Applied Soft Computing 35 (2015) 398–410

Contents lists available at ScienceDirect

Applied Soft Computing

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

Review Article

Soft computing methods for the prediction of protein tertiary structures: A survey

Alfonso E. Márquez-Chamorro [a,*], Gualberto Asencio-Cortés [a], Cosme E. Santiesteban-Toca [b], Jesús S. Aguilar-Ruiz [a]

[a] School of Engineering, Pablo de Olavide University, Seville, Spain
[b] Centro de Bioplantas, University of Ciego de Ávila, Cuba

ARTICLE INFO

Article history:
Received 4 March 2015
Received in revised form 17 June 2015
Accepted 20 June 2015
Available online 7 July 2015

Keywords:
Protein structure prediction
Soft computing
Protein contact map
Support vector machines
Neural networks
Evolutionary algorithms

ABSTRACT

The problem of protein structure prediction (PSP) represents one of the most important challenges in computational biology. Determining the three dimensional structure of proteins is necessary to understand their functions at molecular level. The most representative soft computing approaches for solving the protein tertiary structure prediction problem are summarized in this paper. These approaches have been categorized following the type of methodology. A total of 90 relevant works published in last 15 years in the field of protein structure prediction have been reported, including the best competitors in last CASP editions. However, despite large research effort in last decades, a considerable scope for further improvement still remains in this area.

Contents

* Corresponding author. Tel.: +34 954062449.
E-mail address: amarcha@upo.es (A.E. Márquez-Chamorro).

http://dx.doi.org/10.1016/j.asoc.2015.06.024
1568-4946/

1. Introduction: background and purpose

Proteins are an important class of macromolecules present in all biological organisms. They form the basis of cellular and molecular life and significantly affect the structural and functional characteristics of cells and genes [13]. Numerous functions, as structural support, mobility, protection, regulation or transport, are developed by proteins in the cells. Proteins are formed by union of simpler substances called amino acids. Amino acid sequence determines the structure of proteins and is the link between the genetic message in DNA and the three-dimensional structure which is associated to a biological function. Therefore, the knowledge of the sequence is essential to discover the protein functionality [8].

A protein can be seen on four different levels depending on which structures of the protein are considered. Essentially, primary structure of proteins consist of linear sequences of twenty natural amino acids joined together by peptide bonds. The secondary structure of a protein refers to the interactions due to a regular arrangement of hydrogen bonds between CO and NH groups (carboxyl and amino) of its amino acids, forming different motifs (α-helix, β-sheet, loops and turns). The tertiary structure is a description of the complex and irregular folding of the polypeptide chain in three dimensions. These complex structures are held together by a combination of several molecular interactions (e.g. ionic, hydrophobic or hydrogen bonds) that involve the amino acids of the chain. The quaternary structure is the final dimensional structure formed by all the polypeptide chains making up a protein [13].

A protein spontaneously folds into a 3-dimensional structure after having been manufactured in the ribosomes. A specific protein will fold in the same way and will end up with the same 3D structure. This phenomenon is called the native state of the protein [8]. Protein folding represents the process whereby higher structures are formed from the primary structure. A folded protein can have more than one stable folded state or conformation. Each conformation has its own biological activity. Anfinsen's experiment discovered that the amino acid sequence determines the native structure of a protein [1].

Sometimes, a protein can fold into a wrong shape. A single missing or incorrect amino acid could cause such a misfold. As already stated, protein function is determined by its structure, which can be inferred from the sequence of amino acids, therefore a misfold implies that a protein can not fulfill its function correctly. Alzheimer's disease, Cystic fibrosis and other neurodegenerative diseases are now attributed to protein misfolding. The knowledge of the misfolding factors and understanding the protein folding process, would help in developing cures for these diseases. Therefore, the knowledge of the structure of the protein provides a great advantage for the development of new drugs and the design of new proteins.

Despite of having a huge number of protein sequences as result of the genome sequence projects [86,101], the number of known protein structures is significantly lower than the proportion of known sequences. The difficult determination of these structures, using experimental methods such as X-ray or nuclear magnetic resonance (NMR), contributes to increase this gap between sequence and protein structures. Therefore, it is necessary the use of computational methods which predict 3D protein structures in a cheaper and faster way.

Different soft computing approaches have been developed to deal with the PSP problem. The main soft computing paradigms for the application of protein structure prediction are artificial neural networks (ANNs), evolutionary computation (EC) and support vector machines (SVMs). Furthermore, protein structure prediction methods can be further classified according to a biological approximation: homology-based methods, threading methods and *ab initio* methods.

Homology methods are based on the comparative of protein sequences with known structures. These methods are based on the hypothesis that similar protein sequences determine similar 3D structures.

Threading methods, also called sequence-structure alignment or fold recognition methods, try to align a protein sequence to a 3D structure. Threading methods are based on the idea that evolution conserves the structure rather than the sequence. On the other hand, *Ab initio* methods try to find a 3D model of the protein exclusively using the amino acid sequence, according to the laws of physics and chemistry.

In this paper, we present a survey on relevant methods of protein tertiary structure prediction based on soft computing techniques (neural networks, support vector machines and evolutionary computation). The remaining of this paper is structured as follows. Section 2 summarizes some basic concepts of PSP. The following sections describe all the different employed techniques (neural network, support vector machines, evolutionary algorithm, statistical methods and other predictive methods). Finally, last section summarizes the main conclusions of the study.

2. Preliminary concepts

In this section, several basic concepts related to PSP are briefly explained. First, protein structure prediction workflow is described. In the second place, the most relevant data structures used to represent the tertiary structure of a protein, such as the 3D models, *e.g.* torsion angle and lattice models, distance maps (DM) and contact maps (CM), are identified. Additionally, we specify the performance metrics for the validation of the prediction algorithms in this area. This is a particularly important issue, given the variety of methods and different measures employed, for the comparative analysis among algorithms.

2.1. Protein structure prediction workflow

The general steps for the PSP methodology are summarized in this section. Given a new protein sequence with an unknown structure, homology modeling can be considered as first step. This template-based methodology is based on the assumption that similar sequences encode similar 3D structures. If the target sequence does not have homologous proteins with known structure, fold recognition methods can be required. This type of protein modeling methods are based on the threading alignment. Structure templates are aligned with the target sequence by optimizing a scoring function based on statistical knowledge. If these two approaches turn out to be insufficient, we finally need *ab-initio* methods, which exclusively use the sequence information to predict the structure. First, a data input selection is required. Data input can consist of different features obtained from amino acid sequence. For instance, the frequency of appearance of amino acids in the sequence, physico-chemical properties of the residues, evolutionary information extracted from the sequence, such as position specific scoring matrices or correlated mutations, or 1D features prediction as secondary structures (SS) or solvent accessibility (SA), are usually employed as input data. The following step considers the selection of a data output model. Different representation models for the predicted structure, such as contact map or torsion angle model, are classified in the following section. The selection of an algorithmic technique for the prediction is also required. This survey is focused on the classification of the different algorithmic methodologies (statistical and soft computing approaches) for PSP. The quality assessment of the generated model is also crucial to understand the validity of the methods. Performance metrics section specify which are the most common assessment measures