Predicting the Structure of Life's Essential Function

Crissan Harris
University of California, Santa Cruz
Summer Undergraduate Research Fellowship – Information Technology

Abstract

Why do we care about predicting protein structures? Proteins are the building blocks of living cells. Their folded 3D shape is different for different proteins and is essential to the function of the protein and the cell. Knowledge of the structure of proteins is used in drug design, design of synthetic proteins, and reengineering of defective proteins. Because protein structures are expensive to determine experimentally (both in dollars and in time), the availability of a computational method of determination has become a necessity.
CASP is a community wide experiment where leading researchers in the field of bioinformatics use computational methods to predict the three-dimensional structure of a protein given only the amino-acid sequence. The goal of CASP is to obtain an accurate and objective assessment of the current abilities in the field of computational structure prediction. This is the research of one such lab of predictors.

1. Overview

CASP (Critical Assessment of Techniques for Protein Structure Prediction) is a community wide experiment where leading researchers in the field of bioinformatics use computational methods to predict the three-dimensional structure of a protein given only the amino-acid sequence. The structures of these target proteins have been solved experimentally by either crystallography or NMR spectroscopy, but the results are kept from being released until after CASP is over. This provides a truly blind environment for the predictions and allows predictors to use their tools to try to come up with the correct structure without already knowing what the structure is. The goal of CASP is to obtain an accurate and objective assessment of the current abilities in the field of computational structure prediction.

This summer I worked on protein structure predictions in Professor Kevin Karplus' lab for the CASP7 community wide experiment. Given a sequence of amino acids, I used programs and tools that have already been implemented in conjunction with some hand tweaking of parameters in order to obtain a protein structure that was considered the most likely to be correct. Many of the tools are recent additions, while some of them have been used for many years and in many of the previous CASP experiments.

CB_burial, near-backbone, dssp-ebghstl, notor2, sep, str2, stride-ebghtl, alpha, bys, and dssp-ehl2 are all secondary structure predictors that have been created by various members of Kevin Karplus' lab group. They use neural nets to predict local secondary structure properties, like helix, sheet, turn, h-bond, etc., with residue properties that are conserved through evolution. Some of these are implemented in the SAM server and some are used in the hand tweaking of parameters. These predictions allow us to add in specific structure constraints to improve the overall correctness of the protein structure.

The SAM server is the premier suite of Hidden Markov Model tools, originally created by Anders Krogh, extended and maintained by Richard Hughey. It uses HMMs, database query, secondary structure predictions, and more. It produces remote homology detection and sequence alignments which are then used in modeling the structure.

Undertaker is a fragment-packing program, created by Kevin Karplus. It gives a prediction of tertiary structure using conformation generation and scoring. It takes into account the secondary structure predictions as well as alignments from the server to produce an automatic three-dimensional protein structure prediction.

ProteinShop is an interactive tool for protein manipulation, created by the Visualization Group in Berkeley Lab's Computational Research Division. It has the ability to move pieces of the protein around by hand in an interactive 3-D environment. This can be a very powerful tool, but can also be very tricky to use and thus is not used as often as the other tools.

These tools, along with other viewing software and other leading servers, are used to predict the protein structures that are submitted to CASP7 for assessment.

2. Methods and Procedures

Using the sequence of amino acids and a library of proteins with known 3D structures, know as the "template library", we use our tools to align the parts of the proteins that are similar. This results in one of three cases.

2.1. Comparative Modeling

When the majority of the protein's structure can be determined from the template libraries and servers then we use comparative modeling techniques. In comparative modeling, Undertaker is able to determine the majority of the structure and all that is left to determine are some minor details. Usually we focus on closing breaks in the backbone chain, eliminating soft clashes of atoms, and improving the packing of the protein. This is done by altering the cost-function, which is an input file to Undertaker.

In addition to altering the cost-function, we can look through the top scoring alignments from the template library and choose to let Undertaker read in a selected number of them, rather than all of them. This can result in a slightly different model that follows the top scoring alignments more closely.

Another way for us to tweak the model slightly is to browse the secondary structure predictions looking for clues as to where each specific residue should be. For example, a hydrogen bond between two residues may help to form a hairpin turn between beta sheets or may give us a better idea about how two alpha helices pack together.

These small alterations in the position of the residues, rather than the actual structure and fold of the protein, are the focus of comparative modeling.

3. Fold Recognition

When a portion of the protein's structure can be determined from the template libraries and servers then we use fold recognition techniques. In this case, there is usually a significant potion of the protein that is similar in many of the top scoring alignments, but another portion which is very different. These structures and folds can be tricky to determine.

One method for determining the structure of the unknown parts is to create a sub-domain of the unknown part of the protein. We can then take this shorter amino acid sequence and run it through Undertaker again to see if any new results are found. Often times it is able to find alignments that were not found in the first run and from this we can get a very useful model. This sub-domain is then reattached to the first part of the protein, creating a chimera, and then optimizing the protein from there. This can be a very useful technique but often harbors the question of how the two (or sometimes more) sub-domains interact. This is where the local secondary structure predictions or ProteinShop can become very useful.

Many of the methods that were used in comparative modeling are also used in fold recognition. Tweaking the cost-function can allow Undertaker to change the conformation of the protein and can also be used to add specific secondary structure constraints such as hydrogen bonds, alpha helices, beta sheets and more.

It is important to come up with many different possible models with different conformations. This way, we can provide a variety of possibilities that we think may be the correct conformation.

3.1. New Fold

When there is not enough information in the template libraries to determine the structure or fold of the protein then we use new fold or ab-initio techniques. These cases are the most difficult in terms of making an accurate prediction. The predictor is essentially in the dark about how the protein folds into its 3-D structure.

In the best case scenario there is some information in the local secondary structure predictions that can give some clue as to structures, hydrogen bonding, and distance constraints between residues. However, often times the predictor has very little information and very weak predictions.

In this case it is even more important to generate a variety of possible models. Looking into distant alignments and secondary structure predictions is a necessity for creating likely protein structures.

4. Targets

I worked on 4 comparative modeling targets, 6 fold recognition targets, and 1 new fold target. With each target I worked with one or more other researchers in the lab and was supervised by Professor Kevin Karplus.

I was able to become familiar with all of the tools and all of the prediction methods as described above. I used cost-function parameter tweaking, as well as adding specific alpha helix, beta sheet, and hydrogen bond constraints that I deduced from the local secondary structure predictions. I also did some work with sub-domains and chimeras in conjunction with ProteinShop. Two of the targets were also dimers, and were optimized as such. This doesn't involve any new methods but it does take a lot more processing time since the protein is twice as large as normal. The reason that we optimize them as dimers rather than monomers is that it improves the structure of the interface between the two monomers which is hard to identify correctly without this kind of optimization.

In all, I was able to work with each of the elements that were involved in this CASP7 experiment, and hopefully, my work will help with the improvements of the tools and predictions in the future.

5. Conclusions

CASP7 ended in early August, but many of the protein structures have not yet been released. We are currently scoring our top models to the released models with a scoring method that was created by Professor Kevin Karplus' lab, but we will not be able to find out how we did in comparison to other predictors until November of this year.

We have noted, however, that it seems that the field of predictions is moving away from the hand tweaking implementations that we have been using and more toward using server models. In view this, the lab will be trying to improve the SAM server so that we will be ready to compete in CASP8 in 2008. The improvements will take into consideration all of the work that was done during CASP7, trying to implement an automatic system for what we did by hand.

Of course, the ultimate goal is to be able to create a system by which an amino acid sequence can determine the three-dimensional structure of a protein using purely computational methods. The results of this research during CASP7 will help to further that goal.

6. Acknowledgements

I would like to acknowledge the help and expertise of Professor Kevin Karplus and his Protein Bioinformatics Lab group at UCSC, as well as the SURF-IT group.

This work was completed as part of UC Santa Cruz's SURF-IT summer undergraduate research program, an NSF

CISE REU Site. This material is based upon work supported by the National Science Foundation under Grant No. CCF-0552688.

7. References

[1] Preuss, Paul. "'ProetinShop': solving protein structures from scratch." Science Beat Berkeley Lab. 28 February, 2003. http://www.lbl.gov/Science-Articles/Archive/CRD-proteinshop.html

[2] Kaplus, Kevin. et al., "Combining local-structure, fold-recognition, and new fold methods for protein structure prediction." *Proteins: Structure, Function, and Genetics.* 53.S6 (2003): 491-496.

[3] Karplus, Kevin. et al., *Origami with strings: protein folding by computer*. http://cubic.bioc.columbia.edu/meetings/casp-6-5/slides/casp65_karplus.pdf

[4] Karplus, Kevin. et al., "SAM-T04: What is new in protein-structure prediction for CASP6?" *Proteins: Structure, Function, and Genetics.* 61.S7 (2005): 135-142.

[5] Karplus, Kevin. et al., "What is the value added by human intervention in protein structure prediction?" *Proteins: Structure, Function, and Genetics.* 45.S5 (2002): 86-91.