Chapter 34
Protein Structure Prediction Using Homology Modeling: Methods and Tools

Akanksha Gupta
Netaji Subhas Institute of Technology, India

Pallavi Mohanty
Netaji Subhas Institute of Technology, India

Sonika Bhatnagar
Netaji Subhas Institute of Technology, India

ABSTRACT

Sequence-structure deficit marks one of the critical problems in today's scenario where high-throughput sequencing has resulted in large datasets of protein sequences but their corresponding 3D structures still needs to be determined. Homology modeling, also termed as Comparative modeling refers to modeling of 3D structure of a protein by exploiting structural information from other known protein structures with good sequence similarity. Homology models contain sufficient information about the spatial arrangement of important residues in the protein and are often used in drug design for screening of large libraries by molecular docking techniques. This chapter provides a brief description about protein tertiary structure prediction and Homology modeling. The authors provide a description of the steps involved in homology modeling protocols and provide information on the various resources available for the same.

INTRODUCTION

Knowledge of the protein structure is an important step towards rational structure based drug design and virtual screening of large molecular libraries. Sequence-structure deficit marks one of the critical problems in today's scenario where high-throughput sequencing has resulted in large datasets of protein sequences, but their corresponding 3D structures still need to be determined. Currently there are about 2 million protein sequences in Swissprot and TrEMBL, among them about 30,000 proteins have had

DOI: 10.4018/978-1-5225-1762-7.ch034

their structures solved experimentally. Although the rate of experimental structure determination will continue to increase, the number of newly discovered sequences grows much faster than the number of structures solved. Since experimental structures can only be determined for a small fraction of proteins, computational methods for protein structure modeling play an increasingly important role yielding models suitable for a wide spectrum of applications.

The term "homology modeling", also called comparative modeling refers to modeling of 3D structure of a protein by exploiting structural information from the known configurations of similar proteins. The method is based on the fact that structural conformation of a protein is more highly conserved than its amino acid sequence, and that small or medium changes in sequence normally result in little variation in the 3D structure. The necessary condition for successful homology modeling is a sufficient similarity (>30%) between the protein sequences. The importance and applicability of homology modeling is steadily increasing with the increase in the number of known structures in the protein structure database. Homology models contain sufficient information about the spatial arrangement of important residues in the protein, enabling us to study the binding site and design/dock drugs suitable for binding to the molecule. Stable and reliable repositories have been developed to give access to these annotated and evaluated models.

With the advent of new modeling softwares and algorithms, homology modeling is rapidly becoming the first choice for obtaining the 3D structure coordinates as well as other valuable structural and functional insights. The recent advancements in homology modeling, particularly in detecting and aligning sequences with Template structures, loop and side-chain modeling, model validation and evaluation have largely contributed to the prediction of more accurate models of protein structure, which was not possible several years ago. Improvement in sequence search/analysis, scoring systems and tertiary structure prediction methods allows development of models robust, error free models with high statistical significance. Publicly available, easy to use modeling servers and structure validation tools have further assisted the development of this technique.

Homology modeling is a multistep process, that can be summarized in following steps 1) Template identification and alignment, 2) alignment correction, 3) model generation, that includes backbone, loop and side chain generation 4) model optimization, and ; 5) validation. The first and foremost step is to fetch the Target protein sequence from various freely available databases. This is followed by Template identification, where, a protein Template whose 3D coordinates is known and that shares the maximum similarity with the Target sequence is selected. Similarity check between the Target and Template protein sequence is done by sequence alignment. The most widely used algorithm to run alignment is BLAST (Basic Local Alignment Search Tool). For a good alignment between both Target and Template sequence, the Template should not possess an E-value larger than 1. Sometimes it may be difficult to align two sequences in a region where the percentage sequence identity is very low. To resolve this issue, multiple Templates homologous to the Target protein are used, and an alignment correction is performed manually. After obtaining a good alignment, model generation starts. Quality of the model generated largely depends upon the correctness of the Template identified and accuracy of the alignment generated. If the alignment is accurate, there are less chances of the model being inaccurate. In Model generation, if the sequence alignment is good, the coordinates of those Template residues that show up in the alignment with the model sequence can be copied. If two aligned residues differ, only the backbone coordinates (N, Cα, C and O) are copied. The Target-Template alignment may contain gaps due to insertions and deletions of amino acids or bases. These loops are modeled using commercially available or manually generated loop libraries. There are generally two methods of loop modeling: 1) by