Project Title:

Protein Folding Prediction Using X-ray Diffraction Data as Constraints

Name:

Kam Zhang
Rojan Shrestha
Francois Berenger
Ryo Takahashi

Affiliation:
Zhang Initiative Research Unit, Advanced Science Institute, Wako Institute

1. Background and purpose of the project, relationship of the project with other projects

The protein-folding problem of how the primary sequence determines its tertiary structure is one of the great challenges in computational biology. It is known that all the information required specifying the tertiary structure of a protein is encoded in its primary sequence. Moreover, "structure determines function" is a well-established paradigm. Our ability to predict the structures of proteins from their sequences will greatly facilitate our understanding of the important biological functions that proteins play in their biological systems.

There are two fundamental challenges in protein structure prediction. One is the construction of precise energy functions that could be used to assess the thermodynamic stability of a protein at a given conformation state. The other is to find the global minimum energy conformation in the complex energy landscape. Both are formidable tasks to tackle. First, the basic physical forces that govern atomic interactions are incompletely and poorly understood. Secondly, it is computationally prohibitive to search for the global minimum energy conformation even if the precise energy function were available.

X-ray crystallography is the principle method of determining 3D structures of proteins. It uses the diffraction phenomenon caused by the interaction of crystals with X-rays. Only the amplitudes of the diffraction can be measured experimentally but not their phases. However, both the amplitudes and phases are required to reconstruct the crystallographic image in order to reveal the atomic positions in the crystal. Phase retrieval is therefore of crucial importance in macromolecular structure determination.

We propose to develop a computational method to solve the protein crystallographic phase problem. We plan to use the powerful conformation search capability of Rosetta to sample the conformation space and to use the X-ray diffraction amplitude as a pseudo "energy function" to identify the correct conformation. In this formulation, we have separated the energy function from conformation search in protein folding. In principle, the correct structure can be identified using the X-ray amplitudes as pseudo energy as long as that conformation can be generated or sampled. In this way, we have reduced the complexity of protein structure prediction under the constraints of X-ray amplitudes to a conformation search problem. Our research goals are:

1. To develop a method and create software tools to enable the prediction of globular protein structures from their sequences using X-ray diffraction data as additional constraints.
2. To improve upon the state of the art structure prediction methods to enable routine molecular replacement solution in X-ray crystallography using predicted protein structures.
3. To develop methods and tools to efficiently assess the fit of a given structural conformation to the observed X-ray diffraction data.

Our research efforts will be focused on developing methods that could efficiently calculate the fit of a given protein conformation to the X-ray diffraction data. In conventional protein structure prediction, the free energy is calculated as the sum of the enthalpic and entropic terms. The enthalpic term consists of the free energies from the bond, angle, dihedral, van der Waals, charge-charge interactions, H-bond, hydrophobic as well as solvation. We propose to add to the above free energy components an extra term that comes from the X-ray diffraction data.

We will use the Rosetta software suite as a starting point of method development and as a tool for conformation search. Since the real space all atom energy function has already been developed and incorporated into the Rosetta software suite, the key missing component that needs to be developed is the evaluation of the fit for a given structural conformation to the X-ray diffraction data. Both the enthalpic and entropic free energy terms depend on the atomic coordinates expressed in the real or Cartesian space. However, the free energy term from the X-ray data is expressed in the reciprocal or Fourier space. The real and reciprocal spaces can be inter-converted by a Fourier

transform. Since the predicted structure can be at an arbitrary position and arbitrary orientation with respect to the X-ray data, a computationally expensive six-dimensional search is needed in order to place the predicted molecule in the correct position and orientation before the fit between predicted and observed data could be measured. This is the so-called molecular replacement problem in X-ray crystallography. Most of our research effort will be focused on this area.

2. Specific usage status of the system and calculation method

The computational approach to protein folding has been notoriously time-consuming due to the astronomical number of conformations that have to be searched and evaluated. The IBM BlueGene project was initiated with one of its goals stated as solving the protein folding problem. Although employing various techniques to speed-up or "short-circuit" the conformation search have made the protein folding computationally feasible, it still remains as a computationally intensive endeavor.

When the protein folding method is used to solve the X-ray crystallographic phase problem, the computational challenge is further increased. This is due to the nature of the pseudo "energy function" from X-ray diffraction amplitudes are measured in Fourier or spectrum space, whereas the protein conformation is in Cartesian or real space. One Fourier transform is needed to assess the fit of each conformation to the experimentally measured amplitudes. Moreover, since the predicted structure can be at an arbitrary position and arbitrary orientation with respect to the X-ray data, a computationally expensive six-dimensional search need to be conducted in order to place the predicted molecule in the correct position and orientation before the fit between predicted and observed data could be assessed.

3. Result and Conclusion

Rosetta algorithm is the most powerful method for the prediction of high resolution structures at atomic level using only the amino acid sequence. However, we have found that the Rosetta free energy cannot always identify the best structure among many others generated. Furthermore, conformational space search is another bottleneck inside the Rosetta protocol. We have proposed to drive the Rosetta score function to trap the lowest energy conformation by integrating the score from the match of the predicted structure against X-ray data. We have designed classes to incorporate the molecular replacement libraries within the structure Rosetta integrated package. We have generated a large numbers of structures using Rosetta for specific amino acid sequences to determine minimum accuracy required to find the molecular replacement solution.

To utilize X-ray data in Rosetta requires the correct orientation and position of the structure model in diffraction space. This can be achieved by molecular replace methods, such as EPMR and PHASER. We have transformed these two standalone programs into libraries we can reuse. We have been evaluating how and within which limits these tools can be used in order to work on Rosetta-generated protein models.

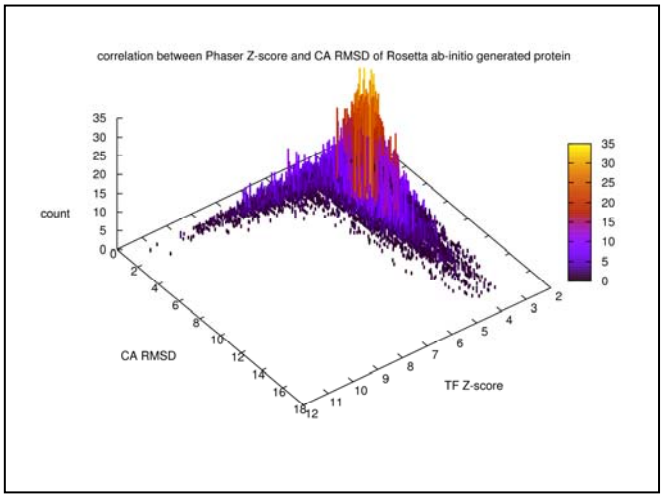
We have selected 16 protein sequences with the number of residues ranging from 51 to 128. We have generated millions of atomic models using the massively parallel supercomputer at RICC. The result is shown in the

Model sequence	Structure factors	Space group	No. of residues in model	No. of molecules in ASU	No. of models, large-scale	No. of models selected for molecular replacement	No. of successful molecular replacement models	Highest TFZ	RMSD (CA / All Atom) *
1bq9	1bq9	$P2_12_12_1$	51	1	1.2×10^5	-	-	-	
1pgx	2igd	$P2_12_12_1$	59	1	4.2×10^5	665	83	8.4	4.12 / 4.62
5cro	5cro	H32	55	4	7.4×10^5	681	45	23.5	2.42 / 2.66
1hz6	1hz6	$P2_12_12_1$	63	3	7.3×10^5	-	-	-	
1a32	1a32	$P2_12_12_1$	65	1	2.8×10^5	-	-	-	
1ctf	1ctf	$P4_32_12$	68	1	3.2×10^5	-	-	-	
1ubi	1ubq	$P2_12_12_1$	71	1	7.0×10^4	-	-	-	
1dtj	1dtj	C2	74	4	3.1×10^5	-	-	-	
1ig5	1ig5	$P4_32_12$	75	1	7.5×10^4	-	-	-	
1opd	1opd	P1	85	1	7.5×10^4	-	-	-	
1a19	1a19	$I4_1$	89	2	6.0×10^4	-	-	-	
1bm8	1mb1	$P4_12_12$	99	1	9.2×10^5	-	-	-	
1aiu	2hsh	C2	105	1	4.4×10^5	-	-	-	
256b	256b	P1	106	2	1.5×10^5	-	-	-	
1elw	1elw	$P4_1$	117	2	1.1×10^5	-	-	-	
2chf	2fka	F432	128	1	3.5×10^6	69	17	9.3	3.92 / 4.17
2chf	3chy	$P2_12_12_1$	128	1	3.5×10^6	200	184	8.8	2.82 / 3.23
2chf	6chy	$P2_12_12_1$	128	2	3.5×10^6	223	18	11.9	3.22 / 3.39

table below. A subset of these models with the lowest free energy was selected for molecular replacement using PHASER. A high percentage of Rosetta generated atomic models with low free energy can be successfully used for molecular replacement with PHASER. The highest all-atom RMSD for models that could be used for molecular replacement successfully ranges from 2.66Å to 4.62Å.This is well within reach of Rosetta, which can generate models with significantly lower RMSD for a typical small protein. It also shows that PHASER is quite tolerant with respect to the quality of the model for a successful molecular replacement.

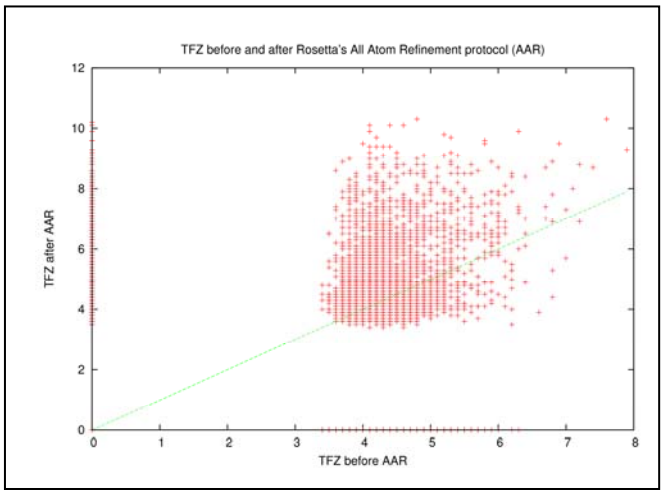
As we want to use X-ray information to help protein structure prediction, we are interested in knowing if there is a correlation between the molecular replacement score (TFZ) from the PHASER software and the root mean square difference to the true structure . To see if there is any correlation, we generated 15000 molecules on RICC using Rosetta ab-initio protocol. Then, molecular replacement was tried on each produced molecule. The result is shown in the 3D histogram figure below.

The 3D histogram shown is a numerical integration of the 2D plot, hot color show that a given region is explored often while cold color means the region is hardly explored. We are interested by the region where the TFZ-score is higher than 6, which means PHASER has a high degree of confidence in the molecular replacement solution. In such regions, the RMSD value spread is quite low (between 2 and 4 CA RMSD). It seems that there is a good correlation between TFZ and RMSD. Therefore, the TFZ-score could be used during protein folding in order to drive the search towards higher quality in-silico generated protein models.



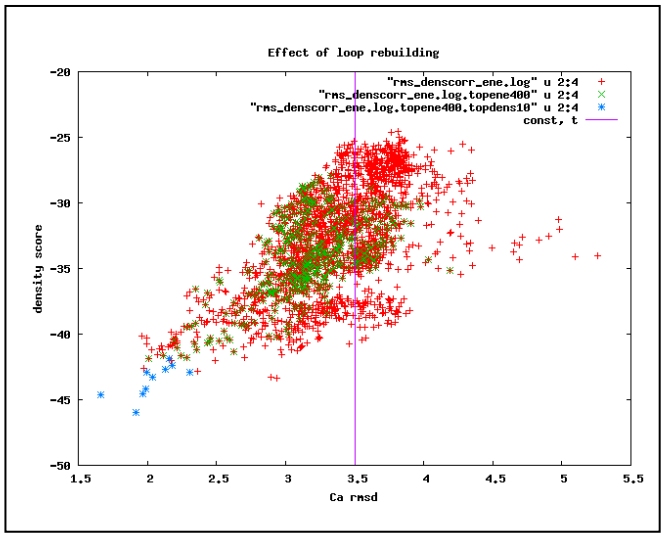
We also tried to see if the TFZ score from PHASER can be used to select good models during protein folding. This experiment needed 20000 models to be generated on RICC using the Rosetta ab-initio protocol. PHASER was also run two times during folding. On the x axis, we can see the TFZ score after only 20% of the total folding time for one molecule before the all atom refinement (AAR) step. On the y axis we see the same score but after the folding process is finished. Points on the x-axis mean molecular replacement was unable to find a solution once the protein was completely folded. Points on the y axis mean the reverse: a molecular replacement solution was not found during folding but one was found once it was completed (this is highly expected in fact). The green line (identity function) separates models being improved from models being degraded from the perspective of molecular replacement.

Very interesting points are those under the green line and with a TFZ score higher than 6. They are special models for which even they are still in an early stage of the folding process, they can already be selected as very promising structures for molecular replacement. The all atom refinement in Rosetta has improved the TFZ-scores for the majority of models. However, the TFZ-scores were decreased for the minority of models. It is necessary to perform the all atom refinement step in order to increase the success rate of molecular replacement. However, it is also necessary to perform molecular replacement on models before the all atom refinement step in order to capture those models that are good enough to have a successful molecular replacement solution.



We would like to include additional experimental constraints into Rosetta, such as the Small-angle X-ray scattering (SAXS) data. The SAXS give us an envelope of the target structure. For this aim, we have checked an algorithm which utilizes the electron density map to refine the predicted structures. This algorithm inside Rosetta is divided into two steps. First is loop rebuilding, and second is all-atom refinement. The procedure in one of our experiments using this algorithm is as follows. Here, 256B protein was selected as a target, and a synthetic 5Å map obtained from experimental structure was used. First, 10 models were selected from the prediction results by Rosetta with target sequence. By applying the loop rebuilding, each

model generated 204 models. From these models, 10 models were picked up. After applying all-atom refinement with map to these 10 models, each model generated 1000 models.



The figure above shows the energy landscape using the results after loop rebuilding. Density score is a part of total energy used in Rosetta, which utilizes the map data. RMSD becomes lower for better refinement. Vertical line represents the best RMSD in the starting models. From this figure, it is found that this algorithm can refine the structure very well. Best case was changing RMSD from 3.5Å to 1.66Å. From these results, the experimental SAXS data which can restrict the Cartesian space like the density map can be expected to refine the predicted structures. We are modifying Rosetta to utilize SAXS envelope data, which will be tested on the high-performance supercomputer.

4. Schedule and prospect for the future

The understanding of sequence-structure relationships as well as more accurate measurement of the energetics that govern protein folding will not only facilitate the *ab inito* prediction of structures from sequences but also enable the design of novel proteins with desired functions. The prediction of protein structure from sequence not only addresses one of the fundamental problems in biology, i.e., the protein folding code, but also meets the structural challenges presented by the vast amount of DNA sequence data generated by various genome projects. The ability to predict the protein structure from the sequence alone will greatly facilitate our understanding of how proteins function in biological systems. The understanding of protein functions will empower us in the maintenance of health and the treatment of disease.

Our research will develop an improved method of predicting protein structures with higher accuracy. The higher accuracy of the predicted structures will have broadened utility from the functional understanding of the overall folding architecture to the computation design of small molecule drugs.

Our research will also develop a new method that can be used to solve protein structures given an X-ray diffraction data using computers only without any additional experimental data. This will enable the determination of some difficult structures for which experimental phasing are not feasible. This will also reduce the cost associated with protein crystal structure determination by saving the expenses needed for experimental phase determination. This will significantly increase the number and type of protein structures available to the scientific community.

5. If you wish to extend your account, provide usage situation (how far you have achieved, what calculation you have completed and what is yet to be done) and what you will do specifically in the next usage term.

We don't need to extend our account at this time since we have started using RICC from Sept, 1, 2009. Our account will expire by Sept., 30, 2010.

6. If you have a "General User" account and could not complete your allocated computation time, specify the reason.

We haven't used all of our allocated computation time since our account will be valid until Sept., 30, 2010. The reason that our usage time is below average is because we have spent most of our time on programming at the beginning stage. We anticipate that our usage will increase since we are entering the testing stage of our project.