African Journal of Biotechnology Vol. 11(53), pp. 11625-11629, 3 July 2012
Available online at http://www.academicjournals.org/AJB
DOI:10.5897//AJB12.849
ISSN 1684–5315 ©2012 Academic Journals

Full Length Research Paper

Are specialized servers better at predicting protein structures than stand alone software?

B. N. Saeed[1]* and H. Q. Rabail[2]

[1]Department of Biosciences, SZABIST, Karachi, Pakistan.
[2]Department of Computer Science, SZABIST Karachi, Pakistan.

Accepted 1 June, 2012

This research study answers the question that technology is the best for predicting protein structures. Stand-alone software only depend on protein structure prediction algorithms, while web servers consult a number of other sources such as meta servers and protein data banks to produce a protein structure achieved through consensus.

Key words: Protein structure prediction, web servers, meta servers, stand-alone software, insulin.

INTRODUCTION

DNA is the code of life. It is the basic building block of universal existence. Since the beginning of 21st century, scientists have been occupied with the study of DNA and how to decipher it. Understanding DNA holds the key to solving many problems in not only molecular biology but also biotechnology, genetic engineering and DNA computing. It has a very complex structure and although it was deciphered many years ago, it still contains a lot of information that has not been decoded yet. DNA is made up of amino acids and proteins. The actual challenge is to understand how a DNA folds itself, what rule or pattern does it follows. This question has haunted scientists for quite some time and it is believed that if the answer to this question is discovered, it will be one of the major discoveries of the 21st century.

Protein structure prediction is defined as predicting 3D structures from protein sequences (Baker and Sali, 2001). It is considered as being the holy grail of molecular biology. The traditional lab methods of protein structure prediction are expensive and time consuming, hence computing algorithms and tools have been developed to aid in protein structure prediction (Baker and Sali, 2001). There are thousands of bio-molecular software and tools available on the Internet for the purpose of protein structure prediction and it is a daunting task to choose the best amongst them. There are no proper standards and guidelines available for the development of such tools and hence it becomes even more difficult for a microbiologist to choose which software or tool will suit the purpose in the best possible way. This research report focuses on two technologies being used for protein structure prediction. One is stand alone software and the other is web server. This research explores both these technologies and by experiment and quantitative analysis, answers the question that which one of the two technologies is the best at protein structure prediction.

Literature review

Mihansan (2010) has classified protein structure prediction tools into 3 categories: Stand alone program, a server and meta server. He believes that web servers have made a biologist's life easy (Mihasan, 2010). The biologist no longer has to go through the tiresome process of searching for appropriate software, downloading it and installing it (Mihasan, 2010). Web servers have freed the biologist from using his own computational resources and save his time and money. By using a web server for protein structure prediction, all the computations are done elsewhere. The biologist only has to input a protein sequence using a web browser and the results will be emailed to him after few hours or days (Mihasan, 2010).

*Corresponding author. E-mail: bushra406@gmail.com.

According to Mihansan's results, Meta servers perform even better than simple web servers (Mihasan, 2010). Meta servers produce more accurate results as they consult various different servers and databases and combine the best possible results which are then mailed back to the biologist. According to Mihansan's paper, Swiss model is the best and most widely used protein structure prediction server followed by 3D- Jigsaw (Mihasan, 2010). He also suggests that the stand alone program modeller is also a good option as its open platform and can run smoothly on Linux, Windows and Mac (Mihasan, 2010). Bujnicki and Fischer (2004) suggested that it would be a better idea if scientists use a combination of various different models and methods (Bujnicki and Fischer, 2004). By doing so, they will have better and more accurate results (Bujnicki and Fischer, 2004). Hence they suggest that scientists and biologists should not rely on the results of just one server because most servers cannot differentiate between weak hits and wrong hits (Bujnicki and Fischer, 2004). They did an analysis of five Meta servers which are: PMOD, PCON, ALEPHOJURY, Rosetta and 3D SHOTGUN.

Their results suggested that Meta servers are much more accurate at predicting structures than the simple primary servers. They concluded that the higher the n in Meta n, the better the results would be (Bujnicki and Fischer, 2004). From all the five Meta servers the duo analyzed, they concluded that Rosetta is the best and produces efficient and better results than others (Bujnicki and Fischer, 2004). Ginalski et al. (2003) suggested that Meta predictors are much more accurate than individual protein structure prediction algorithms (Ginalski et al., 2009). They conducted an experiment using the meta server 3D Jury to show how it produces high quality, accurate models using sets of models created with the help of various methods (Ginalski et al., 2009).

Zhang (2009) concludes that an algorithm I- Tasser, produces accurate models because it incorporates various templates from other servers (Zhang, 2009). In another paper, he suggests that LOMETS, which is a local threading Meta server; produces accurate models within less period of time if run on a local computer cluster (Zhang, 2009). He suggests that experiments show that models produced by Meta servers are at least 7% more accurate than the models produced by simple servers (Zhang, 2009). Kim et al. (2004) stated in their study that their experiment showed that the Robetta server can produce very good quality accurate protein models (David et al., 2004). Kelley and Sternberg (2009) conducted an experiment to show that a server called Phyre can do protein structure prediction of a protein sequence that has 250 residues in only 30 min (Kelley and Sternberg, 2009) which is a big achievement as it shows that the time factor can be lowered in case of web servers and Meta servers.

Eyrich and Rost (2003) suggest that Meta- PP is another web server which produces quality models (Volker and Rost, 2003). Similarly, Kaufmann et al. (2009) illustrated that the Rosetta protein modelling suite performs good quality De Novo protein structure prediction and has successively performed well in Critical Assessment of Structure Prediction (Kristian et al., 2010). Most of the research papers show that web servers and Meta servers are better at protein structure prediction than stand alone software. But the major problem faced by web servers is how to minimize time taken to produce results. Lee et al. (2009) have developed a protein structure prediction pipeline for computing clusters (Lee et al., 2009). This pipeline is a standalone protein structure prediction software package that performs all three types of protein structure prediction and allows the users to submit unlimited number of queries (Lee et al., 2009).

Problem domain

Protein structure prediction is one of the most challenging problems in today's world. A lot of scientists believe that it holds the key to finding cures for deadly diseases like acquired immune deficiency syndrome (AIDS) and cancer. It can also aid in making a DNA computer and unleash great computing power. Various algorithms and software are used to predict a protein sequence of any given molecule, but all the techniques have their pros and cons (Wikipedia, 2011). Most algorithms and software can easily do homology modelling of proteins, but a majority of them fail at Ab initio modelling. The potato dextrose broth (PDB) template database is said to hold enough protein structure templates to predict protein structure of any protein whose protein structure is not known (Zhang and Skolnick, 2005). But still this is debatable since some researchers feel that the PDB library is still not sufficient.

Two kinds of computing technologies are used in protein structure prediction; specialized web servers and stand alone software (Baker and Sali, 2001). Specialized web servers are connected to special online databases and PDB libraries whereas the stand alone software only uses modelling algorithms without consulting any database or any web server (Baker and Sali, 2001). In our first independent study we focused on two stand alone software. We showed how the two software; Abalone and Biodesigner can be used to do protein structure prediction of HIV/AIDS. Although, the software did produce a 3D model of the virus and also incorporated drugs into it, the question was that how reliable and accurate the 3D model produced by the stand alone software is? Another fact that discovered was that the stand alone software failed to do protein structure prediction of amino acid sequences which had more than 120 residues. The stand alone software could only do homology modelling based on PDB template files obtained from the PDB library. Having read several

Table 1. Names of 5 stand alone software and 5 web servers.

Stand alone software	Web servers
Abalone	Robetta
Biodesigner	Swiss model
Easy modeller	3D jigsaw
Modeller	Phyre
Gamess	Jpred3

research papers regarding this problem, the aim of this study was to find out whether specialized web servers are better at protein structure prediction than stand alone software.

RESEARCH METHODOLOGY

The research methodology used for this research was a combination of empirical research, comparative research and quantitative research along with methodological literature review. Recent research papers related to the topic were reviewed to study what other researchers have discovered and analyzed. Five (5) stand alone software and 5 web servers were chosen for this research.

The names are given in Table 1. Protein sequence of two known protein structures that is, Hepatitis A and Insulin was obtained from NCBI BLAST and used as an input. The models generated by the 5 web servers and 5 stand alone software were verified and analyzed by structural analysis and verification server of UCLA. The models were compared to good quality validated models obtained through traditional methods. Three (3) statistical methods have been used to check the quality of the model; Ramachandran plot, overall quality factor, 3D-1D plot and time taken. The results obtained were analyzed quantitatively and based on the results obtained a conclusion was drawn as to which one is the best technological method for protein structure prediction.

What is protein structure prediction?

Protein structure prediction is defined as predicting the 3D structure of a protein molecule from its protein sequence or amino acid sequence (Baker and Sali, 2001). It is a set of techniques used to derive a three dimensional structure of any kind of protein molecule (Baker and Sali, 2001). Protein structure prediction is used by biologists and scientists to study protein molecules of any element or entity. It is used to study molecular structure of various diseases to discover cures. It is believed that if protein structure prediction is perfected, it can help scientists in finding a cure for even deadly diseases like HIV/AIDS and cancer (Baker and Sali, 2001). It is also used in the field of medicine to develop medicines and drugs for various kinds of viruses and diseases (Baker and Sali, 2001). Apart from its extensive use in medicine, protein structure prediction is also used in fields like biotechnology and DNA computing to study and design various enzymes that can help scientists in building a DNA computer.

A scientist named Ainfinsen in 1950 presented a theory known as the 'Thermodynamic Hypothesis' (Mihasan, 2010). He stated that all the information required determining the structure of a protein molecule is contained in its protein or amino acid sequence (Mihasan, 2010). He suggested that the three dimensional structure of any kind of protein molecule can be predicted from its amino acid sequence only (Mihasan, 2010). This was a revolutionary idea which encouraged many scientists to take up the challenge of developing methods and tools to help in protein structure prediction. Protein structure prediction is known as the Holy grail of molecular biology. It is one of the most challenging problems of 21st century and various researchers have tried various methods to solve this problem. The traditional lab methods and techniques such as high resolution electron microscopy, X-ray crystallography and nuclear magnetic resonance spectroscopy are extremely expensive and time consuming (Baker and Sali, 2001). Hence, since the past 50 years, scientists have been busy in developing computational methods and techniques that can accurately predict the protein structure of any kind of protein molecule.

Methods and approaches used in protein structure prediction

There are 4 main approaches used in Protein Structure prediction, which are: Comparative or homology modelling, fold recognition and threading methods, *De novo* or *Ab initio* methods and hybrid or integrative methods. Comparative modelling is when the three dimensional structure of a target protein sequence is determined by using the structures of proteins which belong to the same family as the target sequence (Schwede et al., 2008). The three dimensional structure of the similar proteins is used as templates. For successful homology modelling, it is important that the most suitable template is searched for (Schwede et al., 2008). Fold recognition and threading methods are used when the target protein sequence has no similarity with any other protein (Schwede et al., 2008). The most challenging protein structure prediction approach is the *De novo* or *Ab Intio* method (Schwede et al., 2008). In this method, the protein structure is predicted using only the primary amino acid sequence. Hybrid or integrative methods use a combination of experimental and computation techniques to predict the protein structure of any given sequence (Schwede et al., 2008).

PRACTICAL WORK

Our first step was to download all the stand alone software and install them on our computer. The next step was to obtain the protein sequence of Hepatitis A. While using the protein sequence of Hep A we faced some problems. Most web servers and software cannot do protein structure prediction of residues above 120 in case of software and 1200 in case of web servers. Hence, protein sequence of insulin was used as input instead of Hep A. The protein sequence of Insulin was obtained from the NCBI (2001) website. After installing and downloading all the stand alone software and obtaining the Insulin sequence, we started the process of protein structure prediction. We gave the insulin sequence as input in all the stand alone software and saved the models produced as PDB files on my computer. Our next step was to use the insulin protein sequence as an input for all the web servers chosen for this research.

All the web servers emailed the results back to us within 2 days. We saved all the models on our computer as pdb files. The next step was to analyze and verify all the protein models using statistical methods. For this purpose we chose three methods, which are: Ramachandran plot, overall quality factor and 3D-1D plot.

Ramachandran plot

Ramachandran plot or the Ramachandran diagram is a graph that represents the backbone torsion angles, phi and psi. The Ramachandran plot is used to validate and verify a protein model by plotting a graph that shows all the possible angles of an amino acid residue in a protein structure.

Table 2. Results of Ramachandran plot, overall quality factor and 3D-1D PLOT.

Standalone Software	Ramachandran plot (favoured regions percentage (%))	Overall quality factor (%)	3D – 1D plot (percentage of residues that had an averaged 3D-1D score > 0.2) (%)	Time
Abalone	100	0.000	0.90	5 min
Biodesigner	90	0.000	0.46	40 s
Modeller	100	0.000	0.59	95 s
Easy Modeller	84.2	75.906	69.77	15 min
Gamess	80.21	52.857	52.87	19 min
Webserver				
Robetta	84.9	72.34	72.87	73 h
Swiss Model	82.1	80.198	83.19	28 h
3D Jigsaw	65.7	54.286	51.72	2 h
Phyre	80.8	86.000	86.85	1 h
Jpred3	83.7	79.000	75.51	5 min

Overall quality factor

The overall quality factor shows the statistics of non bonded interactions between different types of atoms in a protein structure and then compares it with good quality refined structures.

3D – 1D plot

The 3D – 1D plot shows how compatible the three dimensional structure is with its protein sequence. A set of good quality models is used as a reference to obtain a score for each of the amino acid residues.

STATISTICAL ANALYSIS AND EVALUATION OF RESULTS

During this research, a fact was realized that the Ramachandran plot is not an appropriate measure for analyzing the quality of a protein model. The Ramachandran plot does not compare the phi and psi angles of a protein model with other refined models produced by traditional lab methods. The results in Table 2 show that even if the overall quality factor of model is close to 0, the Ramachandran plot can have a percentage of about 100% of the points that are in the favored region. Hence, for this research paper where we are comparing the models with good quality refined models, Ramachandran plot should not be taken into account (Table 2). According to the overall quality factor, the results show that Swiss Model and Phyre predicted the best and most accurate protein structure. In case of standalone software, Easy Modeller did a fairly good job but that is only because of the fact that Easy Modeller connects online to a protein data bank to produce a more accurate result. The 3D – 1D plot also shows that Swiss Model and Phyre predicted the best models. But if you consider the time factor, standalone software take less time in predicting a model than web servers.

Conclusion

The results clearly show that web servers are better at predicting protein structures than standalone software. Although web servers take a lot of time to produce results, their quality of models is far better than those produced by stand alone software. The only reason why web servers take time is because they have thousands of protein structure prediction jobs and queries lined up at a time. Secondly, the web servers consult protein data banks and other Meta servers before producing results to increase the accuracy and quality of results. So in my opinion, quality of results is much more important than time. Even though the web servers are producing much accurate results than stand alone software, they still need to address few issues. Web servers cannot do protein structure prediction of protein sequences which have more than 1200 residues.

The web servers simply refuse to take queries that have above 1200 residues. There are still a lot of protein sequences whose protein structures have not been predicted yet because they are extremely long. And predicting their structures through traditional lab methods is expensive and time consuming. Hence, the web servers need to refine their algorithms so that they are able to predict protein structures of sequences with more than 1200 residues. Another approach that can improve web servers even more is the use of neural networks. Hence to conclude, web servers are better at protein structure prediction than stand-alone software. They still need to improve a few things such as accepting protein sequence with above 1200 residues, and optimizing their algorithms to reduce the time taken to produce results.

REFERENCES

Baker D, Sali A (2001). Protein structure prediction and structural genomics. J. Sci., 29493(6): 10-12.

Bujnicki JM, Fischer D (2004). Meta Approaches to Protein Structure Prediction. J. Nucleic Acids Mol. Biol., 16:178-182.
David EK, Chivian D, Baker D (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res., 32: 14-16.
Ginalski K, Elofsson A, Fischer D, Rychlewski L (2009). A simple approach to improve protein structure predictions. J. Bioinforma., 19: 1015-1018.
Kelley LA, Sternberg MJE (2009). Protein structure prediction on the Web: a case study using the Phyre server. J. Nature Prot., 4: 3.
Kristian WK, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J (2010). Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. J. Biochem., 49: 2987-2998.
Lee MS, Bondugula R, Desai V, Zavaljevski N, Yeh IC, Wallqvist A, Reifman J (2009). PSPP: A Protein Structure Prediction Pipeline for Computing Clusters. PLOS one, p. 4.
Mihasan M (2010). Basic Protein structure prediction for the biologist. J. Biol. Sci. Belgrade, 62(4): 857-871.
NCBI (2011). Available:http://blast.ncbi.nlm.nih.gov/Blast.cgi
Schwede T, Sali A, Eswar N, Peitsch MC (2008). Computational Structural Biology- Methods and Applications. USA. University of California at San Francisco.
Volker AE, Rost B (2003). META-PP: Single interface to crucial prediction servers. J. Nucleic Acids Res., 31: 13.
Wikipedia (2011). List of Protein Structure Prediction [Online]. Available:http://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software
Zhang Y (2009). I-TASSER: Fully automated protein structure prediction in CASP8 J. Proteins. 77: 100-113.
Zhang Y, Skolnick J (2005). The protein structure prediction problem could be solved using the current PDB library. USA, P. Natl. Acad. Sci., 102(1): 1029-1034.