Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 95-106

http://www.ripublication.com/jcib.htm

A New Algorithm for Reconstruction a Protein 3D Structure from Contact Map

Hamed J. Al-Fawareh

Faculty of Science and Information Technology
Zarqa University, Zarqa, Jordan
E-mail: fawareh@zpu,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 algorithm and experimental result.

Keyword: Distance Geometry, Embedded Algorithm, Protein structure.

Introduction

Bioinformatics have been applied in many difficult, complex applications, and in different environments (Plero et al., 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 and Helge, 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 et al. 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 matlab". A result and discussion also provided.

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 and Zhijun, 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 and Zhijun, 1999).

Research Background

Vassura et al. (Vassura et al. 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. (Hu et al., 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, 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, (Jorge and zhijun, 1999), 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.

Algorithm Description

Reconstructing a protein from its contact map using Matlab is a method 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 algorithm which finds a set of three dimensional coordinates consistent with some given contact map of threshold.

The algorithm contains three modules. The SCANNER 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 this method shown in algorithm 1. 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 et al, 2008, Gutpa et al., 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. Shortest path function is used to obtain the accurate set of distance which satisfies the triangle inequality. FSOLVE function from MatLap system is used generate accurate three dimensional coordinates D_{ijk}.

Correct procedure Algorithm 2 takes the set of coordinate as input to find the possible radius mobility of some not yet well placed residues. Furthermore the procedure, move these residues to new position with new coordinates using mobility and correct direction procedures. This process iteratively applies until control parameter Q becomes zero (Q is number of tray to correct coordinates) or until ε percentage of error becomes acceptable.

Algorithm(1): Scanning Procedure

```
Scan_CM(CM)
%take native CM of n node as input
for i=1 to n

 for j=i+1 to n
 count=0
 for k=1 to n
  if i and k contact && k and j contact
   count=count+1
     if i and j is contact && count< 10
 or i and j is not contact && count> 20
   then node i and j is contact in NCM
            Break and take anther node
Return NCM
```

Algorithm(2): Correct procedure

```
coorect_coordinat (CM,C,T)
For i=1 to n
 For j=1 to n
  If CM (i, j)contact and D(i, j)
greater than Threshold
          OR
    CM(i, j)not contact and D(i, j)less
than or equal Threshold
  R = mobility (i)
    New Coordinate (i)
=correct_direction (i)
End
```

Algorithm(3): Mobility

```
Mobility (i)
For j=1 to n
 If CM (i, j) contact and D (i, j) less
than or equal Threshold
         D1=max (D1, D (i, j))
      Else
If (CM (i, j) not contact and D (i,j)
greater than Threshold
      D0=min(D0, D (i, j))
      M (i) =min (D0-T, T-D1)
End
```

Algorithm(4): Correct Direction

```
correct_direction(i)
for j=1 to n
   if CM(i,j)contact and D(i,j)greater
              than Threshold
           OR
        CM(i,j)not contact and
D(i,j)less than or equal Threshold

     if CM(i,j)contact
           V= V - C(i)-C(j)/D(i,j))
              else
           V=V + C(i)-C(j)/D(i,j))
      k =C(i)+ ( V *(r/norm(V)))
End
```

The main objective of this module is to detect the dense areas that form the basic functional areas in the contact map. Looking for the dense area is an important step that will improve the performance of the predicting 3D structure of protein from it CM.

Scanning module used SCAN_CM Procedure to reprocess all contact residues. This module 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 less than 10 neighbors that are closest to a specific point or C(i , j)=0 are consider in contact if they share with greater than 20 neighbors that are closest to a specific point. Find a number of neighbors will increase the probability of selected contact residues. In other cases; this approach will decrease the probability of wrongly predicted contact pairs outside the dense area.

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 by count_distance procedure in the otherwise the distance of nodes which are not contact set as random number depending on threshold value. Shortest_path_dist procedure is used to obtain the best set of distance which satisfy the triangle inequality, (i.e. for all i , j , k node from 1 to n, distance [i , j] <= distance [i , k] + distance [j , k]).

Algorithm(5): count distance

```
count_distance(Thresholde,i,j)
%the set of distances in this procedure
are taken from literature survey
if i equal j
   x=0
if |i-j| equal 1
   x=3.8
if |i-j| equal 2
   x=6+ random(1,1)
if |i-j| equal 3
   x=7 +random(1,1)
if  |i-j| greater than 3
   x=(0.91-(T/100))*T
return X
End
```

To compute a 3D point the Producer Module used a consistent distance matrix D with supported by nonlinear_coordinat procedure. This procedure used FSOLVE function from MATLAB tools, FSOLVE finds a root (zero) of a system of nonlinear equations, FSOLVE calls compute distance function which accepts random set of coordinates (vector x) as starting points also a distance matrix D as parameter to solve

nonlinear system using Euclidian distance equation between every pairs of amino acid protein sequence are selected to solve the nonlinear system. This process applies iteratively until the best set of three dimensional coordinates fit for distance matrix D.

Algorithm(6):nonlinear coordinate

```
nonlin_coordinat(D)
C is coordinate matrix
For a=1:10
x0=random set of coordinate take as
starting point
C = FSOLVE(@(x)
compute_coordinat(C,D),x0)
If set of coordinates accept
Break and Return C
End
```

The procedure accepts the results if the root(zero) of the system is found, otherwise a new random initial set of coordinate is generated and the procedure restart from beginning by using MATLAB tool.

The Producer Module used new_contact_map and compare_contact maps procedures to the current set of coordinate to extract new contact map and compare two contact map (native CM with predict CM) to find error percentage.

Algorithm(7): New contact map

```
New_contact_map(CM,Coordinate,Thres
hold)
for i=1 to n
   for j=i to n
         if i equal j
               D(i,j)=zero and
NCM(i,j)contact

          Else
       D(i,j)= sqrt ((C(1,i)-
C(1,j))^2+(C(2,i)-C(2,j))^2+(C(3,i)-
C(3,j))^2);
if D(i,j) less than or equal Threshold
    NCM(i,j)is contact
else
 NCM(i,j)is not contact
End
End
```

Algorithm(8): compare contact maps

```
compar_contact_maps(NCM,CM)
  e=0
  for i=1 to n
   for j=1 to n
     if NCM(i,j)not equal CM(i,j)
        e=e+1
Return error=e/n*n
End
```

Corrector Module takes the producer module output then applies Correct_coordinate procedure to find the possible radius mobility of some not yet well placed residues and move these residues to new position with new coordinates.

Mobility procedure takes maximum distance between residue i and j if the two nodes are contact and minimum distance otherwise, to calculate radius of mobility of residue i.

$$D_0 = \min\{dij \mid dij > t \text{ and } CM[i, j] = 0\}$$
$$D_1 = \max\{dij \mid dij \leq t \text{ and } CM[i, j] = 1\}.$$

Then takes minimum distance between that i.e $Mi = \min\{D_0 - t, t - D1\}$. Correct_direction procedure is used to determine the direction of move of residue i without do any effect of correct residues. The module extracts new contact map depending on new correction and compare the two contact maps. In order to get an accurate result we process this step iteratively until control parameter Q becomes zero (Q is number of tray to correct coordinates) or until ε percentage of error becomes acceptable. If the consistent set of coordinates is found the module used plot 3D function from Matlab. The plot3 function displays a three-dimensional plot of a set of data points. plot3(X, Y, Z), where X, Y, Z are vectors or matrices, plots one or more lines in three-dimensional space through the points whose coordinates are the elements of X, Y, and Z. Plot 3D maps a protein 3D structure and allows to make some rotation on 3D structure to obtain the accurate structure. This approach focuses on choosing the threshold value for computing the contact map, which is effect on connects between the contact map and its 3D structure, not any threshold give accurate contact map which provides exact 3D structure. The experimental results show that the contact maps computed using threshold values (12-18) Å allow better 3D structure recovery than those computed at thresholds (7-9) Å. 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 it with the native 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.

Experimental Results

This paper presents experimental results that show the efficiency of our proposed method for predicting a protein structure. We took the list of proteins of different lengths related to the most popular classes from the PDB. For each protein in the selected list we generate different contact maps by changing the threshold value and

analysis the result when we scanning the contact map with a pre assigned threshold to show the accuracy of extract contact map from dense area instead of whole area. Protein 2IGD has a single chain and contains one α_helix and two parallel β_sheet. As shown in figure1 the result of 12 different contact map generated by changing the contact threshold from 7 to 18 Angstrom and percentage of error before and after used the correct procedure with average time. The analysis of the result show that the correct procedure try to decrease the percentage of error when it applies iteratively to find the best set of coordinates consist with the native contact map to predict the best fit 3D structure .

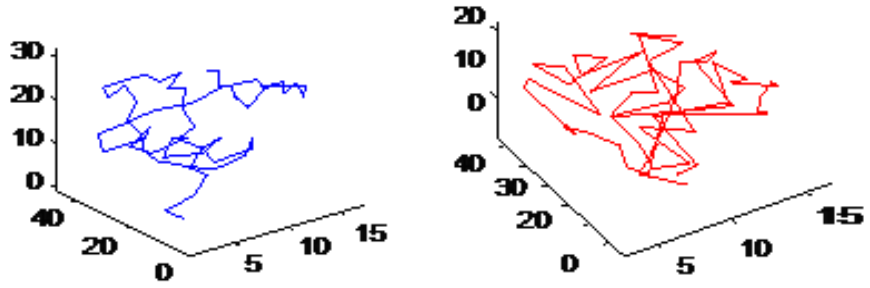

Figure 1: Recovery of 3D structure of protein 2IGD: native structure (blue) compared to a recovered structure (red), at threshold value 7.

The contact map computed with a threshold equal to 7 Angstrom does not contain enough global information of the protein structure and it similar to a huge number of protein contact map. The prediction structure is not clear compare with the native structure as shown in Figure 1. When the contact map is computed at a threshold of 16 Angstrom more features appear and the recovered 3D structure is more similar to the native one. This finding prompted us to do a search in the threshold space to optimize the percentage of error. We find that a better 3D reconstruction is obtained when a high threshold value is adopted (12 Angstrom or higher) when contact maps are computed.

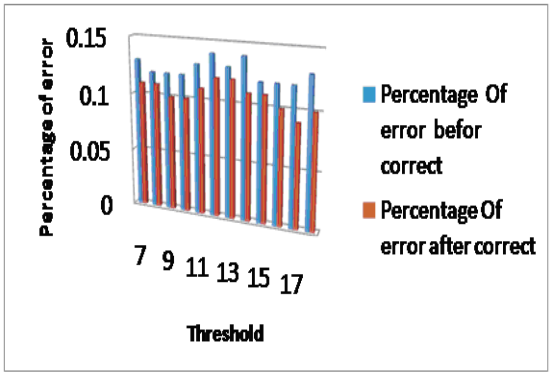

Figure 2: Threshold & percentage of errors of 2IGD protein.

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 the proposed algorithm.

References

[1] Plero, F., O. Osavaldo, C. Rtta and V. Alfonso, 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.

[2] Vassura, M., L. Margara, P. Di Lena, F. Medri, P. Fariselli and R. Casadio, (2008), "Fault tolerance reconstruction of 3D structure from protein contact maps. CS Department, Bioinformatics Group, Univ. Bologna, Italy.

[3] http://bioinformatics.oxfordjournals.org/content/early/2008/04/01/bioinformatics.btn115.full.pdf

[4] M., Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli and R. Casadio,2008), "Reconstruction of 3D Structures From Protein Contact Maps", CS department, bioinformatics group, Univ. Bologna, Italy, 2008.357-363.

[5] Wikipedia, 2008. The free encyclopedia. Bioinformatics Web, Modified on 26 March 2008.

[6] Philip, E.B. and W. Helge, 2003, Structural Bioinformatics (Hand Book). Supercomputer Center, Pharmacology Department, University California San Diego, Wiley-Liss Publisher, San Diego,.

[7] Mireille, G., 2006. Distance Geometry, Helix Packing, and Contact Map Congruency Advisors". Queen's University; Australia.

[8] Jorge, M. and W. Zhijun, 1999, "Distance geometry optimization for protein structure. J. Global Optimization, 15:219-234,1999.

[9] 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.

[10] Hu, J., X. Shen, Y. Shao, C. Bystroff and M.J. Zaki, 2002. "Mining Protein Contact Map", BIOKDD02: workshop on data mining on bioinformatics, with SIGKDD02mConference 2002: pp 3-10.n.

[11] 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.

[12] Pollastri, G. and P. Baldi, 2002. Prediction of contact maps by recurrent neural networks architectures and hidden context propagation from all cardinal corners. Bioinformatics, 1:1-19.
[13] http://gruyere.ucd.ie/papers/2002_ISMB.pdf