Computational Biology and Chemistry 84 (2020) 107192

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A benchmark of optimally folded protein structures using integer programming and the 3D-HP-SC model

Leandro Takeshi Hattori*, Matheus Gutoski, César Manuel Vargas Benítez, Luiz Fernando Nunes, Heitor Silvério Lopes

Bioinformatics and Computational Intelligence Laboratory, Federal University of Technology Paraná (UTFPR), Av. 7 de Setembro, 3165, 80230-901 Curitiba (PR), Brazil

ARTICLE INFO

Keywords:
Biological sequences
Hydrophobic-polar model
Integer programming
Protein structure problem

ABSTRACT

The Protein Structure Prediction (PSP) problem comprises, among other issues, forecasting the three-dimensional native structure of proteins using only their primary structure information. Most computational studies in this area use synthetic data instead of real biological data. However, the closer to the real-world, the more the impact of results and their applicability. This work presents 17 real protein sequences extracted from the Protein Data Bank for a benchmark to the PSP problem using the tri-dimensional Hydrophobic-Polar with Side-Chains model (3D-HP-SC). The native structure of these proteins was found by maximizing the number of hydrophobic contacts between the side-chains of amino acids. The problem was treated as an optimization problem and solved by means of an Integer Programming approach. Although the method optimally solves the problem, the processing time has an exponential trend. Therefore, due to computational limitations, the method is a proof-of-concept and it is not applicable to large sequences. For unknown sequences, an upper bound of the number of hydrophobic contacts (using this model) can be found, due to a linear relationship with the number of hydrophobic residues. The comparison between the predicted and the biological structures showed that the highest similarity between them was found with distance thresholds around 5.2–8.2 Å. Both the dataset and the programs developed will be freely available to foster further research in the area.

1. Introduction

The Protein Structure Prediction (PSP) problem is an active field of research in Bioinformatics. One of the many issues studied in this field comprises forecasting the three-dimensional native structure of proteins using only their primary structure information (Dill and MacCallum, 2012).

Proteins have key functions in the living cell, such as transmembrane receptors (Vinogradova et al., 2000; Chua et al., 2011), storage (Reinhard et al., 1999), cellular processes (Mortishire-Smith et al., 1995), and signaling (Grace et al., 2007). However, when some proteins fail to fold into their functional form, they are associated to some human diseases, such as Alzheimer (Benaki et al., 2005, 2006). They are also associated with viruses, like the Hepatitis C Virus (HCV) (Gouttenoire et al., 2009), Human Immunodeficiency Virus (HIV) (Amodeo et al., 2017), Bovine viral diarrhea virus (Sapay et al., 2006), and bacteria, e.g., *Escherichia coli* (Duarte et al., 2007) and *Acholeplasma laidlawii* (Lind et al., 2019).

Experimental techniques are not trivial for unveiling protein structures. This is evidenced by the number of protein sequences that have been discovered along time (more than 160 million sequences[1]) compared to the number of known protein structures (155,618 structures[2]). This huge gap shows that it is still quite important to invest efforts in methods for unveiling protein structures. In this sense, computational approaches in Bioinformatics have been explored in the literature (Dorn et al., 2014a; Ovchinnikov et al., 2018) for the PSP problem.

According to the Levinthal's paradox, a protein can assume an astronomic number of possible conformations (Karplus, 1997). Therefore, if a protein was supposed to sequentially try all possible conformations sequentially until finding its native form, this possibly would take untold time. However, the paradox is that most proteins fold spontaneously on less than a millisecond time scale. To the computational point of view, finding the native structure of a protein is an open challenge. Atkins and Hart (1999), using the simplest HP model, demonstrated that an algorithm for exhaustive search of all conformations would take time that grows exponentially, as the size of the protein

* Corresponding author.
E-mail addresses: lthattori@gmail.com (L.T. Hattori), nunes@utfpr.edu.br (L.F. Nunes), hslopes@utfpr.edu.br (H.S. Lopes).

[1] As in September, 10th 2019 at the site https://www.uniprot.org/.

[2] As in September, 10th 2019 at the site https://www.rcsb.org/.

https://doi.org/10.1016/j.compbiolchem.2019.107192
Received 6 July 2019; Received in revised form 9 December 2019; Accepted 10 December 2019
Available online 23 December 2019
1476-9271/

grows linearly. Therefore the protein structure prediction problem is reputed as NP-complete, that is, it cannot be solved in polynomial time. To overcome such a computational complexity, several approaches have been proposed, such as methods based on mathematical programming (Carr et al., 2003; Yanev et al., 2011, 2017) and those based on heuristic approaches (Parpinelli et al., 2014; Li et al., 2015, 2017; Kaushik and Sahi, 2017). The Integer Programming (IP) optimization is a mathematical method that can present the optimal result given a set of constraints and an objective function (Nunes et al., 2016). This approach has been poorly explored in the literature when applied to the PSP problem (Yanev et al., 2011, 2017). However, since it produces optimal results, it can be very useful for establishing the ground truth for comparison with other heuristic approaches.

Due to the computational power required for the PSP problem, simpler but non-simplistic models named Coarse-Grained (CG) have been explored in the literature (Kmiecik et al., 2019). These models can represent many biological behaviors at a meso-scale (Tozzini, 2005), such as the hydrophobic core formation, presented in many protein domains (Kalinowska et al., 2017), and the protein aggregation process, which is related to proteinopathies (Frigori, 2017). Among the many variants of CG models, the 3D-HP-SC lattice (Pierri et al., 2008) is a representation where the conformation of the protein is contained in a tri-dimensional (3D) lattice, and each amino acid is represented by two beads, a backbone and a Side Chain (SC). The backbone elements represent the C_α and the binding structure between the amino acids sequence and the SC represents their radicals, which can be either hydrophobic (H) or polar (P). This model is a variation of the traditional 3D-HP lattice approach, with the addition of the SC, making it a more detailed model, with more degrees of freedom and more similarity to the biological form, however at the cost of increased computational complexity (Benítez and Lopes, 2010; Dubey et al., 2018).

The objective of this work is to propose a baseline result for the PSP problem, based on real amino acids sequences extracted from the Protein Data Bank (PDB). This baseline is intended to foster researchers to create or improve algorithms for PSP with CG models. The IP method was used to find the native 3D-HP-SC structure. The constraints and the objective function of the IP were previously presented in Nunes et al. (2016). All data and software used in this work are freely distributed, encouraging other researchers to compare results with those achieved in this work. In addition, we also present a comparison between the predicted and the real biological structures.

This paper is organized as follows: Section 2 introduces some background about the Protein Structure Prediction problem. Section 3 presents the method for extracting and converting the biological sequences used in this work. Section 4 describes the IP model. Section 5 presents a comparison of the structures predicted by our method and the biological structures from the PDB. Section 6 shows the results obtained and the corresponding analysis. Finally, in Section 7 the conclusions and future directions are pointed out.

2. Protein structure prediction

Lattice models can help to predict a protein structure in a less computationally expensive way, than all-atom models that have to consider a huge number of possible conformations, given the high degree of freedom of the model. Therefore, depending on the size of the protein, its computational simulation turns out to be prohibitive. Hence, several simplified models, also called Coarse-Grained (CG), have been studied for the PSP problem in the past years (Blaszczyk et al., 2019). Many CG models, either continuous or discrete, have been explored in the literature, for instance: the Hydrophobic-Polar model (Lau and Dill, 1989), Lattice Polymer Embedding (Unger and Moult, 1993), Charge Graph Embedding (Fraenkel, 1993) and Contact Map model (Wang and Xu, 2013), among others (Lopes, 2008; Benítez, 2015). Among these models, the HP is a representation that has become widespread, since it considers the hydrophobicity of the protein sequence, which is one of the main factors driving the protein folding process (Mirsky and Pauling, 1936; Ben-Naim, 1994; Brylinski et al., 2006; Baldwin and Rose, 2016; Kalinowska et al., 2017). Subsequently, many variants of the HP model have been explored, such as 2D-HP, 3D-HP, and more recently 3D-HP-SC (Benítez and Lopes, 2010, 2010b; Dorn et al., 2014b; Dubey et al., 2018).

Mathematical approaches have also been proposed to tackle the PSP problem with CG models, for instance: Mann et al. (2009), Yanev et al. (2011), Wang and Xu (2013). In Mandal and Jana (2012), a two-dimensional (2D) HP model based on IP was used. Similar IP approaches were presented by Carr et al. (2003), Yanev et al. (2011) and Yanev et al. (2017). In general, using synthetic sequences, the IP approaches presented better results when compared with methods based only on meta-heuristics.

A step forward was proposed by Nunes et al. (2016), who used the IP method for solving the PSP with a more complex representation model, the 3D-HP with Side Chain (3D-HP-SC). Comparing the results obtained in that work with those obtained using a Parallel Genetic Algorithm (pGA) (Benítez and Lopes, 2010b), the IP method performed better results for most synthetic proteins. Also, IP was run only once per protein sequence, since it is deterministic, and needed a computational architecture with multiple threads. Despite of the advantages given by the IP approach, the meta-heuristics are still widespread in the literature, mainly using evolutionary computation methods (Bošković and Brest, 2016; Li et al., 2015, 2017; Kaushik and Sahi, 2017).

There are several protein databases in the literature, as shown in Table 4. However, only few of these databases contain structural information, such as the PDB. Along time, as the PDB has grown and become more popular, several works have been made feasible, and such information has been used for the computational methods, for instance: Yu et al. (2004), Pierri et al. (2008), Vilar et al. (2008), Hoque et al. (2009), Nardelli et al. (2013).

In this work, we use the IP approach and biological sequences the method, differently from other works that use synthetic sequences. It is important to notice that the use of mathematical methods in general, with the 3D-HP-SC model has been poorly explored in the literature, and, to date, no work was yet found that uses biological data to validate the IP results. However, this endeavour is much more challenging, when compared to the lattice models without side chain. In order to show that the 3D-HP-SC model is more appealing than other simplistic lattice models, a comparison between the predicted structure and the real biological structure was done, following the procedure suggested by Pierri et al. (2008).

3. Protein sequence dataset

The proteins selected for the benchmark were extracted from the PDB provided they met the two following requirements:

a. First, the number of amino acids (AA) has to be less than 30 AAs to be computationally feasible, according to Nunes et al. (2016).
b. Second, only proteins containing the 20 proteinogenic AAs (standards encoded in the genetic sequence) were considered so that a translation to the HP alphabet would be possible (Alberts, 2002).

In this work, we used small proteins, with less than 100 AAs, due to the high computational power required by the IP approach Nunes et al. (2016). Despite the small size of the proteins, Su et al. (2013) pointed out they can play important biological functions in both for macro (Benaki et al., 2005, 2006) and micro (Duarte et al., 2007; Lind et al., 2019) organisms.

Table 1 shows the information of the 17 protein sequences selected for the benchmark, including the PDB identification (PDB ID), the number of AAs in the sequence, the HP sequence, the organism to which it belongs, the secondary structures contained in the conformation, the protein classification according to the PDB file, and the

Table 1
Protein sequences extracted from the PDB and used in the experiments.

PDB ID	# AA	HP sequence	Organism	SS.	PDB class	Reference
1M23	13	$PHP_2H_2PHP_2H_2P$	*Oryctolagus cuniculus*	$\alpha - h$	MEMBRANE PROTEIN	Reinhard et al. (1999)
1JBL	14	$(HP)_2P_2H_7P$	*Helianthus annuus*	$\beta - s$	PROTEIN BINDING	Korsinczky et al. (2001)
2B0Y	17	$P_4HP_2H_2P_2H_3P_3$	*Homo sapiens*	$\alpha - h$	CELL ADHESION	Grace et al. (2007)
1DPQ	20	$PH_4P_4H_3P_6HP$	*Homo sapiens*	$\alpha - h$	CELL ADHESION	Vinogradova et al. (2000)
1Z2T	23	$(PH)_2HPH_5P_3HP_2HP_2HPH$	*Acholeplasma laidlawii*	$\alpha - h$	BINDING PROTEIN	Lind et al. (2019)
2GD3	24	$H_3PH_2PH_5(PH)_3H_2P_3H$	*Homo sapiens*	$\alpha - h$	UNKNOWN FUNCTION	Benaki et al. (2006)
1Y32	24	$H_3PH_2PH_5P_3HPH_3P_3H$	*Homo sapiens*	$\alpha - h$	UNKNOWN FUNCTION	Benaki et al. (2005)
2LKE	24	$PH_4P_6H_2P_2H_6PHP$	*Homo sapiens*	$\alpha - h$	CELL ADHESION	Chua et al. (2011)
2NVJ	25	$PH_3PH_2P_3HP_2HPH_4PH_2PHP$	*Saccharomyces cerevisiae*	$\alpha - h$	HYDROLASE	Duarte et al. (2007)
1PLP	25	$HPPHP_2HP_3H_2P_2HP_2HPH_2P_2H_2$	*Homo sapiens*	$\alpha - h$	MEMBRANE PROTEIN	Mortishire-Smith et al. (1995)
1HZ3	26	$(HP)_2(PH)_2HPH_6PH_4PHP_3$	*Homo sapiens*	$\beta - s$	BINDING PROTEIN	Zhang et al. (2000)
2IWJ	27	$PH_2P_3H_3PH_2P_2HPH_3PH_2P_3H_2$	*Human immunodeficiency virus 2*	$\alpha - h$	TRANSFERASE	Amodeo et al. (2017)
1FSD	28	$P_4(HP)_3PHP_5HP_2H_2P_2(HP)_2$	synthetic	$\alpha - h$	NOVEL SEQUENCE	Dahiyat and Mayo (1997)
1NMJ	28	$(HP)_2HP_3H_2PHP_4H_5P_2H_2P_3$	*Rattus norvegicus*	$\alpha - h$	MEMBRANE PROTEIN	Huang et al. (2004)
2AJJ	28	$PHP_2H_2PH_2P_2HP_3HP_2H_2P_2H_6$	*Bovine viral diarrhea virus 1*	$\alpha - h$	MEMBRANE PROTEIN	Sapay et al. (2006)
1PSV	28	$PHP_2(HP)_3PHP_5HP_2H_2P_2(HP)_2$	synthetic	$\alpha - h$	DESIGNED PEPTIDE	Dahiyat et al. (1997)
2JXF	30	$P_3HP_2HPH_4P_2H_2PH_2PH_2P_2H_4P_2$	*Hepatitis C virus*	$\alpha - h$	MEMBRANE PROTEIN	Gouttenoire et al. (2009)

reference in the literature. Among the selected proteins, seven are found in humans, two in other mammals, one in a plant, five in microorganisms and two were synthetically designed.

The conversion of a protein sequence to the Hydrophobic-Polar model is shown in Algorithm 1. It was implemented in the Python programming language together with the Biopython[3] framework. Starting from the PDB ID, the program downloads the PDB file and the corresponding AA sequence is obtained. Next, to convert the AA sequence to HP sequence, an auxiliary file with the hydrophobicity conversion table is used. The mapping of the 20 amino acid types to either H or P uses the hydrophobicity scale proposed byAlberts (2002), and shown in Table 2. This same mapping was used in a previous work (Benítez, 2015). Next, the HP sequence information is saved in a file as a binary string, where '1' and '0' represents hydrophobic and polar AAs, respectively.

Algorithm 1. Protein sequence conversion procedure.

```
Input PDB ID
Download PDB File
Read PDB File
for i = 0: N do
Extract Amino Acid AA_i ∈ SEQRES
Add AA_i to Sequence[i]
end for
Read HP Classification File
for i = 0: N do
if Sequence[i] = = 'H' then
HP_S equence[i] ⟵ 'H'
else
HP_S equence[i] ⟵ 'P'
end if
end for
Save HP_S equence
```

4. Integer programming approach for protein folding

In this section, the representation of the 3D-HP-SC model in a lattice is described, followed by the definition of the variables, objective function and the constraints of the IP method.

4.1. Lattice representation

In the 3D-HP-SC model, each amino acid is represented by two elements: the backbone and a side-chain. The protein is contained in a fixed-size cubic lattice, and all edges between vertices are unitary

[3] http://biopython.org/.

Table 2
Hydrophobicity conversion table by Alberts (2002).

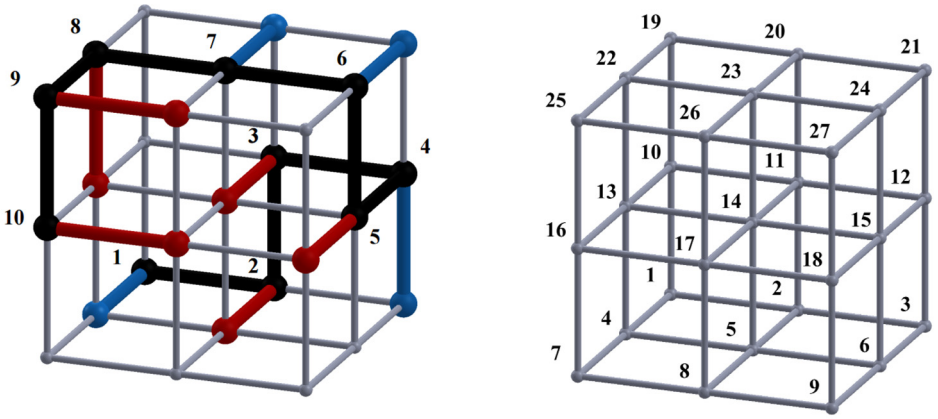
Amino acid	Hydrophobicity classification	Amino acid	Hydrophobicity classification
ALA	H	MET	H
CYS	H	ASN	P
ASP	P	PRO	H
GLU	P	GLN	P
PHE	H	ARG	P
GLY	H	SER	P
HIS	P	THR	P
ILE	H	VAL	H
LYS	P	TRP	H
LEU	H	TYR	P

(Pierri et al., 2008). Backbone and side-chain allocation must meet the following constraints:

a. All elements must respect the cubic lattice frontier;
b. Elements are allocated at the vertices of the lattice, and they cannot overlap each other;
c. The protein chain cannot be broken, so there is a neighborhood dependence between backbone–side-chain, as well as and backbone–backbone.

For instance, Fig. 1(a) shows the backbone element 4 positioned at the lattice index 12 (refer to Fig. 1(b) for the numbering). Thus, the allocation indices of the corresponding side-chain element and the previous and following backbone elements should respect the neighborhood of position 12. Thus, they should be allocated at lattice indices 3, 11, 15 or 21, provided they are not used by other elements. These neighborhood indices in the lattice are called the feasible edges of the lattice (Nunes et al., 2016). Consequently, the next and the previous backbones (3 and 5) and the side-chain of backbone 4 need to be allocated in one of the neighborhood position lists and cannot overlap each other.

The IP method optimizes the protein structure by maximizing the number of (hh) contacts, the counting of pairs of nonconsecutive hydrophobic side-chains are in contact with each other. A hh contact exists when a hydrophobic side-chain has another hydrophobic side-chain in its neighborhood lattice. Such approach has been used in several investigations (Dubey et al., 2018; Nunes et al., 2016) guided by the clue that the interactions between hydrophobic amino acids play the most important role in the protein folding process in several domains, comprising the secondary and tertiary conformations (Kalinowska et al., 2017).



(a) Sample of a 3D-HP-SC protein model with amino acid number.

(b) Sample of a 3D lattice 3x3x3 with lattice position number.

Fig. 1. Tri-dimensional Lattice with $3 \times 3 \times 3$ dimension. The black color represents the backbone, blue and red represent respectively the hydrophobic and polar side-chains, and grey the lattice structure. Based on Nunes et al. (2016).

4.2. Parameters and variables

The proposed IP method for maximizing the number of hh contacts of a protein structure in a lattice has several parameters, described here.

- S is a binary vector that describes the hydrophobicity of the side-chains of all the n of amino acids of the protein. Hydrophobic side-chains are represented by zeros, while polar side-chains are ones.
- $I = \{1, ..., n\}$ specifies the index set of S. This parameter is divided into two subsets, I_e, the subset of even indices, and I_o, the subset with odd indices. The subset I_e, in turn, is divided again into H_e and P_e, which represent, respectively, the hydrophobic and polar amino acids with even indices in the subset. Conversely, the same applies to I_o, which is divided into H_o and P_o, representing hydrophobic and polar with odd indices in the subset.
- $L = \{1, ..., m\}$ is the lattice index, and m is the number of vertices of the lattice. Similarly to I, L is divided into two subsets: one for even indices (L_e) and other for odd indices (L_o).

Another parameter is the set that stores the neighborhood index for each vertex v of the lattice, which is represented by $N(v)$. Hence, the neighborhood $N(v)$ of of vertex v is determined as follows $N(v) = \{ \forall\, t \in L \mid d(v, t) = 1\}$. Where $d(v, t)$ is the Euclidean Distance between vertices t and v.

Eq. (1) shows the binary variable that represents the backbone (x) and side-chain (y) positions in the lattice.

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

where $i \in I_o$ and $v \in L_o$ or $i \in I_e$ and $v \in L_e$, which intended where the backbone/side chain i is allocated at the vertex v.

Eq. (2) show the binary variable that represents where contacts between hydrophobic side chains are found.

$$\mathrm{hh}_{(\mathrm{iv})(\mathrm{jw})} \in \{0, 1\} \tag{2}$$

where $i \in H_e$, $j \in H_o$, $(v, w) \in E$, this indicating where there are hydrophobic contacts between the side-chain elements i and j, on the edge (v, w).

4.3. Objective function and constraints

As mentioned before, the objective function aims at maximizing the number of contacts between hydrophobic side-chains (hh), according to Eq. (3), and the following equations represent the model constraints.

$$(\mathrm{MAX})\mathrm{z} = \sum_{(v,w)\in E} \sum_{i\in H_e} \sum_{j\in H_o} \mathrm{hh}_{(\mathrm{iv})(\mathrm{jw})} \tag{3}$$

Constraints (4 and 5) guarantee that each backbone is assigned to exactly one vertex in the lattice:

$$\sum_{v\in L_o} x_{\mathrm{iv}} = 1 \qquad \forall\, i \in I_o \tag{4}$$

$$\sum_{v\in L_e} x_{\mathrm{iv}} = 1 \qquad \forall\, i \in I_e \tag{5}$$

Eqs. (6) and (7) ensure that the side-chain will be at least in one lattice position:

$$\sum_{v\in L_e} y_{\mathrm{iv}} = 1 \qquad \forall\, i \in I_o \tag{6}$$

$$\sum_{v\in L_o} y_{\mathrm{iv}} = 1 \qquad \forall\, i \in I_e \tag{7}$$

Eqs. (8) and (9) ensure that even and odd elements do not overlap each other:

$$\sum_{i\in I_o} x_{\mathrm{iv}} + \sum_{j\in I_e} y_{\mathrm{jv}} \leqslant 1 \qquad \forall\, v \in L_o \tag{8}$$

$$\sum_{i\in I_e} x_{\mathrm{iv}} + \sum_{j\in I_o} y_{\mathrm{jv}} \leqslant 1 \qquad \forall\, v \in L_e \tag{9}$$

Eqs. (10) and (11) ensure that the backbone sequence dependence will respect the neighborhood position in the lattice:

$$\sum_{w\in N(v)} x_{(i+1)w} \geqslant x_{\mathrm{iv}} \qquad \forall\, i \in I_o - \{n\},\, v \in L_o \tag{10}$$

$$\sum_{w\in N(v)} x_{(i+1)w} \geqslant x_{\mathrm{iv}} \qquad \forall\, i \in I_e - \{n\},\, v \in L_e \tag{11}$$

Eqs. (12) and (13) ensure that backbone side-chain dependence will respect the neighborhood position in the lattice:

$$\sum_{w\in N(v)} y_{\mathrm{iw}} \geqslant x_{\mathrm{iv}} \qquad \forall\, i \in I_o,\, v \in L_o \tag{12}$$

$$\sum_{w\in N(v)} y_{\mathrm{iw}} \geqslant x_{\mathrm{iv}} \qquad \forall\, i \in I_e,\, v \in L_e \tag{13}$$

Eqs. (14) and (15) ensure that values equal to 1 can only be obtained where there are hh contacts:

$$\sum_{j \in H_o} \text{hh}_{(\text{iv})(\text{jw})} \leqslant y_{\text{iv}} \quad \forall\, i \in H_e, (v, w) \in E \tag{14}$$

$$\sum_{i \in H_e} \text{hh}_{(\text{iv})(\text{jw})} \leqslant y_{\text{iv}} \quad \forall\, i \in H_o, (v, w) \in E \tag{15}$$

5. Comparison with biological structures

In this Section, we aim at comparing the structure of the crystallized proteins of PDB with the corresponding structure predicted by the IP approach. This is accomplished indirectly, by means of the number of the hydrophobic contacts of both structures.

For this comparison, the membrane proteins shown in Table 1 were excluded, since, we use soluble proteins that follow the hydrophobic driving force in the water. Among the 12 remaining sequences of that Table, we selected the 10 containing only alpha-helices ($\alpha - h$) as secondary structures, since they are more likely to form hh contacts than other protein secondary structures.

Following the method proposed by Pierri et al. (2008), we extracted the Cartesian coordinates of the protein structure from PDB files and, from them, the coordinates of the $C\alpha$ of each amino acid. Then, the Euclidean distances between hydrophobic amino acids was computed and, using a variable threshold, the number of hh contacts was obtained. The distance thresholds were between 3.8 and 9.5 Å, as suggested by Onofrio et al. (2014). Finally, the number of hh was compared to those obtained by the IP model, and a score is computed with Eq. (16).

$$\text{score} = \text{hh}_{\text{PDB}} - \text{hh}_{\text{IP}}, \tag{16}$$

where hh_{PDB} and hh_{IP} are the number of hydrophobic contacts obtained from the real protein found in the PDB, and that predicted by the IP method, respectively.

For evaluating the matching between the number of hh contacts of the real conformation (obtained as shown above) and the proposed IP method, a pairwise comparison was done and the following values were calculated:

- True Positive (TP): the number of hh contacts that appear in both, the structure predicted by the IP method and the PDB structure;
- False Positive (FP): the number of hh contacts that appear in the structure predicted by the IP method but are not present in the PDB structure;
- False Negative (FN): the number of hh contacts that do not appear in the structure predicted by the IP method but are present in the PDB structure;

Using TP, FP and FN, the overall similarity between protein structures generated by the IP method and the real structure was accessed by computing Precision and Recall Eqs. (17) and (18):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

6. Experiments and results

The Integer Programming model was developed using Java language with CPLEX Solver 12.5.[4] All experiments performed in this work were run in a computer with an Intel Core i7-4790 processor at 3.60 GHz with 32 GB of RAM and Ubuntu 14.04 LTS operational system.

The proposed model was used to create a benchmark of 17 protein sequences (see Table 1) which were optimally folded according to the lattice restrictions mentioned before.

For each protein sequence, Table 3 shows the maximum number of the hydrophobic – side-chain contacts (hh), computational time, and the folding sequence found by the method. The coordinates and movements use six letters, uppercase, and lowercase, which represents the spatial movement in the lattice of the side-chain and backbone, as follows:

- Uppercase letters represent the movements of the next backbone element relative to the previous one;
- Lowercase letters represent the movement of the side-chain relative to its backbone;
- There are six neighborhood positions, corresponding to east (E, e), west (W, w), north (N, n), south (S, s), up (U, u) and down (D, d).

Fig. 2 shows samples of the protein structures found, where it is possible to observe the presence of the hydrophobic core, as mentioned before in Section 4.

Table 3 shows the predicted structures of real-world protein sequences, as well as the number of hydrophobic side-chain contacts (hh). It is noticed, however, that the optimal solutions refer to the model, and search space used (including constraints, objective function, variables, and lattice size). That Table also shows the processing time, which ranged between 125 s and 1303176 s (~15 days). As supposed, experiments indicate that processing time increases as the sequence size grows, with a clear exponential trend concerning the number of hydrophobic side-chains (see Fig. 3).

It was also expected a linear relationship between the number of hydrophobic side-chain contacts hh, and the number of hydrophobic residues in the sequence (H). Our experiments have clearly confirmed this relationship (see Fig. 4). The interpolated equation can be used to compute an upper bound for hh, for unknown instances.

As mentioned in Section 5, we proposed a comparison between the structures predicted by the IP method with those of PDB, in terms of hh contacts. Since the actual definition of a contact in the real-world requires a previous definition of a threshold between elements, Fig. 5 shows the score (Eq. (16)) results for each distance threshold. We observe that for short-distance thresholds, around 3.8 Å, the number of hh_{PDB} was smaller than those predicted by IP. On the other hand, for long-distance thresholds, around 9.5 Å, many hh_{PDB} interactions appear, when compared to the hh_{IP}. In the middle range-thresholds, around 5.3–8.2 Å, is the region where the structure predicted by the IP model achieved most similarity with the crystallized structure of the protein, regarding the number of hydrophobic contacts.

Considering this above-mentioned range of highest hh similarity, a deeper analysis was carried out using Eqs. (17) and (18), and results are shown in Fig. 6. In this figure, it is observed that as the threshold increases, the *precision* also increases, but the *recall* decreases. This behavior was expected since, as mentioned before, the higher the threshold, the larger the hh_{PDB}.

A further analysis revealed a convergence point of *precision* and *recall*, that takes place around a threshold of 6.7 Å, suggesting a good trade-off to be used in further experiments.

7. Conclusions

This paper presents a benchmark of biological protein sequences obtained from the PDB and converted to a coarse-grained model with side-chain (3D-HP-SC). For each protein, the folded structure with the maximum hydrophobic side-chain contacts was found by using an Integer Programming (IP) method.

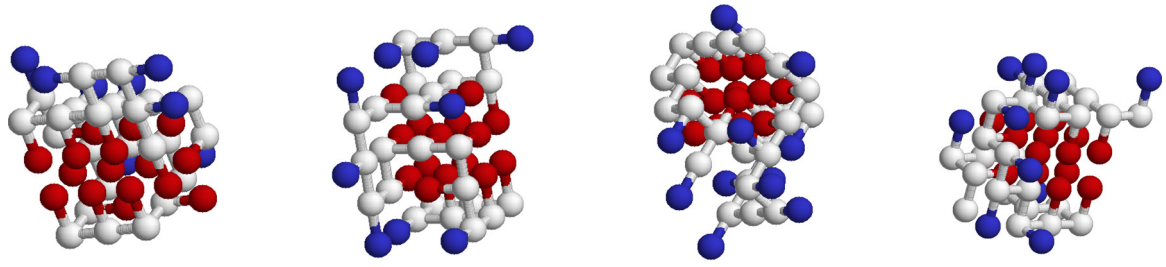
[4] IBM Corporation, Armonk, NY, USA.

Table 3
Predicted structure obtained for 17 real and synthetic protein structure, with the number of hydrophobic contacts (hh_{IP}), number of hydrophobic amino acids in the sequence ($\#H$), and processing time using the IP approach.

PDB ID	# AA	hh_IP	# H	Time (s)	Result
1M23	13	6	6	125	eUwSeSuWnDnWwNuDsEsNuWuN
1JBL	14	10	9	1002	uNwDsDsEsEeUsWsUsUsEsDsEsUu
2B0Y	17	7	6	286	dWdSdSsUeNnEnDsEsUeUnSdSdWdNuNnW
1DPQ	20	12	8	1081	dNeNeDeSeDdEdEnEsUwUwNwEsDwDnWdNwUeWsNd
1Z2T	23	18	12	7471	dSuWwSuEsEnUwSwUwNwNwNnWnDeWsUnWnSeSwDwSeUwE
2GD3	24	24	16	215725	nDnWnWsUnEnUsNeNdWdWdSeSsDnDsNeDdNuEuEuDwEeUeUu
1Y32	24	23	15	198232	sDsDsEeUsUsEeSdSwDwDwNwEnSeSsWwUuWuWnUnSuDeDeNn
2LKE	24	21	13	114041	dNwNwDwSwDnDdSuSsWsUwUnUnWsWwNeNeNeDeSeDeSdEuDd
2NVJ	25	22	14	99448	dEuNuNuEnSuSuEdNdEsUwSsUuWdUsWdWdNdEdNwDdEdEdSwU
1PLP	25	17	11	33281	dSsEsNdEsDsDwSdSsUuWnWnWwDwEnEsDeNuNdWuWuWdUsUeS
1HZ3	26	25	15	754334	eUsNwDwNwUeUuWsUsWsDsWsDnDnSuEuSwSwEeUnWnWnUnEnEnUn
2IWJ	27	23	15	144882	sNwNwUuSuSsWnDwDwEsNwNwNuWuWuWwSeSeUeSsUeNeWuNuDsEeU
1FSD	28	11	9	1185	wDeWwSwSeWsUeUwUeSsDsDeWsNdNdUuEuEsEdEuSwDwDnSwUwUsWwUe
1NMJ	28	21	13	99298	nDeNnDsNeDdSdSdEuNuEeNuNeUwUnUnSdWdSdEdSdEeDeDwNwEnDdSd
2AJJ	28	24	15	92859	wDeDdNdUeUeNeDeDeDnEnSuSuEeNeNuNnUwUsUsEdSuSwDwDwSwUwUw
1PSV	28	14	10	3126	nSuEsNnUwSuSwUsWnSuDeWnDeNeNwNwNwUsUeUuSdSdSeWuDeNuDeSw
2JXF	30	28	16	1303176	nDdNdEeSuEuEnEuDnWnWnDnEdEdNwNwNdUsWsDdWsUsUnUwSdEdEdSdWsWw

Table 4
Current main protein databases.

Database	Description	Web address
PDB	Biological molecular structure	http://www.rcsb.org/
UniProtKB	Information about protein sequence and functionality	http://www.uniprot.org/
PIR	Integrated information that support genomic, proteomic and system biology	http://pir.georgetown.edu/
PROSITE	Information about protein domains, families and functionalities sites	http://www.expasy.org/prosite/
Prints	Information about protein sequence, focus on protein 'figerprinting'	http://www.bioinf.man.ac.uk/dbbrowser/
BLOCKS	A homology database (not update)	http://blocks.fhcrc.org/
eMOTIF	A database from protein motifs, it is derived from BLOCKS and PRINTS database	http://motif.stanford.edu/distributions/
PRODOM	Protein domain families information extracted from Uniprot	http://prodom.prabi.fr/
InterPro	Database with protein families, domain and sites informations	http://www.ebi.ac.uk/interpro/



(a) 2GD3 structure. (b) 1Y32 structure. (c) 2LKE structure. (d) 2NVJ structure.

Fig. 2. Results of the integer programming using the 3D-HP-SC protein model. The white color represents the backbone, blue and red represent respectively the hydrophobic and polar side-chains.

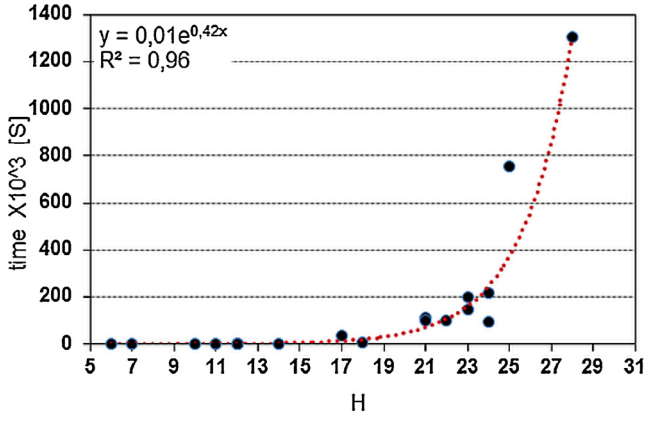


Fig. 3. Processing time (seconds) versus the number of hydrophobic side-chains. The dotted line is the exponential adjustment of the points.

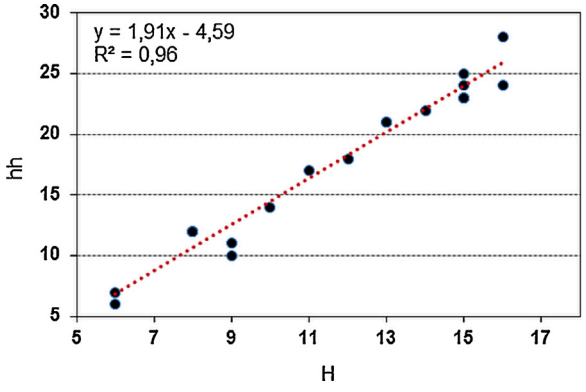


Fig. 4. Number of hydrophobic side-chain contacts (hh) versus the number of hydrophobic side-chains in the sequence. The dotted line is the linear adjustment of the points.

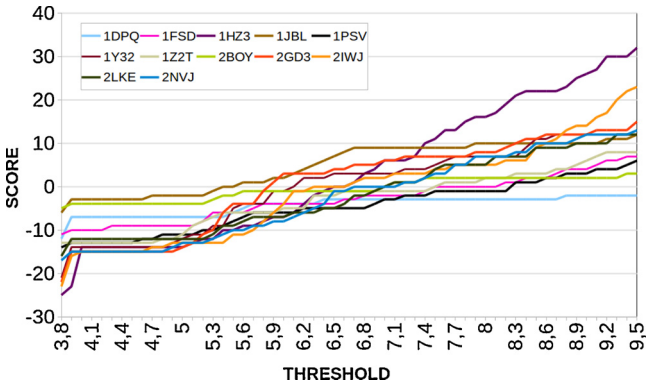


Fig. 5. Difference between hh_{PDB} (from the predicted structure) and hh_{PDB} (from the crystallized protein) at distinct thresholds.

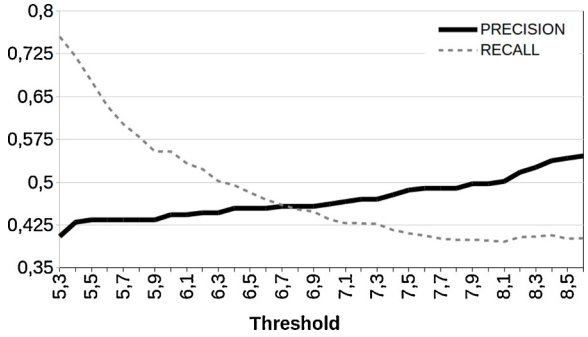


Fig. 6. Precision and Recall of the hydrophobic contact pairs.

Results show that, even using simple representation model, high processing power is required, which grows exponentially with the number of hydrophobic side-chains in the sequence. However, this is the price to pay for optimality. Therefore, the method may not be computationally efficient to be applied to even a small protein with ~100 residues. Considering the current computational resources, the proposed method is a proof-of-concept, but it can be promising in the near future as technology evolves. On the other hand, the exponential trend for solving larger instances suggests that meta-heuristic approaches, such as Evolutionary Computation techniques, may be useful, even not guaranteeing the optimal solution.

Also, the observed correlation between the number of hydrophobic side-chain contacts and the number of hydrophobic amino acids suggests upper-bounds that can be useful in future works.

Additionally, the set of protein sequences presented and the software package developed in these work are publicly available to foster further research (see Appendix A). The package includes a program for downloading protein data from the PDB and converting to the 3D-HP-SC model, in case researchers want to extend the benchmark here presented to other biological proteins of specific interest. Such a benchmark can be a useful baseline for comparing the performance of other methods.

The comparison between the crystallized structure of the proteins with the corresponding structure predicted by the IP method, regarding the hh contacts, suggested a threshold range where the similarity between the approaches is the highest. This finding can establish a parameter for future research.

In the literature, many computational approaches for the PSP are driven by the maximization of the hydrophobic contacts to guide the search for the native structure of proteins. However, the results presented in Section 6 showed that the use of such approach may not be enough to accurately predict the hh contacts that are present in the real biological structure. As a consequence, more sophisticated methods are needed, possibly including other physical and biological aspects of the folding process.

Synthetic data are important for the development of computational methods and are usual in Computational Biology. In special, in the PSP literature most approaches use synthetic data. However, computational results obtained with this sort of data may be of small applicability to real-world proteins. On the other hand, with the growing availability of real biological data (for instance, in the PDB), the use of real data in computational experiments become more attractive, possibly leading to more impactful results. This is a motivation for bringing closer biological data and computer methods.

In future work, force fields could be included in the model, such as the Lennard–Jones energy that appears as amino-acids interact along the folding. Also, a variety of other coarse grain models and lattices could be tested (see Lopes, 2008; Pierri et al., 2008). A study involving a comparative results between IP with other computational methods and models, such as Monte Carlo (Krobath and Faísca, 2013), Molecular Dynamics (Benítez, 2015), and Deep Learning approaches (Hattori et al., 2018), will be considered. Other protein sequences could be considered, such as those presented in the Assessment of protein Structure Prediction (CASP) event.

Acknowledgements

L.T. Hattori would like to thank CAPES for the scholarship. M. Gutoski would like to thank Brazilian National Research Council (CNPq) for the scholarship 141983/2018-3. H.S. Lopes thanks to CNPq for the research grant 311778/2016-0. Both, H.S. Lopes and C.M.V. Benítez thank CNPq, CAPES and Fundação Araucária for the funding grants.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compbiolchem.2019.107192.

References

Alberts, B., 2002. Molecular Biology of the Cell. Garland Science, New York, USA.
Amodeo, P., Castiglione, M., Ostuni, A., Cristinziano, P., Bavoso, A., 2017. Structural Features of the C-Terminal Zinc Finger Domain of the HIV-2 Nc Protein (Residues 23-49). PDB ID: 2IWJ.
Atkins, J., Hart, W., 1999. On the intractability of protein folding with a finite alphabet. Algorithmica 25, 279–294.
Baldwin, R.L., Rose, G.D., 2016. How the hydrophobic factor drives protein folding. Proc. Natl. Acad. Sci. 113, 12462–12466.
Ben-Naim, A., 1994. American Chemical Society, volume 568 of *ACS Symposium Series*, chapter Hydrophobic–Hydrophilic Forces in Protein Folding. pp. 371–380.
Benaki, D., Zikos, C., Evangelou, A., Livaniou, E., Vlassi, M., Mikros, E., Pelecanou, M., 2005. Solution structure of humanin, a peptide against Alzheimer's disease-related neurotoxicity. Biochem. Biophys. Res. Commun. 329, 152–160.
Benaki, D., Zikos, C., Evangelou, A., Livaniou, E., Vlassi, M., Mikros, E., Pelecanou, M., 2006. Solution structure of Ser14Gly-humanin, a potent rescue factor against neuronal cell death in Alzheimer's disease. Biochem. Biophys. Res. Commun. 349, 634–642.
Benítez, C.M.V., 2015. Contributions to the Study of the Protein Folding Problem Using Bioinspired Computation and Molecular Dynamics, PhD Dissertation. Federal University of Technology – Paraná.
Benítez, C.M.V., Lopes, H.S., 2010. Parallel artificial bee colony algorithm approaches for protein structure prediction using the 3dhp-sc model. In: Essaaidi, M., Malgeri, M., Badica, C. (Eds.), Intelligent Distributed Computing IV. Springer, Berlin, Heidelberg, pp. 255–264.
Benítez, C.M.V., Lopes, H.S., 2010b. Protein structure prediction with the 3D-HP side-chain model using a master-slave parallel genetic algorithm. J. Braz. Comput. Soc. 16, 69–78.
Blaszczyk, M., Gront, D., Kmiecik, S., Kurcinski, M., Kolinski, M., Ciemny, M.P., Ziolkowska, K., Panek, M., Kolinski, A., 2019. Protein structure prediction using coarse-grained models. Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes. Springer, Cham, pp. 27–59.
Bošković, B., Brest, J., 2016. Genetic Algorithm with advanced mechanisms applied to the protein structure prediction in a Hydrophobic-Polar model and cubic lattice. Appl. Soft Comput. 45, 61–70.
Brylinski, M., Konieczny, L., Roterman, I., 2006. Hydrophobic collapse in (in silico) protein folding. Comput. Biol. Chem. 30, 255–267.

Carr, R., Hart, W.E., Newman, A., 2003. Discrete Optimization Models for Protein Folding. Technical Report. Sandia National Laboratories.
Chua, G., Tang, l., Amalraj, X.Y., Tan, M., Bhattacharjya, S.M.S., 2011. Structures and Interaction Analyses of the Integrin Alphambeta2 Cytoplasmic Tails.
Dahiyat, B.I., Mayo, S.L., 1997. De novo. Science 278, 82–87.
Dahiyat, B.I., Sarisky, C.A., Mayo, S.L., 1997. De Novo. Science 273, 789–796.
Dill, A., MacCallum, J.L., 2012. The protein-folding problem, 50 years on. Science 338, 1042–1046.
Dorn, M., e Silva, M.B., Buriol, L.S., Lamb, L.C., 2014a. Three-dimensional protein structure prediction: methods and computational strategies. Comput. Biol. Chem. 53, 251–276.
Dorn, M., e Silva, M.B., Buriol, L.S., Lamb, L.C., 2014b. Three-dimensional protein structure prediction: methods and computational strategies. Comput. Biol. Chem. 53, 251–276.
Duarte, A.M., de Jong, E.R., Wechselberger, R., van Mierlo, C.P., Hemminga, M.A., 2007. Segment TM7 from the cytoplasmic hemi-channel from V O-H + -V-ATPase includes a flexible region that has a potential role in proton translocation. Biochim. Biophys. Acta 1768, 2263–2270.
Dubey, S.P., Gopalakrishna, K.N., Balaji, S., Kumar, M.S., 2018. A review of protein structure prediction using lattice model. Crit. Rev. Trade Biomed. Eng. 46, 147–162.
Fraenkel, A.S., 1993. Complexity of protein folding. Bull. Math. Biol. 55, 1199–1210.
Frigori, R.B., 2017. Be positive: optimizing pramlintide from microcanonical analysis of amylin isoforms. Phys. Chem. Chem. Phys. 19, 25617–25633.
Gouttenoire, J., Castet, V., Montserret, R., Arora, N., Raussens, V., Ruysschaert, J.M., Diesis, E., Blum, H.E., Penin, F., Moradpour, D., 2009. Identification of a novel determinant for membrane association in hepatitis C virus nonstructural protein 4B. J. Virol. 83, 6257–6268.
Grace, C.R., Cowsik, S.M., Shim, J.Y., Welsh, W.J., Howlett, A.C., 2007. Unique helical conformation of the fourth cytoplasmic loop of the CB1 cannabinoid receptor in a negatively charged environment. J. Struct. Biol. 159, 359–368.
Hattori, L., Benítez, C., Lopes, H., 2018. A novel approach to protein folding prediction based on long short-term memory networks: a preliminary investigation and analysis. Proc. IEEE World Congress on Computational Intelligence. IEEE Press, Piscataway, NJ, pp. 1–6.
Hoque, T., Chetty, M., Sattar, A., 2009. Extended hp model for protein structure prediction. J. Comput. Biol. 16, 85–103.
Huang, J., Yao, Y., Lin, J., Ye, Y.H., Sun, W.Y., Tang, W.X., 2004. The solution structure of rat Abeta-(1-28) and its interaction with zinc ion: insights into the scarcity of amyloid deposition in aged rat brain. J. Biol. Inorg. Chem. 9, 627–635.
Kalinowska, B., Banach, M., Wisniowski, Z., Konieczny, L., Roterman, I., 2017. Is the hydrophobic core a universal structural element in proteins? J. Mol. Model. 23 205–205.
Karplus, M., 1997. The Levinthal paradox: yesterday and today. Fold. Des. 2, S69–S75.
Kaushik, A.C., Sahi, S., 2017. Biological complexity: ant colony meta-heuristic optimization algorithm for protein folding. Neural Comput. Appl. 28, 3385–3391.
Kmiecik, S., Wabik, J., Kolinski, M., Kouza, M., Kolinski, A., 2019. Protein dynamics simulations using coarse-grained models. Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes. Springer, Heidelberg, pp. 61–87.
Korsinczky, M.L., Schirra, H.J., Rosengren, K.J., West, J., Condie, B.A., Otvos, L., Anderson, M.A., Craik, D.J., 2001. Solution structures by 1H NMR of the novel cyclic trypsin inhibitor SFTI-1 from sunflower seeds and an acyclic permutant. J. Mol. Biol. 311, 579–591.
Krobath, H., Faísca, P.F.N., 2013. Interplay between native topology and non-native interactions in the folding of tethered proteins. Phys. Biol. 10, 016002.
Lau, K.F., Dill, K.A., 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22, 3986–3997.
Li, B., Lin, M., Liu, Q., Li, Y., Zhou, C., 2015. Protein folding optimization based on 3D off-lattice model via an improved artificial bee colony algorithm. J. Mol. Model. 21, 261–269.
Li, T., Zhou, C., Hu, M., 2017. An improved artificial bee colony algorithm for 3D protein structure prediction. Proceedings of the International Conference on Biometrics Engineering and Application. ACM, New York, NY, USA, pp. 7–12.
Lind, J., Barany-Wallje, E., Ramo, T., Wieslander, A., Maler, L. Structure, Position of and Membrane-Interaction of a Putative Membrane-Anchoring Domain of alMGS. PDB ID: 1Z2T.
Lopes, H.S., 2008. Evolutionary algorithms for the protein folding problem: a review and current trends. In: Smolinski, T.G., Milanova, M.M., Hassanien, A.E. (Eds.), Applications of Computational Intelligence in Bioinformatics and Biomedicine: Current Trends and Open Problems. Springer-Verlag, Heidelberg, pp. 297–315 volume 1. chapter 12.
Mandal, S., Jana, N.D., 2012. Protein structure prediction using 2D HP lattice model based on integer programming approach. In: Proceedings of 2012 International Congress on Informatics. Environment, Energy and Applications, Elsevier. pp. 17–18.
Mann, M., Smith, C., Rabbath, M., Edwards, M., Will, S., Backofen, R., 2009. CPSP-web-tools: a server for 3D lattice protein studies. Bioinformatics 25, 676–677.
Mirsky, A.E., Pauling, L., 1936. On the structure of native, denatured, and coagulated proteins. Proc. Natl. Acad. Sci. 22, 439–447.
Mortishire-Smith, R.J., Pitzenberger, S.M., Burke, C.J., Middaugh, C.R., Garsky, V.M., Johnson, R.G., 1995. Solution structure of the cytoplasmic domain of phospholamban: phosphorylation leads to a local perturbation in secondary structure. Biochemistry 34, 7603–7613.
Nardelli, M., Tedesco, L., Bechini, A., 2013. Cross-lattice behavior of general ACO folding for proteins in the HP model. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM, New York, NY, USA. pp. 1320–1327.
Nunes, L.F., Galv ao, L.C., Lopes, H.S., Moscato, P., Berretta, R., 2016. An integer programming model for protein structure prediction using the 3D-HP side chain model. Discrete Appl. Math. 198, 206–214.
Onofrio, A., Parisi, G., Punzi, G., Todisco, S., Di Noia, M.A., Bossis, F., Turi, A., De Grassi, A., Pierri, C.L., 2014. Distance-dependent hydrophobic–hydrophobic contacts in protein folding simulations. Phys. Chem. Chem. Phys. 16, 18907–18917.
Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F., Baker, D., 2018. Protein structure prediction using Rosetta in CASP12. Prot.: Struct., Funct., Bioinf. 86, 113–121.
Parpinelli, R.S., Benitez, C.V., Cordeiro, J.A., Lopes, H.S., 2014. Performance analysis of swarm intelligence algorithms for the 3D-AB off-lattice protein folding problem. Int. J. Multiple-Valued Log. Soft Comput. 22, 267–286.
Pierri, C.L., De Grassi, A., Turi, A., 2008. Lattices for ab initio protein structure prediction. Prot.: Struct., Funct., Bioinf. 73, 351–361.
Reinhard, C., Harter, C., Bremser, M., Brügger, B., Sohn, K., Helms, J.B., Wieland, F., 1999. Receptor-induced polymerization of coatomer. Proc. Natl. Acad. Sci. USA 96, 1224–1228.
Sapay, N., Montserret, R., Chipot, C., Brass, V., Moradpour, D., Deléage, G., Penin, F., 2006. NMR structure and molecular dynamics of the in-plane membrane anchor of nonstructural protein 5A from bovine viral diarrhea virus. Biochemstry 45, 2221–2233.
Su, M., Ling, Y., Yu, J., Wu, J., Xiao, J., 2013. Small proteins: untapped area of potential biological importance. Front. Genet. 4, 286.
Tozzini, V., 2005. Coarse-grained models for proteins. Curr. Opin. Struct. Biol. 15, 144–150.
Unger, R., Moult, J., 1993. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull. Math. Biol. 55, 1183–1198.
Vilar, S., Gonzez-Daz, H., Santana, L., Uriarte, E., 2008. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. J. Comput. Chem. 29, 2613–2622.
Vinogradova, O., Haas, T., Plow, E.F., Qin, J., 2000. A structural basis for integrin activation by the cytoplasmic tail of the alpha IIb-subunit. Proc. Natl. Acad. Sci. USA 97, 1450–1455.
Wang, Z., Xu, J., 2013. Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics 29, i266–i273.
Yanev, N., Milanov, P., Mirchev, I., 2011. Integer programming approach to HP folding. Serdica J. Comput. 5, 359–366.
Yanev, N., Traykov, M., Milanov, P., Yurukov, B., 2017. Protein folding prediction in a cubic lattice in hydrophobic-polar model. J. Comput. Biol. 24, 412–421.
Yu, Z.G., Anh, V., Lau, K.S., 2004. Fractal analysis of measure representation of large proteins based on the detailed hp model. Physica A: Stat. Mech. Appl. 337, 171–184.
Zhang, S., Iwata, K., Lachenmann, M., Peng, J., Li, S., Stimson, E., Lu, a., Felix, Y., Maggio, A., Lee, J.J., 2000. The Alzheimer's peptide AB adopts a collapsed coil structure in water. Struct. Biol. 130, 130–141.