Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Simplified Lattice Models for Protein Structure Prediction: How Good Are They?

Swakkhar Shatabda and **M. A. Hakim Newton** and **Abdul Sattar**
IIIS, Griffith University and QRL, NICTA *
swakkhar.shatabda@nicta.com.au

Abstract

In this paper, we present a local search framework for lattice fit problem of proteins. Our algorithm significantly improves state-of-the-art results and justifies the significance of the lattice models. In addition to these, our analysis reveals the weakness of several energy functions used.

A protein folds into a three dimensional native structure that has the minimum free energy. This structure is unique, stable and essential for its proper functioning. Knowledge about this structure is of paramount importance, since it can have an enormous impact on the field of rational drug design. The *in vitro* methods for Protein structure prediction (PSP) are slow and expensive. Computational methods have been used to solve this problem for more than thirty years. One of the challenges for solving PSP is the unknown nature of the energy function. Moreover, the all-atomic details of structures require huge computational power. For these reasons, researchers preferred to model the problem in a simplified way by restricting the locations of the amino acids of the proteins to discrete lattice points (cubic or face-centered) and search is guided by a simple energy function that considers the contact potentials (Miyazawa and Jernigan 1985; Berrera, Molinari, and Fogolari 2003; Lau and Dill 1989). A lattice $\mathbb{L}$ is a set of points in $\mathbb{Z}^n$ where the points are integral linear combinations of given N basis vectors. Two lattice points $p, q \,\epsilon\, \mathbb{L}$ are said to be in contact or $neighbors$ of each other, if $q = p + \vec{v_i}$ for some vector $\vec{v_i}$ in the basis of $\mathbb{L}$. Every two consecutive amino acid monomers in the sequence are in contact (*connectivity* constraint) and two amino acids can not occupy same point in the lattice (*self avoiding walk* constraint). For any given protein sequence s, the free energy of a structure c is calculated by the following equation:

$$E(c) = \sum_{j \geq i+1} \mathsf{contact}(i,j).\mathsf{energy}(i,j) \qquad (1)$$

where $\mathsf{energy}(i,j)$ is the empirical energy value between two amino-acids i and j and $\mathsf{contact}(i,j)$ is 1 if i and j are in contact and otherwise 0. Given this model, the protein structure problem can be defined as follows: given a sequence s of length n, find a self avoiding walk $p_1 \cdots p_n$ on the lattice that minimizes the energy defined by (1).

The optimal structures found by these models are used as candidate structures or decoys after reconstructing the backbone and adding the side chains. However, the effectiveness of the methods depends on the energy functions. In this paper, we investigate the goodness of lattice fit and effectiveness of such energy functions used for simplified lattice models. We propose a constraint based local search (CBLS) framework to find the lattice fits for real proteins on different types of lattices. Our approach produces significantly improved lattice fits compared to the state-of-the-art methods. The local search framework also allows us to analyze the different energy functions used in simplified models.

Local Search Framework

In the lattice fit problem, the task is to find a self avoiding walk on a discrete lattice that has the minimum root mean square distance (dRMSD) value with the given native structure defined in Eq. 2.

$$\text{dRMSD} = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(d_{ij}^{\text{given}} - d_{ij}^{\text{native}})^2}{n*(n-1)/2}} \qquad (2)$$

where d_{ij}^{given} and d_{ij}^{native} denotes the distances between ith and jth amino acids respectively in the given conformation and the native conformation of the protein. In calculating the RMSD values, the distance between two neighbors in the lattice ($\sqrt{2}$ for FCC and 1 for cubic) is considered to be equal to the average distance (3.8Å) between two α-Carbons on the native structure.

Algorithm 1: localSearch()

```
initialize()
while time ≤timeout do
    selectPoints()
    generateMoves()
    simulateMoves()
    selectBestMove()
    executeSelectedMove()
    updateTabuList()
    if stagnation then
        moveSize++
    if improving then
        moveSize⟵ 1
```

The local search framework that we propose for lattice fit problem, is based on Kangaroo (Newton et al. 2011), a

CBLS system that provides maintains necessary invariants and constraints. The procedure is given in Algorithm 1. At each iteration, amino acid points are selected randomly if they are not in the tabu list depending on $movesize$. We implemented a generalized version of the $kinkjump$ and $crankshaft$ moves used in the literature of simplified PSP. We select the best move that minimized dRMSD. We initialize the structures by a greedy procedure. For each of the amino-acids we keep assigning the points in the lattice that minimizes the dRMSD value. The procedure backtracks whenever there is a violation in the constraints. This initialization procedure produces initial structures with lower dRMSD values when compared to random initialization.

Experiments

We ran our experiments on a cluster of computers. Each node in the cluster is equipped with Intel Xeon CPU X5650 processors @2.67GHz, QDR 4 x InfiniBand Interconnect. In Table 1, we report average dRMSD values of the lattice fits in the *’end’* column for the proteins taken from the PISCES web server for different algorithms. We used the 1198 proteins used in (Mann et al. 2012). We ran our algorithm for 1 hour for each of the proteins. Values for the other algorithms are taken as reported in (Mann et al. 2012). We also show the average initial dRMSD of our approach in the *‘initial’* column. From the reported values it is evident that our method finds significantly better lattice fits and it is also an indication of the goodness of lattice fits of native structures.

We also analyze the correlation of dRMSD value with different energy functions. We take into consider three different energy functions: i) 20×20 energy function, bre (Berrera, Molinari, and Fogolari 2003), ii) 20×20 energy matrix, mj (Miyazawa and Jernigan 1985), and iii) basic hydrophobic-polar model, hp (Lau and Dill 1989). All these energy functions are used extensively in the literature of simplified protein structure prediction. We use our algorithm to minimize the dRMSD of candidate structures and report values of energy functions at each iterations for one hour. From the values of the energy functions of the structures and respective dRMSD values, we then try to find the correlation between the terms. In Fig. 1, a plot is shown for the protein 1A6M. In ideal case, with the decrease in the dRMSD value, there should be a decrease in the energy function value. But we see that they show either negative correlation (negative slope) or no correlation (flat). We perform this analysis for a limited number of proteins from the benchmark set. However, we can conclude that these proteins show a very low correlation of dRMSD values with the energy functions. The implication is these energy functions and the simplified energy model does not provide adequate discriminatory information to find native structures. The reason behind this is mainly due to the fact that these simplified methods do not take into consider the secondary structure informations.

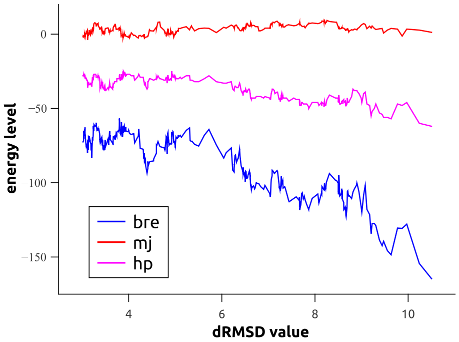

Figure 1: Plot of the values of different energy functions against dRMSD values for the protein 1A6M.

Lattice	(Park and Levitt 1995)	LatFit	Our Approach	
Type		(Mann et al. 2012)	initial	end
Cubic	2.34	2.08	2.86	**1.95**
FCC	1.46	1.34	2.03	**1.28**

Table 1: Comparison of average dRMSD values produced by different approaches with our approach

Conclusion and Future Work

In this paper, we propose a local search framework that produces state-of-the-art results for the lattice fit problem of real proteins. This confirms the effectiveness of using discrete lattices in PSP. In addition to this, we also analyze different simplified energy functions and their effectiveness to find better structures in terms of dRMSD values. In the analysis part, the test of effectiveness of the energy functions is limited to a few proteins only. We wish to provide a detail analysis for all the proteins to get a comprehensive picture. Furthermore, we believe that we can improve the fit by selecting points intelligently rather than randomly.

References

Berrera, M.; Molinari, H.; and Fogolari, F. 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* 4:8.

Lau, K. F., and Dill, K. A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22(10):3986–3997.

Mann, M.; Saunders, R.; Smith, C.; Backofen, R.; and Deane, C. M. 2012. Producing high-accuracy lattice models from protein atomic coordinates including side chains. *Adv. Bioinformatics* 2012.

Miyazawa, S., and Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18(3):534–552.

Newton, M. A. H.; Pham, D. N.; Sattar, A.; and Maher, M. J. 2011. Kangaroo: An efficient constraint-based local search system using lazy propagation. In *CP*, 645–659.

Park, B. H., and Levitt, M. 1995. The complexity and accuracy of discrete state models of protein structure. *Journal of Molecular Biology* 249(2):493 – 507.