Information Sciences 454–455 (2018) 178–199

Contents lists available at ScienceDirect

Information Sciences

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

Protein folding optimization using differential evolution extended with local search and component reinitialization

Borko Bošković*, Janez Brest

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

ARTICLE INFO

Article history:
Received 20 October 2017
Revised 21 April 2018
Accepted 25 April 2018
Available online 25 April 2018

Keywords:
Protein folding optimization
Three-dimensional AB off-lattice model
Differential evolution
Local search
Component reinitialization

ABSTRACT

This paper presents a novel Differential Evolution algorithm for protein folding optimization that is applied to a three-dimensional AB off-lattice model. The proposed algorithm includes two new mechanisms. A local search is used to improve convergence speed and to reduce the runtime complexity of the energy calculation. For this purpose, a local movement is introduced within the local search. The designed evolutionary algorithm has fast convergence speed and, therefore, when it is trapped into the local optimum or a relatively good solution is located, it is hard to locate a better similar solution. The similar solution is different from the good solution in only a few components. A component reinitialization method is designed to mitigate this problem. Both the new mechanisms and the proposed algorithm were analyzed on well-known amino acid sequences that are used frequently in the literature. Experimental results show that the employed new mechanisms improve the efficiency of our algorithm and that the proposed algorithm is superior to other state-of-the-art algorithms. It obtained a hit ratio of 100% for sequences up to 18 monomers, within a budget of 10^{11} solution evaluations. New best-known solutions were obtained for most of the sequences. The existence of the symmetric best-known solutions is also demonstrated in the paper.

1. Introduction

The protein structure prediction represents the problem of how to predict the native structure of a protein from its amino acid sequence. This problem is one of the more important challenges of this century [17] and, because of its nature, it attracts scientists from different fields, such as Physics, Chemistry, Biology, Mathematics, and Computer Science. Within the protein structure prediction, the Protein Folding Optimization (PFO) represents a computational problem for simulating the protein folding process and finding a native structure. Most proteins must fold into a unique three-dimensional structure, known as a native structure, to perform their biological function [2]. A protein's function is determined by its structure. The inability of a protein to form its native structure prevents a protein from fulfilling its function correctly, and this may be the basis of various human diseases [24].

The PFO belongs to the class of NP-hard problems [11] and, with current algorithms and computational resources, it is possible to predict the native structures of relatively small proteins. The reason for that is the huge and multimodal search space. For example, a polypeptide that has only 18 amino acids, will have 31 angles within a simplified AB model (see

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

https://doi.org/10.1016/j.ins.2018.04.072
0020-0255/

Section 3). Using uniform discretization with only 10 values for each angle, there would be 10^{31} possible configurations. To evaluate and select the correctly folded conformation among all these conformations in the time elapsed since the Big Bang, we need the huge computational speed of $10^{31}/(4.32 \cdot 10^{17}) = 2.31 \cdot 10^{13}$ conformation evaluations per second. This is much faster than the speed obtained within our experiment, where we can evaluate only $5.73 \cdot 10^5$ conformations per second. From these numbers, we can see that the search space is huge, even in the simplified model, which makes this problem very hard. However, in reality, the proteins fold into their native conformation on a time scale of seconds, and this contradiction is known as Levinthal's paradox [7]. An optimization algorithm can give good results of a PFO problem only if it can locate good solutions and evaluate solutions efficiently. Here, the approximation techniques, such as heuristic and metaheuristic, with efficient data structures, become the only viable alternatives as the problem size increases.

Some simplified protein models exist, such as HP models within different lattices [5] and the AB off-lattice model [28]. Simplified protein models were designed for development, testing, and comparison of different approaches. The AB off-lattice model was used in the paper for demonstrating the efficiency of the proposed algorithm. This model takes into account the hydrophobic interactions which represent the main driving forces of a protein structure formation and, as such, still imitates its main features realistically [14]. Although this model is incomplete, it allows the development, testing, and comparison of various search algorithms, and offers a global perspective of protein structures. It can be helpful in confirming or questioning important theories [3].

Our algorithm is based on the Differential Evolution (DE) algorithm that was proposed by Storn and Price [29]. It is a powerful stochastic population-based algorithm. Three simple operators, mutation, crossover, and selection, were used inside the DE algorithm to transform real-coded individuals with the purpose to locate optimal or sub-optimal solutions. Because of its simplicity and efficiency, it was used in various numerical optimization problems, such as an animated trees reconstruction [36], an intrusion detection [1], and an image thresholding [25]. An advanced DE variant, such as L-SHADE [30] was also the winner of the recent CEC (IEEE Congress on Evolutionary Computation) competitions. For more details about DE, we refer the reader to [27] and to survey [10].

It has been shown that the PFO has a highly rugged landscape structure containing many local optima and needle-like funnels [16], and, therefore, the algorithms that follow more attractors simultaneously are ineffective. In our recent work [4], to overcome this weakness, we proposed a Differential Evolution (DE) algorithm that uses the *DE/best/1/bin* strategy. With this strategy, our algorithm follows only one attractor. The temporal locality mechanism [35] and self-adaptive mechanism [6] of the main control parameters were used additionally to speed up the convergence speed. When the algorithm was trapped in a local optimum, then random reinitialization was used. This algorithm belongs to the ab-initio PFO methods, which optimize structures from scratch, and do not require any information about related sequences. It showed a very fast convergence speed, and it was capable of obtaining significantly better results than other state-of-the-art algorithms.

Taking into account the finding of the previous paragraph, we propose two new mechanisms, that, additionally, improve the efficiency of our algorithm. A new local search mechanism was designed in order to improve convergence speed and to reduce the runtime complexity of the algorithm. A similar idea was already used within the HP model [5], where it is applied to the cubic lattice. Using a simple local search mechanism, where only one solution's component is changed, can produce a structure whereby a lot of monomers are moved. This means their positions must be recalculated and efficient energy calculation is not possible. In contrast to simple local search, our mechanism improves the quality of conformations using the local movements within the three-dimensional AB off-lattice model. We define a local movement as a transformation of conformation, whereby only two consecutive monomers are moved locally in such a way that the remaining monomers remain in their positions. The described local movement allows efficient evaluation of neighborhood solutions and faster convergence speed.

With the fast convergence speed the algorithm can locate good solutions quickly, but it has a problem locating good similar solutions. For example, if an algorithm locates a good solution that is different from the global best solution in only one or few components, then the random restart, that was used in our previous work, is not an efficient solution. For that purpose, a component reinitialization was designed and incorporated within our algorithm. This mechanism is employed when the local best solution is detected. Instead of the random restart, it produces similar solutions that are different from the local best solution in only a few components.

We called the proposed algorithm DE_{lscr} and it was tested on two sets of amino acid sequences that were used frequently in the literature. The first set included 18 real peptide sequences, and the second set included 4 well-known artificial Fibonacci sequences with different lengths. Experimental results show that the proposed mechanisms improve the efficiency of the algorithm, and the algorithm is superior to other state-of-the-art algorithms. Its superiority is especially evident for longer sequences. With the proposed algorithm, that is stochastic, we cannot prove the optimality of the obtained conformations. However, we can infer about them according to the observed hit ratio. The experimental results show that our algorithm obtained a hit ratio of 100% for sequences that contain up to 18 monomers. For all longer sequences, we can only report the best-known conformations that are almost surely not optimal. Based on these observations, the main contributions of this paper are:

1. The proposed new DE algorithm for the PFO on a three-dimensional AB off-lattice model.
2. The local search mechanism that improves convergence speed and reduces runtime complexity of solution evaluations within the neighborhood.