Expert Systems with Applications 36 (2009) 12446–12453

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An efficient hybrid Taguchi-genetic algorithm for protein folding simulation

Cheng-Jian Lin *, Ming-Hua Hsieh

Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung County 411, Taiwan, ROC

ARTICLE INFO

Keywords:
Protein structure prediction
HP lattice model
Genetic algorithm
Taguchi method
Particle swarm optimization

ABSTRACT

Given the amino-acid sequence of a protein, the prediction of a protein's tertiary structure is known as the protein folding problem. The protein folding problem in the hydrophobic–hydrophilic lattice model is to find the lowest energy conformation. In order to enhance the performance of predicting protein structure, in this paper we propose an efficient hybrid Taguchi-genetic algorithm that combines genetic algorithm, Taguchi method, and particle swarm optimization (PSO). The GA has the capability of powerful global exploration, while the Taguchi method can exploit the optimum offspring. In addition, we present the PSO inspired by a mutation mechanism in a genetic algorithm. We demonstrate that our algorithm can be applied successfully to the protein folding problem based on the hydrophobic-hydrophilic lattice model. Simulation results indicate that our approach performs very well against existing evolutionary algorithm.

1. Introduction

The prediction of protein structure from its amino-acid sequence is one of the most prominent problems in computational biology. A protein's function depends mainly on its tertiary structure, which in turn depends on its primary structure. Mistakes in the folding process create proteins with abnormal shapes, which are the causes of diseases such as cystic fibrosis, Alzheimer's, and mad cow. If we could predict the tertiary structures of proteins from their sequences, we would be able to treat these diseases better. The knowledge of protein tertiary structures also has other applications, such as in the structure-based drug design field (Bui & Sundarraj, 2005).

Currently, protein structures are primarily determined by techniques such as MRI (magnetic resonance imaging) and X-ray crystallography, which are expensive in terms of equipment, computation, and time. Additionally, these techniques require isolation, purification, and crystallization of the target protein. Computational approaches to protein structure prediction are therefore very attractive. The difficulty in solving protein structure prediction problems stems from two major sources: (1) finding good measures for the quality of candidate structures, and (2) given such measures, determining optimal or close-to-optimal structures for a given amino-acid sequence (Krasnogor, Hart, Smith, & Pelta, 1999).

Recently, many researchers (Krasnogor, Pelta, Lopez, Mocciola, & de la Canal, 1998; Krasnogor et al., 1999; Patton, Punch, III, & Goodman, 1995; Pedersen & Moukt, 1997) have used evolutionary algorithms, such as the genetic algorithms (GA), for solving the protein folding problem. Genetic algorithms are stochastic search techniques based on the mechanism of natural selection, which requires information to search effectively in a large or poorly understood search space. The effectiveness of crossover and mutation is weakened in the protein folding problem (Krasnogor et al., 1998, 1999), since by increasing the compact folded structure, the failure of the crossover operation increases due to collisions. Further, in sequences of mutation, there will be often invalid conformations due to collisions within compact conformation. Therefore, some researchers (Bui & Sundarraj, 2005; Jiang, Cui, Shi, & Ma, 2003; König & Dandekar, 1999; Takahashi, Kita, & Kobayashi, 1999) have proposed various hybrid methods to improve GA. The above-mentioned improved GA methods were mainly aimed at the crossover and mutation operations.

Recently, the Taguchi method is a robust design approach. It uses many ideas from statistical experimental design for evaluating and implementing improvements in products, processes, and equipment. The fundamental principle is to improve the quality of a product by minimizing the effect of the causes of variation without eliminating the causes. The Taguchi method is suitable for a wide range of applications (Kaytakoğlu & Akyalçın, 2007; Liu, Fung, & Wang, 2007; Wang & Huang, 2008), including the following practices: quality engineering, experimental design, business data analysis, management by total results, pattern recognition, and so on. The Taguchi method is a series of approaches that predicts and prevents troubles or problems that might occur in the market after a product is sold and used by a

* Corresponding author.
E-mail address: cjlin@ncut.edu.tw (C.-J. Lin).

0957-4174/$ - see front matter Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.04.074

customer under various environment and real-life conditions for the duration of the product life.

In this paper, we focus on the 2D hydrophobic-polar (HP) lattice model. An efficient hybrid Taguchi-genetic algorithm (HTGA) is proposed for solving the protein folding problem in the 2D HP model. The proposed HTGA is mainly aimed to improve the crossover and mutation operators and enhance exploitation capability. In order to improve the crossover operation, we use the Taguchi method to select the better genes. In the mutation operation, we employ the merits of PSO to improve the mutation mechanism. Lin, Liu, and Lee (2008) and Lin and Hong (2007) used the particle swarm optimization to improve the mutation mechanism. Simulation results show that our method has a better performance than those of existing methods in protein folding problem.

The remainder of this paper is structured as follows: Section 2 gives the preliminaries and the formal definition of the protein folding problem in the 2D HP lattice model. Section 3 describes our approach in detail. The proposed hybrid Taguchi-genetic algorithm combining the traditional genetic algorithm, the Taguchi method, and particle swarm optimization is presented. The experimental results obtained by our method and by other methods are compared in Section 4. Finally, the conclusion is given in the last section.

2. Preliminaries

In this section, we briefly present the 2D HP protein folding problem and its free energy calculation.

2.1. The 2D HP protein folding problem

Lattice proteins are highly simplified computer models of proteins which are used to investigate protein folding. Dill (1985) proposes the hydrophobic-polar model. Because proteins are such large molecules, containing hundreds or thousands of atoms, it is not possible with current technology to simulate more than a few microseconds of their behavior in complete atomic detail. Hence, real proteins cannot be folded in a computer simulation. Lattice proteins, however, are simplified into two ways: the amino acids are modeled as single "beads" rather than by every atom, and the beads are restricted to a rigid (usually cubic) lattice. This simplification means they can fold to their energy minima in a time quick enough to be simulated. Lattice proteins are made to resemble real proteins by introducing an energy function (Takahashi et al., 1999), a set of conditions which specify the energy of interaction between neighboring beads, usually taken to be those occupying adjacent lattice sites. The energy function mimics the interactions, which include hydrophobic and hydrogen bonding effects, between amino acids in real proteins. The beads are divided into types, and the energy function specifies the interactions, depending on the bead type, just as different types of amino acid interact differently.

One of the most popular lattice models, the HP model, feature just two bead types: H (hydrophobic or non-polar) and P (hydrophilic or polar). An instance is shown in Fig. 1 for the 2D HP lattice model (Guo, Feng, & Wang, 2007). The black squares denote the hydrophobic amino acid and the white squares denote the hydrophilic. The dotted line denotes the H–H contacts (free energy) in the conformation, which are assigned an energy value of −1. The free energy is minimum value; the number of H–H contact is the maximum. Fig. 1 shows a protein structure with 9 H–H contacts (energy = −9). Since the native state of a protein generally corresponds to the lowest free energy state for the protein, the optimal conformation in the HP model is the one that has the maximum number of H–H contacts which gives the lowest energy value.

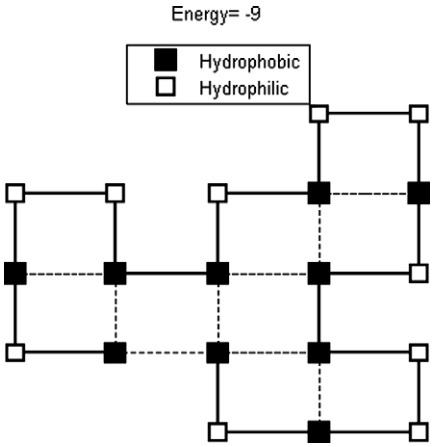

Fig. 1. An optimal conformation for the sequence "$(HP)_2PH(HP)_2(PH)_2HP(PH)_2$" in a 2D lattice model.

2.2. Calculating the free energy

For any sequence in any particular structure, the free energy can be rapidly calculated from the free energy function. For the simple HP model, this is simply an enumeration of all the contacts between the H residues that are adjacent in the structure but not in the chain. Most researchers consider a lattice protein sequence protein-like only if it possesses a single structure with an energetic state lower than in any other structure. This is the energetic ground state, or the native state. The relative positions of the beads in the native state constitute the lattice protein's tertiary structure. Lattice proteins do not have genuine secondary structure, although some researchers have claimed that they can be extrapolated to real protein structures, which do include secondary structure, by appealing to the same law by which the phase diagrams of different substances can be scaled onto one another. By varying the free energy function and the bead sequence of the chain (the primary structure), effects on the native state structure and the kinetics (rate) of folding can be explored. This may provide insights into the folding of real proteins. In particular, lattice models have been used to investigate the free energy landscapes of proteins (Guo et al., 2007), i.e. the variation of their internal free energy as a function of conformation. We present the minimum free energy function of the 2D HP lattice model with calculation conditions as follows:

$$E^* = \left[\frac{\sum_{i=1}^{n}\left[\sum_{j=1}^{n} f(i)S(i,j)\right]}{2}\right] \quad n = \text{length of the protein sequence}$$
$$\text{s.t. } S(i,j) = f(j) * \|d_j - d_i\| = 1 \quad \text{and} \quad \|i-j\| > 2 \tag{1}$$

where f is a mapping function: $f \to \{0,1\}$. That is, $f(i) = 0$ represents the hydrophilic residue and $f(i) = 1$ represents the hydrophobic residue. $\boldsymbol{d} = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ is a vector set, where d_j denotes a projection onto the Cartesian coordinate. If each residue is connected to its sequence neighbor on an adjacent lattice site, then $S(i,j) = 1$. Otherwise $S(i,j) \neq 1$. Each lattice site is only occupied by one amino acid residue, which we call a conformation valid.

3. Methods

In this section, we review the Taguchi method and particle swarm optimization. An efficient hybrid Taguchi-genetic algorithm is also presented.