Computational Biology and Chemistry 53 (2014) 251–276

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Review Article

# Three-dimensional protein structure prediction: Methods and computational strategies

Márcio Dorn[a,*], Mariel Barbachan e Silva[b], Luciana S. Buriol[a], Luis C. Lamb[a]

[a] Federal University of Rio Grande do Sul, Institute of Informatics, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil
[b] Federal University of Rio Grande do Sul, Center of Biotechnology, Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, RS, Brazil

ARTICLE INFO

*Article history:*
Received 12 June 2014
Received in revised form 3 October 2014
Accepted 7 October 2014
Available online 12 October 2014

*Keywords:*
Three-dimensional protein structure prediction
Structural bioinformatics
Ab initio methods
Knowledge-based methods

ABSTRACT

A long standing problem in structural bioinformatics is to determine the three-dimensional (3-D) structure of a protein when only a sequence of amino acid residues is given. Many computational methodologies and algorithms have been proposed as a solution to the 3-D Protein Structure Prediction (3-D-PSP) problem. These methods can be divided in four main classes: (a) first principle methods without database information; (b) first principle methods with database information; (c) fold recognition and threading methods; and (d) comparative modeling methods and sequence alignment strategies. Deterministic computational techniques, optimization techniques, data mining and machine learning approaches are typically used in the construction of computational solutions for the PSP problem. Our main goal with this work is to review the methods and computational strategies that are currently used in 3-D protein prediction.

**Contents**

## 1. Introduction

Structural Bioinformatics is one of the key research areas in the field of Computational Biology (Zhang et al., 2005; Altman and Dugan, 2005; Clote and Backofen, 2000; Pevzner, 2000; Liljas et al., 2001; Gopakumar, 2012). Structural Bioinformatics concerns the analysis and prediction of three-dimensional (3-D) structures of biological macromolecules such as Proteins[1], RNA and DNA (Zhang et al., 2005; Altman and Dugan, 2005). This structural information corresponds to 3-D macromolecular structures obtained through different experimental methods such as protein

* Corresponding author.
*E-mail addresses:* mdorn@inf.ufrgs.br (M. Dorn), mariel.barbachan@ufrgs.br (M.B. e Silva), buriol@inf.ufrgs.br (L.S. Buriol), lamb@inf.ufrgs.br (L.C. Lamb).

[1] In this review proteins and polypeptides are treated as synonymous.

http://dx.doi.org/10.1016/j.compbiolchem.2014.10.001
1476-9271/© 2014 Elsevier Ltd. All rights reserved.

crystallography (X-ray diffraction), electron microscopy or nuclear magnetic resonance (NMR). This information allows one to study folds and local motifs in proteins, molecular folding, evolution and structure/function relationships.

One of the main research problems in structural bioinformatics is the prediction of three-dimensional protein structures. Proteins are long sequences formed out of 20 different amino acid residues that in physiological conditions adopt a unique 3-D structure[2] (Anfinsen et al., 1961). Knowledge of the protein structure allows the investigation of biological processes more directly, with higher resolution and finer detail. The sequence–protein–structure paradigm (also known as the "lock-and-key" hypothesis) says that the protein can achieve its biological function only by folding into a unique, structured state determined by its amino acid sequence (Anfinsen, 1973). Nevertheless, currently it has been recognized that not all protein functions are associated to a folded state (Dunker et al., 2008, 2001; Uversky, 2001; Tompa and Csermely, 2004; Tompa, 2002; Wright and Dyson, 1999). In some cases proteins must be unfolded or disordered to perform their functions (Gunasekaran et al., 2003). These proteins are called intrinsically disordered proteins (IDP) and represent around 30% of the protein sequences. Despite the presence of IDP proteins an important aspect of understanding and interpreting the function of a given protein involves characterizing molecular interactions. These interactions can be intramolecular (ionic bonds, covalent bonds, metallic bonds, etc) or intermolecular (hydrogen bonds and other non-covalent bonds such as van der Waals forces). The knowledge of the 3-D structure of polypeptides gives researchers very important information to infer the function of the protein in the cell (Branden and Tooze, 1998; Laskowiski et al., 2005a,b; Lesk, 2002): structural functions; catalysis in chemical reactions; transport and storage; regulatory functions; gene transcription control; recognition functions. Further details about protein function prediction can be found in Whisstock and Lesk (Whisstock and Lesk, 2003), Rentzsch and Orengo (Rentzsch and Orengo, 2009) and Lee et al. (Lee et al., 2007).

The determination of protein structure is both experimentally expensive (due to the costs associated to crystallography, electron microscopy or NMR), and time consuming (Guntert, 2004). The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the Genome Projects[3] and the number of 3-D structures of proteins which are currently known. Only a tiny portion of protein sequences have experimentally solved three-dimensional structures. These figures not only clearly illustrate the need for, but also motivate further research in computational protein structure prediction methods. Over the last 10 years several computational methodologies, systems and algorithms have been proposed as a solution to the three-dimensional protein structure prediction (3-D PSP) problem (Bujnicki, 2006; Moult, 2005; Osguthorpe, 2000; Tramontano, 2006). These methods are divided into four classes, that shall be described in detail in this review (Floudas et al., 2006): (1) First principle methods without database information (Osguthorpe, 2000); (2) First principle methods with database information (Rohl et al., 2004; Srinivasan and Rose, 1995); (3) Fold recognition and threading methods (Bowie et al., 1991; Jones et al., 1992; Bryant and Altschul, 1995; Turcotte et al., 1998); and (4) Comparative modeling methods and sequence alignment strategies (Martí-Renom et al., 2000; Sánchez and Sali, 1997). The first group of methods aims at predicting new folds only through (computational) simulation of physicochemical properties of the folding process of the proteins in nature. The other groups represent the methods that are able of performing fast and effective prediction of protein 3-D structures when known template structures and fold libraries are available (Kolinski, 2004).

Predicting the correct 3-D structure of a protein molecule is an intricate and arduous task. The 3-D PSP and Protein Folding (PF) problems[4] are classified in computational complexity theory as NP-complete problems (Crescenzi et al., 1998; Fraenkel, 1993; Hart and Istrail, 1997; Levinthal, 1968; Ngo et al., 1997), i.e., they are among the hardest problems in terms of computational requirements. For a formal definition of NP-completeness see Garey and Johnson (Garey and Johnson, 1979). This complexity is due to the folding process of a protein being highly selective. A long amino acid chain ends up in one out of a huge number of 3-D conformations. In contrast, the conformational preferences of single amino acid residues are weak. Thus, the high selectivity of protein folding is only possible through the interaction of many residues. Therefore, non-local interactions play an important role in protein three-dimensional structure, as local sequence–structure relationships are not absolute (Rackovsky, 2010). *Ab initio* methods (first principle methods without database information) can obtain novel and unknown protein folds. Nevertheless, the complexity and the high dimensionality of the search space (Ngo et al., 1997) even for a small protein molecule makes the problem intractable (Levinthal, 1968). The direct simulation of protein folding in atomic details, as used in Molecular Dynamics (MD)[5], is not tractable (van Gunsteren and Berendsen, 1990) (for large proteins of medical and scientific interest) due to high computational costs, despite the efforts towards the development of distributed computing platforms. On the other hand, homology modelling does not lead to such problems; however, it can only predict structures of protein sequences which are similar or nearly identical to other sequences of known structures. Fold recognition via threading, in turn, is limited to the fold library derived from the Protein Data Bank (PDB) structures (Berman et al., 2000).

In order to tackle the computational complexity of the 3-D PSP problem, current 3-D protein structure prediction methods make use of a wide range of optimization algorithms (Klepeis et al., 2003). Metaheuristics are used to provide near optimal solutions. In addition, considering the limitations of the four classes of protein structure prediction methods, researchers have recently developed hybrid methods which combine principles of the four classes, as can be observed in last CASP editions (Moult et al., 2014, 2011). For example, the accuracy presented by homology modeling methods is combined with the capacity of *Ab initio* methods in predicting novel folds (Dhingra and Jayaram, 2013; Dorn et al., 2008; Fan and Mark, 2004). In order to reduce the complexity and the high dimensionality of the conformational search space inherent to *Ab initio* methods, information about structural motifs found in known protein structures can be used to construct approximate conformations. These approximate conformations are expected to be sufficient to allow later refinement by means of Molecular Mechanics (MM) such as MD simulation (van Gunsteren and Berendsen, 1990). In a refinement step, global interactions between all atoms in the molecule (including e.g. non-bond interactions) are evaluated and deviations in the polypeptide main-chain and side-chain torsion angles can be corrected (Fan and Mark, 2004). These in turn

[2] Anfinsen hypothesis states that "all the information that dictates the native fold of protein domain is encoded in their amino acid sequence".

[3] DOE Genomic Science. http://genomics.energy.gov (accessed 01.09.14).

[4] Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil.

[5] MD is a simulation method in which the protein system is placed into a random conformation and then the system reacts to force atoms to exert on each other. The model assumes that, as a result of these forces, atoms move in a Newtonian manner. The trajectory of the system should lead to the native conformation.