International Journal of Biological Macromolecules 85 (2016) 48–62

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac

Review

Current updates on computer aided protein modeling and designing

Faez Iqbal Khan [a], Dong-Qing Wei [a], Ke-Ren Gu [a], Md. Imtaiyaz Hassan [b], Shams Tabrez [c,*]

[a] School of Chemistry and Chemical Engineering, Henan University of Technology, Zhengzhou 450001, China
[b] Centre for Interdisciplinary Research In Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 10025, India
[c] King Fahd Medical Research Center, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 21 August 2015
Received in revised form
17 December 2015
Accepted 21 December 2015
Available online 28 December 2015

Keywords:
Protein modeling
Protein design
Molecular dynamics simulations
In silico mutation analysis

ABSTRACT

Determination of the three dimensional (3D) structure of a protein can provide important details about its biological functions and mechanism of action. However, despite their significance, the precise three-dimensional structures of most of the proteins are not fully determined till date. The main focus of the current review article is to gain a better understanding of the structural features of the proteins using computational techniques, and their relationship with function. Protein modeling and design is the method aimed to fold a primary amino acids sequence into protein structure with the ultimate goal of designing novel function and behavior. Moreover, proteins can also be designed from scratch or by similarity with the known protein structure. In the current article we have tried to cover various computer aided protein modeling and designing *via* homology and *ab initio* modeling, folding study using Molecular Dynamics (MD) methods and *in silico* mutation analysis.

Contents

Abbreviations: RMSD, root mean square deviation; MD, molecular dynamics; GROMACS, groningen machine for chemical simulations.

* Corresponding author at: King Fahd Medical Research Centre, King Abdulaziz University, P.O. Box 80216, Jeddah 21589, Saudi Arabia. Fax: +966 26952076.
E-mail address: shamstabrez1@gmail.com (S. Tabrez).

http://dx.doi.org/10.1016/j.ijbiomac.2015.12.072
0141-8130/

1. Introduction

Proteins are linear chains of amino acids that adopt a distinct three-dimensional (3D) structure in their native surroundings medium. The native 3D structure allows the protein to carry out its biochemical function [26]. It is important to know the mechanism that how a sequence of amino acids can fold into its functional 3D structure. Protein folding is entirely a physical process that depends on the precise amino acid sequence of the protein and the surrounding solvent [25]. Moreover, the prediction method for a protein structure remains complicated since even short amino acid sequences can form an abundant number of geometrical structures. The protein structure with the least free energy values has to be identified and optimised in order to get a functionally active protein molecule [72].

A protein is composed of several stages of structure *viz*, primary, secondary and tertiary. The primary structure of a protein is described by the specific amino acid sequence. The local bonding between amino acids in the protein form secondary structure. The two most common forms of pattern found in the secondary structure are α-helices and β-sheets connected by loops [93]. The tertiary structure *i.e.*, 3D structure is the final actively folded structure of protein. The availability of experimental protein structures has inspired the development of methods for computational structure prediction that are knowledge-based rather than physics based [12].

The goals of protein design are to engineer enzymes to perform functions under a wider range of conditions, entirely new functions, and find out the behavior protein folding in response to mutations [156]. Currently, attention of the scientific community is focused on redesigning portions of existing protein to increase stability and improve function [123]. The mechanisms of the mutational effects are useful to study the nature and stability of enzymes. For this purpose, we have highlighted several stability and flexibility predictors in this current article to assess the relationship between point mutation and enzyme stability [13]. Structure is major determinate of function of a protein [44,46].

In silico protein modeling is helpful in predicting the 3D protein models followed by the prediction of its active sites [45]. Moreover, the presence of critically active residues in the structural framework can be found by computational analysis [64,4]. The molecular dynamics (MD) simulations can be further implemented to assess the conformational preferences of protein [129,63,65]. The MD trajectories at different temperatures can clearly reveal thermostable nature of particular protein [42]. Additionally, a constant pH molecular dynamics simulation at a wide pH range is helpful to establish the optimum activity and stability profiles [41,55]. For this purpose, the MD simulations at fixed protonation states in an explicit water environment can be carried out to evaluate the effect of the physiological pH on particular protein. Further, the different stability and flexibility predictions are helpful to assess the relation of point mutations and enzyme stabilities. The protein modeling and design methods pave the way to engineer new and better thermostable enzymes with desired properties. The overall work flow of protein modeling research methodology have been depicted in Fig. 1.

2. Structure prediction analysis

Structure of a protein can be predicted by using various algorithms of "Homology modeling" and "*ab initio*" methods. The aim of homology or comparative protein modeling is to construct an appropriate 3D replica for the protein with an unknown experimental structure (*i.e.*, the target) by means of sequence similarities to that of a known structure (*i.e.*, the templates) present in databases [56]. For reliable model building, two criteria must be achieved. First, there must be a detectable likeness between the structure of the template and the target of the protein sequence. Secondly, a significant correct alignment between the structure of the template and the target sequence must be computed.

The comparative or homology modeling of the protein 3D structure continues to be the most precise method in contrast to other available methods [70]. The overall accuracy range achieved by comparative models, are similar to nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography techniques [114]. The structural similarity can typically be presumed, if the likenesses among the proteins are noticeable at their sequence level. Furthermore, protein sequences with non-measurable similarities show the presence of analogous structures. There is an assumption that a fraction of all sequences submitted in the databases are noticeably related to not less than one identified protein structure [38]. Thus, homology modeling could potentially be applied to over 150,000 protein sequences among approximately 500,000 submitted sequences [7] which matches up to about 10000 experimentally determined protein 3D structures [10].

However, the importance of homology model prediction is gradually growing as the quantity of distinctive structural folds that proteins assumed are limited due to less number of experimentally determined structures [50,33]. If the resulting model of the query is not satisfactory because of the low sequence identity (<25%) with the template structure, then *ab initio* structure prediction protocols such as I-TASSER [108] and ROBETTA servers [66] are used. The I-TASSER server is based on *ab initio* algorithms, originally produces 3D models from a manifold threading based alignment and repeated structural compiling simulations. It also predicts the function of the query proteins by means of the structural comparisons of 3D predicted models with other proteins of known structures which generates the outputs that contain secondary structures as well as full-length tertiary predictions, including Enzyme Commission (EC) numbers and ligand-binding sites [108]. Similarly, the ROBETTA server [66] uses *de novo* or *ab initio* methods to calculate the structural features of proteins which do not have any struc-