Applied Soft Computing 45 (2016) 61–70

Contents lists available at ScienceDirect

Applied Soft Computing

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

Genetic algorithm with advanced mechanisms applied to the protein structure prediction in a hydrophobic-polar model and cubic lattice

Borko Bošković*, Janez Brest

Faculty of Electrical Engineering and Computer Science, University of Maribor, SI-2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 16 February 2016
Accepted 5 April 2016
Available online 14 April 2016

Keywords:
Protein structure prediction
Hydrophobic-polar model
Cubic lattice
Genetic algorithm
Clustering
Crowding
Repair mechanism
Local movements
Opposition-based mechanism

ABSTRACT

This paper presents a genetic algorithm applied to the protein structure prediction in a hydrophobic-polar model on a cubic lattice. The proposed genetic algorithm is extended with crowding, clustering, repair, local search and opposition-based mechanisms. The crowding is responsible for maintaining the good solutions to the end of the evolutionary process while the clustering is used to divide a whole population into a number of subpopulations that can locate different good solutions. The repair mechanism transforms infeasible solutions to feasible solutions that do not occupy the lattice point for more than one monomer. In order to improve convergence speed the algorithm uses local search. This mechanism improves the quality of conformations with the local movement of one or two consecutive monomers through the entire conformation. The opposition-based mechanism is introduced to transform conformations to the opposite direction. In this way the algorithm easily improves good solutions on both sides of the sequence. The proposed algorithm was tested on a number of well-known hydrophobic-polar sequences. The obtained results show that the mechanisms employed improve the algorithm's performance and that our algorithm is superior to other state-of-the-art evolutionary and swarm algorithms.

1. Introduction

The problem of protein structure prediction (PSP) represents the computational problem of how to predict the native structure of a protein from its amino acid sequence [1]. Protein function is determined by its structure. With the wrong structure, a protein cannot correctly fulfill its function [2]. With current algorithms and computational resources it is possible to predict the native structures of relatively small proteins and to help in drug design and proteomics [1]. The PSP problem is one of the more important challenges of this century [3] and because of its nature it attracts scientists from different fields such as physics, chemistry, biology, mathematics, and computer science. However, despite large research efforts in recent decades, further improvements are still possible for this problem [2].

The PSP problem is a hard combinatorial optimization problem, even under an hydrophobic-polar (HP) simplified model [4] that has been proved to be NP-complete [5]. Researchers have developed a great variety of simplified protein models [6–8]. These models reduce the number of degrees of freedom, take into account only specific interactions and therefore speed up the calculations. Although these models are incomplete, they allow for the development, testing, and comparison of various search algorithms and they offer a global perspective of protein structures. They can be helpful in confirming or questioning important theories [9].

This paper presents the genetic algorithm (GA_{PSP}) for constructing the native tridimensional structure of a given amino acid sequence under an HP model and cubic lattice. This algorithm belongs to the *ab initio* protein prediction type of methods which predict structures from scratch and do not require any information about related protein structures. The proposed GA is extended with crowding and clustering mechanisms, because the PSP problem is multimodal by nature. The crowding is employed for maintaining the diversity of the population and for preserving good solutions to the end of the evolutionary process. The clustering is used to divide a whole population into more subpopulations that can locate different good conformations [10]. Additionally, our algorithm includes a repair mechanism which transforms infeasible conformations to feasible ones. Feasible conformations do not occupy a lattice point with more than one monomer. Thus, the repair mechanism detects collision (a lattice point that is occupied with two

* Corresponding author.
E-mail addresses: borko.boskovic@um.si (B. Bošković), janez.brest@um.si (J. Brest).

http://dx.doi.org/10.1016/j.asoc.2016.04.001
1568-4946/

or more monomers) and tries to repair it or form a contiguous self-avoiding path within the lattice. A backtracking algorithm is used for this purpose, as proposed in [11]. For every detected collision, a repair mechanism tries to move a monomer from an already occupied point to an unoccupied neighborhood point. If there are no such points available, then it repeats this process for the previous monomer.

In order to speed up the convergence of the algorithm and to improve the energies of the generated feasible conformations, the algorithm uses local search that includes two operators of local movement. Local movement is a transformation of protein conformation whereby only a few monomers are locally moved from one point to another while the remaining monomers remain on their points. The introduced operators try to improve the energy of the conformations with local movements of one or two consecutive monomers. Local movements that have been described in the literature are used either as genetic operators that are applied on randomly chosen positions [12] or as macromutations within local searches [9]. In contrast to these operators, our operators try to locally move only one or two consecutive monomers, where at least one of them is an H–monomer, through the entire protein conformation. Used mechanisms inside an algorithm change conformation in a specific direction. Therefore an opposition-based mechanism is introduced to transform conformations of the next generation into opposite directions using inverse amino acid sequence. This mechanism is based on the assumption that inverse sequence allows the algorithm to easily improve conformation of the monomers at the beginning or at the end of the sequence.

The main contributions of this paper are: (1) the proposed new GA algorithm for the PSP problem, (2) the introduced local search or local movement operators that try to improve the energy of the conformations with local movements of one or two consecutive monomers throughout the entire conformation, and (3) a proposed opposition-based mechanism that transforms conformations into the opposite direction and the corresponding amino acid sequence into its inverse sequence. The proposed algorithm is tested on different well-known HP sequences and compared with the state-of-the-art evolutionary and swarm algorithms. Experimental results show that the proposed algorithm is superior to its competitors.

The remainder of this paper is organized as follows. The existing evolutionary and swarm algorithms for the PSP problem are briefly described in Section 2. Sections 3 and 4 describe the hydrophobic-polar model and the proposed algorithm in sufficient detail. The experimental setup and results are presented in Sections 5 and 6. Section 7 concludes this paper.

2. Related work

Evolutionary and swarm algorithms have previously been successfully applied to the HP model on the cubic lattice. In [13] the authors developed a GA algorithm that has a population of conformations. Conformations are changed using mutation, in the forms of conventional Monte Carlo steps, and crossovers, in which parts of the sequence of directions are interchanged between conformations. Both genetic operators iterate until feasible conformations are created. The results show that the GA algorithm is superior to conventional Monte Carlo methods. This work was upgraded in [14] where the authors showed that GA is effective for determining protein structure. In this work, infeasible conformations were used within the evolutionary process and their energies were calculated using a penalty function. Additionally, a crowding mechanism was used for maintaining population diversity.

The study [15] considered that the two operators of multi-point crossover and a local perturbation are required for any GA to be fully effective. A systematic crossover search is proposed in [16] that uses information about differences between two individuals. This mechanism in combination with GA significantly increased the search effectiveness. Another study [17] compared relative and absolute encoding of conformations. As a result of encoding, the search space is different and this is the reason why the GA described in [14] is superior to the GA described in [13]. The authors in this study also proposed a modified energy potential that facilitated the GA search and identified weaknesses in the constraint management strategies.

A repair procedure and evolutionary operators whose functioning is closed in feasible space are used within the evolutionary algorithm with an embedded backtracking algorithm [11]. A memetic algorithm with a self-adaptive local search mechanism is introduced in [9]. The results of that study confirm that the local search improves the algorithm's performance. In [18] GA is used with a modification of the scoring system for improving the model's capacity and to generate more natural-like structures. In this paper an islands' algorithm was also analyzed and it improved performance of the GA algorithm. The same authors in [12] developed a robust and efficient GA that employs a phenotype based crowding mechanism for the maintenance of useful diversity within the population. The GA algorithm that is extended with heuristic repair methods for dealing with infeasible intermediate candidate solutions is analyzed in [19].

A memetic algorithm with the following novel features is introduced in [20]: a modified fitness function, a systematic generation of population that automatically prevents infeasible conformations, a generalized non-isomorphic encoding scheme that implicitly eliminates a generation of symmetrical conformations, population clustering and the identification of a meme according to the genotype, and a 2-stage mechanism for migrating domain knowledge between different basins of attraction. The multi-objective approach to constraint-handling was investigated in [21]. For that purpose the HP model was reformulated as an unconstrained multi-objective problem by treating constraints as an additional objective function. From the obtained results, the authors showed that the multi-objective strategy can be incorporated within state-of-the-art algorithms and can improve their performances.

The following have also been applied to the HP model on a cubic lattice: an ant colony optimization algorithm with local search [22], a particle swarm optimization based algorithm [23], an estimation of distribution algorithm [24], a differential evolution algorithm [25], and an immune algorithm [26].

Unlike those in the literature [12,18], the proposed algorithm does not use elitism for fast convergence. The PSP problem is multimodal by nature and without elitism the algorithm has a greater chance of avoiding local optima. Nevertheless, our algorithm shows fast convergence. The local search is used for this purpose. In contrast to those in the literature [12,22,9], our local search uses local movements that are applied to the whole sequence and not only to the one randomly chosen position within the sequence, and it is not limited to local movements of only one monomer. It also performs local movements of two consecutive monomers. The proposed algorithm additionally includes the clustering mechanism that is designed according to [10] and the repair mechanism according to [11]. The clustering mechanism improves the convergence rate and the diversity of the population while the repair mechanism forms feasible solutions. Thus the algorithm does not need to compute an unnecessary evaluation of infeasible solutions. The local search and repair mechanisms perform their tasks from the beginning to the end of the sequence. After each generation, the introduced opposition-based mechanism transforms all individuals into their opposite directions. The result of this transformation is that the local search and repair mechanisms perform their tasks