Applied Soft Computing 15 (2014) 88–99

Contents lists available at ScienceDirect

Applied Soft Computing

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

A multiple minima genetic algorithm for protein structure prediction

Fábio Lima Custódio*, Helio J.C. Barbosa, Laurent Emmanuel Dardenne**

Laboratório Nacional de Computação Científica, Av. Getúlio Vargas 333, Petrópolis, RJ, Brazil

ARTICLE INFO

Article history:
Received 1 October 2012
Received in revised form 18 October 2013
Accepted 29 October 2013
Available online 6 November 2013

Keywords:
Genetic algorithms
Protein structure prediction
HP model
Multiple minima

ABSTRACT

Protein structure prediction (PSP) has a large potential for valuable biotechnological applications. However the prediction itself encompasses a difficult optimization problem with thousands of degrees of freedom and is associated with extremely complex energy landscapes. In this work a simplified three-dimensional protein model (hydrophobic-polar model, HP in a cubic lattice) was used in order to allow for the fast development of a robust and efficient genetic algorithm based methodology. The new methodology employs a phenotype based crowding mechanism for the maintenance of useful diversity within the populations, which resulted in increased performance and granted the algorithm multiple solutions capabilities. Tests against several benchmark HP sequences and comparative results showed that the proposed genetic algorithm is superior to other evolutionary algorithms. The proposed algorithm was then successfully adapted to an all-atom protein model and tested on poly-alanines. The native structure, an alpha helix, was found in all test cases as a local or a global minimum, in addition to other conformations with similar energies. The results showed that optimization strategies with multiple solutions capability present two advantages for PSP applications. The first one is a more efficient investigation of complex energy landscapes; the second one is an increase in the probability of finding native structures, even when they are not at the global optimum.

1. Introduction

The protein structure prediction (PSP) problem is one of the most interesting challenges of modern computational biology [1]. *Ab initio* protein prediction methods aim at constructing the native tridimensional structure for a given amino acid sequence, without requiring the use of experimental information about related protein structures.

Methods for PSP have a wide range of important biotechnological applications, e.g., the design of new proteins and folds [2,3], structure based drug design projects [4], refinement of theoretical models obtained by comparative modeling [5,6], and obtaining experimental structures from incomplete nuclear magnetic resonance data [7,8]. Furthermore, the function of a protein is a consequence of its structure, and to be able to predict the native structures of proteins would help to take advantage of the large amount of biological information that is being generated by genome sequencing projects.

Most PSP methods follow the thermodynamics hypothesis, i.e., the conformation adopted by the protein under physiological conditions is the conformation with the lowest Gibbs free energy [9]. Thus the problem is formulated as a minimization problem and can be divided in two sub-problems: (i) to define an appropriate energy function that places the native structure on its global minimum and is able to discriminate correct from incorrect folds; (ii) to develop an efficient and robust search strategy capable of dealing with a large number of variables and a highly degenerated and complex energy landscape.

The search for a method capable of predicting the native structure in the absence of known homologous structures is still an open problem. Research efforts are periodically evaluated during the biannual CASP (Critical Assessment of Structure Prediction) meetings. Currently the most promising methods utilize some form of information from known protein structures, nevertheless they are still based on optimizing complex and expensive energy functions [10–12]. This search is often carried out by metaheuristics and, amongst them, Genetic Algorithms (GA) are noteworthy [13] because of their robustness and wide applicability. These advantages can be explained, at least partially, by the GA's stochastic nature and because they work with a population of candidate solutions (that makes this type of method naturally parallel). Another attractive feature is that they do not require differentiability or continuity of the fitness function. Despite their advantages, they usually require a large number of fitness function evaluations in order to reach optimal or near optimal solutions specially when complex models are involved such as those used for PSP with atomic details.

* Corresponding author. Tel.: +55 24 2233 6283.
** Corresponding author.
E-mail addresses: flc@lncc.br, flcustodio@gmail.com (F.L. Custódio), dardenne@lncc.br (L.E. Dardenne).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.10.029

It is a common practice to adopt models with reduced complexity in PSP and protein folding related studies [14], e.g., lattice models. By sacrificing atomic details, lattice models may be used to extract essential folding principles, make predictions, and unify the knowledge on several protein properties without the heavy computational cost associated with all-atom models [15–17]. In addition to reducing the number of atomic "types", one important approximation introduced by lattice models is the discretization of the conformational space. While this makes the construction of an exact protein structure impossible, some key features of the optimization problem are retained. The energy landscape presents massive multimodality, different conformations with the same energy (degeneracy), and large regions with unfeasible conformations. Consequently, methods capable of finding low energy conformation for lattice models have not only the potential to provide insights on the folding process but also can be applied to complex models with atomic details [18]. The first studies of protein structure prediction (PSP) with lattice models considered that the abstract formulations had limited practical applications. However, further studies showed results that could be applied to more detailed models and also that the search methodologies could be adapted to other related problems [19].

The simple well defined energy function of lattice models facilitates a faster and more reliable development of robust optimization methodologies. Reliability is an important aspect because when dealing with complex atomistic models the success of a methodology is subject to inaccuracies in the energy models.

This paper reports a research on a multiple minima genetic algorithm applied to the hydrophobic-polar (HP) simplified protein model in a cubic lattice. The performance of the GA was analyzed and compared to other methods from the literature. The methodology was then adapted and applied to an all-atom model under a classical force field energy function.

1.1. Hydrophobic-polar model in a cubic lattice

Simplified models employed on PSP studies try to reflect general characteristics of protein structures [20]. The HP model's folding process has some behavioral similarities with the folding of real proteic system [21,22]. The hydrophobic–hydrophilic (or hydrophobic-polar, HP) model [23] describes the proteins based on the principle that during the folding process hydrophobic amino acids tend to be hidden from a polar solvent, thus resulting in the formation of a hydrophobic core within the native tridimensional structure. The HP-model abstracts the amino acid sequence to a binary sequence of monomers that are either hydrophobic (H) or polar (P).

Conformations are usually bound to some lattice, with each monomer occupying one site and there have been several works describing different types of lattices [18]. In this work, a tridimensional structure is portrayed as a self-avoiding walk on a cubic lattice. The energy is calculated as the negative of the number of hydrophobic–hydrophobic contacts (HH contacts) which are defined as two non-consecutive (non-bonded) monomers occupying adjacent sites on the lattice.

A conformation is valid only when no lattice site is occupied by more than one monomer. Invalid conformations are said to contain collisions. For a valid conformation c under the HP model, with n HH contacts, the energy E is given by:

$$E(c) = n \cdot (-1) \tag{1}$$

In spite of the model simplicity, finding optimal structures on a cubic lattice has been classified as a $\mathcal{NP}$-hard problem [24,25]. As a result, several nature-inspired metaheuristics have been applied to the problem [26], such as, Immune Algorithm [27], Ant Colony Optimization [28], Differential Evolution [29], Particle Swarm Optimization [30] and Evolutionary Algorithm [31]. The proposed algorithm belongs to the evolutionary class, more specifically genetic algorithms.

The energy landscape of the HP-model has some key features that directed the design of the genetic algorithm implemented in this work. These are (i) considerable multimodality, (ii) high degeneracy, and (iii) large regions of invalid conformations [32,25,33–35]. Another important attribute is that low energy structures may have different topologies which can be very different from the ones observed for global minima [36].

Several algorithms have been applied to the PSP-HP problem on cubic lattices and were tested against the same set of benchmark sequences used in this work. Those include evolutionary algorithms, construction based approaches (chain growth) and other heuristics.

1.2. Evolutionary algorithms applied to the HP model on cubic lattices

Unger and Moult [37] employed a hybrid genetic algorithm/Monte Carlo method, using an encoding where individuals (candidate solutions) were represented in a sequence of absolute directions $s \in \{U, D, L, R, F, B\}$ (Up – Down – Left – Right – Front – Back), for a chain of length n. Only valid conformations were accepted and each genetic operator would iterate until a new valid solution was created. The results showed that the hybrid GA constructed low energy solutions in fewer steps (function evaluations) than a pure Monte Carlo method.

In turn, Patton et al. [38] employed a standard GA with a relative encoding, i.e., $s \in \{U, D, L, R, F\}$ which has the advantage of eliminating return moves, e.g., $\{L, R\}$, that under absolute encoding schemes cause collisions. Invalid conformations were tolerated but suffered penalties, that is, collided monomers did not have their HH contacts computed and the total number of HH contacts was decremented by the number of collisions. A crowding scheme, using energy as the crowding criterion, was used to maintain diversity within the populations. Their results for test sequences from [37] showed that structures with lower energies required fewer function evaluations than the hybrid GA/MC.

Khimasia and Coveney [39] introduced a "simple GA" using absolute encoding with fitness-based selection and elitism. Conformations with collisions were tolerated, but penalized. The GA was tested with the sequences set from [37] and [40] and performed well with the shorter sequences (27 monomers) from [37]. It was suggested that a multi-point crossover could improve the performance for longer sequences. The study by Krasnogor et al. [34] explored the relative merits of absolute and relative encodings, defining a series of isomorphisms amongst genetic operators for both codifications. They suggested that the most useful operators acted as local optimizers, that is, the new structures remained within the neighborhood of the original ones on the conformation space. They also penalized invalid conformations.

A GA was applied with emphasis on local search (a memetic algorithm) following the ideas of Krasnogor et al. [34] and on the implementation of speciation methods based on the genotype (for preserving diversity in the population) [41]. The results on the test sequences set from Yue et al. [40] showed that local search has good potential for increasing the performance when simple operators (one-point crossover and mutation) are used.

Custódio et al. [42] introduced a modification to the scoring system of the original HP model in order to generate more natural structures. Additionally, they showed that a simple genetic algorithm could have its performance improved with the use of a new selection scheme (modified elitism) and application of a multi-point crossover. Mansour and Kanj [31] applied a simple genetic