A Novel Approach to Protein Folding Prediction based on Long Short-Term Memory Networks: A Preliminary Investigation and Analysis

Leandro Takeshi Hattori*, César Manuel Vargas Benítez§, Matheus Gutoski †,
Nelson Marcelo Romero Aquino ††, Heitor Silvério Lopes‡
Bioinformatics and Computational Intelligence Laboratory
Federal University of Technology-Paraná
Curitiba, Brazil
Emails: *lthattori@gmail.com, §cesarbenitez@utfpr.edu.br, † matheusgutoski@gmail.com,
††nmarceloromero@gmail.com, ‡hslopes@utfpr.edu.br

Abstract—**The Protein Folding Problem (PFP) is considered one of the most important open challenges in Biology and Bioinformatics. Long Short-Term Memory (LSTM) methods have risen recently, achieving the state-of-art performance for several Bioinformatics problems such as, protein secondary and tertiary protein structure prediction. This paper describes the application of a novel approach based on the LSTM networks to the PFP using a coarse-grained model of proteins. An specific encoding scheme for representing protein folding states is also presented. The proposed approach was evaluated by means of several experiments with a dataset of protein folding, which was obtained by Molecular Dynamics simulations. We also propose a novel method for evaluating the performance of the approach based on measures used in Bioinformatics. Furthermore, a new analysis method for protein folding pathways is presented. Results suggest that the proposed approach is able to learn the protein fold transitions. Also, it is promising for the research areas related to Bioinformatics and Computational Intelligence.**

I. Introduction

The Protein Folding Problem (PFP) is considered to be one of the most challenging open problems in science. Basically, a PFP consists in determining the sequence of folding events that leads from the primary structure of a protein to its native structure which, in turn, defines its specific biological function.

Notwithstanding, researchers have been focusing on the study of this process and, consequently, a large amount of information is currently available regarding this issue. This is mainly due to its importance for medicine, the several genome sequencing projects being conducted in the world and the development of computational models and approaches for the PFP. For instance, several diseases, known as proteinopathies, are believed to be the result of misfolded proteins (i.e. proteins structurally abnormal), such as Alzheimer's disease, cystic fibrosis and some types of cancer [1]. Here, it is important to know that therapeutic drugs for proteinopathies can be discovered from previous knowledge of polypeptide structures. Also, this problem rises three broad questions: (i) What is the physical code by which an amino acid sequence dictates a protein's native structure? (ii) How can proteins fold so fast? (iii) Can we devise a computer algorithm to predict polypeptide structures from their sequences? [2].

To the best of our knowledge, the Molecular Dynamics (MD) approach (including its variations) is the only computational method that really provides a time-dependent analysis of the folding mechanism [3]. Generally, it involves the three-dimensional coordinates of the particles that form the protein and numerical integration of the classical equations of motion. Despite the great advances in recent years, MD simulations have been limited mainly by their computationally expensive brute force calculation. Due to the lack of methods for solving such class of problems in a reasonable computing time, the need for alternative non-traditional mathematical approaches for reproducing the complex behavior of the folding process has risen.

For decades, Computational Intelligence (CI) has provided a large range of robust optimization methods, capable of successfully dealing with complex optimization problems, such as the Protein Structure Prediction (PSP) [4]. Furthermore, within the scope of CI, Deep Learning (DL) methods have yielded significant results on Bioinformatics [5], [6] during the recent years, including the torsion angles prediction methods proposed by [7], for instance. Among DL approaches, the Long Short-Term Memory (LSTM) networks have excelled results in sequential/temporal problems. Therefore, an alternative non-deterministic way to reduce the inherent complexity of the simulations with three-dimensional structures is proposed in this preliminary work, using a minimalist representation of proteins and a LSTM architecture.

The main highlights of this work are:

- a novel approach based on LSTM networks applied to the protein folding prediction;
- a novel method for evaluating the predictor performance based on measures commonly used in Bioinformatics;
- a novel encoding scheme for representing protein folding states and low-level input/output representation for Deep Learning approaches;

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

- a new benchmark *in silico* dataset for the PFP, using the 3DAB *off-lattice* model of proteins;
- a new analysis method for protein folding pathways based on Heatmap visualization;
- a novel validation method based on the Hold-out strategy, specifically designed to protein folding pathways.

This paper is organized as follows: Section II presents the background about the Protein Folding Problem, the coarse-grained model utilized in this work and the related works. Also, Section III describes in details the LSTM approach. Furthermore, Sections IV and V shows the computational experiments and results, including the protein structures dataset. Finally, in Section VI some conclusions and future directions are pointed out.

A. Related works

Several works in Deep Learning (DL) have been presented in the last decades and gained the attention of the scientific community in several domains, including Bioinformatics [8]. Specifically in proteomics, emergent applications with DL have arisen, such as secondary protein structure classification [5], protein homology prediction [6] and inference of the protein torsion angles [7].A short review of DL applied to protein prediction problems was provided by [9], where some possible applications in this area were pointed, as well as the application of Neural Turing Machines and Memory Networks.

Regarding the DL methods, Recurrent Neural Networks have been extensively used in proteomics problems. For instance, a Bidirectional Long Short-Term Memory (BLSTM) network was employed to learn effective features from *pseudo* macromolecules in [6]. In addition, an application of Deep Recurrent Neural Network with Bidirectional Long Short-Term Memory (DBLSTM) for 8-class secondary structure classification from sequence is presented by [5].

Many DL methods in protein structure problems utilized Contact Map (CM) representation, in which positions of a matrix M indicate if the amino acid pairs are in contact. For example, [10] reported a paper for CM prediction using steps of increasing resolution, where a DL architecture is used to progressively refine the inference of the contacts. More recently, a DL method for predicting amino acids contact by integrating both Evolutionary Coupling and sequence conservation information through two deep residual neural networks were presented by [11]. Moreover, [12] presented a Deep Transfer Learning scheme that predicts amino acids contact and then predicts their structure models. As outlined above, CM is a commonly used protein model representation in DL. However, this model raises a problem: the reconstruction of the protein, which is still an open problem [13].

Other protein representations can be used to avoid the problem cited above. A similar macromolecule representation is based on the dihedral angles, which is applied to Ramachandran plot analysis. With this representation, four different DL architectures to predict polypeptide torsion angles were evaluated in [7]. This study also uses Mean Absolute Error (MAE) for evaluating the prediction of phi and psi angles. Another manner to represent the biomolecule structure is by using spherical coordinates. However, this method is sparsely explored in the literature [14].

Concerning the protein model, the 3DAB *off-lattice* is currently the main coarse-grained model used in the Computational Intelligence area for dealing the PFP. For instance, the application of a parallel ecology-inspired algorithm (pECO) for the polypeptide structure reconstruction from CMs is described in [13]. More recently, an improved Artificial Bee Colony (ABC) was employed by [15] to infer protein structures. Also, a Biogeography Based Optimization with Chaotic Mutation (BBO-CM) algorithm that prevents premature convergence was proposed by [16] for forecasting protein structures.

From the above-mentioned works, we hypothesize that Deep Learning can play an important role in the Bioinformatics and Proteomics problems. To the best of our knowledge, this work presents the first approach based on LSTM networks for the PFP. Also, differently from the commonly applied protein structure representations, we use relative spherical coordinates for dealing with geometrical constraints due to the bonds between amino acids. Moreover, it is possible to observe from the literature that the 3DAB *off-lattice* model is commonly used for the PSP. Therefore, a new benchmark dataset is provided for the PFP in this present study.

II. The Protein Folding Problem

Protein folding is the process by which polypeptide chains are transformed into compact structures that perform biological functions. Under physiological conditions, the most stable three-dimensional structure is called the native conformation and actually allows a protein to perform its function. Despite the considerable theoretical and experimental effort expended to study the protein folding process, a detailed description of the mechanisms that govern the folding process have not been discovered yet.

The Protein Folding Problem (PFP) is the prediction of the protein folding pathways, which consists in determining the sequence of folding events that leads from the primary structure of a polypeptide to its native structure. Moreover, the Protein Structure Prediction (PSP) consists in predicting the protein structure from sequence (i.e. primary structure).

Several computational models have been proposed for representing protein structures with different levels of complexity and, consequently, computational feasibility. Despite their simplicity, they have provided several valuable insights regarding the folding process. Also, it is important to recall that the computational approach for searching a solution for the PFP using the simplest models, the so-called Hydrophobic-Polar (HP) models, was proved to be NP-complete [17].

The prediction of the structure of a protein is modeled as the minimization of the corresponding free-energy, following the Anfinsen's Thermodynamic Hypothesis [18] [1]. It is also known that the native conformation of a protein represents the

[1] Nobel Prize Laureate in 1972

folding state with minimal free-energy. A schematic energy landscape for protein folding, using the 3D *off-lattice* model of proteins (see Section II-A), is shown in Figure 1. In this figure, it is possible to observe that a unique folding state (i.e. native structure) is reached through many independent pathways (represented by blue arrows) starting from different initial folding states.

Moreover, the differences between the many-pathway protein folding model derived from theoretical energy landscape and the pathway model derived from experiments are described by [19].

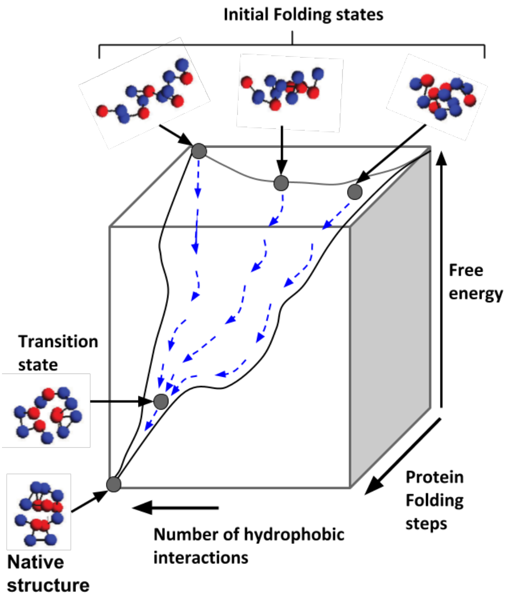

Fig. 1. The Energy Landscape for Protein Folding. Adapted from [20] by using the 3D-AB model of proteins. Red and blue balls are hydrophobic and polar amino acids, respectively

A. The 3D-AB off-lattice model of proteins

The AB *off-lattice* model was introduced by [21] to represent protein structures. In this model, residues are represented by single interaction sites located at the Cα position and are linked by rigid unit-length bonds ($\hat{b_i}$) to form the protein structure.

In this model, the 20 proteinogenic amino acids are classified into two classes, according to their affinity to water (hydrophobicity): ’A’ (hydrophobic) and ’B’ (hydrophilic or polar). Also, the energy function of a folding is given by Equation 1 [22].

$$E = E_{Angles} + E_{Torsion} + E_{LJ} = -k_1 \sum_{i=1}^{N-2} \widehat{b_i} \cdot \widehat{b_{i+1}} - k_2 \sum_{i=1}^{N-3} \widehat{b_i} \cdot \widehat{b_{i+2}} + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} 4\varepsilon(\sigma_i, \sigma_j)(r_{ij}^{-12} - r_{ij}^{-}6) \quad (1)$$

where,
E_{Angles}, $E_{torsion}$ and E_{LJ} are the energies from bond angles, torsional forces and Lennard-Jones potential, respectively. r_{ij} represents the distance between ith and jth residues; $\sigma = \sigma_0$, ..., σ_N from a binary string that represents the protein sequence. $\varepsilon(\sigma_i, \sigma_j)$ is chosen to favor the formation of the hydrophobic core (’A’ residues). Thus, $\varepsilon(\sigma_i, \sigma_j)$ is 1 for AA interactions and 1/2 for BB/AB interactions. Finally, it is important to mention that the model can be explored for different values of k_1 and k_2 as stated by [22].

Figure 1 shows examples of structures using the 3D-AB *off-lattice* model.

III. The Long Short-Term Memory approach

The Long Short-Term Memory (LSTM) network was introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [23] in order to overcome the inherent error signals flowing backward in time of conventional methods such as standard Recurrent Neural Networks (sRNN), which tend to either explode or vanish. LSTM networks include memory cells, that are hidden units that lead to the natural behavior of remembering inputs for a long time. Thus, the gradient can flow for a long time, avoiding the vanishing gradient problem.

Figure 2 [2] shows a memory cell, which is an accumulator that has a connection to itself at the next time step.

Basically, the LSTM process is composed by three gates (i.e. the forgot gate f_t, update gate i_t and output gate o_t). The variables x_t and h_t represent the input and output of the network at time t. The layers with sigmoid and hyperbolic tangent activation functions are represented by σ and $tanh$.

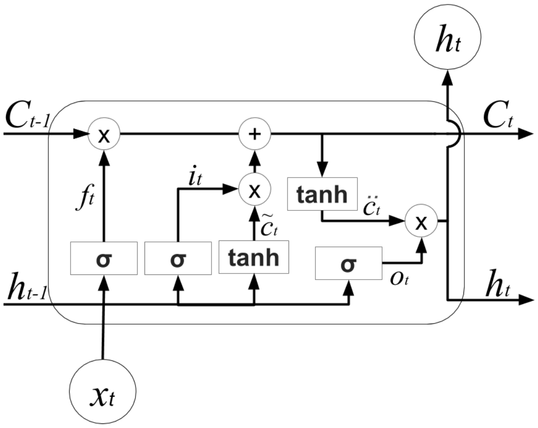

Fig. 2. The internal structure of a Long Short-Term Memory cell (LSTM).

Equations 2–8 present the mechanism of a LSTM.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2) \qquad i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$\widetilde{\mathcal{C}}_t = tanh(W_C[h_{t-1}, x_t] + b_C) \quad (4) \qquad C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{\mathcal{C}}_t \quad (5)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (6) \qquad \ddot{C}_t = tanh(C_t) \quad (7)$$

$$h_t = o_t \circ \ddot{C}_t \quad (8)$$

[2] Based on colah.github.io/posts/2015-08-Understanding-LSTMs/

where, f_t, i_t, C_t, o_t and h_t ($\in \Re^k$) are activations of the forgot gate, update gate (or input gate vector), internal long term memory cell state, output gate vector (candidate) and output vector considering k hidden units, respectively. In addition, W_f, W_i, W_C and W_o ($\in \Re^{3q\times k}$) are the weight matrices; and b_f, b_i, b_C and b_o ($\in \Re^k$) are bias terms. Also, the symbol $\circ$ denotes the Hadamard product operator.

As mentioned before, an alternative non-deterministic approach applied to the PFP based on the LSTM networks is proposed in this work. Figure 3 presents a simplified overview of the proposed approach. The protein folding is simulated by using a trained/tested LSTM. Basically, a *one-step* prediction (**part 4**) is done to ensure that the ith folding state (*output*, $t+1$) from the previous (*input*, t).

For training and testing the LSTM network (**part 3**), the 3D-AB *off-lattice* model is used to represent protein structures and protein folding pathways data are generated by using a Molecular Dynamics (DM) approach (**part 1**, see Section IV-A). Then, the data are processed and split (**part 2** into feature vectors, according to the encoding scheme (see Section III-A).

A. Feature Encoding

It is known that Deep Learning methods are highly based on the features and how they are encoded. This work proposes a feature encoding scheme for dealing with geometrical constraints due to the fixed unit-length bonds between amino acids, when using the 3DAB *off-lattice* model.

Considering the folding of a protein with S amino acids, a one-dimensional feature vector will represent the set of relative spherical coordinates of the amino acids, as shown in Figure 4. The first amino acid of the sequence is located at the origin. Thus, a feature vector has $(2S-2)$ variables, such that positions kth and $k+1$th represent the spherical coordinates θ_k and φ_k of the second to S amino acid of the sequence. The input and output vectors are normalized in the range of $[0:1]$.

B. Cost Function

The cost function applied in this study is the Mean Absolute Error (MAE), shown in Equation 9, which is used to evaluate the prediction of relative spherical coordinates. This measure is the absolute difference between the predicted ($output$) and the $target$ spherical coordinates. Here, both $target$ and $output$ are in the range $[0:1]$ and S is the sequence length.

$$MAE = \frac{\sum_{i=1}^{S-1} |target - output|}{S-1} \tag{9}$$

C. Network Setup and Architecture

In this study, the Gradient Optimizer RMSProp [24] is used for optimizing the gradient descent of the LSTM network and controlling the value of the learning rate (η), which determines the size of the steps taken towards to the opposite direction of the gradient.The RMSProp was selected among others based on a previous analysis of optimizers [5].

The parameters for the LSTM network are: number of epoch equals to 1000, with $\eta = 0.0005$. The architecture of the network, which was inspired by the works presented in [15], [25], is composed of: One Fully Connected (FC) layer, followed by one LSTM layer and two FC layers. The activation function of the last FC layer is a sigmoid function, in order to obtain an output in the range $[0,1]$, whereas in the other FC layers, Rectified Linear Units (ReLUs) were used. With the purpose of avoiding overfitting in the training phase, we apply L2 regularization with weight equals to 0.001 and two dropout layers with probability 0.2 in the two final FC layers.

D. Evaluation Measures

In this work, the Radius of Gyration (R_g) [26] is used to measure the compactness of the residues of the protein. The smaller R_g means that the set of amino acids are more compact. The equation of R_g is present in the Equation 10.

$$R_g = \sqrt{\frac{\sum_{i=0}^{N-1} [(x_i - \bar{X})^2 + (y_i - \bar{Y})^2 + (z_i - \bar{Z})^2]}{N}}, \tag{10}$$

where i and N are the i-th protein residue and the number of residues, respectively. x_i, y_i and z_i are the Cartesian coordinates, and $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ are the average of their respective axis.

This work also proposes a novel method for comparing protein structures ($Output$ and $Target$), using the 3D-AB model, by calculating the RMSD (Root-mean-square deviation). Algorithm 1 presents this method.

Algorithm 1 Predictor performance evaluation algorithm

```
1: **Start**
2: $P_1 \leftarrow$ *decoding($Output$)*
3: $P_2 \leftarrow$ *decoding($Target$)*
4: $RMSD \leftarrow kabsch(P_1, P_2)$
5: **End**
```

Basically, Algorithm 1 has three steps, where the first two steps are decoding procedures and the last one represents a quality assessment.

Algorithm 2 presents the decoding procedure, which consists of a Spherical to Cartesian coordinates conversion. Basically, this procedure returns the Cartesian coordinates of each amino acid of the protein structure.

In line 4, it is measured the similarity between the structures obtained in the previous two steps, using the Root-Mean-Square-Deviation (RMSD), as shown in Equation 11.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{S-1} |P_{1,i} - P_{2,i}|}{S}}, \tag{11}$$

where S, P_{1i}, P_{2i} represents the number of amino acids, Cartesian coordinates of the protein structures (P_{1i}) and (P_{2i}), respectively.

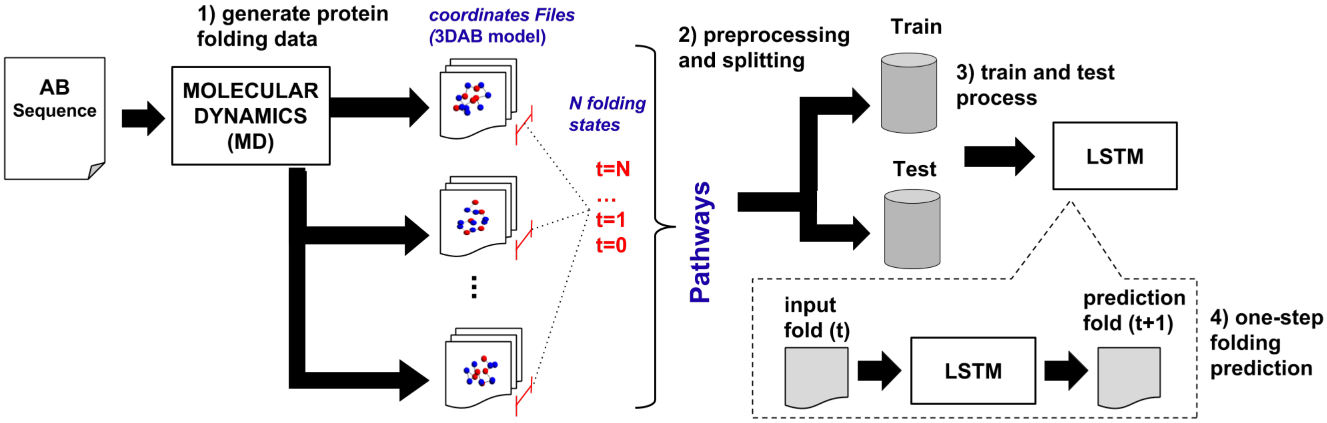
AB Sequence
1) generate protein folding data
MOLECULAR DYNAMICS (MD)
coordinates Files (3DAB model)
N folding states
t=N
...
t=1
t=0
Pathways
2) preprocessing and splitting
Train
Test
3) train and test process
LSTM
input fold (t)
LSTM
prediction fold (t+1)
4) one-step folding prediction

Fig. 3. Overview of the proposed approach.

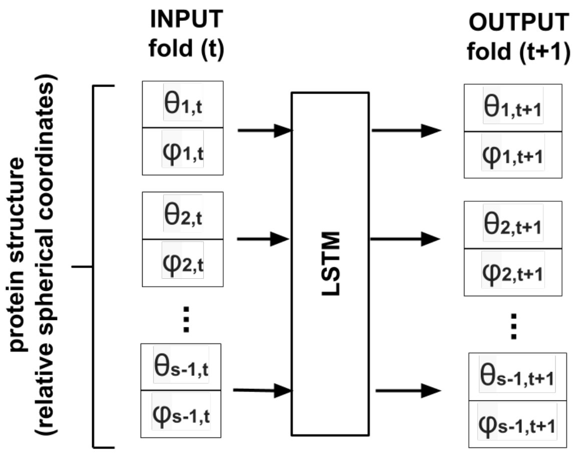

Fig. 4. Input/output Encoding of the LSTM.

Algorithm 2 Decoding procedure – *decoding(p)*

```
Start
Let $S$ be the protein size (number of amino acids)
Let $\theta_i$ and $\varphi_i$ be the input spherical coordinates
Let $a$ be the output Cartesian coordinates
Let $dx$, $dy$ and $dz$ Let $r$ be the unit length bond between $i$ and $(i+1)$ amino acids
$a.x[0] = 0; a.y[0]; a.z[0] = 0;$
for $i = 1 \to S - 1$ do
    $a.x[i] = 0; a.y[i] = 0; a.z[i] = 0$
    $a.x[i] = r \cdot sin\varphi_i \cdot cos\theta_i$
    $a.y[i] = r \cdot sin\varphi_i \cdot sin\theta_i$
    $a.z[i] = r \cdot cos\varphi_i$
    $a.x[i] = a.x[i] + a.x[i-1]$
    $a.y[i] = a.y[i] + a.y[i-1]$
    $a.z[i] = a.z[i] + a.z[i-1]$
end for
return $a$
End
```

Since the RMSD is a rotation-dependent measure, an optimized RMSD is done using the Kabsch method [27] to obtain the lowest RMSD.

IV. Computational Experiments

All experiments done in this paper were run on a computer with an Intel Core i7 processor at 3.30GHz, a GPU Nvidia Titan X and a minimal installation of Ubuntu 14.04 LTS [3]. The software was developed using the Python programming language, the Lasagne 0.9 and Theano 1.0 frameworks [4].

A. Dataset

The protein folding dataset[5] was created in order to train and evaluate our protein folding predictor. This was accomplished by using the Molecular Dynamics (MD) method, which was proposed in a previous work [28].

For simulating the folding pathways, we used a synthetic protein sequence based on the Fibonacci numbers that was proposed by [21]. The AB sequence is $AB^2AB^2(AB)^2BAB$, where A and B represent the hydrophobic and hydrophilic amino acids, respectively.

The dataset is composed of 20 different folding pathways, which start from different initial folding states (i.e. protein structures along the folding process, as shown in Figure 1), where each pathway is composed of 101 equally spaced in time folding states.

As commented in Section I, a new method for analyzing the folding pathways is proposed in this study. This method is based on bi-dimensional Heatmap visualization, where the average RMSD measure of the ith folding state of all pathways is represented as colors. Larger values and lower values are represented by warm and cold colors, where warm colors represent highly different initial folding states (i.e. unfolded states). On the other hand, blue indicates equal structures (i.e. native state).

Figures 5 and 6 present the similarity between initial and

[3] Available in: www.ubuntu.com

[4] Available in: http://lasagne.readthedocs.io/en/latest/user/tutorial.html

[5] https://github.com/bioinfolabic/protein_folding_DL_3DAB_off-lattice.git

final folding states of all pathways, respectively. It is possible to observe that the protein structures are more diverse at the beginning as well as the structures tend to be more similar during the final stages of the folding mechanism, according to the behavior of the *in vivo* protein folding process and the Anfinsen's thermodynamic hypothesis (see Section II). Here, it is important to recall that the energy landscape is characterized by several intermediate folding states and energy barriers between the two significantly folded states, the denatured and the native states. Thus, it is possible to conclude that the pathways generated by the MD approach do not reach the native state, where the entire heatmap of the final pathways (Figure 6) must be blue. Finally, considering that the MD is deterministic, we can conclude that the 20 pathways are different and suitable for training and testing the Deep Learning approaches.

Based on such protein folding data, we created a tuple of folds to generate our dataset, where, the first ith fold and another jth fold are the t $(input)$ and $t+1$ $(target)$ time steps, end up with 100 pairs of folds per pathway. Next, the tuples of folds are split in train and test subsets using a Hold-out validation (70% to train and 30% to test). Seven and three folds at each 10 subsequent tuples were randomly selected for training and testing the LSTM network, respectively. It is important to recall that in order to ensure that the test subset has a representative data of the whole protein folding mechanism, a homogeneous amount of tuples of folds was selected from the pathways.

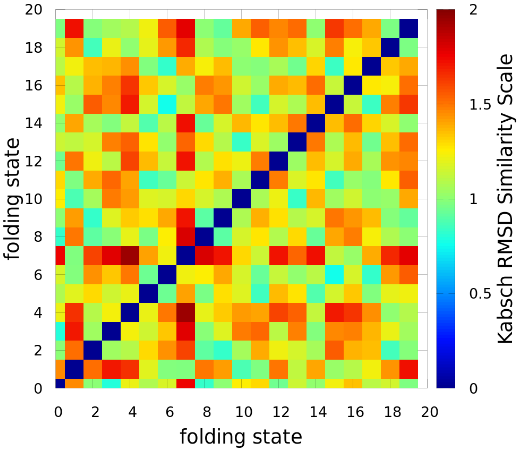

Fig. 5. Heatmap visualization of Similarity (RMSD) between the 20 initial folding states

A preprocessing procedure was also applied to this dataset to generate feature vectors, according to Section III-A. Algorithm 3 presents this procedure. First, the Cartesian coordinates (x, y, z) of the amino acids are converted to relative spherical coordinates (θ, φ). It is important to recall that the $atan2$ function returns a positive value for counter-clockwise angles, and a negative value for clockwise angles.

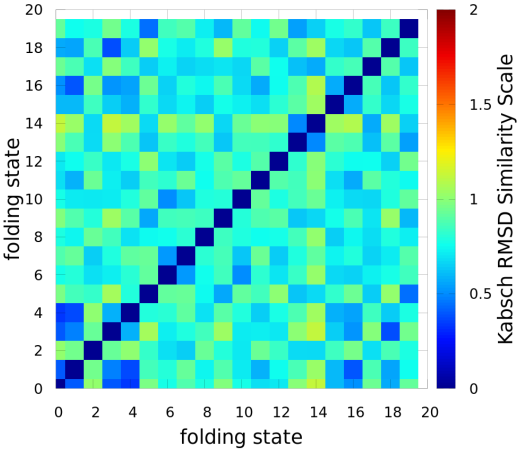

Fig. 6. Heatmap visualization of Similarity (RMSD) between the 20 final folding states

Finally, the normalization of the relative spherical coordinates is done, which are in the range $[0 : \pi]$ and $[-\pi : \pi]$. Thus, the features are normalized between $[0 : 1]$.

Algorithm 3 Encoding procedure – *encoding(p)*

```
Start
Let $S$ be the protein size (number of amino acids)
Let $p$ be the input Cartesian coordinates ( $\overrightarrow{x}_i$, $\overrightarrow{y}_i$, $\overrightarrow{z}_i$)
Let $\overrightarrow{a}$ be the output relative Spherical coordinates
Let $dx$, $dy$ and $dz$
Let $r$ be the unit length bond between $i$ and $(i+1)$ amino acids
for $i = 1 \rightarrow S - 1$ do
    for $j = 0 \rightarrow S$ do
        $x[j] = x[j] - x[i-1]$
        $y[j] = y[j] - y[i-1]$
        $z[j] = z[j] - z[i-1]$
    end for
    $a.r[i-1] = sqrt(x[j]^2 + y[j]^2 + z[j]^2)$
    $a.\theta[i-1] = acos(z[j]/a.r[i-1])$
    $a.\varphi[i-1] = atan2(y[j]/x[j])$
    $normalize(a)$
end for
return $a$
End
```

V. Results and Analysis

This section presents the results obtained by our approach for the protein folding prediction.

The sequence length used for the LSTM is 12, with 2 features in each sequence input, which represent the relative spherical coordinates of the amino acids (see Section III-A).

The performance of the LSTM approach with increasing number of neurons was also evaluated. Table I presents the results obtained. The MAE obtained are equal to 0.339, 0.331 and 0.320 for 200, 400 and 800 neurons, respectively It is possible to observe that the MAE measure decreases when the number of neurons increases, indicating that the quality of the structures obtained is proportional to the number of neurons. Also, we present the prediction error of the relative spherical coordinates (θ and φ). The LSTM with 800 neurons improved the quality of the prediction, achieving the smallest error compared to the LSTM with 200 and 400 neurons.

TABLE I
MEAN ABSOLUTE ERROR (MAE) OF THE TEST AND THE PREDICTION ERROR PER RELATIVE SPHERICAL COORDINATES

Sph. Coord.	**Prediction Error** ($Avg \pm \sigma$)		
	Neurons		
	200	**400**	**800**
θ	0.146 ± 0.118	0.141 ± 0.115	0.138 ± 0.112
φ	0.192 ± 0.191	0.187 ± 0.194	0.180 ± 0.194
MAE Loss (test)	0.339	0.331	0.320

Table II shows the results considering the prediction error per amino acid. The LSTM that achieved the lowest error results, in most cases, is the network with 800 neurons, and also presented the lowest standard deviation (σ). An interesting analysis presented here, is that the major error is during the early stages of the network prediction. Therefore, it indicates that the use of bidirectional information could be an interesting in order to improve the prediction performance of our approach.

TABLE II
MEAN ABSOLUTE ERROR (MAE) PER AMINO ACID

Sph. Coord. per Amino acid	**Prediction Error** ($Avg \pm \sigma$)		
	Neurons		
	200	**400**	**800**
2	0.200 ± 0.166	0.201 ± 0.165	0.199 ± 0.166
3	0.183 ± 0.149	0.180 ± 0.151	0.178 ±0.148
4	0.184 ± 0.174	0.182 ± 0.173	0.179 ± 0.172
5	0.175 ± 0.147	0.170 ± 0.148	0.165 ± 0.149
6	0.172 ± 0.166	0.169 ± 0.162	0.164 ± 0.167
7	0.154 ± 0.151	0.151 ± 0.154	0.139 ± 0.151
8	0.177 ± 0.177	0.169 ± 0.176	0.164 ± 0.178
9	0.149 ± 0.148	0.145 ± 0.155	0.141 ± 0.150
10	0.165 ± 0.164	0.159 ± 0.172	0.148 ± 0.162
11	0.150 ± 0.148	0.139 ± 0.143	0.135 ± 0.142
12	0.165 ± 0.168	0.159 ± 0.167	0.152 ± 0.161
13	0.150 ± 0.150	0.149 ± 0.152	0.143 ± 0.152

Figure 7 presents the best protein folding prediction for each LSTM configuration. For each configuration, the result obtained by the LSTM ($output$) is compared with the $target$, which was previously generated by MD simulations. It is possible to observe that our approach obtained slightly different structures compared to the $target$, as indicated by the MAE measure. Also, the LSTM with 800 neurons obtained protein structure with lower RMSD (0.272) and MAE (0.110) values. In this figure, it is possible to observe the formation of the hydrophobic core, according to the measure of compactness of the entire protein (RG_{all}) and of the hydrophobic residues (RG_h). Due to the nature of the problem, such core was already expected, suggesting that the proposed approach can capture some properties of the protein folding process.

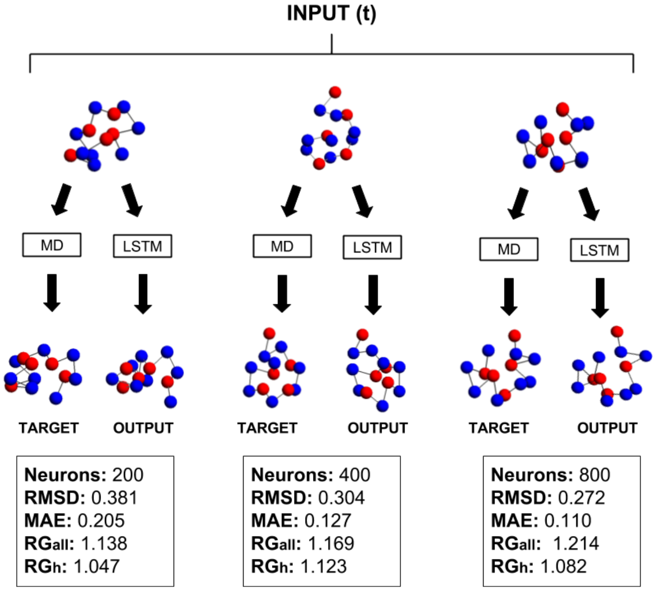

Fig. 7. The best predictions for each LSTM configuration (neurons amount [200; 400; 800]

VI. CONCLUSION

The PFP is still an open problem in Bioinformatics for which there is no closed computational solution. Even using the simplest model, the computational approach for searching a solution for the PFP was proved to be NP-complete.

This paper reports the first results of an ongoing project. Basically, a preliminary investigation of a novel approach based on LSTM networks applied to the Protein Folding Problem is presented. The LSTM network is used to *one-step* prediction of folding states, using the protein 3DAB *off-lattice* model of proteins.

The 3DAB *off-lattice* model has been sparsely explored in current literature, due to the high level of complexity of its energy landscape. In fact, to the best of our knowledge, this paper presents the first application of a Deep Learning method to the PFP using 3DAB *off-lattice* model.

This work also offered new reference values for benchmark protein folding pathways that can be used in the future by other researchers for testing computational approaches applied to the same problem.

Regarding the LSTM network, three number of neurons were employed in order to quantify the significance of increasing the number of them without causing either overfitting or underfitting. Here, it is important to recall that the computational cost was also considered in order to avoid making the network impractical. It is possible to observe that better results can be achieved increasing the number of neurons, according to the MAE and RMSD measures.

The results obtained suggest that the proposed approach is able to learn the protein fold transitions. Lower MAE measures indicate that the obtained structures are similar to the target structures. In addition, the measure of compactness of the protein structure prediction showed the tendency to form a hydrophobic core inside the protein, according to the *in vivo* protein folding process. Also, it is possible to observe that some issues need to be improved, such as the precision of the prediction. Furthermore, the results of prediction error per amino acid showed that the higher errors are in first predictions. This suggests that a bidirectional information of the protein structure could be an important feature for improving the predictor performance.

In a broader sense, it is clear that the processing time of the simulations is an important drawback. Thus, the use of GPUs is essential to allow us to obtain results in a reasonable processing time.

Future works will include the study of self-adjustment of parameters and the use of bidirectional recurrent neural network models. Also, a comparison with classical methods and other machine learning approaches will be assessed.

Further works will also focus on more intensive experiments with these and other benchmarks, and an analysis of the influence of the dataset size on the performance.

Besides the RMSD and MAE measures, we intend to study other metrics in order to contribute to better understanding the approach and process.

Finally, we believe that the use of LSTM networks to the PFP using coarse-grained models is very promising for the research areas related to Bioinformatics and Computational Intelligence. Although there are interesting research directions that suggest the continuity of this work, the initial objectives were achieved satisfactorily.

Acknowledgment

The authors would like to thank the Brazilian National Research Council (CNPq), CAPES and Fundação Araucária for the research grants to H.S. Lopes, C.M.V. Benítez, M. Gutoski and L.T. Hattori; N. M. Romero Aquino thanks the Organization of the American States, the Coimbra Group of Brazilian Universities and the Pan American Health Organization; Also, we thank Nvidia for the GPU TITAN X donation.

References

[1] I. Moreno-Gonzalez, G. Edwards Iii, N. Salvadores, M. Shahnawaz, R. Diaz-Espinoza, and C. Soto, "Molecular interaction between type 2 diabetes and alzheimer's disease through cross-seeding of protein misfolding," *Molecular Psychiatry*, vol. 22, no. 9, pp. 1327–1334, 2017.

[2] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[3] A. Liwo, M. Khalili, and H. A. Scheraga, "Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains," *Proc. of the National Academy of Sciences*, vol. 102, no. 7, pp. 2362–2367, 2005.

[4] C. Benítez, R. Stubs Parpinelli, and H. Lopes, "A heterogeneous parallel ecologically-inspired approach applied to the 3D-AB off-lattice protein structure prediction problem," in *Proc. of the Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI CBIC)*, no. 1. Piscataway, NJ: IEEE Press, 2013, pp. 592–597.

[5] L. T. Hattori, C. M. Benítez, and H. S. Lopes, "A deep bidirectional long short-term memory approach applied to the protein secondary structure prediction problem," in *Proc. of the 4th IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. Piscataway, NJ: IEEE Press, 2017, pp. 1–6.

[6] S. Li, J. Chen, and B. Liu, "Protein remote homology detection based on bidirectional long short-term memory," *BMC Bioinformatics*, vol. 18, no. 1, pp. 443–450, 2017.

[7] H. Li, J. Hou, B. Adhikari, Q. Lyu, and J. Cheng, "Deep learning methods for protein torsion angle prediction," *BMC Bioinformatics*, vol. 18, no. 1, p. 417, 2017.

[8] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[9] K. Paliwal, J. Lyons, and R. Heffernan, "A short review of deep learning neural networks in protein structure prediction problems," *Advanced Techniques in Biology and Medicine*, vol. 5, no. 139, pp. 1–2, 2015.

[10] P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, 2012.

[11] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLOS Computational Biology*, vol. 13, no. 1, pp. 1–34, 2017.

[12] S. Wang, Z. Li, Y. Yu, and J. Xu, "Folding membrane proteins by deep transfer learning," *Cell Systems*, vol. 5, no. 3, pp. 202 – 211.e3, 2017.

[13] C. Benítez, R. S. Parpinelli, and H. S. Lopes, "An ecologically-inspired parallel approach applied to the protein structure reconstruction from contact maps," in *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. New York, NY, USA: ACM, 2016, pp. 1299–1306.

[14] V. M. Reyes, "Representation of protein 3d structures in spherical (ρ, ϕ θ) coordinates and two of its potential applications," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 3, no. 3, p. 161, 2011.

[15] T. Li, C. Zhou, and M. Hu, "An improved artificial bee colony algorithm for 3d protein structure prediction," in *Proceedings of the 2017 International Conference on Biometrics Engineering and Application*. New York, NY, USA: ACM, 2017, pp. 7–12.

[16] N. D. Jana, J. Sil, and S. Das, "Protein structure optimization in 3d ab off-lattice model using biogeography based optimization with chaotic mutation," in *Proceedings of the Fourth ACM IKDD Conferences on Data Sciences*. New York, NY, USA: ACM, 2017, pp. 6:1–6:7.

[17] J. Atkins and W. Hart, "On the intractability of protein folding with a finite alphabet," *Algorithmica*, vol. 25, no. 2, pp. 279–294, 1999.

[18] C. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 96, pp. 223–230, 1973.

[19] S. W. Englander and L. Mayne, "The case for defined protein folding pathways," vol. 114, no. 31, pp. 8253–8258, 2017.

[20] C. Dobson, "Principles of protein folding, misfolding and aggregation," *Cell & Developmental Biology*, vol. 15, pp. 3–16, 2004.

[21] F. Stillinger and T. Head-Gordon, "Collective aspects of protein folding illustrated by a toy model," *Physical Review E*, vol. 52, no. 3, pp. 2872–2877, 1995.

[22] A. Irback, C. Peterson, F. Potthast, and O. Sommelius, "Local interactions and protein folding: A three-dimensional off-lattice approach," *Journal of Chemical Physics*, vol. 107, no. 1, pp. 273–282, 1997.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] T. Tieleman and G. Hinton, "Neural networks for machine learning, lecture 6.5 – rmsprop," 2012.

[25] S. K. Sønderby and O. Winther, "Protein secondary structure prediction with long short term memory networks," arXiv preprint arXiv:1412.7828, 2015.

[26] A. R. Khokhlov, *Statistical Physics of Macromolecules*. New York, USA: AIP-Press, 1994.

[27] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 34, no. 5, pp. 827–828, 1978.

[28] C. M. V. Benítez and H. S. Lopes, *Molecular Dynamics for Simulating the Protein Folding Process Using the 3D AB Off-Lattice Model*. Berlin Heidelberg: Springer, 2012, pp. 61–72.