The Role of a Priori Information in the Minimization of Contact Potentials by Means of Estimation of Distribution Algorithms

Roberto Santana, Pedro Larrañaga, and Jose A. Lozano

Department of Computer Science and Artificial Intelligence
University of the Basque Country, Donostia-San Sebastian, Spain
rsantana@si.ehu.es, pedro.larranaga@ehu.es, ja.lozano@ehu.es

Abstract. Directed search methods and probabilistic approaches have been used as two alternative ways for computational protein design. This paper presents a hybrid methodology that combines features from both approaches. Three estimation of distribution algorithms are applied to the solution of a protein design problem by minimization of contact potentials. The combination of probabilistic models able to represent probabilistic dependencies with the use of information about residues interactions in the protein contact graph is shown to improve the efficiency of search for the problems evaluated.

Keywords: estimation of distribution algorithm, protein design, energy minimization algorithms.

1 Introduction

The goal of protein design is to find sequences of aminoacids with desired structural and functional properties. The problem has been approached by the application of directed search methods which cast the search as an optimization method. The approach requires the definition of a simplified model of the proteins, a fitness function that associates a value to each solution according to its 'quality', and a search procedure to efficiently sample the search space. In the field of protein design, these methods have been called "directed approaches to protein design".

Another class of methods has been covered under the umbrella of "probabilistic approaches to protein design" [14]. They use site-specific aminoacid probabilities rather than specific sequences and are usually employed in domains where the information available about the problem is incomplete. Probabilistic approaches include the use of consensus sequences [8] to determine low energy sequences and other methods where the probabilities learned can be used to guide search algorithms.

In this paper, we present a different approach which is based on the use of estimation of distribution algorithms (EDAs) [7,9,13]. EDAs are evolutionary algorithms that construct an explicit probability model of a set of selected solutions.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 247–257, 2007.

EDAs have been used for protein structure prediction in simplified models [17], protein side chain placement [18] and *de novo* peptide design [2]. Their suitability to deal with protein problems is given by the incorporation of machine learning techniques in the construction of the models. These learning algorithms automatically extract relevant regularities and complex structural patterns shared by promising solutions. The information learned can be compactly stored in the probabilistic model, which is later used to guide the exploration of the search space. EDAs are also different from probabilistic approaches that use probabilities to bias the search (e.g. Monte Carlo based techniques [21]) and where probabilities are unchanged during the search.

The paper is organized as follows. In the following section we present the energy function and introduce the problem of finding the aminoacid sequence with the lowest energy. Section 3 describes the main characteristics of EDAs and introduces the EDAs based on tree models used in our application. Section 4 gives a description of the experimental framework. The numerical results are shown in Section 5. Section 6 presents the main conclusions of our work and discuss future work.

2 Approach to Protein Design: Finding the Sequence with the Lowest Energy

In this section, we introduce the problem of finding the aminoacid sequence with the lowest energy for a given energy function. We use X_i to represent a discrete random variable. A possible value of X_i is denoted x_i. Similarly, we use $\mathbf{X} = (X_1, \ldots, X_n)$ to represent an n-dimensional random variable and $\mathbf{x} = (x_1, \ldots, x_n)$ to represent one of its possible values.

We will approach the protein design problem following a strategy that is based on the optimization of contact functions. Contact potentials or scoring functions [19,20] measure how likely it is for a sequence to fold to a given structure. Although the potential functions have been mainly used to distinguish native from decoy structures [19,20], they can also be employed to study the distribution of native-like features in sequence space [11].

In [10,11], the sequence evolutionary selection mechanisms are analyzed focusing on the stability energy of sequences. Although the 'survival probability' of a protein sequence depends on a number of other factors such as protein function and protein flexibility, the sequence-structure relationship can be analyzed in terms of energy. The analysis assumes that native sequences were selected because they were highly probable as a function of energy.

We will denote the native sequence corresponding to the structure σ as $\mathbf{x}^\sigma$. $E(\mathbf{x}, \sigma)$ is the energy of sequence $\mathbf{x}$ in structure σ and $E_\sigma = E(\mathbf{x}^\sigma)$ is the native energy of sequence $\mathbf{x}^\sigma$ in structure σ. The quantity $N(E_\sigma)$ is the number of sequences whose energy in σ would be no greater than that of the actual native sequence. $N(E_\sigma)$ is called the *evolutionary capacity* of structure σ because it reflects how far the current state of molecular evolution σ is from the possible optimum in terms of energy [11].