Inverse Protein Folding on 2D Off-Lattice Model: Initial Results and Perspectives

David Pelta and Alberto Carrascal

Models of Decision and Optimization Research Group
Depto. de Ciencias de la Computación e I.A.
Universidad de Granada, 18071 Granada, Spain
dpelta@decsai.ugr.es

Abstract. Inverse protein folding or protein design stands for searching a particular amino acids sequence whose native structure or folding matches a pre specified target.

The problem of finding the corresponding folded structure of a particular sequence is, *per se*, a hard computational problem.

We use a genetic algorithm for searching the space of potential sequences, and the fitness of each individual is measured with the output of a second GA performing a minimization process in the space of structures.

Using an off-lattice protein-like 2D model, we show how the implemented techniques are able to obtain a variety of sequences attaining the target structures proposed.

1 Introduction

In very simple terms, a protein consists of a chain of amino acids' residues. The chemical properties and forces between the amino acids residues are such that, whenever the protein is left in its natural environment, it folds to a specific 3-dimensional structure, called its native state, which minimizes the total free energy. This 3D structure is specially relevant because it determines completely how the protein functions and interacts with other molecules. Most biological mechanisms at the protein level are based on shape-complementarity, so that proteins present particular concavities and convexities that allow them to bind to each other and form complex structures, such as skin, hair and tendon [5].

The field of structural bioinformatics is plenty of very interesting problems, and two of them are computationally hard: the protein structure prediction problem and the inverse protein folding problem (or protein design) Both attract the attention of researchers since a long time ago.

Protein structure prediction is one of the most significant technologies pursued by computational structural biology and theoretical chemistry. It has the aim of determining the three-dimensional structure of proteins from their amino acid sequences. In more formal terms, this is expressed as the prediction of protein tertiary structure from primary structure. Given the usefulness of known protein

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 207–216, 2007.

structures in such valuable tasks as rational drug design, this is a highly active field of research.

It is assumed that: a) all the information needed for a protein to fold, is coded in the sequence [1] and, b) the corresponding structure is the one that minimizes the free energy of the system. Under this situation, the problem can be re-stated as minimizing an energy function that may have several definitions as in the all-atom model [16], HP lattice models [12,18], those only considering the backbone [9] or atomic models where torsion (usually restricted) angles are considered [3].

Ab initio protein folding methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. Under simple models, this problem was shown to be NP-hard [6,17].

The progress of the field can be tracked through the results of the *Critical Assessment of Techniques for Protein Structure Prediction* (CASP) wide community contest. See for example, [13] for a review of the last decade.

Inverse Protein Folding (or Protein Design) is the design of new protein molecules from scratch. The number of possible amino acid sequences is infinite, but only a subset of these sequences will fold reliably and quickly to a single native state. Protein design involves identifying such sequences, in particular those with a physiologically active native state.

Inverse Protein Folding requires an understanding of the process by which proteins fold. In a sense it is the reverse of structure prediction: a tertiary structure is specified, and a primary sequence is identified which will fold to it.

Some simplifications are usually made to approach this problem. The set of aminoacids is limited, the fitness of a particular sequence is measured using a threading like approach, target structures are restricted to lattices, etc. [7,8,2]. Pierce and Winfree [14] showed that optimizing the set of rotamers for a specified backbone conformation is NP-Hard. The reader should note that under this model, there is no "folding process". A good overview of the field can be seen in [15,11] and the references therein.

In this work, we focus on this second problem using a genetic algorithm to explore the sequences space (*GA-seq*). In order to measure the fitness of each individual, we need to use a second genetic algorithm that explores the space of structures (*GA-struct*) looking for the one that minimizes an energy function. A remarkably aspect is that, although a 2D space is used, structures are not restricted to a particular lattice geometry.

The hypothesis of this work is that using *GA-seq* and given a particular target structure, *it is possible to obtain a sequence that, when folded with* GA-struct*, the corresponding structure resembles the target one.*

The paper is organized as follows: in Section 2, the models for sequences and structures are presented, and the main characteristics of both GAs implemented. Next, Section 3 is devoted to experiments and results. We divide the experiments in two parts: firstly, we analyze *GA-struct* in terms of convergence and suitability; secondly, *GA-seq* is tested for several target structures. Final comments and future lines of research are outlined in Section 4.