Improving Protein Structure Prediction Utilizing A Reformed Filtering Process

Rachel Davis
Drake University
Des Moines, Iowa
Email: rachel.davis@drake.edu

Jennifer Ogden
St. Mary's College
Moraga, California

Silvia Crivelli
University California Davis,
and Lawrence Berkeley National Lab
Berkeley, California
Email: sncrivelli@lbl.gov

***Abstract*—Knowledge of protein structure grants insight on a protein's function and capabilities within a system. Currently available automated prediction of structures is limited and inconsistent. The Critical Assessment of protein Structure Prediction (CASP) is a worldwide competition to automate the process, and supplementary WeFold allows competing groups to connect with ideas on how to predict proteins. By examining a popular filtering process on the WeFold pipeline, a clearer look into one aspect on why prediction is untrustworthy is presented. Comparing the number of accurate models based on a global distance test to a given template before and after the filtering procedure gives a sense of how well the program is filtering. By adjusting the parameters presented in the filtering code, more accurate results can be yielded.**

***Keywords*—CASP, WeFold, Protein Structure, Filtering**

I. Introduction

Within the human body there are trillions of protein structures [1]. The role they play is essential to many biological systems ranging from structural basis to chemical reactions. Often proteins are associated with disease, as misfolded protein can lead to such diseases. Figure 1 depicts the four separate stages of protein folding. Proteins are typically made up from a set of 20 different types of amino acids arranged in one or more chains [2]. These chains range in length between tens and thousands [2], called a primary structure. The primary structure chain is in the top left corner of Figure 1. Next comes a secondary structure made up of three important forms: alpha helices, beta sheets, and random coils [1]. Figure 1 shows these structures in the lower left. Lastly, the tertiary structure is the three dimensional shape that largely determines the function of that protein. This is the upper right of Figure 1 and depicts a protein that's been folded. A quartenary structure, the last in Figure 1, is a series of proteins linked together.

II. Background

The Critical Assessment of protein Structure Prediction (CASP) is a competition to determine unknown protein structures using automated methods [3]. Currently there exists no reliable system of protein folding, in spite of CASP working towards such a program since 1994 [6]. Stemming from efforts of CASP is WeFold, a coopetition [1]. Those who compete in CASP share pieces of their prediction method on WeFold. Separate pieces of these methods are assembled into pipelines.

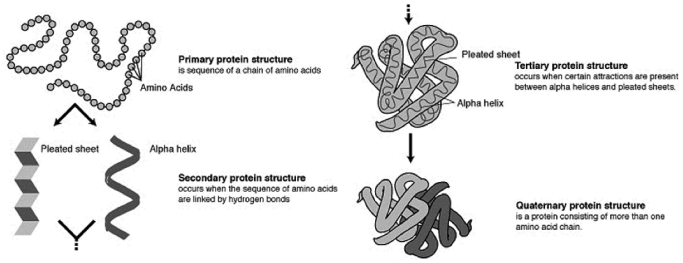

Fig. 1. Progression of Structures [12]

In order to improve the automated prediction a careful examination of pieces of the pipeline is needed. A solid building block is in the filtering of models collected from worldwide program Foldit [3]. Foldit is an interactive game in which players fold proteins into models to try and score the lowest rosetta energy; rosetta energy being the value that determines the likelihood of protein accuracy. Every move from Foldit is saved as a possible structure, which leaves roughly a hundred thousand predicted structures per target protein. This is where the filtering becomes important. If too many accurate models are filtered from the full set, a higher chance for accuracy is lost. Yet if not enough models are filtered the data set is too large to work with. Therefore, an accurate filtering process is highly important.

III. Methods

The main portion of this code is written in various Python [7] scripts. The basis of this Python code was from George Khoury. It was within this code that the important variables for filtering were modified. Namely were two particular scores: SASA and Secondary Structure. The Solvent Accessible Surface Area (SASA) score is one of the main factors in testing a model. There is a program called naccess, which calculates the SASA for a given model.

The second important number involves calculating the number secondary structure elements. STRIDE [9] is the program utilized that generates this secondary structure. The final piece of the project utilizes Zhang Labs TMscore [10]. The TMscore generates a global distance test total score, abbreviated GDT-TS. This is what the CASP scores the models on. Comparing each structure to the officially released crystal structure of a protein determines accuracy.

The last major piece of technology was Hopper Cray

XE6, a supercomputer. The National Energy Research Scientific Computing Center (NERSC) was gracious enough to aid WeFold in their efforts. Hopper has two hundred and twelve terabytes in memory and placed fifth on a list of supercomputers [5].

1) Experimental Numbers: There are many variables when setting filter numbers that make it difficult to fit accurate numbers. The variables were set largely upon percentage, but even that has risk. A larger protein will have a higher percentage of error while a smaller protein is harder to beat.

2) Algorithm: The first step to filter the code was to establish which structures were unique. The C++ IDUniqueConfs program compares the structures and establishes which are too similar to be counted as multiple structures. This code generates a file with a list of all structures within the data set and which are unique and which need to be discarded. After these structures have been discarded comes the largest portion of the research: George Khoury's Python.

The code first looks for the file generated by IDUniqueConfs, here forward to be referred to as Unique, then parses this file to contain only the unique models. Next, each model is separately run through the body of the code. The first test is the naccess code, which generates the SASA for this specific .pdb file, or protein database. This number is then compared against the best energy proteins SASA multiplied by some factor. Note that this is the first experimental number. If the SASA for this model is acceptable, the code continues to the next check.

This next test is the secondary structure checker. The same procedure is repeated, but instead of SASA it is the number of secondary structure elements multiplied by some modifier. These numbers are not evaluated the same way. The SASA number is an energy function–thus the lower the energy the more likely the protein is appropriate. The number of secondary structure elements is not a score to be passed, but is more a guideline. A protein with only one secondary structure element likely will not be the real protein; therefore our projected number will be above our set number. It is only if these two tests pass that we consider it a viable candidate and passes the filter. Figure 2 is the graphic representation of each of these moving parts in comparison to the time commitment to each part.

The tests themselves were run on a particular target, TR722. It is a unique target in that it has two forms: a symmetric form and a normal form. Tests were generated form the normal form, because the data created was more accurate than the symmetric.

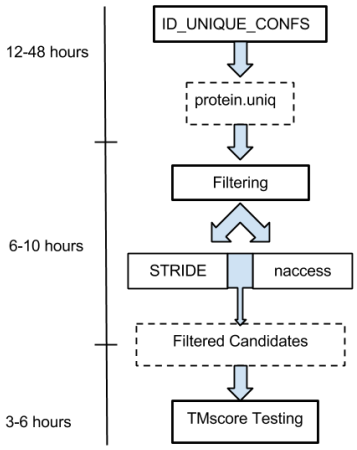

Fig. 2. Structure of the Code

Test Name	SASA	#SS	% Above	#Models
Baselines	-	-	7.78123	94204
Unique	-	-	8.1757	21723
First	+25	-20	8.434	4200
Second	+35	-25	9.82779	6741
Third	+45	-25	8.56	18941
Fourth	+55	-25	8.503	19005
Fifth	+45	-50	8.084	20930
Sixth	+40	-30	8.5475	17215
Seventh	+40	-20	8.6024	16286
Eight	+45	-20	8.5282	16381

Fig. 3. A table showing the associated SASA and secondary structure elements of each test in conjunction with their percentage above the template and the number of models in the final set. The top two rows show the initial data set and the unique set respectively.

IV. Results

Many tests were run with varying results. As the SASA scores increased, more models were left in the data set. Similarly as the secondary structure requirements decreased in strictness there was even less models sorted out. A baseline test with a forty-five percent increase in SASA and a fifty percent decrease in secondary structure elements filtered out only 793 models. Figure 3 is a shows a sampling of the tests.

From this table in Figure 3 we can tell that the baseline has an almost 7.8% above the template rating. This was with a total of 94,204 models. The image above in Figure 4 shows a histogram of such information. The dotted line depicts the template of the figure before any refinement. You can see that many moves hurt the structure more than improved. The lower image in Figure 4 is a figure showing histogram after the unique files had been established. This left 21,723 models for the filtering process to cover. This data set had an 8.2% above template score.

Figure 3 also clearly shows was one score was well above the others with a ranking of 9.8% above template. This nearly 10% accuracy with one thirteenth of the data is a great start to filtering. For interest sake, Figure 5 is the histogram of the best filtered candidates.

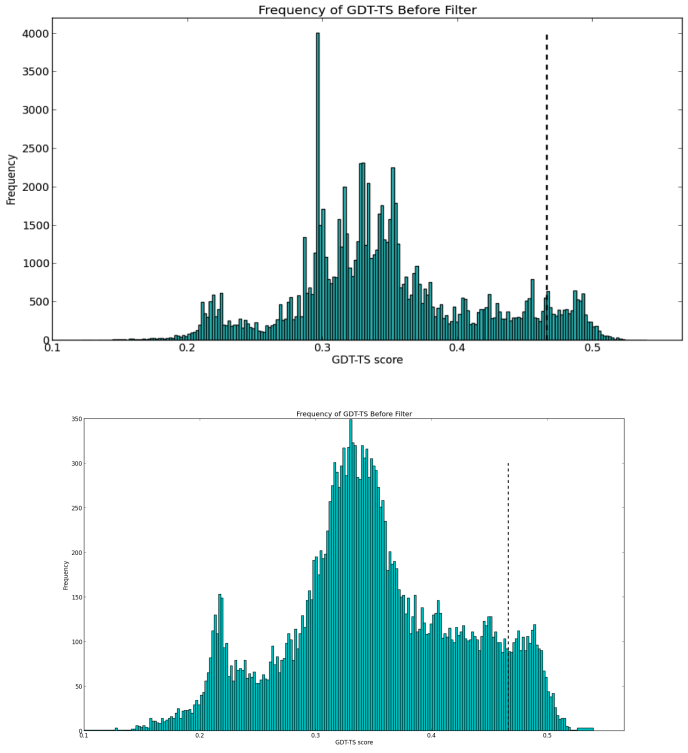

Fig. 4. Histograms of the baseline data and the unique data.

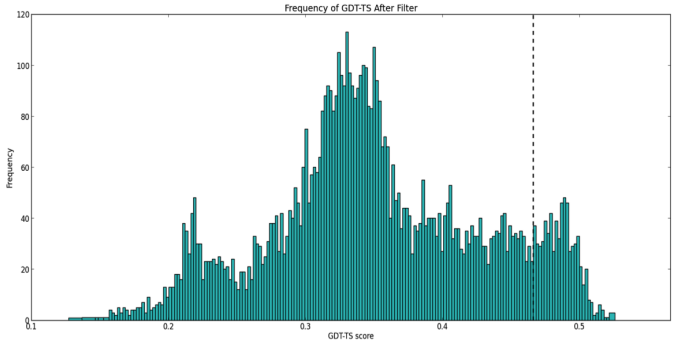

Fig. 5. Histogram of the highest accuracy data set.

V. Conclusion

This code was the first attempt to filter models before working with them. The full data set sometimes had a reasonable amount of accuracy, although often it was less than 10%. After running the Unique code, this accuracy window has potential to grow even smaller. While the percentage could remain the same or grow smaller, proportionally more accurate models would have been lost. This research did not explore with this specific program and no changes were made to improve upon it.

In conclusion, a nearly 10% accuracy was very encouraging. Considering the number of models sorted from the initial set was over ten times in size with an accuracy higher than both baseline sets, this was a significant increase. Data not shown above includes a test that was run on target TR705. A test was run using this 9.8% increase program on the initial data set. Overall, this test showed a nearly 5% increase. With these two sets in particular, a relatively high certainty can be placed on this program to accurately filter to a smaller data set with an increase in above percentages.

A. *Future Work*

With every research project there is always room to improve. Firstly, more tests would give a closer approximation of where the SASA and SS bounds should lie. A tenth of a percent could potentially create a significant difference in the accuracy of prediction. The next biggest change would be the flexibility in code. Where the code was left off had a very static pathway used, and could stand to become much more dynamic. This was a valiant effort, but as always is put at the wayside in favor of generating more structures. Hopefully the successor of this work will take up the mantle of these challenges.

Acknowledgments

The authors would like to thank NERSC for flexible computing power and ever-helpful insight; our program directors, Aimee Tabor and Elizabeth Bautista, for bringing us together in this amazing event; and our families for their unending support. We would also like to extend a special thank you to all collaborators of WeFold.

Another thank you belongs to Silvia Crivelli, Yushu Yao, and Taghrid Samak for ever important guidance as wonderful mentors. Lawrence Berkeley National Lab has supported our endeavors along the way, and we are grateful. Finally, the NSF funded this program and we are eternally grateful for the opportunity.

References

[1] S. Crivelli. Lawrence Berkeley National Lab Lecture. June 26th 2014.

[2] S. Fairchild, R. Pachter, and R. Perrin. "Protein Structure Analysis and Prediction" in the Mathematica-Journal, vol. 5 issue 4, pp. 64-69. Fall 1995.

[3] G. Khoury et al. "WeFold: A Coopetition for Protein Structure Prediction," Accepted Article, doi: 10.1002/prot.24538, 2014.

[4] D. Frishman and P. Argos, "Knowledge-Based Protein Secondary Structure Assignment," Proteins, vol. 23, pp. 566-579, 1995.

[5] https://www.nersc.gov

[6] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round x." Proteins, vol. 82 (Suppl 2), pp. 1-6, 2013.

[7] https://www.python.org

[8] https://www.rosettacommons.org

[9] http://webclu.bio.wzw.tum.de/stride/

[10] http://zhanglab.ccmb.med.umich.edu/TM-score/

[11] http://www.bioinf.manchester.ac.uk/naccess/

[12] http://www.umass.edu/molvis/workshop/imgs/protein-structure2.png