American Journal of Biochemistry and Biotechnology 4 (4): 375-384, 2008
ISSN 1553-3468

Improving Protein 3D Structure Prediction Accuracy using Dense Regions Areas of Secondary Structures in the Contact Map

[1]Amer F. Al-Badarneh, [2]Mohammad A. Khalil and [3]Mo'taz A. Al-Hami
[1]Department of Computer Information Systems,
Jordan University of Science and Technology, P.O. Box 3030, Irbid, 22110, Jordan
[2]Department of Basic Medical Sciences, Cellular and Molecular Neurobiology,
King Fahad Medical City, P.O. Box 59046, Riyadh 11525, Saudi Arabia
[3]Department of Computer Science,
Jordan University of Science and Technology, P.O. Box 3030, Irbid, 22110, Jordan

Abstract: Problem Statement: The protein folding problem is a fundamental problem in structural molecular biology. This problem describes how a protein is transformed from its primary sequence (i.e., amino acid sequence) into the three dimensional structure (3D structure) for this sequence that determines the function of the protein. The 3D structure of a protein can be represented using a square symmetrical binary matrix called contact map. The concept of contact map facilitates the transformation of the folding problem into a computational one, so various computational approaches use the contact map to predict protein secondary structures. Correlation mutation analysis is an approach that tries to study the mutated patterns that appear in the multiple sequence alignments, this approach predicts every pair of protein residues to be in contact or not independently of the other pairs. **Approach:** This study proposed an improvement over correlation mutation analysis to predict the secondary structures that exist in the contact map. The proposed method uses regions of the secondary structures instead of independent pairs as in the typical correlation mutation analysis; also it applies the analysis on the dense regions rather than the whole contact map. **Results:** The proposed method was implemented on proteins related to different classes (i.e., mainly alpha, mainly beta, mixed alpha beta and low secondary structures). The test proteins are extracted from the Protein Data Bank (PDB) of solved structures. The results show improvements of dense regions accuracy over correlation mutation accuracy and random accuracy. **Conclusion:** According to the amount of wrongly predicted contacts, the results show a large decrease in the wrongly predicted contacts in the dense regions analysis over correlation mutation analysis.

Key words: Bioinformatics, correlation mutation analysis, clustering, protein structure, protein folding, molecular biology, secondary structure prediction

INTRODUCTION

Proteins are necessary for us in enormous variety of different ways. Much of the body materials like muscles, cartilage, skin and hair are constructed from protein molecules. Also proteins play a vital role in keeping the body working properly. Undoubtedly proteins are the most important functional unit in the living organisms[1]. The function of a protein molecule is determined by its 3-Dimensional (3D) structure. The 3D structure of proteins can be determined biologically by X-Ray crystallography[2] or Nuclear Magnetic Resonance (NMR) techniques[3]. In X-Ray crystallography, the protein is crystallized, bombarded with electrons, which creates a diffraction pattern that determines the atomic structure of the protein. The electron diffraction pattern is used to calculate the coordinate of atoms based on the measured electron density. Some limitations associated with X-Ray crystallography include, inability to crystallize some molecules, crystallography is being laborious, limitation of resolution (i.e., nearly about 2.9 A°) and poor reproducibility (same sequence produce more than one structure under different experimental conditions).

In NMR spectroscopy, the molecules are exposed to a static magnetic field, causing the nuclei of atoms to vibrate. Then, the molecules are subjected to a second oscillating magnetic field, generating a characterizing

Corresponding Author: Amer F. Al-Badarneh, Department of Computer Information Systems,
Jordan University of Science and Technology, P.O. Box 3030, Irbid, 22110, Jordan

spectrum for all the atoms for each molecule which becomes a spatial atomic map (i.e., 3D structure).

A closer examination of the 3D structure of proteins led biologist to conclude that secondary structures contained within proteins essentially determine their 3D configuration in space[4]. Therefore, biologists directed their attention to characterizing secondary structures contained within proteins. To date, there exists a complete classification system of protein 3D structure based on the type of secondary structures present in proteins[5,6]. This classification system divides proteins into: mainly alpha secondary structures class, mainly beta secondary structures class, mixed alpha beta secondary structures class and low secondary structures class. The secondary structures contained within a protein forms in nature due to the spatial proximity (i.e., distance in space) of the central carbon (C_α) of the amino acid, typically measured in angstroms (A°). This numerical perception of the 3D structure of protein facilitates the transformation of the biological problem into a computational one[7]. Based on the aforementioned information, the question arose: "Can 3D structure of protein be determined by mining protein primary structure (i.e., amino acid sequence)?".

It is pertinent to state that with the completion of the first stage of the human genome project, namely "structural genomics", has lunched the second stage of the human genome project, namely "functional genomics". The later is concerned with the study of "proteomics", which is the study of proteins content of a cell, tissue, or organ any one time, to elucidate a more comprehensive understanding of the cell. Since the dogma "structure determines function" remains at large valid to date, the science of bioinformatics was inevitably pioneered to capitalize on the combined efforts and advances in structural biology and computational mathematics.

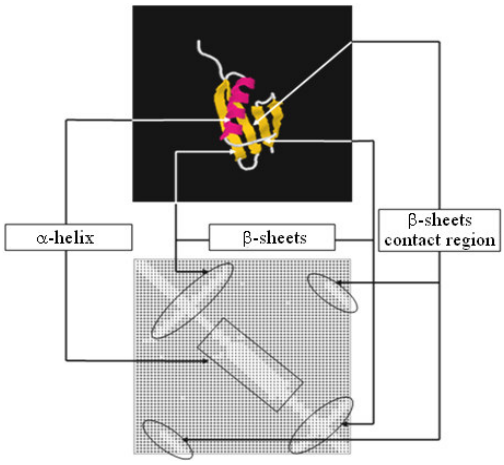

Fig. 1: Contact map for 2IGD protein

A contact map is a representation tool of the protein 3D structure[8]. Two residues (amino acids) a_i and a_j are in contact if their 3D distance (the distance between coordinate of the α carbon) is less than some threshold value t. So using this definition, every pair of amino acids is either in contact or not. A contact map C for a protein sequence with N residues is N×N binary symmetrical matrix. The entry of the contact map C_{ij} is 1 if the two amino acids i and j are in contact and 0 otherwise. Figure 1 shows the structure 2IGD Protein and its contact map.

The contact map shows the overall folding of the structure and gives useful information about protein secondary structures and it captures non-local interaction giving clues to its tertiary structure[9]. α-helices appear in the contact map as clusters along the main diagonal, while β-sheets appear in the contact map as clusters anti-parallel to the main diagonal. Since each residue is in contact with itself, the main diagonal appears in contact state. Neighboring residues with distance less than the predefined threshold t appear in contact beside the main diagonal. The contact map is formed due to the spatial proximity between amino acids in the space. The concept of contact map facilitates the transformation of the folding problem into a computational one, so various computational approaches and mathematical models can participate in analyzing and extracting rules related to this map.

Correlation mutation analysis is an approach that tries to study the mutated patterns that appear in the multiple sequence alignments, this approach predicts every pair of protein residues to be in contact or not independently of the other pairs. This study proposed an improvement over correlation mutation analysis to predict the secondary structures that exist in the contact map. The proposed method uses regions of the secondary structures instead of independent pairs as in the typical correlation mutation analysis; also it applies the analysis on the dense regions rather than the whole contact map. The proposed method was implemented on proteins related to different classes. The results show improvements of dense regions accuracy over correlation mutation accuracy and random accuracy. According to the amount of wrongly predicted contacts, the results show a large decrease in the wrongly predicted contacts in the dense regions analysis over correlation mutation analysis.

Correlation mutation analysis[10,11] is used as an indication of probable physical contact in the 3D structure. The mutated and conserved patterns in the multiple sequence alignments provide evidence of structural constraints. This analysis makes assumption

on that, residues in physical contact have a correlation mutation behavior. This correlation mutation behavior is measured through a correlation coefficient value[12]. The correlation mutation analysis consists of the following four steps that extract the correlation tendency of pairs of residues to be in contact from the multiple alignments (Fig. 2).

Step 1: Building the multiple alignments: To build a good multiple sequence alignments, homologous protein sequences for the test protein were collected by using PSI-BLAST with default parameter and non redundant database. From the homologous protein sequences, the sequences with sequence identity between 25% and 85% were chosen, so more distantly related sequences that belong to the same family and hold enough amount of knowledge about the structure were selected. Then these sequences aligned using ClustalW to generate the multiple alignments[13].

Step 2: Creating the exchange matrix: Exchange matrix is a two dimensional square matrix that describes conservative mutation of an amino acid at each sequence position through the other sequences in the multiple sequence alignments. Using this matrix, all mutations that are created by the residue in a specific position through the sequences, can be recorded. The size of the exchange matrix equals to M×M, where M is the number of sequences in the multiple alignment.

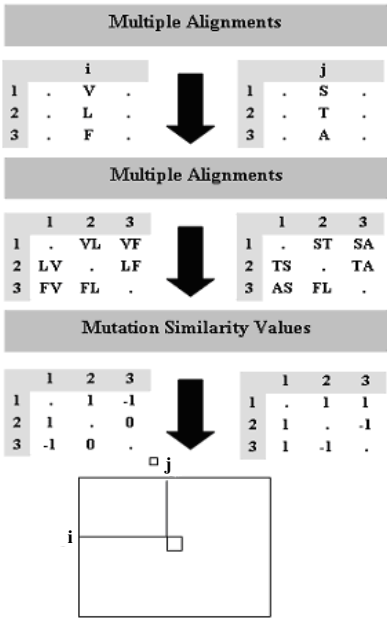

Fig. 2: Correlation mutation analysis example

Step 3: Estimating the mutation matrix: After building the exchange matrix, another two dimensional square matrix is constructed to measure the mutation behavior similarity S(i, k, l) from amino acid k to amino acid l at position i. As the value of the mutation increases to more than zero, it is more likely to consider this mutation happened rather than by chance. The similarities between 20×20 amino acids are taken from statistical scoring matrix Blosume62.

Step 4: Calculating the correlation coefficient: The correlation coefficient measures how strongly one attribute (residue) implies the other attribute, based on the available data in the sequences[14]. The correlation coefficient r_{ij} value between any two positions can be calculated as:

$$r_{ij} = \frac{1}{M}\sum_{kl}\frac{\left(s(i,k,l)-\langle s_i\rangle\right)\left(s(j,k,l)-\langle s_j\rangle\right)}{\sigma_i\sigma_j}$$

Where:
$\langle s_i\rangle$ = The mean value of similarity matrix for residue i, σ_i = The standard deviation of S(i, k, l) value about the mean value $\langle s_i\rangle$.

The higher correlation coefficient value r_{ij} between two residues positions is interpreted as more tendency between the two residues to be in contact[10]. A two-dimensional matrix is created having all possible r_{ij} values. So when the correlation coefficient r_{ij} is above a chosen threshold, the associated pairs of amino acids are in contact. The predicted contact map resulted from all pairs that have correlation coefficient values higher than the predefined threshold.

For calculating the accuracy for this prediction, comparisons are made between the predicted contact map and the experimentally derived contact map. It is worthwhile to mention that completely conserved residues (residues which don't mutate in the multiple alignments) and positions with more that 10% gaps in the multiple alignments are excluded from the correlation mutation analysis, since they represent ambiguous data (i.e., no enough mutation exists or contain many gaps in the alignments) and no clear decision can be made about them.

Correlation mutation analysis suffers from two main drawbacks. The first drawback is the unfair studying of the whole contact map regions. It does not take into consideration the probable dense areas that include the secondary structures like α-helices and β-sheets. Looking for the dense areas that contain secondary structures is an important step that will improve the performance of the correlation mutation analysis. This can be done by weighting the regions

according to the probability of the presence of secondary structures, so the regions that are more probable to contain these secondary structures will contain dense contact clusters, while others will contain sparse contacts. The second drawback in the typical correlation mutation analysis is the increasing amount of non contact pairs predicted as contact (low accuracy), especially for low threshold values for the correlation coefficient r_{ij}. These pairs will decrease the accuracy of the results. So if we concentrate to increase the prediction accuracy to hit true contact pairs and at the same time to decrease the prediction to choose none true contacts, a larger accuracy than what is available in typical correlation mutation analysis can be obtained. In this study, correlation mutation analysis will consider only dense regions that contain secondary structures and use low thresholds values for the correlation coefficient r_{ij}.

MATERIALS AND METHODS

Protein 3-D structure prediction directly from amino acid sequences still remains as an open and important problem in life sciences. The bioinformatics approach first predicts the protein secondary structure which represents an 1D projection of the very complicated 3D structure of a protein. Predicted secondary structure provides information about the regions in the amino acid sequence that contain a certain secondary structure. The goal of secondary structure prediction is to classify a pattern of residues in amino acid sequences to a corresponding secondary structure element: an α-helix, β-sheet, or coil[4]. Table 1 shows the predicted secondary structure for 2IGD protein. Different approaches developed to predict the protein secondary structure[15]; their accuracies range between 70-75%.

The main objective of the proposed method is to detect the dense areas that contain the secondary structures instead of using the whole contact map. These dense areas include regions of α-helices and regions of β-sheets, which form the basic functional areas in the contact map.

The importance of this step is that correlation mutation analysis will be used basically for looking for dense regions (i.e., the regions of β-sheets) and then we use the correlation mutation analysis inside these regions. This process will increase the probability to hit true contact residues and at the same time it will decrease the probability of wrongly predicted contact pairs outside the dense regions. It is worthwhile to mention that some knowledge about the dense areas is extracted from the predicted secondary structure for a protein. From the predicted secondary structure, the positions in the sequence that contain amino acids that form a particular secondary structure (i.e., α-helix or β-sheet) are recorded.

Searching α-helices regions: α-helix was first described by[16]. It is a classical element in the protein structure and one helix can have more influence on the stability and functionality of the protein. The internal structure of α-helix looks like spiral and contains 3.6 residues per complete 360° turn, where there is a hydrogen bond between CO of residue n and NH of residue n + 4. The structure of the α-helix as a spiral is interpreted in the contact map as a thick cluster of contacts along the diagonal, so it is favorable to restrict our searching of the α-helices regions in the contact map to narrow width regions along the diagonal.

The width of these regions is four like the number of residues that can be in contact per turn. The length of α-helix region is taken from its length in the predicted secondary structure. Figure 3 shows the window that captures the α-helix region in the contact map for 2IGD protein. The left part shows the window covers the region that is most candidates to cover the α-helix and the right part shows the original contact map. It is clear to notice that the window covers the α-helix correctly. Also we notice that the length of the window is larger than the length of the α-helix in the original contact map, because we extend window length on both sides to guarantee that α-helix falls inside the window.

Searching β-sheets regions: β-sheets secondary structure is formed by extended consecutive amino acids. In the β-sheet, hydrogen bond occurs between CO and NH groups from residues in strand formed by other parts in the polypeptide[17]. In anti parallel β-sheets adjacent strands are located in opposite directions. Detecting β-sheet is more difficult than detecting α-helix, since it appears anti parallel (orthogonal) to the main diagonal and it has no certain position in the contact map. All we know that it appears between the start position and end position of the β-sheet in the predicted secondary structure. The detection of β-sheet consists of the following steps:

Step 1: Identifying the regions of β-sheets: Predicted secondary structure sequence defines the regions of α-

Table 1: Part of predicted secondary structure for 2IGD protein

Position of amino acid	1	2	3	4	5	6	7	...	28	29	30	…	58	59	60	61
Protein structure	M	T	P	A	V	T	T	...	A	E	T	...	T	V	T	E
Predicted secondary structure	C	C	C	C	C	E	E	...	H	H	H	…	E	E	E	C

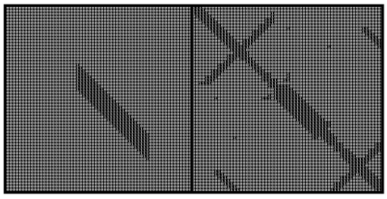

Fig. 3: α-Helices for 2IGD protein. The left part shows the position of α-Helices and the right part shows the original contact map

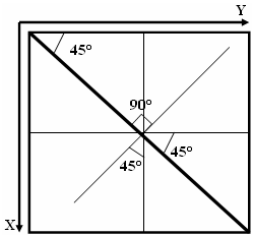

Fig. 4: Creating the initial line for the sliding window

helix and β-sheet, so it is simple to identify the positions at which β-sheet strand starts and finishes in the secondary structure. This region will be subject to the analysis using correlation mutation measurement to find the right position of the β-sheet.

Step 2: Creating initial sliding window: A sliding window is constructed as a sensor that searches the right position of the β-sheet in the contact map. To construct this window, an initial sliding window is constructed in the center of the region defined in Step 1. The center of this window is defined as:

$$\text{Cluster Center (cc)} = \frac{(\text{Start Pos} + \text{End Pos})}{2}$$

From the position (cc), we initiate a line orthogonal on the diagonal. This line is restricted in the area between start position and end position with some extension on both sides to guarantee that the β-sheet inside this region. The goal of this step is to create a sliding window that has a shape similar to the β-sheet. Figure 4 shows the orthogonal line on the diagonal, the line deviates from x-coordinate by 45° down the coordinate, so the slope of the line is -1. This line is initiated in the center of the β-sheet candidate region as shown in the Step 1. An equation of the line with slope m going through a given point $p(x_1, y_1)$ is $y - y_1 = m (x - x_1)$. The point of the Cluster Center (cc) has the coordinate (cc, cc) since it is located on the diagonal

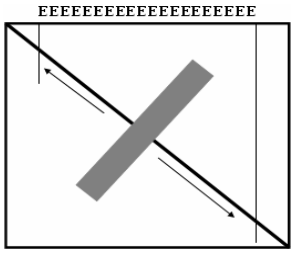

Fig. 5: Looking for the β-sheet

and the equation of the line that goes through (cc, cc) point and has slope value equal to -1 is given by:

$$y + x - 2cc = 0 \quad (1)$$

Any point located between start position and end position satisfying Equation 1 will participate in the orthogonal line. The constructed line forms the bases of the sliding window. Using this line, we will add all neighbors' points, which have distance less than a given threshold d, to the line to create the sliding window. The distance between the point $p(x_1, y_1)$ and the line L of equation $Ax+By+C = 0$ is given by Eq. 2 and the result of compensating Equation 1 in Eq. 2 is shown in Eq. 3:

$$d = \frac{|Ax_1 + By_1 + C|}{\sqrt{A^2 + B^2}} \quad (2)$$

$$d = \frac{|x_1 + y_1 + 2cc|}{\sqrt{2}} \quad (3)$$

In this study, all the neighbors with d<2 were added to the initiated line to form the sliding window. Form the experiments threshold 2 is found very suitable to cover the region of β sheet, other thresholds like 3, 4 can be used but 2 is enough to cover the region of β sheet. The sliding window now can be used as a sensor for the β-sheet.

Step 3: Looking for the β-sheet: The constructed sliding window is now used to look for the β-sheet in the candidate sheet region. The sliding window is used to scan the region from start position to the end position of the β-sheet in the predicted secondary structure. The sum of correlation coefficient values for the residues pairs inside the sliding Window is calculated to get the region of highest correlation coefficient sum (Fig. 5). The region that has the highest correlation coefficient sum is chosen to be the region that represents the real β-sheet.

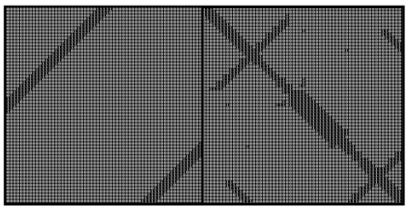

Fig. 6: β-sheet regions for 2IGD protein. The left part shows the position of region that has the highest correlation coefficient sum and the right part shows the original contact map

Fig. 7: 2IGD protein information. CATH classification (Alpha Beta Protein Size: 61 Amino Acids) and its 3D structure with associated contact map. It is amino acid sequence is "MTPAVTTYKLVINGKTLKGETTTKAVDAETAEKA FKQYANDNGVDGVWTYDDATKTFTVTE"

The importance of Step 3 is that it employs the correlation mutation analysis on the cluster level, while the traditional correlation mutation works on pairs level. The tendency of correlation mutation analysis on cluster level is better than its tendency on pairs level. As the correlation mutation analysis achieves more improvements over random prediction, distinguishing the position of real β-sheet through sliding window becomes simpler.

In the literature, most of the previous research concentrated on predicting protein contact maps on the residues pairs level[10] and it neglected distinguishing the regions that have secondary structures. It is worthwhile to mention that the research efforts must focus toward predicting the dense regions as a first step and then focusing on pairs level inside these regions. Figure 6 shows the final sliding window which is used to look for β-sheet for 2IGD protein. The left part shows the position of region that has the highest correlation coefficient sum, while the right part shows the original contact map. The sliding window detects the right positions of β-sheets in this protein.

RESULTS AND DISCUSSION

We used different proteins related to different classes in our experiments. The selected proteins were extracted from the PDB database of solved structure proteins. The accuracy of the prediction is defined as the fraction of the predicted contacts which are correctly predicted. The Random Accuracy (RA) corresponds to placing the predicted contacts randomly in the contact map; it is equal to the percent of contacts derived from experimental method to the contact map size (the set of all pairs). Random Distribution (RD) corresponds to the percent of true contacts predicted using given correlation coefficient threshold to the predicted contacts for the same threshold.

2IGD protein analysis: The 3D structure of the 2IGD protein contains two anti parallel β-sheets and one α-helix between them. Figure 7 shows the β-sheets, α-helix secondary structures, and the correlation mutation analysis of the 2IGD protein. The results in Fig. 8 reflect the improvements of correlation accuracy (Cor A) over random accuracy and random distribution. As we go toward higher correlation coefficient values, the accuracy of correlation mutation increases accordingly. When reaching to correlation coefficient value greater than one, all the measurements go down to zero accuracy since there are no predicted contacts in the map.

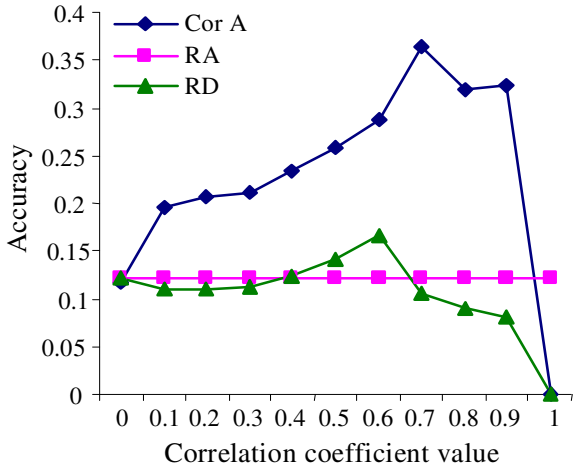

Fig. 8: Prediction accuracy using correlation accuracy (Cor A), Random Accuracy (RA) and Random Distribution (RD)

Figure 9 shows the dense areas accuracy through cluster accuracy (Clus A) measurement. The gap between the cluster accuracy and correlation accuracy is resulted from the orientation process of the sliding windows toward dense areas regions.

Figure 10 shows huge improvements especially for correlation coefficient zero, where the typical correlation mutation analysis grows about six times of the observed contact as wrongly predicted.

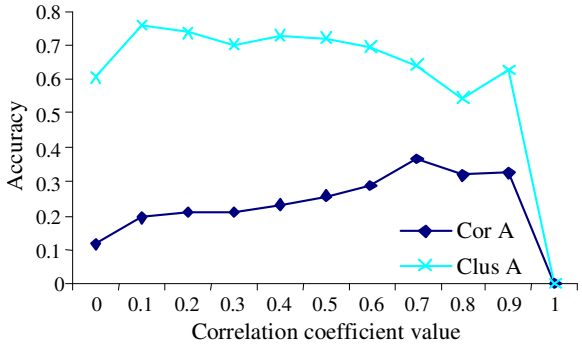

Fig. 9: Prediction accuracy using correlation accuracy (Cor A) and Cluster Accuracy (CA)

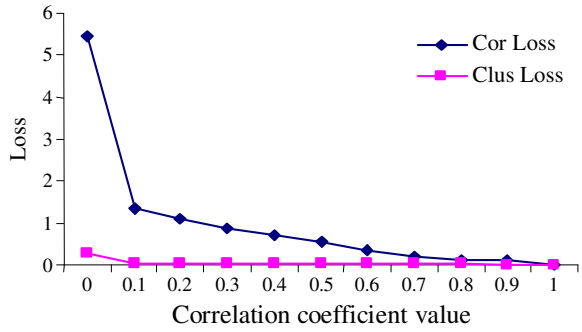

Fig. 10: Loss amount using correlation loss (Cor Loss) and cluster loss (Clus Loss)

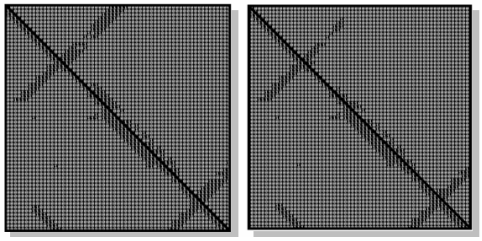

Fig. 11: Prediction analysis using dense areas improvements for 2IGD protein

While using dense areas, nearly about 0.5 times are as wrongly predicted. In Fig. 11, the result of the analysis is shown as predicted contact map. In the left part, the lower part shows contact map observed through experimental methods (i.e., X-Ray crystallography) and the upper part shows the predicted contact map using the dense areas analysis with correlation coefficient value equal to zero. The right part shows the observed contact map (the lower part) and the true predicted contacts (the upper part) using the dense areas analysis with correlation coefficient value equal to zero.

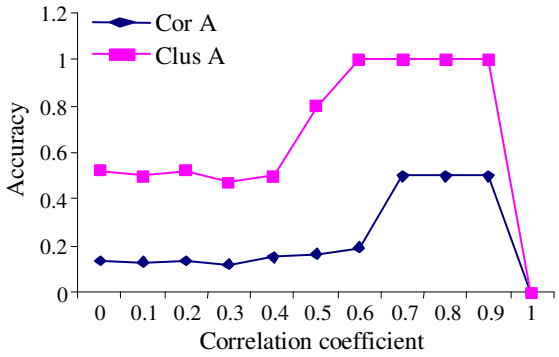

Fig. 12: Prediction accuracy for 6PTI protein, using correlation accuracy (Cor A) and Cluster Accuracy (CA)

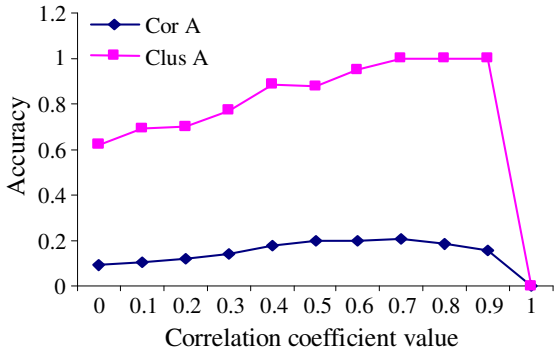

Fig. 13: Prediction accuracy for 451C protein, using correlation accuracy (Cor A) and Cluster Accuracy (CA)

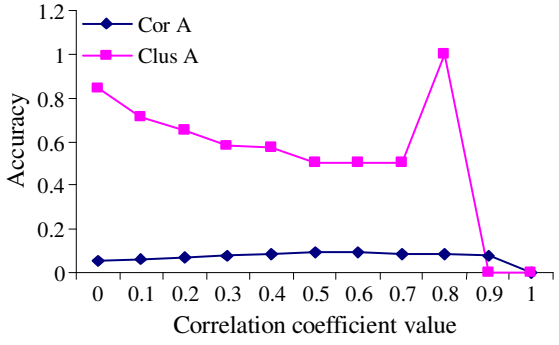

Fig. 14: Prediction accuracy for 1EYH protein, using correlation accuracy (Cor A) and Cluster Accuracy (CA)

Other proteins analysis: Different proteins related to different classes were used in the analysis. Figure 12-15 show the improvement results for these proteins through cluster accuracy measurement. These improvements resulted from using the dense regions analysis instead of the whole contact map.

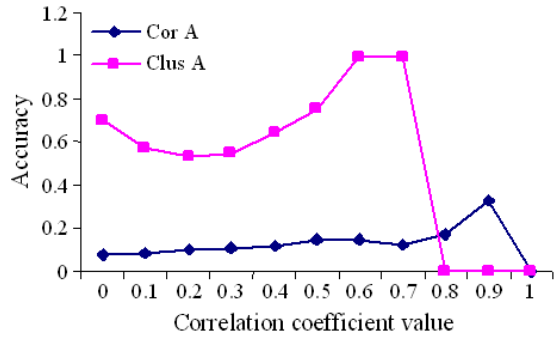

Fig. 15: Prediction accuracy for 1DFU protein, using correlation accuracy (Cor A) and Cluster Accuracy (CA)

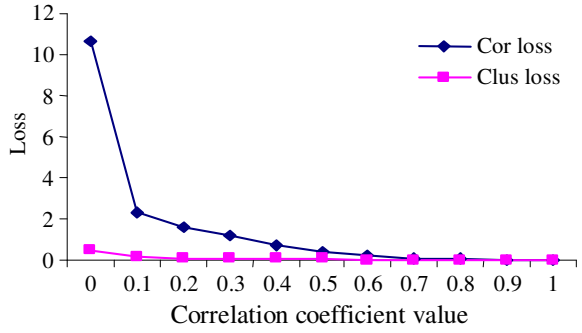

Fig. 16: Loss amount for 1DFU protein, using correlation loss (Cor Loss) and cluster loss (Clus Loss)

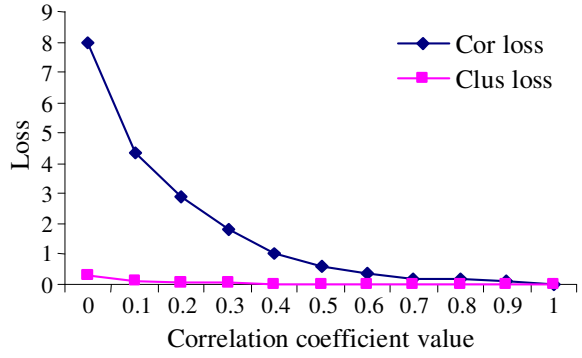

Fig. 17: Loss amount for 451C protein, using correlation loss (Cor Loss) and cluster loss (Clus Loss)

Figure 16-19 show the loss amount for different proteins compare with the typical correlation mutation analysis loss amount. In case of using low correlation coefficient values, the results show a large decrease using dense regions (Clus Loss) which increase the opportunity to hit a true contacts.

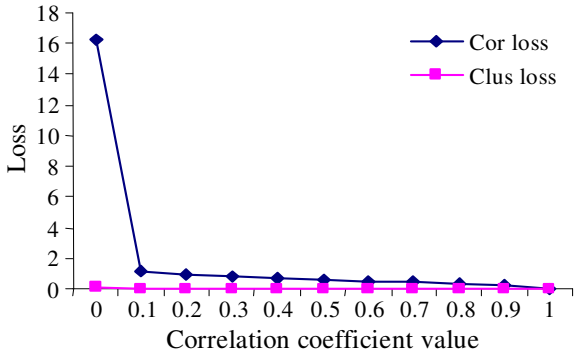

Fig. 18: Loss amount for 1EYH protein, using correlation loss (Cor Loss) and cluster loss (Clus Loss)

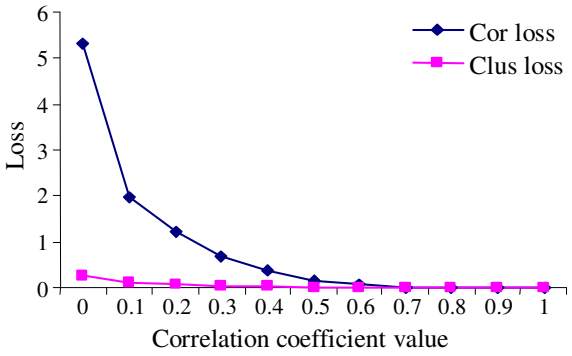

Fig. 19: Loss amount for 6PTI protein, using correlation loss (Cor Loss) and cluster loss (Clus Loss)

CONCLUSION

The protein structure prediction is a fundamental problem in molecular biology and any improvement in prediction accuracy will support the overall prediction process which results in better understanding for the cell. Machine learning approaches strongly participated in the field of prediction protein structure. Correlation mutation analysis was introduced in this study as a feature and entrance to the protein 3D structure prediction. This analysis predicts every pair of residues to be in contact or not independently of the other pairs. This study proposed an improvement over correlation mutation analysis to predict the secondary structures that exist in the contact map using protein primary structure and predicted secondary structure.

The proposed method makes the decisions in the prediction process based on regions of the secondary structures (dense regions) instead of independent pairs

as in the typical correlation mutation analysis, also it applies the correlation mutation analysis to predict the contacts that exist in the dense regions rather than applying it on the whole contact map. The proposed method was implemented on proteins related to different classes (i.e., mainly alpha, mainly beta, mixed alpha beta and low secondary structures), the test proteins are extracted from the Protein Data Bank (PDB) of solved structures. The results show improvements of correlation mutation analysis accuracy over random accuracy, at the same time the results show improvements of dense regions accuracy over correlation mutation accuracy. According to the amount of wrongly predicted contacts, the results show a large decrease in the wrongly predicted contacts in the dense regions analysis over correlation mutation analysis.

Several points have to be taken into account for the future study. First, it is worthwhile to mention that the other research efforts must be guided to increase the amount of knowledge in other features like solvent accessibility, build mathematical models simulating the biological interactions and integrate physical and chemical theories and properties. This will result in features that are valuable with knowledge, influence the learning process and finally increase the prediction accuracy. Second, it is possible to increase the detection accuracy of β-sheets through the sliding windows by measuring not only the sliding window as sum of correlation coefficient values inside the sliding window, but also integrating measurements through other features that distinguish β-sheets regions from the other regions.

REFERENCES

1. Ganapathiraju, M.K., J. Klein-Seetharaman, N. Balakrishnan and R. Reddy, 2004. Characterization of protein secondary structure. IEEE Signal Process. Mag., 21: 78-87. DOI: 10.1109/MSP.2004.1296545.
2. Ladd, M.F. and R.A. Palmer, 2003. Structure Determination by X-Ray Crystallography. 4th Edn., Springer, USA., ISBN: 10: 0306474549, pp: 864.
3. Cavanagh, J., W. Fairbrother, A.G. Palmer, N. Skelton and M. Rance, 2006. Protein NMR Spectroscopy: Principles and Practice. 1st Edn., Academic Press, San Diego, USA., ISBN: 10: 012164491X, pp: 854.
4. Nguyen, M.N. and J.C. Rajapaksi, 2003. Multi-class support vector machine for protein secondary structure prediction. Genome Inform., 14: 218-227. http://www.jsbi.org/journal/GIW03/GIW03F022.pdf.
5. Richardson, J., 1981. The anatomy and taxonomy of protein structure. Adv. Prot. Chem., 34: 167-339. http://www.ncbi.nlm.nih.gov/pubmed/7020376.
6. Orengo, C., A. Michie, S. Jones, D. Jones, M. Swindells and J. Thornton, 1997. CATH-a hierarchic classification of protein domain structures. Structure, 5: 1093-1108. http://www.ncbi.nlm.nih.gov/pubmed/9309224.
7. Tan, C. and D. Jones, 2008. Using neural networks and evolutionary information in decoy discrimination for protein tertiary structure prediction. BMC Bioinform., 9: 94-94. http://www.ncbi.nlm.nih.gov/pubmed/18267018.
8. Singer, M.S., G. Veriend and R.P. Bywater, 2002. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Prot. Eng., 15: 721-725. http://www.ingentaconnect.com/content/oup/proeng/2002/00000015/00000009/art00721.
9. Gupta, N., N. Managal and S. Biswas, 2005. Evolution and similarity evaluation of protein structures in contact map space. Prot.: Struct. Funct. Bioinformat., 59: 196-204. http://www3.interscience.wiley.com/journal/109930518/abstract.
10. Gobel, U., C. Sander, R. Schneider and A. Valencia, 1994. Correlated mutations and residue contacts in proteins. Prot. Struct. Funct. Genet., 18: 309-317. http://www.ncbi.nlm.nih.gov/pubmed/8208723.
11. Shindyalov, I., N. Kolchanov and C. Sander, 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Prot. Eng., 7: 349-358. http://peds.oxfordjournals.org/cgi/content/abstract/7/3/349.
12. Sander, M.C., G.L. Moore and C.D. Maranas, 2003. Using multiple sequence correlation analysis to characterize functionally important protein regions. Prot. Eng., 16: 397-406. http://www.ingentaconnect.com/content/oup/proeng/2003/00000016/00000006/art00397.
13. Thompsom, J.D, D.G. Higgins and T.J. Gibson, 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrichoic. Nucl. Acids Res., 22: 4673-4680. http://nar.oxfordjournals.org/cgi/content/abstract/22/22/4673.

14. Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann, USA., ISBN: 1558609016, pp: 770.
15. Hu, J., X. Shen, Y. Shao, C. Baystroff and M.J. Zaki, 2002. Mining protein contact maps. ACM BIOKDD Workshop on Data Mining in Bioinformatics, July 2002, ACM Press, New York, USA., pp: 3-10. http://www.cs.rpi.edu/~zaki/BIOKDD02/02-hu.pdf.
16. Pauling, L., R.R. Corey and H.R. Branson, 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci., 37: 205-211. http://authors.library.caltech.edu/6990/.
17. Knight, R.D. and L.F. Landweber, 2000. The early evolution of genetic code. Cell, 101: 569-572. DOI: 10.1016/S0092-8674(00)80866-1.