Evolutionary Algorithms for the Protein Folding Problem: A Review and Current Trends

Heitor Silvério Lopes
Bioinformatics Laboratory
Federal University of Technology – Paraná
Av. 7 de setembro, 3165, 80230-901 Curitiba – Brazil
hslopes@utfpr.edu.br

12.1 Introduction

Proteins are complex macromolecules that perform vital functions in all living beings. They are composed of a chain of amino acids. The biological function of a protein is determined by the way it is folded into a specific tri-dimensional structure, known as native conformation. Understanding how proteins fold is of great importance to Biology, Biochemistry and Medicine. Considering the full analytic atomic model of a protein, it is still not possible to determine the exact tri-dimensional structure of real-world proteins, even with the most powerful computational resources. To reduce the computational complexity of the analytic model, many simplified models have been proposed. Even the simplest one, the bi-dimensional Hydrophobic-Polar (2D-HP) model (see Sect. 12.2.2), was proved to be intractable due to its NP-completeness. The current approach for studying the structure of proteins is the use of heuristic methods that, however, do not guarantee the optimal solution. Evolutionary computation techniques have been proved to be efficient for many engineering and computer science problems. This is also the case of unveiling the structure of proteins using simple lattice models.

In this work the nature of the models used for the protein folding problem is reviewed, with special emphasis on discrete models. Also, we analyze how evolutionary computation techniques have been applied to solve it. Amongst these techniques, there are many different variants of genetic algorithms, besides ant colony optimization, differential evolution and artificial immune systems.

This chapter is structured as follows: the remaining of this section introduces some basic aspects of amino acids and proteins, and presents the protein folding problem. Sect. 12.2 presents the several models for protein folding with special emphasis on a specific discrete model: the hydrophobic-polar. Sect. 12.3 is dedicated to the several computational approaches for the protein folding problem, from molecular dynamics and approximation algorithms to several evolutionary computation algorithms. Next, Sect. 12.4 presents challenging issues that limit current research. Finally, in Sect. 12.5 current trends for future research and the conclusion are presented.
12.1.1 Amino Acids and Proteins

The basic structure of an amino acid consists of a carbon atom ($C_\alpha$) connected with an amino group ($\text{NH}_2$), a carboxyl group (COOH) and a side-chain. The only difference between amino acids is due to the composition of their side-chain. There are 20 standard amino acids. According to the physical properties of the side-chain, amino acids can be classified according to its polarity and acidity/basicity. Such classification leads to a hydrophilic (polar) or hydrophobic (nonpolar) character of the amino acid. The distribution of hydrophilic and hydrophobic amino acids along the protein ultimately determines structure of the protein.

The sequence of amino group, $C_\alpha$ and carboxyl group of an amino acid bounded with the following is known as backbone of a protein. There are three main levels of organization of the structure of a protein: primary, secondary and tertiary structures. The primary structure of a protein or polypeptide chain is its linear sequence of amino acids, represented by a string of letters. Some specific regions of the primary structure can fold into known tri-dimensional structures, such as $\alpha$-helices or $\beta$-sheets. These structures are known as secondary structures. The spatial representation of the protein is called tertiary structure. The shape into which a protein naturally folds is known as its native state, or native conformation. For some particular proteins, tertiary structures can be combined to form a super-structure known as quaternary structure.

The tertiary structure of a protein, or the quaternary structure of its complexes, is of particular interest, since it defines the biological function of the protein. The most effective method for unveiling the structure of real proteins is using nuclear magnetic resonance spectroscopy or X-ray crystallography. It is estimated that the human body has around 100,000 different proteins, but a only a small portion of them have its structure known. The Protein Data Bank (PDB) [http://www.pdb.org] is the repository for structural data of proteins. Currently, it holds structural information of almost 50,000 proteins. However, the amount of known proteins which structure is unknown is much larger, thus justifying the use of computational methods for this purpose. Therefore, this is an important research area in Bioinformatics and Computational Biology.

12.1.2 Protein Folding

The Protein Structure Prediction (PSP) problem can be defined as determining the final tri-dimensional structure of a protein by using only the information about its primary structure. On the other hand, the Protein Folding Problem (PFP) is understood as being the discovery of the pathways by which a protein is folded into its natural conformation, during its synthesis [34]. However, in the current literature those two terms are frequently used with no distinction, usually meaning only the first issue. A computational approach to predict the structure of a protein demands a model that represents it abstractly, in a given level of details. Basing on well-established thermodynamical laws, the prediction of the structure of a protein is modelled as the minimization of the corresponding free-energy with respect to the possible conformations that a protein is able to attain.