding is rough[1], it can be trapped in a energy well during its visit in the conformational space. Even with the advanced method like replica exchange molecular dynamics the three dimensional structure may be trapped in a local minima for a considerable time. Due to various reasons it can also be misfolded and trapped in a local energy minima. And one can mistakenly identify the misfolded state as the native state (give the example of human pin1ww domain from Benout Roux and Klaus Schultan).

Calculation of free energy has been successfully attempted by a number of groups[2] -[4]. In recent years methods like Refrence system method, Decativated Morphing[3], Orthogonal Space Random Work, Confinement Method have came out for calculation of free energy method. Some of the groups relies on one of the great strength of free energy calculation, that it is a state

function and thus it is independent of path.

In this work we have applied the confinement method which was originally devoled by Tyka et. al.[5] and Cecchini et. al.[6]. This method relies on a thermodynamic cycle where first, non-harmonic degrees of freedom are removed by applying a series of restraint 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. Then we have applied this method to larger proteins for the first time. The free energy of conformation of pair of proteins with similar sequence but completely different fold was calculated. Encouraged by the result, we applied the method of different targets of CASP9 competition.

2 Results and Discussion

2.1 The confinement method produces correct results in control experiments

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[6]. 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[6,7] 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[6] (see Supporting Information for further details).

Arijit, Al: did we do any other control experiments on small peptides that

we should report here?

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. One of the 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 10, 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 experimental structure (once it was available from the PDB). As Table 4 shows, in X of Y cases, the confinement method correctly assigns a lower free energy to the experimentally determined structure than to any of the decoys. Although differentiating between experimentally determined structures and computer generated predictions is not a stringent test of a scoring method—for example, examining the residue-residue packing can easily distinguish between the two [ref Rosetta Holes]—the results none the less serve as a useful "sanity check".

[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 series of 56-residue proteins (based on Protein G) that adopt one of two different folds depending on small changes in sequence (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 and differ in seven positions. Another pair (GA95/GA95) 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[8−11].

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). There is, however, a serious problem with this analysis: we are comparing an experimentally determined structure with a computer generated model. 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 the relative free energy 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. The confinement method identifies the correct structure for all six sequences. The free energy differences range from around 2.9 to 5.7 kcal/mol, which is consistent with small changes in sequences and with estimates made by Orban and co-workers[9,10].

[Figure 1 about here.]

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

Next, we have applied the confinement method to several targets from the CASP9 and CASP10 experiments. CASP is a world-wide competition aiming at finding the state of the art methodologies in protein folding. Groups, both from knowledge based methods and physics based methods participate in this blind competition and for each sequence target they can submit five models. The predictors were also asked to calculate their best model among the set of five they submitted. However, in many cases predictors were unable to predict/rank their models. Using the confinement method, we aim to rank protein structures using free energy as a scale. In recent CASP experiments Global distance test scose (GDT_TS) method is used for final analysis of different submitted structures with the NMR/crystallographically solved protein native structure. GDT_TS represent the average percentage of residues that are in close proximity in two structures optimally superimposed using four different distance cutoff. Although GDT_TS alone is not a reliable measure for the difficult modeling cases, it has been well accepted method in recent years among the community. We have also compared our free energy result with the GDT_TS score for different cases we studied.

2.3.1 T0559

The native structure is a 67 residue NMR solved structure. We first reviewed all the five models submitted by the group "BAKER-ROSETTASERVER", which was the best predictor group for this particular target. Among them we choose only the three structures, model 1, 3 and 5 to calculate the free energy and subsequently identify the most stable structure among them. The rest of the two structures are close to either of the selected structure and thus they are ommited from our calculation. According to the GDT_TS score, the model 1 was picked up correctly from the set of structure generated . However, the order of model 3 and 5 was not correct. It will be interesting to know whether confinement method can order the strcutures in terms of free energy scale. The result is presented in Figure 5.

2.3.2 T0560

Target T0560 is a 74 amino acid residue target from CASP9. However residues 3-66 were kept for final result. For this target we have compared the free energy of the model 2 and model 4 submitted by the group ”Splicer” along with the crystal structure. The main difference between the model 2 and 4 is the orientation of the .... The final result is presented in Figure 7.

2.3.3 T0538

Target T0538 was a 54 residue target. The native structure was generated using NMR. The best model that is closest to the native structure was generated by the group PconsR. However, their model 3 was better than their model 1. This structure was quite close to the NMR structure and had GDT_TS value of 96.23. We have calculated the free energy between this structure and the best structure submitted by the group ”Shell”, ”FOLDIT” and ”BAKER-ROSETTASERVER”. The calculated free energy value is presented in Figure 6.

2.3.4 T0540

Target T0540 was a 90 residue target. We have taken the best models from the group ”LTB” which was the best performing group for this target and also from the group Mufold and calculated the free energy difference between these models with that of the native crystal structure. As expected, the crystal structure has the lowest free energy value, followed by Model 1 with GDT_TS score of 69.72 and model 2 with GDT_TS value of 53.789. The final result is presented in Figure 8.

2.3.5 T0531

This is a 65 amino acid residue target with residues 6-63 were kept for final analysis. We have calculated the free energy difference between all the five models submitted by the group ”MUFOLD-MD”, which was one of the best performed group for this target. The result is presented in Figure 9.

2.3.6 T0605

This is a 72 amino acid residue protein but for the final calculation only the residues 18-66 amino acids were kept for the final analysis.

The best performing group was ”Baker” for this target. We have carried out our calculation with the model first, second and fifth for the final calculation. The rest two model had closer GDT_TS value compared to the previously mentioned models and thus they were omitted from furthe calculation. The result is presented in Figure 10. (This is a case where we can demand the confinement method is better than GDT_TS).

[Figure 2 about here.]

[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.]

3 Conclusion

4 Method

The confinement method has been described in details in ref. by Tyka et.al. and Cecchini et. al. Here we briefly describe the methodology. We compute

free energy ΔG_{AB} between A and B conformation of the protein. For this purpose we use a thermodynamic cycle.

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

2. Backbone dihedral angle of the whole protein is restrained by a harmonic potential to define the configurational state.

3. In order to calculate the free energy $\Delta G_{A,A*}$ and $\Delta G_{B,B*}$ A and B are gradually transformed in A* and B* by applying a harmonic restraint potential to all atom. For this purpose around 21 simulation were run each of 20 ns with increasing harmonic restraint potential 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.

4. Finally the thermodynamic cycle is closed by calculating the free energy between the final restrained state using normal mode analysis.

In all the calculation amber 10 program is used with ff99SB with GB/SA implicit solvent.

References

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

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

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

[4] 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.

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

[6] 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).

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

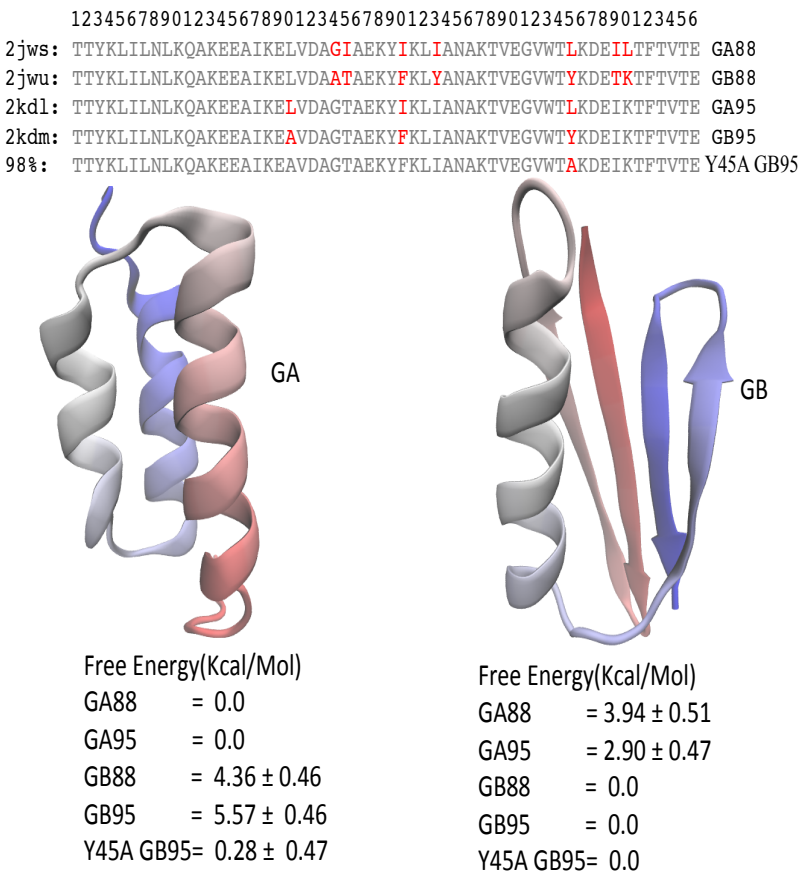


Figure 1: The confinement method correctly predicts the structural preferences of six chameleon sequences. (A) The six sequences used in this study. (B) Each sequence adopts either a Protein G-like fold (denoted GB) or a three-helix bundle fold (denoted GA). The relative free energies of the two folds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

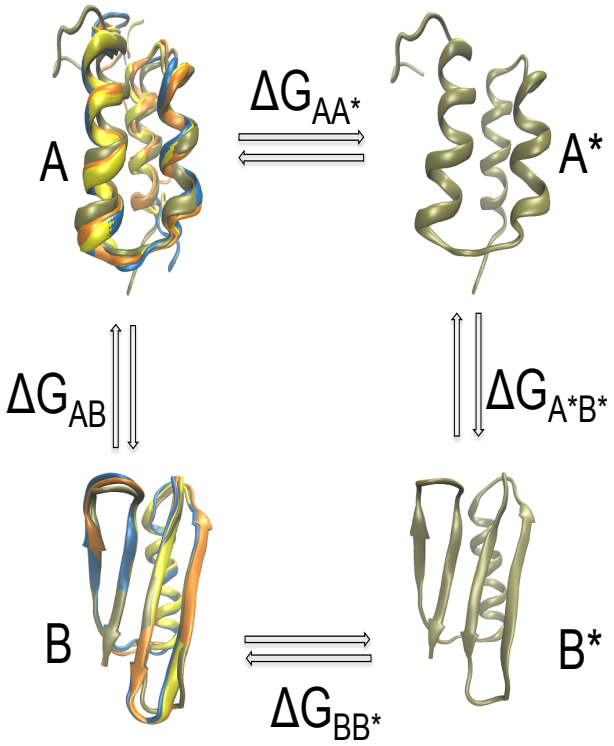


Figure 2: Graphical representation of the thermodynamic cycle involving confinement method.

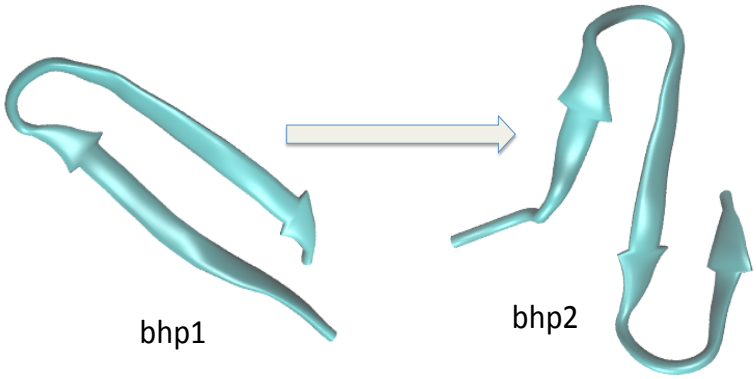
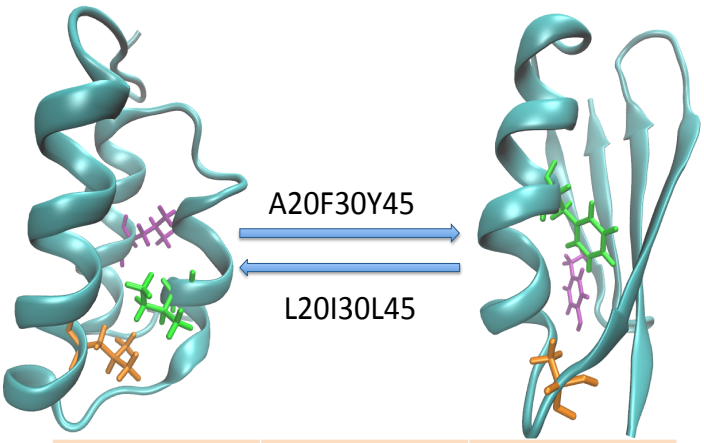


Figure 3: Two conformations from β hairpin from protein G, bhp1 and bhp2. The two stranded β sheet, bhp1 is the native structure and the three stranded β sheet is known as bhp3.



Change in Sequence	ΔG (3α fold)	ΔG (4β+α fold)
L20I30L45	0.0	2.90±0.47
A20F30Y45	4.36±0.46	0.0

Figure 4: Two protein with 95 % similar sequence but different folds

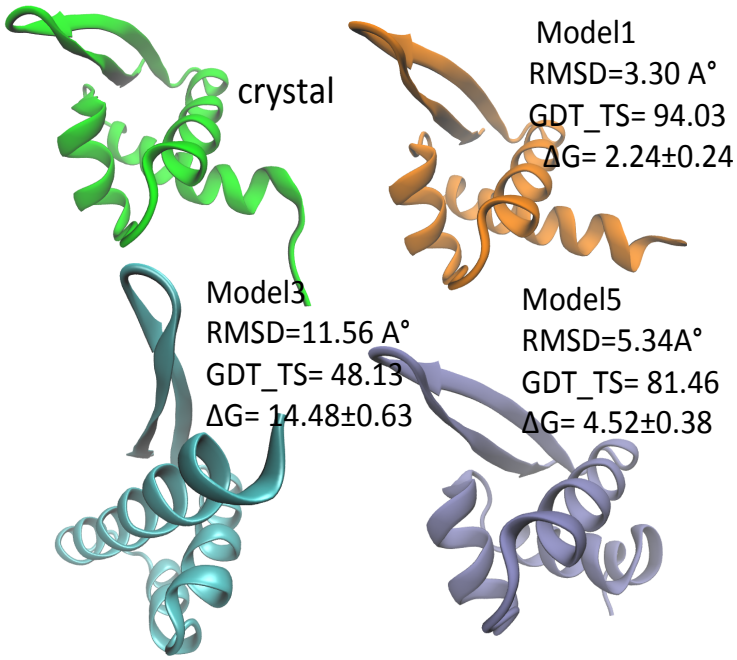


Figure 5: Target T0559 from CASP9

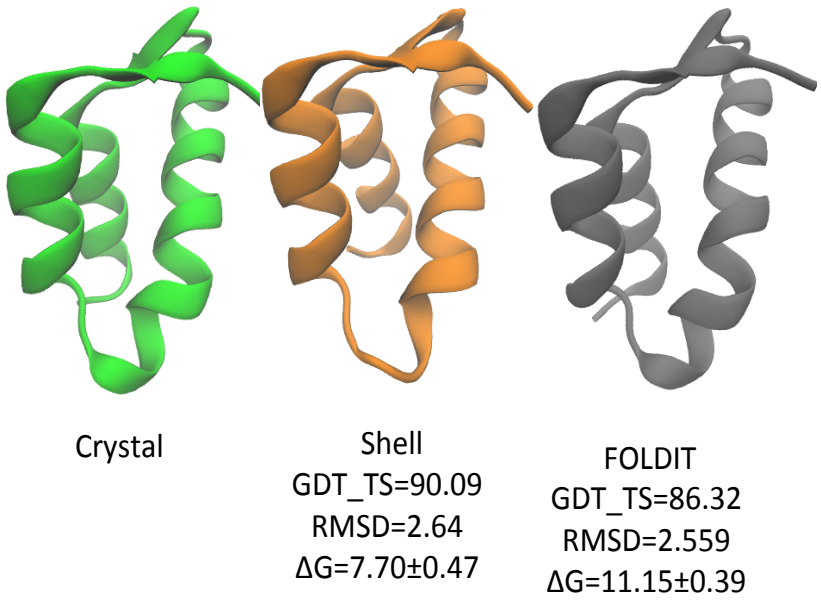


Figure 6: Target T0538

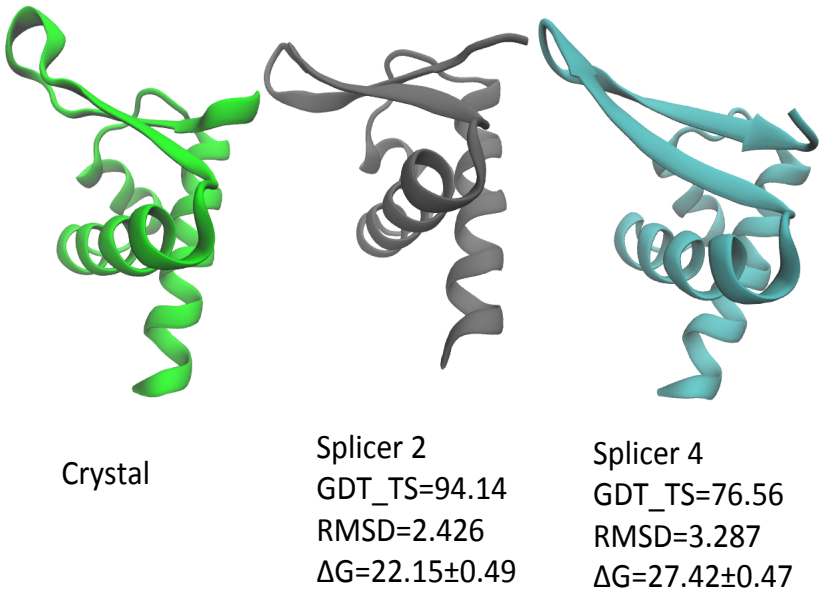


Figure 7: Target T0560

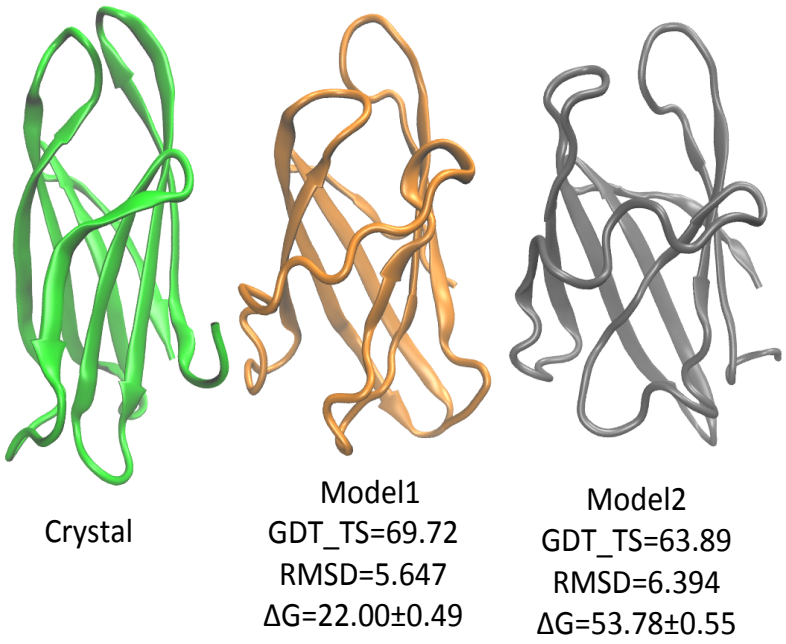


Figure 8: T0540

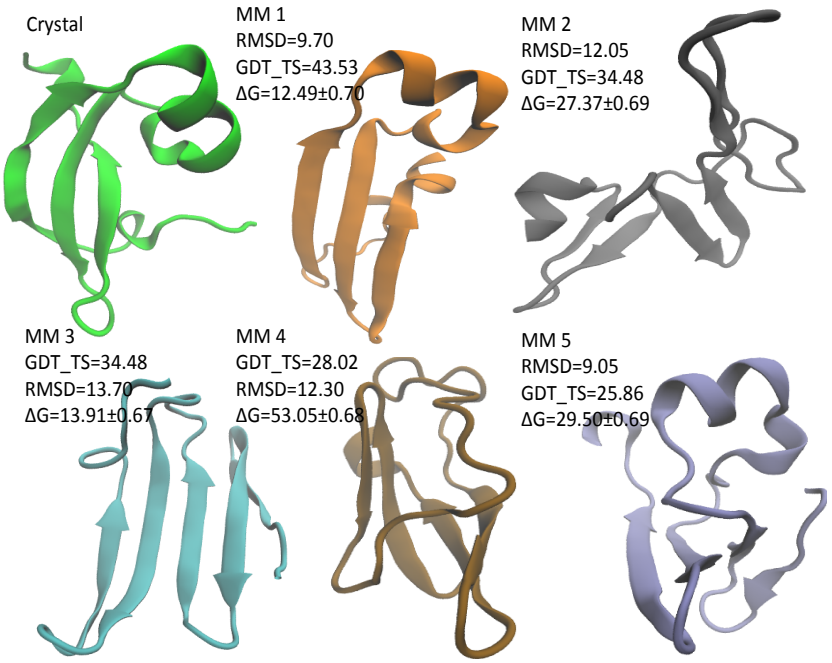


Figure 9: T0531

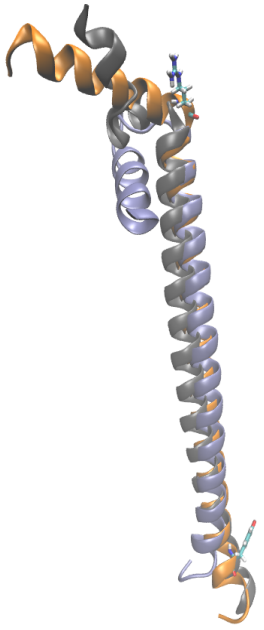


Figure 10: T0605

CASP Target	PDB Identifier	$\Delta\Delta G_{native \rightarrow bestdecoy}$ (kcal/mol)
T0531	2KJX	12.49 ± 0.70
T0538	2L09	7.70 ± 0.47
T0540	3MX7	22.00 ± 0.49
T0559	2L01	2.24 ± 0.24
T0560	2L02	22.15 ± 0.49

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.