Utilizing A Reformed Filtering Process to Improve Protein Structure Prediction

Rachel Davis[1,4], Jennifer Ogden[2,4], Silvia Crivelli[3,4]
[1]Drake University, [2]St. Mary's College, [3]UC-Davis, [4]Lawrence Berkeley National Laboratories

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 results are inconsistent. The Critical Assessment of protein Structure Prediction (CASP) is a worldwide competition, 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, before and after the filtering procedure gives a sense of how carefully the program is filtering. By adjusting the parameters presented in the filtering code, more accurate results can be yielded.

Introduction

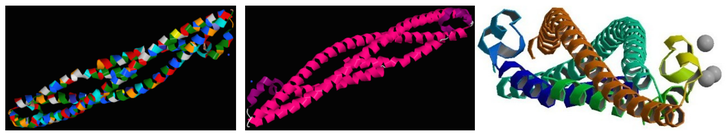
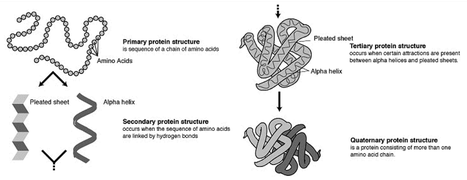


Figure 1: Left shows the amino acid sequence, or primary structure, as a contrast of colors. In the middle is the outlines of the secondary strcture components made up of Alpha Helices and Beta Sheets. Secondary structures also sometimes have Random Coils. Right is the officially released structure. Images of TR722 taken from predictioncenter.org

Within the human body there are trillions of protein structures [1]. They are essential to life, but often proteins are associated with diseases because a misfolded protein can lead to disease[4]. Proteins are typically made up of chains ranging in length between tens and thousands of amino acids [2]. The Critical Assessment of protein Structure Prediction (CASP) is a competition to determine unknown protein structures using automated methods [3]. Currently no reliable automation exists, in spite of the 11 seasons of this project. Stemming from CASP is WeFold, a coopetition [1]. Some who compete in CASP share pieces of their methods on WeFold. In order to improve the automated prediction a careful examination of pieces of the pipeline is needed. Due to the large number of submitted structures, a solid building block is in the filtering of models [3,8]. We need to eliminate enough models, but if too many accurate models are filtered from the full set, a higher chance for accuracy is lost. Therefore, a need to improve the filtering process is important.

Important Values



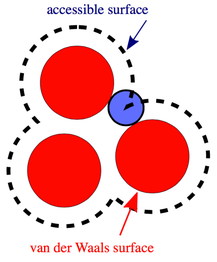


Figure 2: Figure upper right is a depiction of the multiple steps to protein folding. The lower left is the secondary structure step and an important number for filtering. Image source: www.umass.edu

Lower left shows how the Solvent Accessible Surface Area is obtained. SASA is one of the important scores for the filtering process. Image created by: Keith Callenberg.

Tools and Technologies

-Filtering code by George Khoury written in Python [7] which includes:
- naccess, a SASA establishing code [11]
- IDUniqueConfs, generates unique models
- STRIDE, a protein editing program [9]

-TMscore code from Zhang Lab to provide GDT-TS measurements [10]
written in FORTRAN
-Models collected from Foldit[6]
-Hopper, a supercomputer

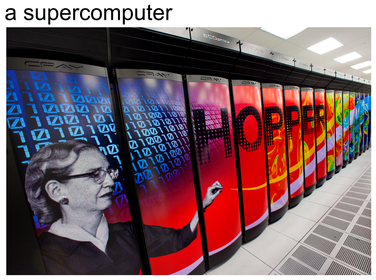


Figure 3: Supercomputer Hopper is a Cray XE6. She has 212 terabytes of memory, and over one hundred and fifty thousand computing cores[5]. Image Source: nersc.gov/nersc-40/galleries/historical-systems

Algorithm

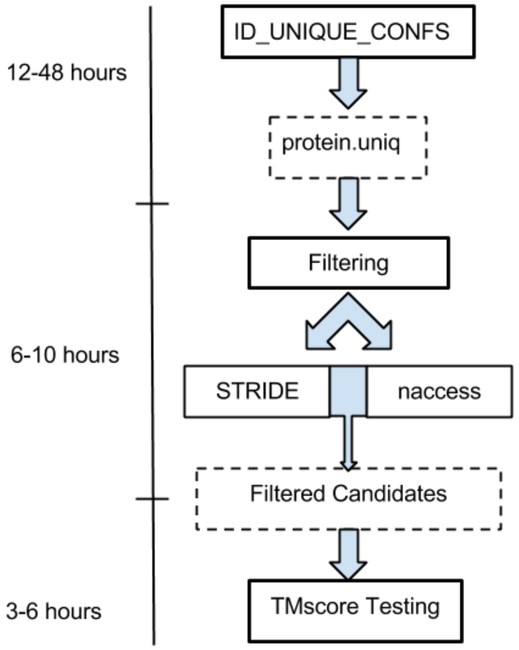
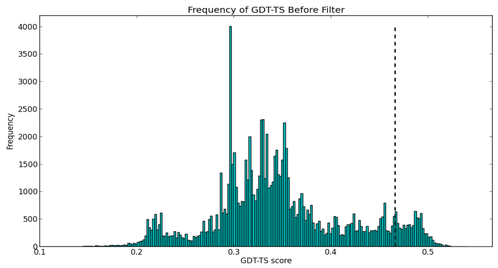


Figure 3: The image left depicts the filtering procedure and the timeline. The first step involves determining unique structures from the full set, then filtering the code using STRIDE and naccess. Finally, the code is tested using Tmscore.

After the initial unique code was run, because it only needs 1 run per target, an average test took around 9 hours.

Visual Results



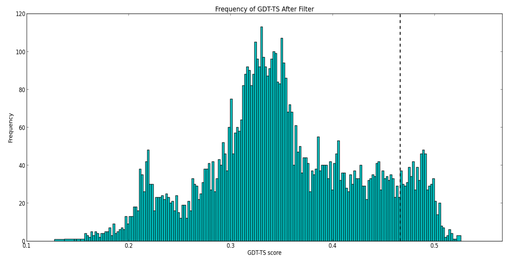


Figure 4: The above left figure is the histogram of data before filter. Right is the histogram of the after filter. The dotted line is the initial template.

Table 1: The table below is a sampling of the datasets and tests run.

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

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/

Conclusion

Overall the filtering was successful. Choosing a structure as an accurate model had a 2% increase with an drastic decrease in number of models. Changing the data set size by a quarter of the starting models can have a large computational impact in both time and accuracy. Proportionally, it seems to be the ID_UNIQUE_CONFS code that has the last amount of filtering success.

Future Directions

Make the code:
- More Dynamic
- More Efficient
- More accurate

As always, we hope to make improvements for CASP12

Acknowledgments

The National Energy Research Scientific Computing Center was a great help in this research. NERSC is supported by the Office of Science at the US Department of Energy under contract number DE-AC02-05CH11231. Thanks also belong to George Khoury, Zhang Lab, CASP participants, & especially those who collaborate in WeFold.