Editorial Article

Protein Structure Prediction: A Viable Tool for Drug Design

Dr. Pandian Sokkar
Assistant Professor in Computational Chemistry and Biophysics

Dr. Pandian Sokkar, Faculty of Allied Health Sciences, Chettinad Hospital and Research Institute, Chettinad Academy of Research and Education, Kelambakkam, Tamilnadu.

Corresponding author - Dr. Pandian Sokkar– (pandian.sokkar@gmail.com)

Chettinad Health City Medical Journal 2020; 9(4): 217 - 219

DOI: https://doi.org/10.36503/chcmj9(4)-03

Abstract

Understanding structure-function relationship of protein molecules is central to biological research. Experimental methods of protein structure determination are often expensive and therefore alternative approaches are in high demand. Computational methods are highly useful in modelling protein structures, however, there is a trade-off between accuracy and computational cost. Even with low computational cost, there exists plenty of methods to model protein structures. Such methods are highly useful in drug discovery research.

Introduction

Major goal of structural biology is to identify the structure function relationship of biomacromolecules. Such information is not only helpful for the deduction of molecular mechanisms behind the complex, it also helps immensely to design novel therapeutics with improved outcomes. This methodology is often termed as structure-based drug design. Recent advances in structure determination methods, such as high throughput procedures in Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography techniques have resulted in the deposition of huge number of protein structures in Protein Data Bank (PDB). However, this number is far less when compared the number of sequences being deposited in protein sequence database, owing to the inherent bottlenecks (time, cost and complexity) associated with protein structure determination techniques. This resulted in a big gap between number of protein sequences and available protein structures. It is impossible to fill this gap by accelerating the experimental methods of protein structure determination alone. Therefore, there have been high demands in alternative approaches in acquiring structural information, including both experimental methods (such as low-resolution cryo-electron microscopy) as well as computational predictions (such as homology modelling).

Development of advanced computer components such as multi-core processors, graphical processing units, memory chips and solid-state hard disks has revolutionized the modern computers. This massive growth in the production of faster computers led to the development of computational approaches in every field, including biological and medicinal research. In structural biology, computers are increasingly used to predict the native structure of protein, simulation of their dynamical behaviour and Quantitative Structure Activity Relationship (QSAR). All these methods have been successfully used in drug design. Here, we review the computational methods for protein structure prediction, which find application in drug research.

Protein structure prediction method

Homology modelling or comparative modelling uses a known protein structure (template) information to predict the structure of target protein, which shares significant sequence similarity with the template.[1] This method is based on the fact that proteins that have similar sequence are usually similar in structure and therefore they are functionally similar. This method is robust and it often results in high quality structures for drug design applications.

First step in homology modelling is to look for a suitable template for the target sequence, typically using a sequence alignment tool such as BLAST. Template is a protein whose structure is already known and shares at least 40% sequence similarity with the target. The limit of 40% similarity is imposed to ensure the sequence similarity has evolutionary origin, rather than random matching of amino acids. After the identification of template protein, an alignment between target sequence and template structure is created, which defines the

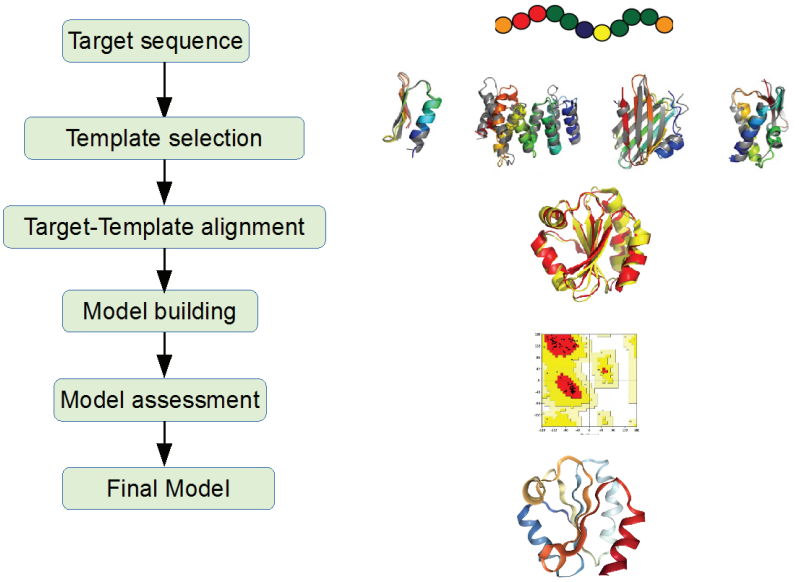

Figure 1: Steps involved in homology modelling

positions of target sequence on the template structure. Using the alignment information, a few models are generated and these models are often subjected to extensive optimization. The models are then assessed, typically based on atomic contacts, packing quality, hydrogen bonding and burial of hydrophobic amino acids. The reliable models can be used to study their interaction with drug candidates.

In the cases where templates could not be identified or target-template sequence similarity falls below 30%, other modelling strategies (such as threading or ab initio protein modelling) could be used. However, these methods often produce low-resolution models. Although the low-resolution models are mostly used for functional annotations, their use in drug research is rather limited.

Recent trends

Much of the current research in protein structure prediction is focused on the development of new techniques to produce high quality models even if the target-template sequence similarity is significantly lower. It has been proposed by many that inclusion of multiple templates or multiple sequence alignment enhances the accuracy of the models.[2,3] Further improvements in the alignment can be attained by the inclusion of secondary structure prediction data for the target sequence. This is very useful if the sequence similarity falls in twilight zone.

In addition, much of the research is devoted to the optimization protocols to refine protein structure models resulting from poor alignment (i.e., low sequence similarity).[4] The optimization methods include simple energy minimization and extensive conformational sampling such as simulated annealing, Monte-Carlo search, and molecular dynamics (MD) simulation. Of these, MD simulation is of particular interest in molecular modelling community. MD simulation couples Newton's equations of motion and classical force field potential energy functions to simulate the dynamical behaviour of molecules at the atomistic level. Albeit the inaccuracy in the force field energy functions, this method has many success stories in biomolecular modelling. Many variants of MD simulation are employed in protein structure optimization such as replica exchange MD simulation (REMD), accelerated MD (aMD), multi-scale MD and steered MD. In many cases, a protein is modelled as individual fragments and the fragments are assembled together, much like assembling a motorbike using individual components.

The performance of different modelling schemes are often evaluated in an annual meeting called "Critical Assessment of Structure Prediction (CASP)". Once in two years, researchers are invited to submit their models for a set of proteins for which the experimental structures are not publicly available. Assessments and results are published in a special issue of the journal PROTEINS. CASP12 was the recent assessment carried out.[5]

Software resources

Plenty of online as well as offline resources available now a days to model protein structures. Modeller is a standalone software, mostly written in Python language, which has many built-in features.[6] One will only need a template PDB file and target sequence file in FASTA format to model protein structures. However, the interface is command-line only and the input methods for Modeller program is not intuitive. Even a simple typographic error or missing ":" character might be very challenging to diagnose the

```
1. vim
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ()

env.io.atom_files_directory = ['.', '../atom_files']

# Create a new class based on 'loopmodel' so that we can redefine
# select_loop_atoms
class MyLoop(loopmodel):
    # This routine picks the residues to be refined by loop modeling
    def select_loop_atoms(self):
        # Two residue ranges (both will be refined simultaneously)
        return selection(self.residue_range('1:A', '22:A'),
                         self.residue_range('227:A', '232:A'))

a = MyLoop(env,
           alnfile  = 'cxcr4-mult.ali',      # alignment filename
           knowns=('4rwsA','3oduA','2k04A'),             # codes of the templates
           sequence = 'cxcr4',               # code of the target
           loop_assess_methods=assess.DOPE) # assess each loop with DOPE
a.starting_model= 1                 # index of the first model
a.ending_model  = 5                 # index of the last model

a.loop.starting_model = 1           # First loop model
a.loop.ending_model   = 5          # Last loop model
a.loop.md_level = refine.slow
a.make()                            # do modeling and loop refinement
```

Figure 2: Sample input file for Modeller program

error. A sample input file is given in Figure 2. Because of this, Modeller is not very popular among the non-specialist researchers. However, the documentation for Modeller is very clear and highly detailed, which makes it easier to utilize all the features of Modeller program. To simplify the job, the Modeller program is also available as an online tool.[7] The online interface is very simple and it requires a sequence file and structure from the user, although not all features are available in the online version and jobs can take longer time to complete depending on the queue status.

These days many other online resources available for the end users. SWISSMODEL is a very intuitive web interface to perform homology modelling in different levels of difficulty.[8] I-TASSER is another web based tool to model protein structures. I-TASSER uses replica exchange Monte Carlo conformational sampling to optimize protein structures. The accuracy of I-TASSER in modelling protein structures is evident from the CASP results, which show the tool is top ranked in several assessments.[9]

In recent years, there has been tremendous effort from researchers to improve the accuracy of the protein structure prediction. Such initiatives result in the reliable prediction of protein structures with minimal information from templates. Advances in computer hardware reduce the computational cost associated with the methodology and therefore the methods are available at everyone's disposal. Accurate protein modelling, together with protein-ligand docking methods, can be very useful in drug discovery research (virtual screening of bioactive compounds), as it reduces cost and time.

References

1. Munsamy G, Soliman ME. Homology modeling in drug discovery-an update on the last decade. Letters in Drug Design & Discovery. 2017, 14:1099-1111.

2) Li J, Cheng J. A Stochastic Point Cloud Sampling Method for Multi-Template Protein Comparative Modeling. Scientific reports. 2016, 6:25687.

3) Sokkar P, Mohandass S, Ramachandran M. Multiple templates-based homology modeling enhances structure quality of AT1 receptor: validation by molecular dynamics and antagonist docking. Journal of molecular modeling. 2011, 17:1565-77.

4) Kumar V, Meyer D, Lall N. Comparative Investigation of Techniques Used for Mtr Homology Model Refinement. J Proteomics Bioinform. 2017, 10:284-97.

5) Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. Proteins: Structure, Function, and Bioinformatics. 2018, 86:7-15.

6) Webb B, Sali A. Comparative protein structure modeling using MODELLER. Current protocols in bioinformatics. 2014, 47:5-6.

7) Kuntal BK, Aparoy P, Reddanna P. EasyModeller: A graphical interface to MODELLER. BMC research notes. 2010, 3:226.

8) Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research. 2014, 42:252-8.

9) Zhang Y. I-TASSER server for protein 3D structure prediction. BMC bioinformatics. 2008, 9:40.