Discrete Applied Mathematics 198 (2016) 206–214

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

An integer programming model for protein structure prediction using the 3D-HP side chain model

Luiz Fernando Nunes [a,b,*], Lauro Cesar Galvão [a,b], Heitor Silvério Lopes [a], Pablo Moscato [b,c,d], Regina Berretta [b,c]

[a] Bioinformatics Laboratory, Federal University of Technology Paraná, Av. 7 de setembro, 3165, Curitiba, Brazil
[b] Centre for Bioinformatics, Biomarker Discovery and Information-based Medicine, The University of Newcastle, Newcastle, Australia
[c] Hunter Medical Research Institute, Australia
[d] ARC Centre of Excellence in Bioinformatics, Australia

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 28 May 2015
Accepted 21 June 2015
Available online 9 July 2015

Keywords:
Integer programming
Protein folding
Lattice models
Bioinformatics

ABSTRACT

In spite of the fact that many simplified model variants of protein structure prediction have been widely studied in the past years, few attention has been given to discrete models with side chains, for which there is no specific benchmark. In this paper, we propose an integer programming model for the 3D-HP side chain protein structure prediction problem. The model accounts for the energy resulting from all types of interactions, between pairs of backbone elements, hydrophilic side chains and hydrophobic side chains. Three sets of instances, modified from the literature, were used in the experiments, and the maximum number of non-local hydrophobic contact was found using the ILOG CPLEX optimization package. We offer the optimal solution found for several instances of the benchmark. It is expected that the mathematical model allow further studies of the protein structure prediction with side chains and may, for some cases, provide new optimal values or new bounds that would rekindle the interest to this fascinating problem domain.

1. Introduction

Proteins consist of chains of amino acids (also called residues) and perform several vital functions in living organisms. Proteins are primarily formed in the ribosome. Amino acids are sequentially added to the chain by chemical bonds, called peptide bonds. During the assembling of a protein, it continuously folds over itself, achieving a final specific three-dimensional structure known as native conformation. This process is known as protein folding. Considering that many diseases are associated with failures in the folding process of proteins, it is generally conjectured that a better understanding of this process may, eventually, contribute to the development of new drugs to treat such diseases [2].

In a protein, each amino acid or residue is represented by a backbone and an associated side chain. A side chain can be either hydrophobic or hydrophilic, depending on its affinity or not to water molecules. The Protein Structure Prediction problem considered in this work, aims at finding the native protein conformation, such that the interactions between hydrophobic side chains are maximized, as explained below.

The Protein Structure prediction problem is one of the most challenging problems in computational Biology [7,15,16]. Considering the complexity of a real protein (analytical model), several discrete and continuous models have been proposed

* Corresponding author at: Bioinformatics Laboratory, Federal University of Technology Paraná, Av. 7 de setembro, 3165, Curitiba, Brazil. Tel.: +55 41 3310 4649; fax: +55 41 3310 4683.
E-mail address: nunes@utfpr.edu.br (L.F. Nunes).

http://dx.doi.org/10.1016/j.dam.2015.06.021
0166-218X/

in order to simplify the computational and mathematical treatment of the folding process. Among these models there are: Hydrophobic–Polar (HP) Model [13], Lattice Polymer Embedding [26], Charged Graph Embedding [9], Perturbed Homopolymer [19], Helicoidal–HP Model [23], and AB Toy Model [21]. Despite these simplifications, the exhaustive search of the conformational space of a protein using the simplest model (HP in two dimensions—2D) leads to a problem which complexity was proved to be NP-hard [9,26]. Consequently, many heuristic methods have been proposed for solving instances of 2D and 3D discrete models [15,22,24,27], as well as for continuous models [20,21,33,18]. However, there are few studies in the literature involving methods to solve the protein folding problem considering models with side chains, such as those suggested by Bromberg and Dill [6] and Hart and Istrail [11]. Some methods can be cited, such as Monte Carlo variants [14], Parallel Genetic Algorithms [3,4], Parallel Artificial Bee Colony Algorithm [5], and Greedy Algorithm [10].

Similarly, only few studies in the literature deal with the above mentioned problem by means of mathematical programming. Mandal and Jana [17] worked in two dimensions and limited the size of the lattice depending on the amount of amino acids. Türkay et al. [25] used integer linear programming to classify proteins according to their secondary structure. Both Carr et al. [7] and Yanev et al. [28,29] proposed similar integer programming methods applied to the HP model. Our work was inspired by those models, but with a quite different approach and applied to the 3D-HP-SC model. Yoon [30] presented two integer programming approaches and five constraint programming (CP) models. His research focuses on *ab-initio* mathematical models to find provably optimal solutions to the 2D-HP protein folding model. Ahn and Park [1] suggested a mathematical formulation of the HP model using a 2D square lattice and provided an upper bound on the optimal value using LP relaxation. It is important to recall that none of the above-mentioned works considered proteins using side chains in their mathematical models. Kingsford et al. [12] presented an integer linear programming formulation to position side chains in a fixed backbone (side-chain positioning problem). They relaxed the integrality constraints to give a polynomial-time linear programming heuristic. They still applied linear programming to position side chains on native and homologous backbones and to choose side chains for protein design.

In this paper, an integer programming model is proposed to deal with the protein folding problem. We used a three-dimensional representation of proteins, based on the Hydrophobic–Polar (HP) model, but using side chains. The use of side chains in protein models is found very sparsely in the literature, possibly due to the complexity involved. However, this feature aggregates a higher level of realism to the simulations. We also propose in this paper a set of benchmarks, derived from other models, so that researchers can further test their algorithms and compare results.

2. The 3D-HP-SC integer programming formulation

Proteins are composed by chain of amino acids. Using simple discrete models, each amino acid can represented by two elements: a backbone and a side chain. All, but the amino acids at the extreme of the chain, are connected to two other amino acids (its predecessor and antecessor) through the backbone. For real-world amino acids, the side chain is the main responsible for defining their chemical and physical features. In this work, the main feature represented by the model is its hydropathicity, that is, its affinity to water. Therefore, amino acids can be either hydrophobic (repel water) or hydrophilic (interact with water). The last one is more usually known as polar.

The simplest discrete model for representing a protein chain was proposed by Dill [8], and it is known as 2D-HP (Hydrophobic–Polar in two dimensions). Amino acids are embedded in a square lattice, such that a self-avoiding path represents a folding. Despite of being far away from reality, this is, possibly, one of the most studied models for this purpose. In this model, the quality of a folding is measured by means of a free energy function that takes into account the number of hydrophobic "contacts" between amino acids. A contact is defined as the unit distance between two non-successive amino acids of the chain. When a protein is folded to its native state, the number of contacts is maximal and the free energy is minimal. Therefore, the protein folding problem can be understood as an optimization problem [16].

The 3D-HP model that we consider assumes that each amino acid of a protein is represented by a backbone and a side-chain. All elements (either backbone or side chain) are embedded in a cubic lattice such that they occupy only one lattice point. In the immediate neighborhood of each backbone element, its respective side chain is positioned. Successive backbones are represented in successive neighborhood points of the lattice to maintain the sequence.

The objective is to define the position of each backbone and its side chain in the lattice, in such a way to maximize the number of hydrophobic interactions (contacts), i.e., to maximize the number of hydrophobic side chains positioned at neighboring vertices of the lattice. The free-energy of a given 3D conformation is inversely proportional to the number of nonlocal hydrophobic side chain contacts, according to Thachuk et al. [22]. Consequently, an algorithm that maximizes the number of contacts, conversely, minimizes the free-energy.

Fig. 1 shows an example with 10 elements, where the backbones are represented by black dots, hydrophobic side chains by red dots and hydrophilic side chains by blue dots. The numbers inside the dots are used just to indicate the position of each backbone and its side chain in the sequence. The specific conformation shown in this figure displays three nonlocal hydrophobic side chain contacts, namely, between side chains 2–3, 3–10 and 5–6.

2.1. Mathematical notation

Let S be a string with n positions where each element belong to set $S = \{0, 1\}$. Each element in this string indicates the hydrophobicity of the side chain associated with the corresponding backbone element. Each hydrophobic side chain is