Reconstruction a Protein 3D Structure From It Contact Map: Tools Comparison

Hamed J. Al-Fawareh[a], Nawal T.Alomoush[b], Jehad Q. Odeh[b]

[a]Faculty of Science and IT
Zarqa Private University, Zarqa
Jordan
Tel: 05-3821110
E-Mail: fawareh@zpu,edu.jo

[b]Faculty of Information Technology
Al al-Bayt University
Jordan
Tel: 02-6297000
E-Mail:jehad@aabu.edu.jo

ABSTRACT

Reconstruction a protein 3D structure using its contact map is not less than revolutionizes molecular biology. Recently, there are many research efforts that provide guidelines for protein contact map prediction; these efforts used machine learning approaches such as neural network and distance geometric. One of the approaches to help biologists is applying a software technique. As the consequence there are many categories of tools that have been developed to incorporate this technique. This paper analyses several predicting protein 3D structure tools. These tools are built to help to understand and predict a protein 3D structure. The paper briefly discusses the advantages of these tools; it also, gives the disadvantages of the existing tools, and, finally, talks about the proposed reconstructing a protein 3D tool.

Keyword:
Distance Geometry, Embedded Algorithm, Protein structure.

1.0 INTRODUCTION

Bioinformatics have been applied in many difficult, complex applications, and in different environments (Plero, 2001). Traditional experimental techniques for deriving macromolecular structure data are X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and electron microscopy, this method give data as a set of Cartesian coordinates representing the position of the atoms in these structure (Philip 2003, Wikipedia 2008),, But these methods remain slow and laborious and don't scale up to current sequencing speeds. Furthermore, using experiments to determine how protein function is daunting task, so that predicting the 3D structure of protein from liner sequence of amino acids is an interesting topic for computer scientists a lot, because it is fundamental open problem in computational molecular biology.

Each protein may contain thousands of atomic in different shapes, a fact which makes it helpful to automatically predict a protein through software tools. These tools for replacing a tradition experiment technique. This problem becomes even more complicated when the developer uses a complicated protein. Contact map help developers by giving them information about the protein system. This information includes, distance map, which is created by contact map, where a distance matrix computed to produce the Boolean values by used a pre assigned threshold value T. Distance map D is a N×N matrix where N is the number of residues in a protein and Dij is the distance between coordinate of the α carbon in two residues i and j which measured in Angstroms Å. Two residues i and j in a protein are come in contact with each other if the 3D distance Dij is less than or equal to some threshold value (Vassura 2008).

For the past few years, several tools have been developed in order to help predicting a protein 3D structure to understand protein functionality. In this paper we will highlight several tools. Developers build these tools for predicting the protein structure, and, each tool manipulates the protein under the protein structure activity. This paper will, also, give a brief discussion about the advantages of the tools; in addition, it talks about the disadvantage of the exiting tools, and proposes a new prediction tool called "reconstructing a protein from its contact map using matlap".

2.0 PROTEIN STRUCTURE AND ITS CONTACT MAP

Most of essential structure and functions of cells is refereed to Proteins. Proteins plays a vital role in keeping the body working properly. For example, they are used to support the skeleton, control sense, move muscles, digest food, and defend against infections and process emotions. There are more than 100,000 proteins that come in all shapes and sizes; however, they are all made up of the same set of 20 amino acids order in different way, its

primary sequence. The structure of a protein is determined by the folding of this primary sequence (Mireille 2006).
Any consideration of protein function must be grounded in an understanding of protein structure. A fundamental principle in all protein science is that protein structure leads to protein function, and protein functions are divers, so its no surprise that protein structures are also divers (Jorge 1999).

Contact map is a great interest for its application in fold recognition and 3D structure determination. A contact map is representation tool of the protein 3D structure. Traditionally, contact map is creates from distance map where a distance matrix computed to produce the Boolean values by used a pre assigned threshold value. Contact map C for a protein sequence with N residues is N×N asymmetric Boolean matrix whose element Cij=1 if residues i and j are contact and Cij=0 otherwise (John 1999). The contact map provides useful information, contacts represent certain secondary structure and it captures non local interaction giving clues to its tertiary structure (Jorge 1999).

3.0 TOOLS FOR RECONSTRUCTING A PROTEIN 3D STRUCTURE FROM CONTACT MAP

Vassura et al. (Vassura 2008) produce a software tool for reconstructing a protein 3D structure form contact map. The tool based on distance geometry which, finds a set of three dimensional coordinates consistent with some given contact map of threshold t. The contact map of a given protein is a binary matrix CM such that CM[i,j] = 1 iff the Euclidean distance between residues i and j is less than or equal to a pre-assigned threshold t. The tools divide system into two phases, the first phase; to generate a random initial set of 3D coordinates. This phase used metric matrix embedding algorithm to obtain good starting coordinates, before that they splitting the initial contact map in sub matrices. The sub matrices are then separately used to create sets of coordinate then merged it to give an initial solution. The merging procedure used rotation and translation to decrease number of error. While the second phase refines the set of coordinates by applying correction and perturbation procedure. The refinement applies until the set of coordinates is consistent with the given contact map or until a control parameter ε becomes 0. The control parameter ε has initial a positive value and it is decremented every some amount of refinement steps. If it reaches the 0 value before a consistent set of coordinates is found, then a new random initial set of coordinates is generated; ε is initialized again to a strictly positive value and the refinement procedure re-starts from the beginning. This phase applies iteratively two local techniques to obtain a new set of coordinates more consistent with the given CM in this step correction procedure doesn't add new errors to the coordinates set but eventually reduces the possibility to move some coordinate not yet well placed residue.

The tool shows that contact maps computed using threshold values greater than those commonly used for distances allow better 3D structure recovery than those computed at lower thresholds (7-9 Å). Repeated application of their method show that the contact map thresholds rang from 10 to 18 Angstrom allow to reconstruct 3D models that are very similar to the protein native structure. The disadvantage of this method apply on just a set of protein and ignore others protein in PDB which may be its more important.

Jing hu et al. (Jing 2002) present techniques describe how data mining can be used to extract valuable information from contact map and focus on discover an extensive set of non local dense patterns and compile a library of such non local interaction, and cluster patterns based on their similarities and evaluate the quality.
This tool used contact map to discover 3D structure by test each two amino acid to determine 3D distance by coordinate of α carbon atom. A pairs of amino acid in contact if distance less than threshold value =7 Å. The method used in this tool is divided into four stages:

- mining dense patterns
- pruning mined patterns
- clustering the dense patterns
- Integration of these patterns with biological data.

In the first stage they scan the DB of CM with 2D slide window. The tool used different window size to capture denser contact close to diagonal. The second stage extracted and isolated the pattern less dense and less distance from the diagonal by weighted the minimum density and verifying window size. Also this stage pruned redundant pattern by using slide window to capture all possible area in a matrix.
In clustering stage, the pattern generated into groups of similar interaction by used agglomerative clustering method. To find non local interaction it calculated a distance between each pair of pattern and between each pair of cluster, before they start clustering. This stage determined threshold for cluster. Then compare all pair of cluster and mark the closest. If the distance between two clusters is less than threshold t merged them into a single cluster. Finally, return to the first stage to continue the clustering. If the distance between the closest pair is greater than certain threshold, the clustering stops.
Their experiments used non redundant set of 2702 proteins from PDB, binary contact maps were generated using several contact thresholds. They discovered 9929 dense patterns in sliding window. The tool results showed that they can give 35% accuracy and 37% coverage for protein structure. The results are encouraging, but it's still far from providing sufficient accuracy for reliable 3D structure prediction.

Pollastri and Baldi, (Pollastri 2002) used a Neural Network to predict protein contact map and find its 3D structure. The tool focus in grained contact map prediction. The approach concentrates on find a 3D structure from liner sequences of protein. The major task in this approach is to

propose and verify precise and robust adaptation rule to predict contact map.
The approach taken was to extract data from PDB. Then choose the proteins have a single chain with number of amino acid less than 50, because of the difficulty in Neural Network to training with long chain of protein.
This tool used distance formula to compute distance matrix and normalize the distance matrix by convert all the distance into (0, 1). In addition, it used a set of threshold value to extract a pair node is in contact. We can summarize the approach described in this tool by four different neural networks to get contact map as follows:

- Back propagation neural network
- Learning vector quantization neural network
- Radial basis function neural network
- Reinforcement network
-

The tool used 20 amino acids as inputs and output scheme. It proposed an easy input encoding scheme which used 5 bit to encode each amino acid and used fixed length of protein. The approaches keep global information to get better prediction. The disadvantages of this approach are time expensive and limitation on the length of protein sequences. The advantages of this approach it has higher resolution than just one contact map.

Jorge and zhijun, (Jorge1999), developed a tool based on Gaussian smoothing to develop an efficient and reliable code to solve the distance geometry problem in protein structure. The algorithm in this tool work with the sparse set of distance constraints while other algorithm work for distance geometry which tend to work with dense set of constraints.
The problem in this approach is the distance geometry for determination of protein structures. The distance geometry is specified by a subset of all atom pairs. The distance between i,j atoms in a subset determine the lower and upper bounds to find a set of positions of the specified atoms. This problem is formulated in terms of finding the global minimum of the function.
The approach in this tool used Gaussian smoothing to transform function F into smoother function with fewer minimizers. The optimization algorithm applied to the transformed function and continuation techniques. The optimizations are used to trace the minimizers of the smooth function back to the original function. The advantage of this approach is work per iteration and proportional to a subset for sparse distance. The computational experiments show that the tool provides an efficient approach to the solution of the distance geometry problem and show an interesting issue is the dependence of the structures on the distance data.

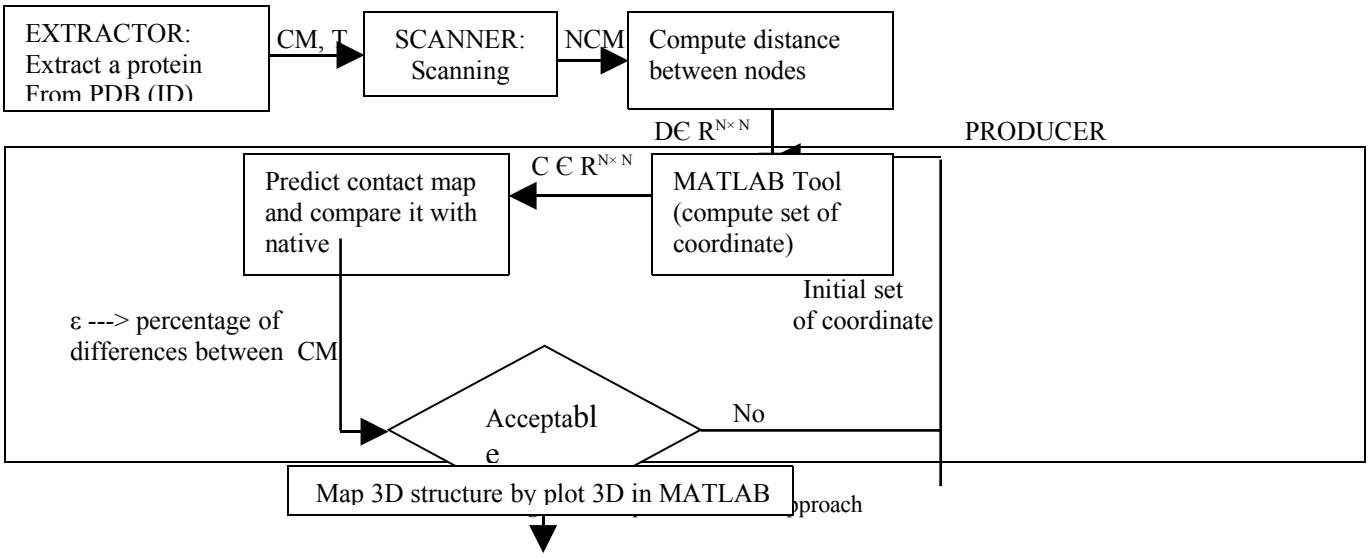
3.1 Proposed Reconstructing Tool

Reconstructing a protein from its contact map using matlap is proposed tool to assists in enhancing constituents and predicts a protein 3D structure. Further more, its, provides information that helps users to correct faults in the protein 3D structure by shifting and rotation. Thus, it helps to make a protein 3D structure is less fault. This section briefly describe the proposed tool finds a set of three dimensional coordinates consistent with some given contact map of threshold.

The proposed reconstructing tool contains three modules as shown in Figure1. The EXTRACTOR MODULE reads the protein from the PDB, and constructs a protein contact map table. The SCANNER MODULE accepts the contact map CM as an input, and produces a new contact map NCM. Scanning the contact map for a protein is much more reliable to predict the more important areas of the contact map which we call it NCM. This process based on prediction quality more than quantity of contacts. This process helps to predict a protein 3D more reliable. In all previous studies shows that predict 50% of the contact map with 5% errors much reliable than predicting 100% of contact map with 25% errors (Vassura, 2008, Gutpa 2004).

The proposed algorithm reprocessing all contact residues and assumes that two atoms i and j are in contact if and only if they share a high number of neighbors, i.e C(i , j)=1 are in contact and share with K neighbors that are closest to a specific point. The PRODUCER MODULE produces distance matrix procedure, which find a possible set of distance between nodes Dϵ $R^{N\times N}$ depending on threshold value range from 10 to 18 A˚ consistent. In addition by using some literature survey about the physical conformation of the proteins this module can know the average distance between adjacent alpha carbons D[i,j] which is 3.84 A˚ i.e | i – j |=1. also, the other distance can be obtained from classified protein.

To compute a 3D point this module used a consistent distance matrix D with supported by MATLAB tools. These tools are used to compute a set of three dimensional coordinates. These coordinates are the best 3D representation for the distance matrix D. The module predicts the initialize starting coordinates randomly. This module iteratively applies some procedure to the current set of coordinate to extract new contact map and compare two contact map (native CM with predict CM) to find number of differences. The module accept the result with error percentage ε is less than 0.3 otherwise a new random initial set of coordinate is generated and the procedure restart from beginning by using MATLAB tool. Finally, when the consistent set of coordinates is found the module used plot 3D function from MATLAB to predict a protein 3D structure. Also, the module does some translate or rotate prediction 3D structure to obtain the most accurate protein 3D structure.



4.0 CONCLUSION

Reconstructing a protein structure is one of the approaches that have been used in folding a protein 3D structure. Reconstructing a protein 3D structure uses distance geometry and neural network approaches to achieve the predictions activity. Furthermore distance geometry is the process of mathematical properties that can be derived from distance value between pairs of point. The distance geometry method is used for extracting information from contact map systems in order to help prediction of a protein 3D structure. In the past few years, several protein prediction tools have been produced. In this paper we have compared the existing predicting protein 3D structure tools. In addition, we have given the disadvantage of these tools. Furthermore, we have discussed about the proposed tool, which is called "reconstructing a protein from its contact map using matlap".

REFERENCES

Plero., F, Osavaldo. O, Rtta.C and Alfonso. V, (2001)"Prediction *of Contact Maps With Neural Network and Correlated mutation*", biology department, Univ. Bologna, Italy, 2001, protein engineering vo.14, no.11, pp.835-843.

Vassura. M, Margara.L, and others, (2008), "*Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps*", CS department, bioinformatics group, Univ. Bologna, Italy, 2008.

Vassura. M., Luciano. M, Filippo. M, Pietro. D and piero.F, (2008), "*Reconstruction of 3D Structures From Protein Contact Maps*", CS department, bioinformatics group, Univ. Bologna, Italy, 2008.

Wikipedia,(2008), "*The free encyclopedia*", Bioinformatics web, modified on 26 March 2008.

Philip and Helge. W,(2003), "Structural Bioinformatics"(hand book), san Diego supercomputer center, Pharmacology department, Univ. California san Diego, wiley-liss publisher; 2003.

Mireille G (2006). "Distance geometry, helix packing, and contact map congruency advisors". Queen's university; Australia.

Jorge.M, zhijun .W. (1999), "Distance geometry optimization for protein structure". Journal on Global optimization, 15,pp.219-234,1999.

Jing. He, Ming. Z and Zhen. (2004),"*CS 6890 Project Report* ", 28/4/2004, pp 2-23

John. M, (1999) " *Predicting protein three-dimensional structure*", Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850, USA; 1999,vo.10, pp583–588.

Jing. H, xiaolan. S, Mohammed. Z, yu. S and Chris. B, (2002) "*Mining Protein Contact Map*", BIOKDD02: workshop on data mining on bioinformatics, with SIGKDD02mConference 2002: pp 3-10.n.

Gutpa N, Managal N, and Biswas S. (2004) "*Evolution and similarity Evaluation of protein structure in contact map space. Proteins: structure, function*", Bioinformatics 2004;95920:196-204.

Pollastri G, and Baldi P. (2002) "*Prediction of contact maps by recurrent neural networks architectures and hidden context propagation from all cardinal corners*". Bioinformatics, 2002;1:1-19.