Residue Contact Prediction for Protein Structure using 2-Norm Distances

Nikita V. Mahajan
Department of Computer Science &Engg.
G.H. Raisoni College of Engineering, Nagpur

L.G.Malik
Department of Computer Science &Engg.
G.H. Raisoni College of Engineering, Nagpur.

ABSTRACT

Bioinformatics is a field which uses technology and is rapidly growing, to solve the problem related to biological area. Study of protein, predicting its structure to know its function is an important field in bioinformatics. Protein unit that is twenty amino acids have total information for converting linear sequences of amino acid into its unique and globular structures. Protein folding problem is process to determine that protein is folding into its exact tertiary structure. Protein folds into matter of seconds to its stable 3-D structure; once it is stable it may perform proper functions. Contact map is an intermediate step for converting 1-d structure to 3-d structure. Contact map is the representative graphical view, how the protein folds into its proper structure. Here, 6 parameters are set, which consider 2-norm distances for generating the contact map. It will consider the co-ordinate data, only of the residue having alpha carbon as contact type and map the contact as 1 if differences is greater than threshold value or -1 if less than threshold value.

Keywords

Protein Structure Prediction, Amino Acids, Contact Map, Non-local Contact Map, 2-Norm Distances.

1. INTRODUCTION

The most emerging field undergoing rapid, exciting growth is Bioinformatics. This has been mainly fuelled by advances in DNA sequencing, Protein sequencing and mapping techniques. Key topic in computational structural proteomics is Protein structure prediction [5] [6].

Protein is nothing but a20 linear amino acids chain which is covalently bounded with other amino acids (referred as residues). Length of these monomers ranges from ten to thousand and they are very important part of mostly biochemical process [3].Amino acid is made up of organic compounds which contain two functional groups that are amino group and carboxyl group. Amino group is represented as - NH_3 which is basic in nature. While carboxyl group is acidic in nature and it is represented as, - COOH. Amino acid is also termed as α amino acid if both amino and carboxyl group is attached to some carbon atom. An alpha-amino acid has the generic formula H2NCHRCOOH, where R is an organic substituent. R Group only changes in each amino acid

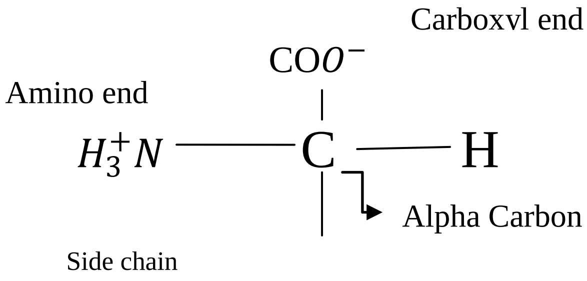
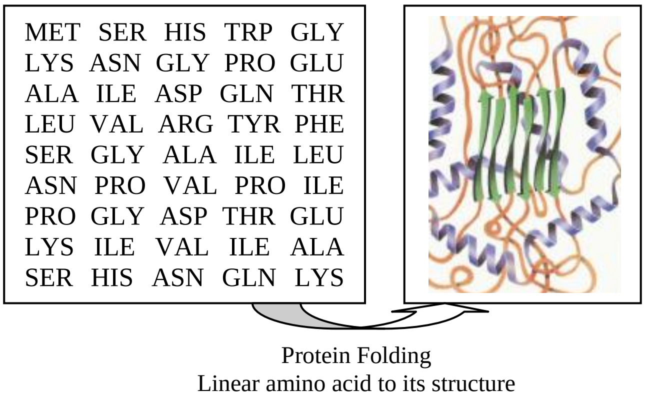


Figure 1: Amino Acid Structure

There are twenty different amino acids; they are represented as three letter codes, such as Alanine by ALA, Glycine as GLY, Proline as PRO, and Cytosine as CYS etc, which are grouped in such a fashion to represent the primary structure of protein.As Proteins play a central role in nearly all biological processes at the cellular level [7].

Protein folding is the process by which proteins fold itself into 3D structure. Protein chains fold into unique, tightly packed, globular structures called folds [3]. The early work of Anfinsen and Levinthal [6] established that a protein chain folds spontaneously and reproducibly to a unique three dimensional structure when placed in aqueous solution. The folding information process is present totally in linear sequences of amino acids, determines its unique fold, and the geometry of a protein fold largely determines its specific biological function. Predicting the tertiary structure of a protein by using its residue sequence is called the Protein Folding Problem [4] [7].



MET SER HIS TRP GLY
LYS ASN GLY PRO GLU
ALA ILE ASP GLN THR
LEU VAL ARG TYR PHE
SER GLY ALA ILE LEU
ASN PRO VAL PRO ILE
PRO GLY ASP THR GLU
LYS ILE VAL ILE ALA
SER HIS ASN GLN LYS

Protein Folding
Linear amino acid to its structure

Figure 2: Protein 3-D Structure Prediction

Indeed, proteins can be characterized by: (i) their primary structure, i.e. the alphanumerical amino acid sequence; (ii) their secondary structure, i.e. the spatial in- formation related to amino acid residues, (iii) the it tertiary structure which describes the spatial coordinate for each atoms composing the whole protein. [5].Protein don’t convert directly from primary structure to tertiary structure, it goes from secondary structure which can be associated to one ofthe two possible states, namely α-helix, β- strand.Alpha Helix (α) is the common motif in the secondary structure of proteins. It’s a right-handed coiled or spiral conformation. Every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid. One turn of the helix represents 3.6 amino acid residues. A single turn of the a-helix involves 13 atoms from the O to the H of the H bond [10].

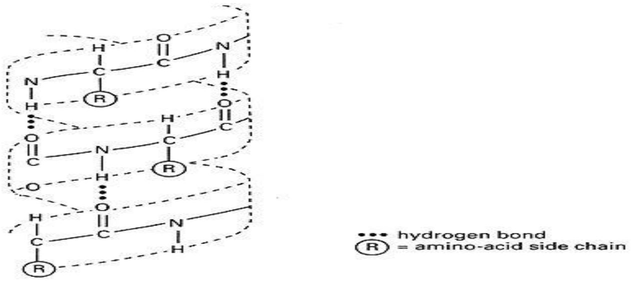


Figure 3: Alpha Helix Structure [13]

Beta sheets consist of the beta strands connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. In the parallel b-pleated sheet, adjacent chains run in the same direction (N ® C or C ® N). In the anti-parallel b-pleated sheet, adjacent strands run in opposite directions [8] [10].

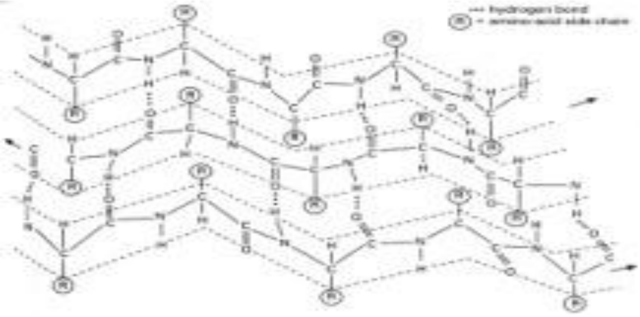


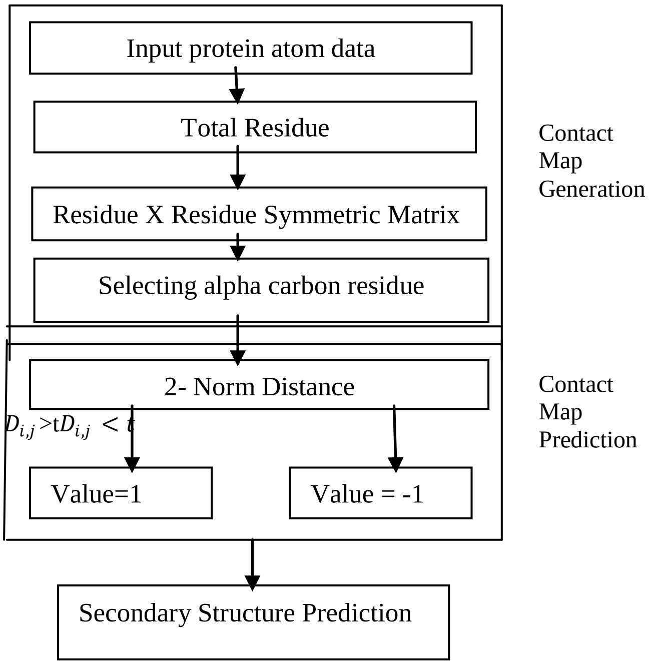
Figure 4: Beta Sheet Structure [13]

Recently a lot of effort has been put in solving the Protein Folding Problem (PFP).Most of them are based on heuristic methods, since the computational complexity of the underlying models makes their complete computational simulation is very expensive and, sometimes, unfeasible [1].Reinforcement Learning (RL) is an approach to machine intelligence in which an agent can learn to behave in a certain way by receiving punishments or rewards on its chosen actions [11].SVM was applied to the features to predict the structural relevance (i.e. if in the same fold or not) [12]. But still the problem remains unsolved [5]. One problem with protein data is that, it is very noisy.An approach to solve this problem is representing residue as contact. The principle behind contact map generation is just to show how the residue interacts with one another. These interactions will eventually lead to generation of tertiary structure of protein. Thus using such representation will reduce the dimensionality of problem down to more manageable point [5].

It is organized as follows. Section II provides definitions of contact maps, non-local contacts. In section III, the proposed model and applied input features are described in more details. Section IV presents experimental results. Finally, Section Vprovides some concluding remarks.

2. METHODOLOGY

The proposed structure calculation procedure contains three stages: contact map generation, contact map prediction and secondary structure prediction.



Flow Chart 1: Schematic overview of the structure prediction procedure

2.1. Contact Map

Contact map is an intermediate step showing how amino acid can be converted from its linear form to its tertiary structure. There for it can be said that Contact map is guide for protein structure prediction. Contact map is the graphical representation of amino acid residue. Therefore for predicting the 3D structure contact map plays vital role [6]. The residues are said to be in contact if the carbon atom of one residue is in contact with carbon atom of other residue. When amino acids come into contact they give rise to various non-covalent bonds such as hydrogen bond, hydrophobic bond [6] [2].

2.2.Generating Contact Map

Input to contact map generation will be the file containing all information about the amino acid. Given a protein sequence P with M residue such as P = $\{R_1, R_2 \ldots\ldots\ldots\ldots\ldots. R_m\}$.

And it can be said that two residue are in contact that is residue$R_i$is in contact with residue R_j if the distance between them are greater than that of specified threshold (t). That is$R_j - R_i < t$

Protein fold is normally defined on 3D Cartesian co-ordinates of amino acid atoms [3]. To find the contact between residue and residue 2-Normdistance is consider

$$.D = |R_j - R_i| = \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2 + (Z_j - Z_i)^2}$$

Where $R_i = (X_i, Y_i, Z_i)$ are the three dimensional co-ordinate of residue R_i [3].

For plotting the contact map M x M symmetric matrix is considered. This matrix will have number of rows and number of column equal to the residue number in protein sequence.

A pair wise residue distance matrix can be defined as

D = $D_{i,j}$ for each protein, where each element of contact map matrix $M_{i,j}$ is set to 1 if the residue (R_i, R_j) are in contact otherwise it is set to -1.

In other words, the contact map of protein with length N can be displayed in matrix whose elements are defined as [4]

$$D_{i,j} = \begin{cases} M_{i,j} = 1 & \text{If } D_{i,j} < t \\ M_{i,j} = -1 & \text{Otherwise} \end{cases}$$

2.3.Contact Map Prediction

Distance between two residuescan be defined in different ways:

- The distance between alpha carbon i.e. $C_a - C_a$
- The distance between beta carbon i.e. $C_b - C_b$
- The distance between All atoms

Distance between alpha carbon atoms$C_a - C_a$ is considered.

The first step in calculating a contact map between proteins is to calculate the alpha carbon distance between pair of their residues. It will first take the entire atom having the alpha carbon.Then in second step will subtract residue1 from all n residues to find distance that is. This subtraction is nothing but calculating 2-Normdistance of one residue with all residues present in protein.The third and final step is to compare the calculated distance with the threshold that is set, if greater then will plot the dot that is will set the matrix as 1 and if not then set as -1.

$$residue_{1,ca} - residue_{2,ca} > t\text{Then } M_{i,j} = 1$$
$$residue_{1,ca} - residue_{2,ca} < t \text{ Then } M_{i,j} = -1$$

.

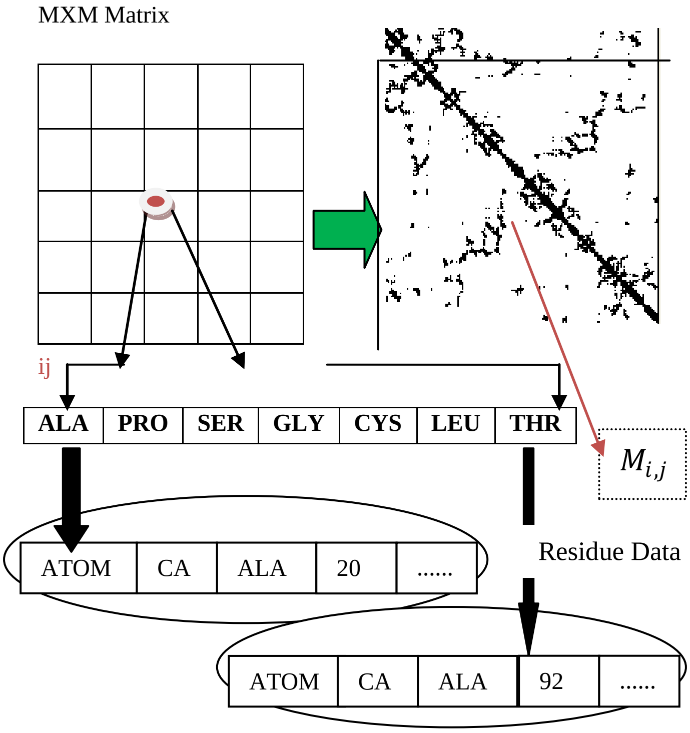


Figure 5: Process for Plotting Contact Map

Figure 5 provide an example, how the contact map for protein is plotted

2.4.Input Parameter

There are six types of input parameters for each pair of positions in a protein sequence, thatcapture the different aspects of the amino acids and the positions i and j

- Sequence Length
  Sequence length (SeqLgt) will be for the total number of amino acid residue in the input protein file.
- Minimum Sequence Separations
  Minimum sequence separation (MinSeqSep) will the minimum sequences separation filter used for the creation of contact map.
- Maximum Sequence Separations
  Maximum sequence separation (MaxSeqSep) will the maximum sequences separation filter used for the creation of contact map
- Distance Cutoff
  Distance cutoff (DisCutoff) will be the threshold specified in angstroms unit, used for creating Contact Map.
- Chain Identification
  Chain Identification (ChainID) will tell that protein data consider as input is perfect data or not
- Contact Type
  Contact type (ConType) is used for creation of contact map. It will specify that which residue carbon atom to be considered for calculating the 2-Normdistance.

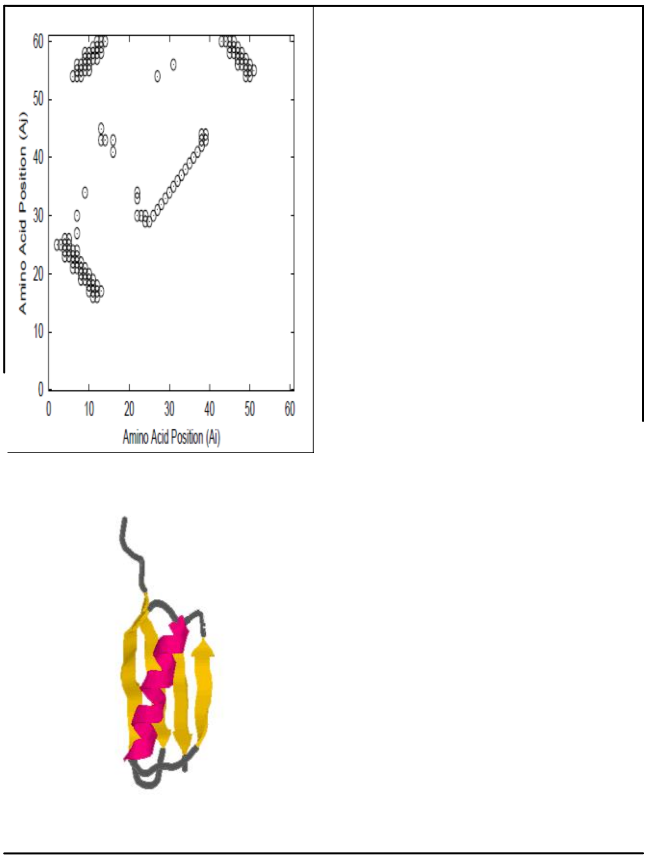
ALL	AL All atoms
C_a	The C-alpha atom of the backbone
C_B	The C-beta atom of the side-chain (C-alpha for glycine)

Table 1: Contact type Supported

2.5. Significance of Contact Map

A host of useful information is provided by contact map. Contact map not only calculate the contact distance of atom but also help to find which atom will contribute for formation of secondary structure [6].

Secondary structures consist of alpha helix and beta sheets. To localize the secondary structure from contact map it can be said ~~that. Diagonal area is alway~~s one as it assumes those residues are always in contact with themselves [3]. The amino acid which form alpha helix will always appear as a thick band near the diagonal as the contact is between single amino acid and its four successors. Amino acid which forms beta sheets will always appear as thin bands which would be parallel or anti parallel to the main diagonal.



(a) (b)

Figure 6: Contact Map of Amino acid (a) [6].
3 Dimensional Protein Structures from Contact map (b). [4]

Figure 6(a) is the perfect example showing the contact map of protein and figure 6(b) shows that how the contact map is being converted into tertiary structure.

In figure 7(a) there is only single band near the diagonal, that amino acid had lead for formation of only one helix as represented into figure6 (b).There are four bands parallel to diagonal, that had lead for formation of four beta sheet in figure 6 (b).

3. EXPERIMENTAL RESULT

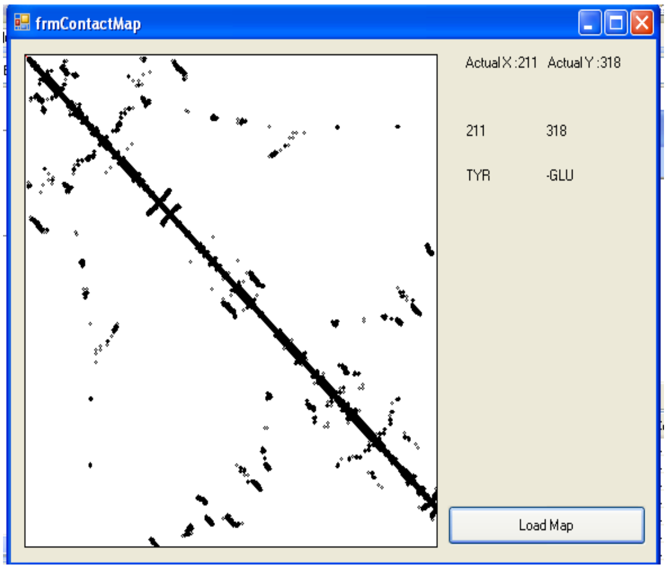
3.1. Data Set

Drs. Edgar Meyer and Walter Hamilton had founded Protein Data Bank (PDB) in 1971 at Brookhaveni National Laboratory. According to PDB selected list PDB contains 3-D structure of Protein and nucleic acid, which has non redundant protein structure with sequence identity lower than 25% [2] [9].The PDB format has 12 sections, in which 46 different fields are represented. The following fields are relevant for our model: SEQRES (defines the amino acids sequence of the protein), HELIX and SHEET (identify the amino acids that form these secondary structures), and ATOM (represents the spatial distribution of the atoms of the protein in its native conformation)

The selected PDB input file for methodology is of Mycobacterium tuberculosis. This file has various sections such as date, author name, Atoms, Chain identification, residue number, residue Co-ordinates (x,y,z) in angstrom unit [2] [9].

3.2. Result

Present the result for predicting the contact map.The input PDB file contains 489 amino acids, thus the sequences length is of 489. The contact map matrix will be of 489x489. Theminimum Sequence separation is 1 and maximum Sequence Separation is 400. Contact type selected asC_a,

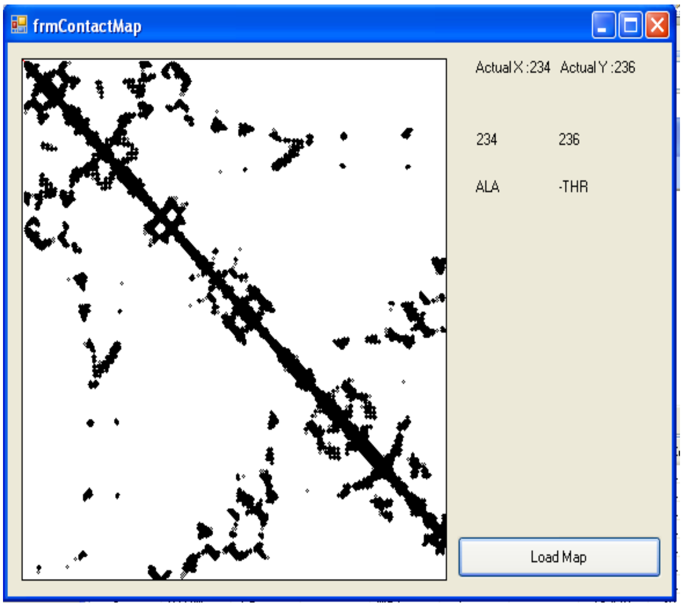


Actual X :211 Actual Y :318

211 318

TYR -GLU

Figure 7: Contact Map distance cutoff is 8 Angstroms



Actual X :234 Actual Y :236

234 236

ALA -THR

Figure 8: Contact Map distance cutoff is 12 Angstroms

Figure 7, Figure 8 shows contact map with distance cutoff of8, 12 Angstroms. The figure also shows the selected amino acidresidue number, residue name and also tell that amino acid belongs to helix as they are close to diagonal value and beta sheet as they are parallel or anti parallel to diagonal.

4. CONCLUSION

As the definition of contact map, contact map provide that which amino acid residue are in contact by assigning it 1 and which is not by assigning it -1. Contact map take protein sequence, align the map to proper templates and map is then consider as folding pathway for protein structure prediction.

	8 Angstrom	12 Angstrom
Features	Less Features available off main diagonal	More feature available off main diagonal
Recovery	It is not so easy	It is easy
Structure prediction	Structure prediction is little bit hard	Structure prediction is easier

Table 2:Comparison of Contact Map with Different Angstrom Value

Thus it can be said that if the threshold value of contact map is increased, then the structural constraint number also increases, and it will be easier to extracted more detail for 3D structure prediction.

5. REFERENCE

[1]Fernanda Hembecker, HeitorSilvério Lopes (2010), “A Molecular Model for Representing Protein Structures and its Application to Protein Folding” IEEE.

[2]NarjesKhatoonHabibi, Mohammad HosseinSaraee (2009), “Protein Contact Map Prediction Based On an Ensemble Learning Method”, In:International Conference on Computer Engineering and Technology.

[3]YosiShibberu and Allen Holder (2011), “A Sepctral Approach to Protein Structure Alignment”. In:IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, NO. 4.

[4]Nitin Gupta, NitinMangal and SomenathBiswas (2004), “Evolution and similarity evaluation of Protein Structures in Contact map space.

[5]Giuseppe Tradigo (2009), “On the Integration of Protein Contact Map Predictions”, In:IEEE

[6]JingjingHu ,XiaolanShen , Yu Shao , Chris Bystroff , Mohammed J. Zaki (2002), “Mining Protein Contact Maps”, In:Workshop on Data Mining in Bioinformatics

[7]Kibeom Park, Michele Vendruscolo, and EytanDomany (2000), “Toward an Energy Function for the Contact Map Representation of Proteins”, In: PROTEINS: Structure, Function, and Genetics 40:237–248

[8]ZaferAydin, YucelAltunbasak, and HakanErdogan (2011), “Bayesian Models and Algorithms for Protein B-Sheet Prediction”.In: Transactions on computational biology and bioinformatics, vol. 8, no. 2

[9]The Protein Data Bank,” http://www.rcsb.org/pdb, 2009

[10] Introduction to protein structure and structural bioinformatics, “secondary-sructure.html”

[11] Gabriela Czibula, Maria-IulianaBocicor and Istvan-GergelyCzibula (2000), “A Reinforcement Learning Model for Solving the Folding Problem".

[12] Muhammad Asif Khan, Muhammad A. Khan, Zahoor Jan, Hamid Ali and Anwar M. Mirza (2010). “Performance of Machine Learning Techniques in Protein Fold Recognition Problem”. In: IEEE.

[13] http://www.ics.uci.edu/~baldig/betasheet_data.html, 2009.