ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

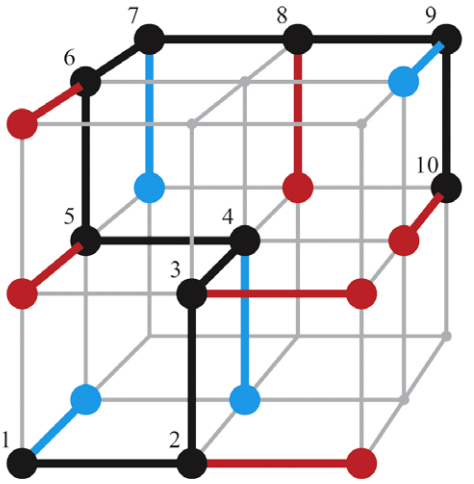

Fig. 1. An example of a conformation with 10 elements, where the backbones are represented by black dots, the hydrophobic side chains by red dots and hydrophilic side chains by blue dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

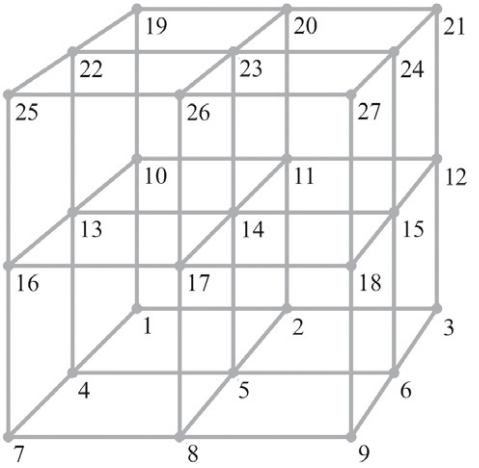

Fig. 2. The nodes of the lattice are numerated so that the neighborhood of an odd node in the lattice is formed only by even nodes and vice-versa.

represented by 1, while 0 represents hydrophilic side chains. We need to assign these amino acids in a 3D square lattice where each vertex can receive one backbone or one side chain. The vertices of this lattice are numerated from 1 to m, where m must to be appropriately chosen to bear all the elements.

Let I be the set of indices in S, i.e. $I = \{1, \ldots, n\}$, and L be the set of indices in the lattice, i.e. $L = \{1, \ldots, m\}$. We break down I in the subsets, as follows: I_e as the set of even indices in I and I_o as the set of odd indices in I, such that $I = I_e \cup I_o$. We can also break down I in the following four subsets: H_e as the set of indices of hydrophobic side chain (ones) in the even positions in I; H_o as the set of indices of hydrophobic side chains (1s) in the odd positions in I; P_e as the set of indices of hydrophilic side chains (0s) in the even positions in I; and P_o as the set of indices of hydrophilic side chains (0s) in the odd positions in I. Thus, $I = H_e \cup H_o \cup P_e \cup P_o$. We break down L as follows: L_e as the set of even elements in L, and L_o as the set of odd elements in L. Thus, $L = L_e \cup L_o$.

Let $N(v)$ represent the set of adjacent vertices to v in the lattice (neighborhood of v). So, $N(v) = \{t \in L | d(v, t) = 1\}$, where $d(v, t)$ is the Euclidean Distance between v and t. The vertices of the lattice are numerated so that the neighborhood of an odd vertex is formed only by even vertices and the neighborhood of an even vertex is formed only by odd ones. For example, consider the lattice shown in Fig. 2. The neighborhood of the vertex 14 is the set formed by the vertices 11, 17, 13, 15, 5 and 23. Likewise, the neighborhood of the vertex 2 is the set formed by the vertices 1, 3, 5 and 11. We call the set of feasible edges (v, w) in the lattice by E, such that $v \in L_o$ and $w \in L_e$, $w \in N(v)$. The set of feasible edges in the lattice can still be represented by F, which is the set of (v, w) such that $v \in L_e$ and $w \in L_o$, $w \in N(v)$.

We consider the following types of interactions: interactions between backbones, interactions between side chains (hydrophobic or hydrophilic) and interactions between backbones and side chains (hydrophobic or hydrophilic). Each type of interaction has an associated value, given by the *Energy* matrix shown in (1), where h represents a hydrophobic side chain,

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

p a hydrophilic side chain and b a backbone. Therefore, ε_{hh} is the energy of an interaction between two hydrophobic side chains, ε_{hp} is the energy associated with an interaction between a hydrophobic side chain and a hydrophilic side chain, and so on.

$$Energy = \begin{bmatrix} \varepsilon_{hh} & \varepsilon_{hp} & \varepsilon_{hb} \\ \varepsilon_{ph} & \varepsilon_{pp} & \varepsilon_{pb} \\ \varepsilon_{bh} & \varepsilon_{bp} & \varepsilon_{bb} \end{bmatrix}. \tag{1}$$

It is possible to consider which type of interaction to take into account and the respective weight by choosing appropriated values to the matrix (1). If there is no interest to consider some types of interactions, just set zeros in the respective components of the matrix.

It is important to note that there are only non-local interactions (no peptide bonds) between backbone elements and side chains if they are of the same parity, i.e., between an even backbone and an even side chain or between an odd backbone and an odd side chain. However, the interactions between backbone elements or between side chains occur only between elements of different parities.

2.2. Variables for the proposed model

We consider that even backbones are placed only on even lattice vertices, and, similarly, odd backbones are placed only on odd lattice nodes. Side chains associated to even backbones are placed only on odd lattice vertices, while side chains associated to odd backbones are only placed on even lattice nodes (see Fig. 1).

Variables x_{iv} are defined for $i \in I_o$ and $v \in L_o$ or $i \in I_e$ and $v \in L_e$, indicating whether or not the backbone element i is placed at lattice node v. So, x_{iv} is 1 if the backbone i is placed at lattice point v or 0, otherwise.

Similarly, variables y_{iv} are defined for $i \in I_o$ and $v \in L_e$ or $i \in I_e$ and $v \in L_o$, indicating whether or not the side chain i, associated to the backbone element i, is placed at lattice node v. So, y_{iv} is 1 if the side chain i is placed at lattice node v, and 0 otherwise.

Variables $bb_{(iv)(jw)}$ are defined for $i \in I_o, j \in I_e - \{i-1, i+1\}$, $(v, w) \in E$, indicating whether or not there is a contact between backbone elements i and j on edge (v, w). Therefore, $bb_{(iv)(jw)}$ is 1 if there is a contact, and 0, otherwise. Notice that $j \in I_e - \{i-1, i+1\}$ is to consider only interactions between non-consecutive backbone elements in the chain.

Variables $hh_{(iv)(jw)}$ are defined for $i \in H_e, j \in H_o$, $(v, w) \in E$, indicating whether or not there is a contact between hydrophobic sides chain element i and j on edge (v, w). Therefore, $hh_{(iv)(jw)}$ is 1 if there is a contact, and 0, otherwise.

Variables $pp_{(iv)(jw)}$ are defined for $i \in P_e, j \in P_o$, $(v, w) \in E$, indicating whether or not there is a contact between hydrophilic side chain elements i and j on edge (v, w). Therefore, $pp_{(iv)(jw)}$ is 1 if there is a contact, and 0, otherwise.

Variables $hp_{(iv)(jw)}$ are defined for $i \in H_e, j \in P_o$, $(v, w) \in E$, or $i \in H_o, j \in P_e$, $(v, w) \in F$, indicating whether or not there is a contact between hydrophobic and hydrophilic side chain elements, respectively, i and j, on edge (v, w). Therefore, $hp_{(iv)(jw)}$ is 1 if there is a contact, and 0, otherwise.

Variables $hb_{(iv)(jw)}$ are defined for $i \in H_e, j \in I_e - \{i\}$, $(v, w) \in E$, or $i \in H_o, j \in I_o - \{i\}$, $(v, w) \in F$, indicating whether or not there is a contact between hydrophobic side chain element i and backbone j on edge (v, w). Therefore, $hb_{(iv)(jw)}$ is set 1 if there is a contact, and 0 otherwise.

Variables $pb_{(iv)(jw)}$ are defined for $i \in P_e, j \in I_e - \{i\}$, $(v, w) \in E$, or $i \in P_o, j \in I_o - \{i\}$, $(v, w) \in F$, indicating whether or not there is a contact between hydrophilic side chain i and backbone j on edge (v, w). Therefore, $pb_{(iv)(jw)}$ is set 1 if there is a contact, and 0, otherwise.

2.3. The proposed integer programming formulation

The integer programming formulation proposed in this work, described below, is an extension of the models presented in Carr et al. [7] and Yanev et al. [29]. The main difference is that we have included variables to represent the side chains (y_{iv}) and other variables that take into account all possible types of interactions between elements ($hh_{(iv)(jw)}$, $pp_{(iv)(jw)}$, $hp_{(iv)(jw)}$, $hb_{(iv)(jw)}$ and $pb_{(iv)(jw)}$).

$$\begin{aligned} \max f = {} & \varepsilon_{bb} \sum_{(v,w)\in E} \sum_{i\in I_o} \sum_{j\in I_e-\{i-1,i+1\}} bb_{(iv)(jw)} + \varepsilon_{hh} \sum_{(v,w)\in E} \sum_{i\in H_e} \sum_{j\in H_o} hh_{(iv)(jw)} + \varepsilon_{pp} \sum_{(v,w)\in E} \sum_{i\in P_e} \sum_{j\in P_o} pp_{(iv)(jw)} \\ & + \varepsilon_{hp} \sum_{(v,w)\in E} \sum_{i\in H_e} \sum_{j\in P_o} hp_{(iv)(jw)} + \varepsilon_{hp} \sum_{(v,w)\in F} \sum_{i\in H_o} \sum_{j\in P_e} hp_{(iv)(jw)} \\ & + \varepsilon_{hb} \sum_{(v,w)\in E} \sum_{i\in H_e} \sum_{j\in I_e-(i)} hb_{(iv)(jw)} + \varepsilon_{hb} \sum_{(v,w)\in F} \sum_{i\in H_o} \sum_{j\in I_o-(i)} hb_{(iv)(jw)} \\ & + \varepsilon_{pb} \sum_{(v,w)\in E} \sum_{i\in P_e} \sum_{j\in I_e-(i)} pb_{(iv)(jw)} + \varepsilon_{pb} \sum_{(v,w)\in F} \sum_{i\in P_o} \sum_{j\in I_o-(i)} pb_{(iv)(jw)} \end{aligned} \tag{2}$$

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

subject to:

$$\sum_{v\in L_o} x_{iv} = 1 \qquad \forall i \in I_o \tag{3}$$

$$\sum_{v\in L_e} x_{iv} = 1 \qquad \forall i \in I_e \tag{4}$$

$$\sum_{v\in L_e} y_{iv} = 1 \qquad \forall i \in I_o \tag{5}$$

$$\sum_{v\in L_o} y_{iv} = 1 \qquad \forall i \in I_e \tag{6}$$

$$\sum_{i\in I_o} x_{iv} + \sum_{j\in I_e} y_{jv} \le 1 \qquad \forall v \in L_o \tag{7}$$

$$\sum_{i\in I_e} x_{iv} + \sum_{j\in I_o} y_{jv} \le 1 \qquad \forall v \in L_e \tag{8}$$

$$\sum_{w\in N(v)} x_{(i+1)w} \ge x_{iv} \qquad \forall i \in I_o - \{n\},\ v \in L_o \tag{9}$$

$$\sum_{w\in N(v)} x_{(i+1)w} \ge x_{iv} \qquad \forall i \in I_e - \{n\},\ v \in L_e \tag{10}$$

$$\sum_{w\in N(v)} y_{iw} \ge x_{iv} \qquad \forall i \in I_o,\ v \in L_o \tag{11}$$

$$\sum_{w\in N(v)} y_{iw} \ge x_{iv} \qquad \forall i \in I_e,\ v \in L_e \tag{12}$$

$$\sum_{j\in I_e - \{i-1,i+1\}} bb_{(iv)(jw)} \le x_{iv} \qquad \forall i \in I_o,\ (v,w) \in E \tag{13}$$

$$\sum_{i\in I_o} bb_{(iv)(jw)} \le x_{jw} \qquad \forall j \in I_e - \{i-1, i+1\},\ (v,w) \in E \tag{14}$$

$$\sum_{j\in H_o} hh_{(iv)(jw)} \le y_{iv} \qquad \forall i \in H_e,\ (v,w) \in E \tag{15}$$

$$\sum_{i\in H_e} hh_{(iv)(jw)} \le y_{jw} \qquad \forall j \in H_o,\ (v,w) \in E \tag{16}$$

$$\sum_{j\in P_o} pp_{(iv)(jw)} \le y_{iv} \qquad \forall i \in P_e,\ (v,w) \in E \tag{17}$$

$$\sum_{i\in P_e} pp_{(iv)(jw)} \le y_{jw} \qquad \forall j \in P_o,\ (v,w) \in E \tag{18}$$

$$\sum_{j\in P_o} hp_{(iv)(jw)} \le y_{iv} \qquad \forall i \in H_e,\ (v,w) \in E \tag{19}$$

$$\sum_{i\in H_e} hp_{(iv)(jw)} \le y_{jw} \qquad \forall j \in P_o,\ (v,w) \in E \tag{20}$$

$$\sum_{j\in P_e} hp_{(iv)(jw)} \le y_{iv} \qquad \forall i \in H_o,\ (v,w) \in F \tag{21}$$

$$\sum_{i\in H_o} hp_{(iv)(jw)} \le y_{jw} \qquad \forall j \in P_e,\ (v,w) \in F \tag{22}$$

$$\sum_{j\in I_e - \{i\}} hb_{(iv)(jw)} \le y_{iv} \qquad \forall i \in H_e,\ (v,w) \in E \tag{23}$$

$$\sum_{i\in H_e} hb_{(iv)(jw)} \le x_{jw} \qquad \forall j \in I_e - \{i\},\ (v,w) \in E \tag{24}$$

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

$$\sum_{j \in I_o - \{i\}} hb_{(iv)(jw)} \leq y_{iv} \qquad \forall i \in H_o, (v, w) \in F \tag{25}$$

$$\sum_{i \in H_o} hb_{(iv)(jw)} \leq x_{jw} \qquad \forall j \in I_o - \{i\}, (v, w) \in F \tag{26}$$

$$\sum_{j \in I_e - \{i\}} pb_{(iv)(jw)} \leq y_{iv} \qquad \forall i \in P_e, (v, w) \in E \tag{27}$$

$$\sum_{i \in P_e} pb_{(iv)(jw)} \leq x_{jw} \qquad \forall j \in I_e - \{i\}, (v, w) \in E \tag{28}$$

$$\sum_{j \in I_o - \{i\}} pb_{(iv)(jw)} \leq y_{iv} \qquad \forall i \in P_o, (v, w) \in F \tag{29}$$

$$\sum_{i \in P_o} pb_{(iv)(jw)} \leq x_{jw} \qquad \forall j \in I_o - \{i\}, (v, w) \in F \tag{30}$$

$$x_{iv}, y_{iv} \in \{0, 1\} \tag{31}$$

$$hh_{(iv)(jw)} \in \{0, 1\} \tag{32}$$

$$bb_{(iv)(jw)}, pp_{(iv)(jw)} \in \{0, 1\} \tag{33}$$

$$hp_{(iv)(jw)}, hb_{(iv)(jw)}, pb_{(iv)(jw)} \in \{0, 1\}. \tag{34}$$

The objective function (2) represents the energy considering the number of the non-local interactions. It considers the interactions between non consecutive backbone–backbone, hydrophobic side chain–hydrophobic side chain, hydrophilic side chain–hydrophilic side chain, hydrophobic side chain–hydrophilic side chain, hydrophobic side chain–backbone and hydrophilic side chain–backbone.

Constraints (3) and (4) guarantee that each backbone i is assigned to exactly one vertex v in the lattice. Constraints (5) and (6) guarantee that each side chain i is assigned to exactly one vertex v in the lattice. Constraints (7) and (8) guarantee that each vertex v in the lattice contains at most one backbone or one side chain. Constraints (9) and (10) are used to force that consecutive backbones on the string are placed on adjacent lattice points. Constraints (11) and (12) are used to force that each side chain on the string is placed on a neighbor to its correspondent backbone. Constraints (13) to (30) are used to force that each element (backbones or side chains) are placed on lattice nodes v and w if there is a contact between these elements on the edge (v, w). Constraints (31)–(34) enforce that all variables are binary.

The constraints that contain variables related to the interactions which are not considered (those with zero value in the Energy matrix (1)) should be eliminated from the model. Thus, the model can be significantly reduced, for instance, when only the interactions between hydrophobic side chains are considered. In this particular case, we have the following mathematical model where only the non-local interactions between the hydrophobic side chains are considered:

$$\max f = \varepsilon_{hh} \sum_{(v,w) \in E} \sum_{i \in H_e} \sum_{j \in H_o} hh_{(iv)(jw)} \tag{35}$$

subject to: (3)–(12), (15), (16), (31) and (32).

3. Computational experiments

For the computational experiments done in this work, we used the model (2) and the above-mentioned constraints. Also, we used $\varepsilon_{hh} = 1$ and all remaining elements of matrix (1) equal to zero, so that the model is reduced to (35).

We have used ILOG CPLEX optimization package,[1] version 12.4, to solve the integer programming problem in a high-performance computing cluster with Dual Xeon 5550 2.67 GHz. In Table 1, 25 benchmark sequences from Yue and Dill [31], [26] and Yue et al. [32] are presented. These instances were also used in a previous work based on a greedy approach [10]. The instances are divided into three groups. For the first and second groups we have used a cubic lattice with side 5. For the third group, of larger instances, we have used a cubic lattice with side 7.

In Table 2, the computational results are presented. Column "NGHPSP" shows the results obtained with the greedy algorithm presented in Galvão et al. [10], while column "CPLEX" shows the results obtained using the current model with CPLEX. In some instances it is shown a percentage number, indicating the duality gap given by CPLEX after 30 days running, when we decided to stop the process. Notice that, for some hard instances, CPLEX could not find any integer solution for some instances after processing for that time (represented by "–").

In the first group, where the lengths of the sequences range from 27 to 36 amino acids, in all but one case CPLEX was able to find the optimal solution. For the instance Dill.4 the processing was stopped when the dual gap was 68%, after 30 days

[1] IBM Corporation, Armonk, NY, USA.

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

Table 1
Benchmark instances for the 3D-HP-Side Chain.

Instance	N	Protein sequence
Dill.1	27	$HP_4H_4P(PH)_3H(HP)_2PH_2P_2H$
Dill.2	27	$HP3H_4(PH)_2HP_3HPH(HP)_2P_2HP$
Dill.3	27	$HPH_2(PPHH)_2H(HPPP)_2H_3P_2H$
Dill.4	31	$(HHP)_3H(HHHHHPP)_2H_7$
Dill.5	36	$PH(PPH)_{11}P$
Unger273d.1	27	$(PH)_3H_2P_2(HP)_2P_{10}H_2P$
Unger273d.2	27	$PH_2P_{10}H_2P_2H_2P_2HP_2HPH$
Unger273d.3	27	$H_4P5HP_5H_3P8H$
Unger273d.4	27	$H_3P_2H_4P_3(HP)_2PH_2P_2HP_3H_2$
Unger273d.5	27	$H_4P_4HPH_2P_3H_2P10$
Unger273d.6	27	$HP_6HPH_3P_2H_2P_3HP_4HPH$
Unger273d.7	27	$HP_2HPH_2P_3HP_5HPH_2(PH)_3H$
Unger273d.8	27	$HP_{11}(HP)_2P_7HPH_2$
Unger273d.9	27	$P7H_3P_3HPH2P_3HP2HP_3$
Unger273d.10	27	$P_5H(HP)_5(PHH)_2PHP_3$
S48.1	48	$HPH_2P_2H_4PH_3P_2H_2P_2HPH_3(PH)_2HP_2H_2P_3HP_8H_2$
S48.2	48	$H_4(PHH)_2H_3(PPH)_2HP_2HP_6(HPP)_2PHP_2H_2P_2H_3PH$
S48.3	48	$(PH)_2HPH_6P_2(HP)_2(PH)_2(HP)_3(PPH)_2HP_2H_2P_2(HP)_2PHP$
S48.4	48	$(PH)_2HP_2HPH_3P_2H_2PH_2P_3H_5P_2HPH_2(PH)_2P_4HP_2(HP)_2$
S48.5	48	$P_2HP_3HPH_4P2H_4(PHH)_2HP(PH)_3P_2HP_5(PHH)_2PH$
S48.6	48	$H_3P_3H(HP)_2(HHP)_3HP_7(HP)_2PHP_3HP_2H_6PH$
S48.7	48	$PHP_4HPH_3(PH)_2H_3(PHH)_2P_3(HP)_2P_2H_3(PPHH)_2P_3H$
S48.8	48	$(PHH)_2HPH_4P_2H_3P_6HPH_2P_2H(HP)_2P_2H_2(PH)_3HP_3$
S48.9	48	$(PH)_2P_4(HP)_3(PH)_2H_5P_2H_3PHP(PH)_2HP(PH)_2H_2P_4H$
S48.10	48	$PH_2P_6H_2P_3H_3PHP(PH)_2(HPP)_3H_2P_2H_7P_2H_2$

Table 2
Comparison of the maximum number of hydrophobic contacts for the benchmarks.

Instance	N	NGHPSP	Time NGHPSP (s)	CPLEX	Time CPLEX (s)	Lattice dimensions
Dill.1	27	21	24	**22**	57,593	$5 \times 5 \times 5$
Dill.2	27	19	23	**21**	157,157	$5 \times 5 \times 5$
Dill.3	27	**23**	25	**23**	339,921	$5 \times 5 \times 5$
Dill.4	31	43	36	68%	[a]	$5 \times 5 \times 5$
Dill.5	36	18	34	**20**	83,262	$5 \times 5 \times 5$
Unger273d.1	27	12	20	**13**	3,810	$5 \times 5 \times 5$
Unger273d.2	27	**13**	21	**13**	4,144	$5 \times 5 \times 5$
Unger273d.3	27	**13**	20	**13**	2,868	$5 \times 5 \times 5$
Unger273d.4	27	**22**	23	**22**	35,790	$5 \times 5 \times 5$
Unger273d.5	27	**13**	21	**13**	2,869	$5 \times 5 \times 5$
Unger273d.6	27	**15**	21	**15**	32,227	$5 \times 5 \times 5$
Unger273d.7	27	16	22	**18**	578,700	$5 \times 5 \times 5$
Unger273d.8	27	**6**	19	**6**	995	$5 \times 5 \times 5$
Unger273d.9	27	**10**	20	**10**	2,138	$5 \times 5 \times 5$
Unger273d.10	27	14	22	**15**	3,701	$5 \times 5 \times 5$
S48.1	48	36	64	–	[a]	$7 \times 7 \times 7$
S48.2	48	36	64	242%	[a]	$7 \times 7 \times 7$
S48.3	48	35	64	272%	[a]	$7 \times 7 \times 7$
S48.4	48	36	63	414%	[a]	$7 \times 7 \times 7$
S48.5	48	35	63	262%	[a]	$7 \times 7 \times 7$
S48.6	48	35	63	461%	[a]	$7 \times 7 \times 7$
S48.7	48	34	62	–	[a]	$7 \times 7 \times 7$
S48.8	48	34	63	–	[a]	$7 \times 7 \times 7$
S48.9	48	36	64	–	[a]	$7 \times 7 \times 7$
S48.10	48	36	64	671%	[a]	$7 \times 7 \times 7$

[a] indicates 30 days of processing after which the process was stopped. The optimal values found are shown in boldface.

running. In other cases the computational time ranged between 57,593 and 339,921 s. In the second group, the lengths of all sequences are 27 amino acids and, in all cases, the optimal solution was achieved. The processing time varied between 995 and 578,700 s (approximately 6.7 days). In the third group, the lengths of all sequences are 48 amino acids. For these instances, no optimal solution has been found and, again, processing was stopped after 30 days running. In six cases the gaps varied between 242% and 671%. In four cases no integer solution was found. It is important to note that, for all instances, the optimal solutions refer to the search space used, that is, the lattice dimensions. Larger or smaller lattices may lead to different results.

Fig. 3 shows the result obtained for a particular instance, Dill.1. As expected, in the native state, the formation of the hydrophobic core is evidenced. The hydrophobic amino acids (whose side chains are represented by red dots) are grouped

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

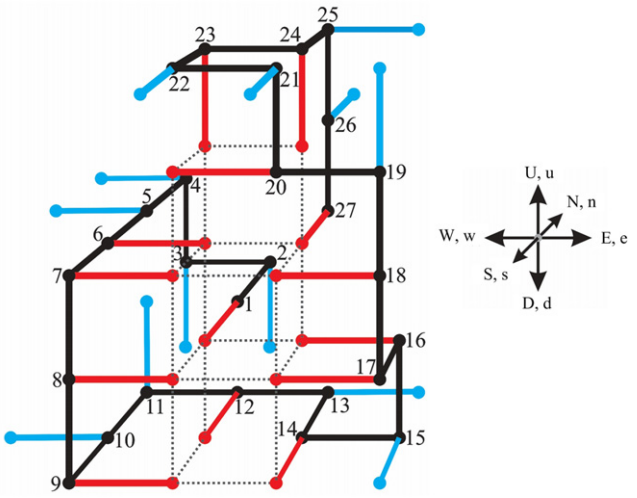

Fig. 3. Best folding obtained for sequence Dill.1. Backbones, hydrophobic and hydrophilic side chains are represented, respectively, by black, red and blue dots. Non-local hydrophobic interactions are indicated by dotted lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Best folding found for some of the benchmarks shown in Table 1.

Instance	N	Best folding found
Dill.1	27	sNdWdUwSwSeSeDeDeNwNuEsEeSsEsUwSwUwUuWwUsWsNdEdNeDnDs
Dill.2	27	wUsUeSeWnWnDnEnDeWnWuNuNeNdUnUsEsEnDsWsDdEsEnSeUnUwNn
Dill.3	27	nWwDeDeWuNnEeUeNwNdEsDsEsUsEnUuWnSwSsEuSuDwDwNwNnUuSw
Dill.5	36	eNnEuDeWnDeWsUnSeUuNnUeNeNdNeDeDsEeDeWsDeSeSuSsEsNuEeNeUwEsUuWwSeUuNwUe
Unger273d.1	27	eDdSwEnSwDnWnDnEsDeNuWsNuWwSwSuSsUuEdUuEsEuNdNdNwDwDs
Unger273d.2	27	uDsEsEeUwEuSsUsUwNuWwDwSwDwSwSwDnWnWnSwUwNnWuNuNeNuDs
Unger273d.3	27	wDwNwUwNeDeDdWuWdSuWsNdUeUeUeSdEdSdWdSsEsEeDwSuWwDwNn
Unger273d.4	27	wUwSwSuDnWnWnUnUnWwUuNwNeDwDeNnUnEsEsEnSuSwUsUsNnWdSd
Unger273d.5	27	eUeSeDeDnSuEdNdEnSeUnUnEuNuDeNwUwEuSuSuDwDdSwUsUuSwUw
Unger273d.6	27	wNwDeWwDeWwSwEuSeSuWuNuWsUnUeSeUeNnEnEdSuEsNeDeDwSdUw
Unger273d.7	27	dEuNeDsDwEsUsEuSuSeDwDeNuEuNuWuWsWwSuSeWnUuSsEnEsUnWn
Unger273d.8	27	dNwNuWuWuDsEsEsEdUuSuDsDsWnWsDnSsEnSsEnUuSdWwNnUsWnUn
Unger273d.9	27	nWuDeNnEnEsUsUuNuWsWsNnUuSeUeSeDeSuEdEeUnEeNeNwNeNwEd
Unger273d.10	27	nSsWsWsNeWdSdWsDnSeDnDdNnDnEuEsUsUwNwNuWuWsDnEsNdEdUw

in the inner part of the molecule, while the hydrophilic amino acids (side chains in blue) are in the outer part. Dotted lines indicate the 22 non-local interactions between hydrophobic side chains.

Using the mathematical model proposed in this paper, Table 3 shows the best solution found for the benchmarks of Table 2.

To describe a folding, we used an alphabet composed by twelve letters: $\{N, S, E, W, U, D, n, s, e, w, u, d\}$. Capital and minuscule letters represent, respectively, the movements of the backbones and the side chains. Therefore, movements are defined in six directions in the space, corresponding to north, south, east, west, up and down. The spatial representation of possible directions for the folding of the backbone and the side chains are shown on the right side of Fig. 3. Recall that, for a sequence of N amino acids, there are $2N - 1$ letters for representing a complete folding (backbone and side chains), since the first backbone of the sequence lies, by default, in the origin of the coordinates system.

4. Conclusions

In this paper we proposed a new integer programming model for 3D-HP protein structure prediction considering side chains. Studies of HP models with side chains are sparsely found in the literature due to its elevated complexity, when compared with simple HP models.

We use a set of benchmark instances and CPLEX to evaluate the integer programming model. It is important to note that the results obtained using CPLEX are optimal for the lattices used. If larger lattices were used (with more clearance for positioning of backbone elements and side chains) better results could, eventually, be found. However, this is unlikely, as there is a tendency (in the model we used) of hydrophobic side chains being positioned towards the inner part of the folding, thus protected from the solvent by the hydrophilic side chains. Most of the instances confirmed the consistency of the model, although, for larger instances the computational power available was not enough.

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021

The proposed mathematical model is as complete as possible, since it considers all types of interactions between the elements (backbone, hydrophilic and hydrophobic side chains). For specific studies it can be significantly simplified if few interaction are considered.

It is important to mention that, in some cases, the results also show the quality of the heuristic previously presented in Galvão et al. [10], where the optimal results were obtained for several instances and very close of the optimal in others cases.

We believe that, providing here the best solutions we found to date is important to foster further research. Making such information available, the proposed model can be improved and other algorithms may come up aiming at finding even better solutions, when possible.

Future work will go in some directions. First, the evaluation of possible relaxations of constraints as well as some strategy to reduce the number of variables may lead to the simplification of the model and, hopefully, to a faster convergence. Also, it may be interesting to investigate the theoretical bounds for the protein structure prediction problem with side chains problem, similarly to what was done by Yanev et al. [28] for regular HP models. Finally, the biological plausibility of the resulting foldings will be accessed when variants of the model will be experimented, for instance, using different values for the *Energy* matrix (1).

Acknowledgments

H.S. Lopes and L.F. Nunes would like to thank the Brazilian National Research Council – CNPq for the research grant 308803/2013-2 and the postdoctoral fellowship grant 202312/2011-9, respectively. P. Moscato would like to thank the ARC Centre of Excellence in Bioinformatics.

References

[1] N. Ahn, S. Park, Finding an upper bound for the number of contacts in hydrophobic–hydrophilic protein structure prediction model, J. Comput. Biol. 17 (2010) 647–655.
[2] O. Aichholzer, D. Bremner, E.D. Demaine, H. Meijer, V. Sacristán, M. Soss, Long proteins with unique optimal foldings in the HP model, Comput. Geom. 25 (2003) 139–159.
[3] C.M.V. Benítez, H.S. Lopes, Hierarchical parallel genetic algorithm applied to the three-dimensional HP side-chain protein folding problem, in: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, IEEE Computer Society, Piscataway, USA, 2010, pp. 2669–2676.
[4] C.M.V. Benítez, H.S. Lopes, Protein structure prediction with the 3D-HP side-chain model using a master–slave parallel genetic algorithm, Journal of the Brazilian Computer Society 16 (2010) 69–78.
[5] C.M.V. Benítez, R.S. Parpinelli, H.S. Lopes, Parallelism, hybridism and coevolution in a multi-level ABC-GA approach for the protein structure prediction problem, Concurr. Comput. 24 (2011) 635–646.
[6] S. Bromberg, K.A. Dill, Side chain entropy and packing in protein, Protein Sci. 3 (1994) 997–1009.
[7] B. Carr, W.E. Hart, A. Newman, Discrete Optimization Models for Protein Folding, Tech Report SAND2002, Sandia National Laboratories, Livermore, USA, 2002.
[8] K.A. Dill, Theory for the folding and stability of globular proteins, Biochemistry 24 (1985) 1501–1509.
[9] A.S. Fraenkel, Complexity of protein folding, Bull. Math. Biol. 55 (1993) 1199–1210.
[10] L.C. Galvão, L.F. Nunes, H.S. Lopes, P. Moscato, A new greedy heuristic for 3DHP protein structure prediction with side chain, in: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine Workshops, IEEE Computer Society, Los Alamitos, USA, 2012, pp. 77–81. http://dx.doi.org/10.1109/BIBMW.2012.6470381.
[11] W. Hart, S. Istrail, Lattice and off-lattice side chain models of protein folding, J. Comput. Biol. 4 (1997) 241–259.
[12] C.L. Kingsford, B. Chazelle, M. Singh, Solving and analyzing side-chain positioning problems using linear and integer programming, Bioinformatics 21 (2005) 1028–1036.
[13] K. Lau, K. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules 22 (1989) 3986–3997.
[14] M.S. Li, D.K. Klimov, D. Thirumalai, Folding in lattice models with side chains, Comput. Phys. Comm. 147 (2002) 625–628.
[15] J. Liu, G. Li, J. Yu, Y. Yao, Heuristic energy landscape paving for protein folding problem in the three-dimensional HP lattice model, Comput. Biol. Chem. 38 (2012) 17–26.
[16] H.S. Lopes, Evolutionary algorithms for the protein folding problem: a review and current trends, in: T.G. Smolinski, M.M. Milanova, A.-E. Hassanien (Eds.), in: Applications of Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications, vol. 151, Springer-Verlag, Heidelberg, Germany, 2008, pp. 297–315.
[17] S. Mandal, N.D. Jana, Protein structure prediction using 2D HP lattice model based on integer programming approach, in: Proceedings of International Congress on Informatics, Enviroment, Energy and Applications, IACSIT Press, Singapore, 2012, pp. 171–175.
[18] R.S. Parpinelli, C. M. V. Benítez, J. Cordeiro, H.S. Lopes, Performance analysis of swarm intelligence algorithms for the 3D-AB off-lattice protein folding problem, J. Mult.-Valued Logic Soft Comput. 22 (2014) 267–286.
[19] E.I. Shakhnovich, A.M. Gutin, Engineering of stable and fast-folding sequences of model proteins, Proc. Natl. Acad. Sci. USA 90 (1993) 7195–7199.
[20] F.H. Stillinger, T. Head-Gordon, Collective aspects of protein folding illustrated by a toy model, Phys. Rev. E 52 (1995) 2872–2877.
[21] F.H. Stillinger, T. Head-Gordon, C.L. Hirshfeld, Toy model for protein folding, Phys. Rev. E 48 (1993) 1469–1477.
[22] C. Thachuk, A. Shmygelska, H.H. Hoos, A replica exchange monte carlo algorithm for protein folding in the HP model, BMC Bioinformatics 8 (2007) 1–20.
[23] P.D. Thomas, K.A. Dill, Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation, Protein Sci. 2 (1993) 2050–2065.
[24] L. Toma, S. Toma, Contact interactions method: a new algorithm for protein folding simulations, Protein Sci. 5 (1996) 147–153.
[25] M. Türkay, F. Üney, Özlem Yilmaz, Prediction of folding type of proteins using mixed-integer linear programming, Comput. Aided Chem. Eng. 20 (2005) 523–528.
[26] R. Unger, J. Moult, Finding the lowest free energy conformation of a protein is a NP-hard problem: proof and implications, Bull. Math. Biol. 55 (1993) 1183–1198.
[27] R. Unger, J. Moult, Genetic algorithms for protein folding simulations, J. Mol. Biol. 231 (1993) 75–81.
[28] N. Yanev, P. Milanov, I. Mirchev, Integer programming approach to HP folding, Serdica J. Comput. 5 (2011) 359–366.
[29] N. Yanev, P. Milanov, I. Trenchev, I. Mirchev, Integer programming approaches to HP folding, in: Proceedings of VIII European Workshop in Drug Design, University of Siena, Certosa di Pontignano, Italy, 2011.
[30] H. Yoon, Optimization approaches to protein folding (Ph.D. Thesis), School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, USA, 2006.
[31] K. Yue, K. Dill, Sequence–structure relationships in proteins and copolymers, Phys. Rev. E 48 (1993) 2267–2278.
[32] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, K.A. Dill, A test of lattice protein folding algorithms, Proc. Natl. Acad. Sci. USA 92 (1995) 325–329.
[33] X. Zhang, T. Li, Improved particle swarm optimization algorithm for 2d protein folding prediction, in: Proceedings of 1st International Conference on Bioinformatics and Biomedical Engineering, IEEE Press, Piscataway, USA, 2007, pp. 53–56.

Please cite this article in press as: L.F. Nunes, et al., An integer programming model for protein structure prediction using the 3D-HP side chain model, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.06.021