

Protein Structure Prediction Using Rosetta

Deandre Q. Baynham, Mert Karakaş, and Jens Meiler

KEYWORDS. Protein, structure prediction, Rosetta

BRIEF. The accuracy of protein structure prediction was evaluated using a Rosetta program.

ABSTRACT: Proteins have various functions in the human body that can be better understood with an accurate model for their structure. There are several methods to determine the structure of a protein experimentally, but these methods are not applicable to all proteins. If the structure of a protein cannot be determined experimentally, computational tools can be applied to predicted structure. In this project, the Rosetta protein structure prediction program was tested on several proteins to determine the accuracy of this protocol for predicting protein structure. The primary sequence of the proteins were input to several programs for secondary structure prediction., then Rosetta created models for tertiary structure using this information. Success of the method was determined by computing the root mean square distance (RMSD) between atoms in the model and in the experimental structure. It appears that smaller proteins have lower RMSD values than the larger ones. This indicates that the protocol is most effective at modeling small proteins, normally less than 150 amino acids in length.

INTRODUCTION.

Since the early years of biochemistry proteins have been the focal point of the field. After the discovery of diastase, the first known enzyme found by Anselme Payen in 1833, many scientists made huge progress in this field, trying to understand the chemical processes inside organisms. Proteins are biological macromolecules that are very important to functions in the body, and are still the main focus at the center of biochemical studies today, including the folding of proteins how folded proteins interact with one another, substrates, drugs, DNA or RNA. Proteins are synthesized in our body from monomeric units called amino acids. Therefore amino are also energy metabolites and essential nutrients.

To understand proteins, one must first consider how the genetic code memorized in a DNA molecule becomes a protein. The first step of this process is known as transcription and begins with the unwinding of DNA by a protein from a group called helicase. Next, a strand of RNA is attached to the unwound DNA, and starts to replicate its nucleotides. After this process is finished, the new messenger RNA molecule has the nucleotides that correspond with those that were on the original DNA molecule. This messenger RNA molecule that is then read out in the ribosome while the transfer in a process known as translation. During this process, RNA recruits amino acids to the messenger RNA. Each group of three nucleotides (codon) of the messenger RNA determines a specific transfer RNA and thereby a specific amino acid. Typically between 80 and 300 amino acids are translated to form one protein. However, smaller proteins with as few as 30-40 amino acids are known, as well as larger proteins with more than 1000 amino acids [1].

Most proteins are made of 20 standard amino acids, except few cases where non-natural counterparts. Although all of the standard amino acids differ a little, they all have the same principal core structure. In all of these amino acids, except proline, there is a carboxylic acid (COO^-) group, a primary amino group (NH_3^+), and a variable R group bonded to a central carbon atom. Since amino acids have a carboxylic group and the primary amino groups, they can act as either an acid or a base.

These amino acids react to form a polypeptide which than folds into secondary and tertiary structure. The most important types of secondary structure include α -helices β -strands. Other regions of the polypeptide stay mostly flexible forming loop sections. The secondary structure elements then come together to form the tertiary structure, or folded structure, of the protein.

The tertiary structure of a protein dictates its function. Therefore, obtaining an atomic resolution tertiary structure model of a protein is crucial in order to get a better understanding of its dynamics and continue further biological studies. However, many proteins of interest evade experimental methods such as X-Ray crystallography and Nuclear Magnetic Resonance (NMR). In such cases, computational programs are used to predict a structural model for these proteins. There are several ways to predict the tertiary structure of a protein computationally. In this project, the accuracy of tertiary structure prediction was evaluated using Rosetta, a commonly used program, via benchmarking over a set of proteins of variety of topologies [2].

MATERIALS AND METHODS.

The primary sequence forms the starting point for protein tertiary structure prediction. By convention the primary sequence is represented as a FASTA file which contain one-letter codes of all amino acids in the sequence. Since Rosetta will be tested on proteins for which the structures were determined experimentally, the experimental structure for each protein in the benchmark set was obtained from the Protein Data Bank (PDB). It, represent the native structure of a protein by individual coordinates of each atom.

Once the FASTA files were obtained, the BLAST program was used, which aligns the sequence to all known sequences and calculates for each position in the sequence how frequently that amino acid was substituted with another amino acid in a similar sequence. The end product of this program is a profile for each position in the sequence, giving the likelihood for observing each of the 20 natural amino acids in that position.

Once the BLAST profile was generated for the sequence of interest, several secondary structure prediction programs were run on each sequence to predict which stretches of the sequence are likely to be an α -helix or β -strand. These programs were PSIPRED and JUFO, which utilize artificial neural networks, in addition to SAM which uses Hidden Markov Models. The BLAST profile is an input to these programs.

After the completion of secondary structure prediction, the last task required before running Rosetta is the generation of "fragments". The program iterates over each overlapping three and nine residue stretch of the sequence of interest, then looks for similar stretches of sequences, thus fragments, from proteins with experimentally determined structures, and picks 200 of such fragment conformations for each position in the sequence. These fragments are collected in a large file that forms the fragment database.

Following this, the Rosetta program was used to predict 50,000 structural models for each protein in the benchmark set using the secondary structure predictions and fragment files.

The structural models generated by Rosetta were evaluated by looking at root mean square deviation (RMSD) values. The RMSD value of a model indicates how close it is to the native structure. First the RMSD values for 50,000 models were converted to histograms in order to look at not only how close the models got to the native structure, but also how frequently such good models were obtained. In addition, by comparing the histograms for different proteins, it can be analyzed if for proteins of a certain size and topology, Rosetta generated more accurate models.

One would also expect Rosetta to rank good models by RMSD with good scores, in order to be able pick such accurate models out of the 50,000 models. In order to see if this happened, RMSD values of all 50,000 models were plotted against the score that was calculated for that model by Rosetta.

RESULTS.

The protocol was completed for six benchmark proteins. The figures below show the native structure, the best Rosetta generated model, and an RMSD distribution plot for each protein (Figure 1-3). 50,000 models exhibit a large range of RMSD and energy values which is expected for protein structure prediction.

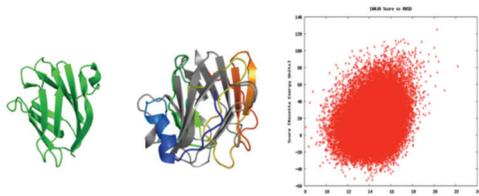


Figure 1. The native structure of 1AAJ, (green), the model with the best RMSD, 8.2Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

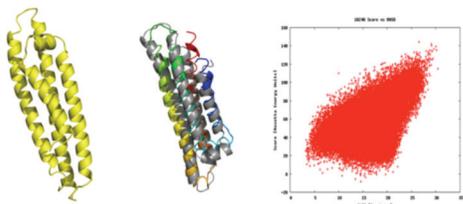


Figure 2. The native structure of 1BZ4 (yellow), the model with the best RMSD, 3.9Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

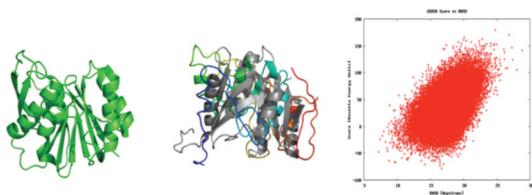


Figure 3. The native structure of 1DUS (green), the model with the best RMSD, 7.3Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

DISCUSSION.

The results indicate that Rosetta can accurately model the topology of smaller proteins. The accuracy of the models generated correlates with the complexity of the protein topology and the sequence length. This is an expected behavior, since for larger proteins more conformations have to be sampled by Rosetta. The accuracies ranged from as low as 3.9Å for 1BZ4 to as high as 12.5Å for 1CHD. Although very accurate models were generated for the smaller proteins, the remainder of models had a wide range of RMSD values. At the same time, the plots indicate that most of these high RMSD value models can be removed using the energy as a filter.

In the future, a more detailed benchmark will be executed including 54 proteins and the evaluations are done also using measures other than RMSD. An additional analysis would be to evaluate how likely good RMSD models can be selected through clustering or energy in the absence of information regarding the native structure. This analysis will also be completed for an alternative protein structure prediction program BCL::Fold, which is being currently developed in Meiler lab, in order to assess the strengths and weakness of both programs.

ACKNOWLEDGMENTS. Thanks to Mert Karakaş for the mentorship and guidance throughout the project. Extended thanks to Dr. Jens Meiler for allowing me to work in his laboratory. I thank the reviewers of my manuscript for many excellent suggestions and the whole Meiler lab at Vanderbilt University.

SUPPORTING INFORMATION.

Figure S1. The native, best model by RMSD and RMSD versus energy plot for 1BGC.

Figure S2. The native, best model by RMSD and RMSD versus energy plot for 1BJ7.

Figure S3. The native, best model by RMSD and RMSD versus energy plot for 1CHD.

REFERENCES.

1. K.T. Simons, et al. *J. Mol. Biol.* **268**, 209(1997).
2. P. Bradley, et al. *Proteins*, **61**, 128(2005).



DeAndre Baynham is a student at Hillsboro High School and enrolled in the School for Science and Math at Vanderbilt.