Summary of the Hand-Modification Protocol Submitted by the SAM-T06 Team to CASP7

By Cynthia Hsu
University of California, Berkeley
Summer Undergraduate Research Fellowship in Information Technology
University of California, Santa Cruz

Abstract

The seventh semiannual Critical Assessment of Protein Structure Prediction (CASP7) is a means of analyzing the accuracy of the protein structure prediction methods conducted by a variety of different teams. The SAM-T06 team at the University of California, Santa Cruz submitted two sets of predictions: an automated set from their SAM-T06 server, and a series of models that had been modified and selected by hand. The result of these submissions showed that the hand modifications were a slight improvement over the automated ones, but not significant enough to make a difference. The main identifiable detriment to this was in the cost function used to evaluate the accuracy of the models. This paper outlines the procedure by which the different models were generated.

1. Introduction

The Critical Assessment of Protein Structure Prediction (CASP) is a community-wide experiment comparing the success of various protein structure prediction protocols. These included automatic server predictions, in which various algorithms search iteratively through databases and generate models accordingly, and "hand" predictions, in which a team of researchers analyze the output of the automatic server predictions and make modifications accordingly. For the purposes of the experiment, the sequences of a hundred proteins, whose structures had recently been deciphered from crystallography, were released to the CASP teams several weeks before the structures were released. From the sequences, all teams attempted to generate an appropriate representation of the three-dimensional structure. This year was the seventh time that the semi-annual CASP experiment had been performed.

The SAM-T06 team, headed by Professor Kevin Karplus at the University of California, Santa Cruz submitted two sets of predictions: those automatically generated by their server, and those which had been hand-modified by the team. The goal of this was to determine what improvements human intuition made on the server, and how these improvements could be automated by hand. This paper summarizes the results of the procedure by which the automated models were modified by hand and the conclusions that can be drawn regarding the protocol used by the CASP7 team based on preliminary analysis of the results.

2. General Procedure

2.1 Automated Modeling

Our submissions to CASP7 used the same protocol as did CASP6, but with an improved Hidden Markov Model (HMM), the SAM-T06, which had an improved ability to identify the top alignment [1, 2]. As with previous CASPs (namely 5 and 6), these HMMs would search through protein templates and fragments in an effort to recognize similar folds and secondary structure, after which the Undertaker fragment-packing algorithm would then use to generate a PDB file, which defined the placement of individual atoms in relation to each other in an x, y, z coordinate plane [2, 3].

The Undertaker algorithm also made additional computations for its initial model based on the local secondary structure alphabets, which were derived from alignment with similar amino acid sequences [2, 3, 4, 8, 9, 10]. These included the O_SEP and N_SEP, which dealt with the separation of two amino acids based on hydrogen bonds in the backbone, and the N_NOTOR and O_NOTOR logos, which described the separation of amino acids in hydrogen bonds with respect to their secondary structure. These supplemented the five alphabets used by Undertaker in the CASP6 protocol to predict backbone properties: DSSP, a simple calculation of probable secondary structure based on alignments; STRIDE, which calculated probabilistic secondary structure based on hydrogen bond energies and sidechain torsion angles; α-pseudotorsion angles, which divided the torsion angles of four successive carbon alpha atoms into eleven classes; STR2, an extension of DSSP that divided beta strands into six classes and differentiated between normal helices, 310 helices, turns, coils, and loops; and BYS, based on Bystroff's partition of the Ramachandran plot of the possible torsion angles between the planes of the backbone [2, 3, 4]. In addition, two burial properties were used, one that coordinated the beta carbon with the approximate center of a sphere of 14 Å, and the near-backbone alphabet, which counted how many residues were within a 9.65 Å radius to determine the level of burial [3, 8].

2.2 Hand Modifications

After generating the initial model, the coordinates in the PDB file can be loaded into the program RasMol, a three-dimensional molecular graphics viewer invented by the University of Edinburgh, in the United Kingdom, in 1989 [5]. Once in RasMol, residues can be viewed in various representations: wireframe, spacefill, alpha-carbon backbone, strands, and ribbons, and Richardson-style cartoons. Scripts can also be utilized to color-code the residues according to the lettering in the local alphabets: "ehl2', based on the dssp script, was the most basic one, in which probable helices were colored pink, probable sheets yellow, and unknown or unstructured coils were colored gray. Other scripts included the "near", "burial", "conserved", and "rr". The "rr" script, which identified residues that were most likely to be in close proximity, was one of the tools added for this season of CASP.

Using the scripts, several analyses could be performed on the initial PDB output of Undertaker. Most significant were the sheet constraints, an added feature to the CASP7 protocol. Often, Undertaker would produce sheets whose beta sheets were out of phase, meaning that the individual strands would buckle or twist in order for the appropriate hydrogen bond donors and acceptors to line up, as these would alternate on either side of the strand. The appropriate alignments could generally be derived by viewing the protein in cartoon representation, with selected residues of the strands in question shown in spacefill, then changing the specifications regarding the span of the strands and the first hydrogen-bond forming residue. In addition, sheets that were too short would often result in the formation of a helix, instead of a turn, whereas sheets that were too long would leave fewer than three residues for the hairpin turn and subsequently cause a break in the chain. By comparing the local structure alphabets such as N_SEP and O_SEP, as well as observing which residues were

identified as having which secondary structures using the "ehl2" script, it was possible to derive conclusions as to what the appropriate secondary structure of these residues were. The "ehl2" script was also useful in identifying how probable helices were, as often breaks were caused when a helix formed when it was not supposed to.

Another frequent problem with the initial alignment was in the placement of exposed and buried residues. Often, residues predicted be exposed by the near script were actually packed tightly against other regions of the protein, and likewise vice versa. For the buried regions, pockets of exposed residues that were predicted to be buried could often be brought closer together by selecting carbon beta atoms protruding from the backbone and including distance constraints for Undertaker to consider when generating future models. The same procedure could be used to identify and pack together certain residues whose R-groups should be in close proximity as indicated by the near script. However, this method would sometimes close the distance by rotating the molecule in the other direction, distorting the chain or possibly an entire subdomain of the protein. In these cases, ProteinShop was used to manually manipulate the orientation of a particular subdomain with the respect to the rest of the protein; this was a method that was novel to CASP7 [6].

Oftentimes, a full-length alignment with enough sequence identity was difficult to find. By examining the similarities and differences between the top alignments in the superimposed PDB file generated by a superimpose Undertaker script, portions of the sequence which were similar to a particular subdomain could be identified. These sequences could be isolated in their own file, after which the hidden Markov Models and the Undertaker algorithm would generate a new PDB file pertaining only to that portion of the sequences. These subdomains were then assembled into a complete model by cutting and pasting the coordinates from all the PDB files.

2.3 Polishing

When a reasonably acceptable model was produced, or in some cases, in which there would be a relatively high sequence identity between the sequence and the templates (>60%, known as comparative modeling), a procedure was enacted that was referred to as polishing. This meant rerunning the Undertaker algorithm, but with slight changes to the parameters entered into its cost function. The most important ones were the dry weights, which would increase the packing of the protein so that more atoms would be fit into a sphere of a particular radius. The weight on sidechains, breaks, and soft clashes were also increased, to increase the packing of the sidechains, to preserve the continuity of the backbone, and to prevent the electron shells of the atoms from clashing into each other. Unlike previous CASPs, these optimizations were typically generated by running the PDB file through GROMACS, a package for molecular simulation and trajectory analysis [7]. GROMACS optimized models were typically much more favored according to the Rosetta server, as they minimized clashes and breaks. These generally did not score as well with according to Undertaker's cost function; however, the skewed torsion angle of the backbone and the orientation of the peptide's R-group were typically corrected by raising the weight on these values in Undertaker's cost function. Additional modifications to the cost function typically included turning off the "maybe_ssbond" cost, if there were no cysteine residues in proximity, and turning off the "maybe_metal" cost, given the absence of a cluster of cysteines or histidines.

In instances when there was a high sequence identity between the target sequence and the templates, the secondary structure and fold of the first model generated by Undertaker would be identical to those models that came from the Hidden Markov Models. In such cases, restricting the sequence alignment to a single source alignment was necessary to create slightly different orientations or coil orientations.

3. Results and Conclusions

In general, when comparing our submissions to the fifty-eight targets whose structures have been released, the models that were generated by the hand modification procedure performed slightly better than those automatically generated by the server. However, our automated SAM-T06 server did much worse than the server predictions submitted by other teams, and the hand modifications were not enough to compensate for this.

One principle drawback to our methods was the failure of our cost function, by which we scored the accuracy of various models, to represent the flaws in both our models and those of other servers. The top scoring server in the CASP7 competition, the Zhang server, produced models that were accessible to us during the time of the competition. However, we generally failed to recognize the accuracy of these models, as they were rated as inferior according to Undertaker's cost function. This was true of many other server models.

The flaws in our cost functions also led to many other mistakes: often, a more accurate model we had generated was discarded in favor of a less accurate one; the majority of these decisions were, in cases in which there was a fairly high sequence identity (40% or more), based upon the scores indicated by our cost function.

4. Future Work

Because the main flaws in our protocol were due to the failure of the Undertaker cost function to adequately select the best model, the main focus of future work is to improve this cost function so that it focuses less on local details and more on the overall arrangement of the backbone. The main focus of this is to include an extra cost for burial, as this is the key means by which Undertaker identifies the relative location of the residues of the backbone in space.

Further analyses need to be run to analyze the accuracy of models that have been optimized by GROMACS in comparison to our own models, and whether running GROMACS-optimized models through polishing runs in Undertaker improves them.

5. Acknowledgements

The protocol for the submissions to the CASP7 experiment were designed and supervised in their entirety by Professor Karplus. This research would also not have been possible without the guidance from the other members of the SAM-T06 team: Martin Madera, George Shackleford, Grant Thiltgen, Firas Khatib, Pinal Kanabar, Zack Sanborn, Chris Wong, Navya Swetha Davuluri, Sylvia Do, and Crissan Harris. This work was completed as part of UCSC'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. I am also very grateful to the people who coordinated the SURF-IT project: Richard Hughey, Roberto Manduchi, and

Colt Hangen. In addition, I would like to thank Christoph Rau, Elisabeth Kain, Alex Williams, and Martina Kovea for their company and support.

References:

1.K arplus K, Sjölander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. Predicting protein structure using hidden Markov models. Proteins 1997;Suppl 1:134–139.
2. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R. SAM-T04: What is New in Protein Structure Prediction for CASP6? Proteins 2005;Suppl 7:135-142.
3. Karplus, K. and Karchin, R. and Draper, J. and Casper, J. and Mandel-Gutfreund, Y. and Diekhans, M. and Hughey, R. ``Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction'' Proteins: Structure Function and Genetics53(S6):491-496,15 Oct 2003
4. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
5. Sayle R, Milner White E J. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. 20 (1995);9:374-376 .
6. Crivelli S, Kreylos O, Hamann B, Max N, Bethel W. ProteinShop: A tool for interactive protein manipulation and steering. Journal of Computer-Aided Molecular Design 18 (2004);4:271-285.
7. Lindahll E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. of Mol. Modeling 7 (2001);8:306-317.
8. Rachel Karchin, Melissa Cline, and Kevin Karplus. Evaluation of local structure alphabets based on residue burial. Proteins: Structure, Function, and Genetics, 55(3):508--518, 5 March 2004.
9. Rachel Karchin, Melissa Cline, Yael Mandel-Gutfreund, and Kevin Karplus. ``Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry.'' Proteins: Structure, Function, and Genetics, 51(4):504--514, June 2003.
10. Karplus, K. and Karchin, R. and Barrett, C. and Tu, S. and Cline, M. and Diekhans, M. and Grate, L. and Casper, J. and Hughey, R. ``What is the value added by human intervention in protein structure prediction?'' Proteins: Structure Function and Genetics 45(S5):86-91,2001.