A Differential Evolution Approach for Protein Folding Using a Lattice Model

Heitor Silverio Lopes and Reginaldo Bitello

Bioinformatics Laboratory, Federal University of Technology – Paraná, Av. 7 de setembro, 3165 80230-901 Curitiba, Brazil

E-mail: hslopes@pesquisador.cnpq.br; regi.bitello@gmail.com

Received November 22, 2006; revised May 22, 2007.

Abstract Protein folding is a relevant computational problem in Bioinformatics, for which many heuristic algorithms have been proposed. This work presents a methodology for the application of differential evolution (DE) to the problem of protein folding, using the bi-dimensional hydrophobic-polar model. DE is a relatively recent evolutionary algorithm, and has been used successfully in several engineering optimization problems, usually with continuous variables. We introduce the concept of genotype-phenotype mapping in DE in order to provide a mapping between the real-valued vector and an actual folding. The methodology is detailed and several experiments with benchmarks are done. We compared the results with other similar implementations. The proposed DE has shown to be competitive, statistically consistent and very promising.

Keywords bioinformatics, differential evolution, evolutionary computation, protein folding

1 Introduction

Proteins are composed by amino acids chains where there is all the information necessary for generating a unique tri-dimensional structure. The exact way proteins fold just after being synthesized in the ribosome is unknown. As consequence, many computational approaches of different levels of complexity have been proposed to simulate the folding of proteins. However, to date, even simple models are still computationally expensive. Recently, several methods have been proposed in the quest of solving the protein folding problem (PFP), such as Monte Carlo simulation\cite{1}, genetic algorithms\cite{2} and ant colony optimization\cite{3}. Despite being an important issue in bioinformatics, there is still no efficient method for solving the PFP.

The objective of this work is to evaluate the applicability of the Differential Evolution algorithm to the PFP using the 2D-HP model, and to compare its performance with other similar algorithms recently published.

2 2D-HP Model

Amongst the several discrete models used to simulate how a protein folds, the hydrophobic-polar (HP) is, possibly, the most simple and most widely studied model. The HP model was proposed by Dill\cite{4}, who demonstrated that some behavioral properties of real-world proteins could be inferred by using this simple model. In this model, the amino acids of a protein are considered either hydrophobic (aversion to water) or polar (affinity to water, the same as hydrophilic). Despite the simplicity of this lattice model, exact algorithms to solve the problem were proved to be NP-hard\cite{5}.

Fig. 1 shows the 2D-HP model for a hypothetical

![2D-HP Model](image)

Fig. 1. Example of H-H contacts in the 2D-HP model.
18 amino acids-long protein, folded in such a way that 6 non-local H-H contacts occur. In this figure, black and white dots are, respectively, hydrophobic and polar amino acids. The square dot is the first element of the chain, and H-H contacts are represented by dotted lines.

The free energy function of a conformation, suggested by [6], is represented in (1):

\[ E = \sum_{i<j} e_{i,j} \Delta (r_i - r_j) \]  

(1)

where: \( \Delta (r_i - r_j) = 1 \) if amino acids \( r_i \) and \( r_j \) are not consecutive in the chain, and \( \Delta (r_i - r_j) = 0 \), otherwise. Depending on the type of contacts between amino acids \( r_i \) and \( r_j \), the energy \( e_{i,j} \) will be \( e_{HH}, e_{HP} \) or \( e_{PP} \), corresponding to H-H, H-P and P-P contacts, respectively. According to [6] this model satisfies the following physical limitations:

1) Compact conformations have a smaller energy value than any other non-compact conformation.

2) Hydrophobic amino acids will be buried inside the conformation, as most possible. This idea is expressed by the relationship \( e_{HP} > e_{PP} > e_{HH} \), that decreases the energy of conformations in which the Hs are hidden inside.

3) Different types of amino acids tend to get apart. This is expressed by the relationship: \( 2e_{HP} > e_{PP} > e_{HH} \).

3 Methodology

3.1 Differential Evolution

Differential Evolution (DE) is an evolutionary computation method invented by Storn and Price[7] for numerical optimization. The central idea of this algorithm is the use of difference vectors for generating perturbations in a population of vectors. This algorithm is conceptually simple and, at most times, converges fast to a good solution. Besides, DE is robust and has few parameters do be tuned. Consequently, it has drawn attention of researchers who have studied its utility for complex optimization problems[8,9]. Fig.2 shows graphically how the vector operators take place in a 2-dimensional space, at a given generation. A detailed description of DE can be found in [10].

3.2 Vector Encoding

In DE, the variables of the problem are encoded in a vector and, usually, the meaning of its elements to the real-world is straightforward. Consequently, the concept of genotype, as in genetic algorithms, is not applicable in the original DE. However, for the specific problem dealt in this work, the adaptation devised to represent possible solutions to the PFP in a real-valued vector requested the establishment of a genotype-phenotype mapping. Individuals in DE are real-valued vectors which, in turn, are decoded into a specific fold of an amino acid chain in a square lattice. The reason for this approach was to use the original DE algorithm, without significant changes.

Basically, there are three ways of representing an amino acid chain in a lattice using the HP model: Cartesian coordinates, internal coordinates and geometrical distances. The proposed implementation of DE uses relative internal coordinates. This coordinates system implicitly assures that the connectivity of the amino acids chains will be preserved when a given conformation is drawn in the lattice. This property avoids loosing time checking the validity of a given conformation.

Using the relative internal coordinates system in a bi-dimensional space, there are three possible movements: \( \{F\} \)orward, \( \{R\} \)ight and \( \{L\} \)eft. Therefore, the phenotypical representation of a solution is defined over the alphabet of movements \( \{F,R,L\} \). A given folding of \( N \) amino acids, represented by an \( N \)-dimensional vector, is defined by a string with \( N-1 \) movements. The genotypical representation is the usual real-valued vector of a regular DE. Considering \( x_{ij} \) the \( j \)-th element of vector \( X \), \( P \) the string representing the sequence of movements of the folding, and \( \alpha < \beta < \delta < \gamma \) arbitrary constants in \( \mathcal{R} \), the genotype-phenotype mapping is defined as follows:

\[
\begin{align*}
\text{If } \alpha < x_{ij} < \beta \text{ then } P_j = L, \\
\text{If } \beta < x_{ij} < \delta \text{ then } P_j = F, \\
\text{If } \delta < x_{ij} < \gamma \text{ then } P_j = R.
\end{align*}
\]  

(2)

Notice that the proposed mapping allows to privilege some movements by enlarging the corresponding range in which it is defined (or narrowing the other