Extraction of Contact Maps for Cervical and Breast Cancer Proteins 3d-Structures

Pokuri Deepika[1], **Dr. K. Suvarna Vani**[2]

[1, 2]Department of Computer Science and Engineering, V.R. Siddhartha Engineering College, Vijayawada, India

Abstract: ***Protein contact map prediction is one of the major research problems in bioinformatics. By this prediction contact map and feature extraction of the primary sequence is known which is useful for determining contact maps. Many researchers have worked on predicting contact maps from protein sequences using different techniques and proposed different prediction algorithms but the accurate result is not determined. In the proposed method the contact map prediction problem based on the primary sequence is performed. Feature extraction method is proposed to extract different features from the protein sequences. After that different machine learning methods are applied to predict contact maps from that features of contact maps like number of atoms. Machine learning algorithms efficiently handle the prediction. The results are better than the existing methods which validate the simple analysis on different Breast Cancer and Cervical Cancer Proteins with Contact maps and feature extraction fields.***

Keyword: Contact Map Prediction, Feature extraction, Breast Cancer and Cervical Cancer Protein Sequences, Protein Data Bank

1. Introduction

Bioinformatics become important of many areas of biology. It is an emerging field undergoing rapid growth in the past few decades. Every protein has some specific function within the body. Few proteins are involved in bodily movement, while others are involved in structural support. Proteins distinct are in functions as well as structures. One of the important goals persists by bioinformatics and theoretical chemistry is protein structure prediction.

Proteins are being classified according to sequence and structural similarity. The four stages of protein structure are primary, secondary, tertiary, and quaternary structure. Each single protein molecule may contain few of these protein structure types. The structure of protein describes the protein function. The primary structure of protein is derived from the amino acid sequence of a protein and it is the most fundamental form of information available about the protein. Proteins perform most of the functions in the cells of living organisms, acting as enzymes to perform complex chemical reactions, recognizing foreign particles, conducting signals, and building cell scaffolds to name just a few. Their function is dictated by their three-dimensional structure, which can be quite involved, even though proteins are linear polymers composed of only 20 different types of amino acids.

Machine learning to target is on prediction, based on known properties learned from the training data. In, the field of biology various applications extensively use methods which are based on machine learning algorithms. Machine learning approaches have found immense importance in numerous bioinformatics prediction methods.

Most functional-prediction methods, both sequence and structure based, rely on inferring relationships between proteins that permit the transfer of functional annotations and binding specificities from one to the other. A notable challenge here is deciphering the connection between the detected similarities (structural or in sequence) and the level of functional relatedness. Function is often associated with domains, and another problem is the identification of functional domains from sequence alone.

The accurate annotation of protein is key to understanding life at the molecular level and has great bio medical and pharmaceutical implications. However, with its inherent difficulty and expense, experimental characterization of function cannot scale up to accommodate the vast amount of sequence data already available. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology.

Contact Maps are used in protein superimposition and for predicting protein similarity studies as they provide a reduced representation than the 3D coordinate and improve the template-target alignment, thereby increasing the accuracy of structure prediction using primary sequence information. The accuracy of present using methods for predicting domain boundaries is not yet completely satisfactory. Several methods provide reliable predictions if a structural template for the protein is available, one is left with the problem of whether the experimental annotation used for the inference refers to the same domain for which the sequence similarity/motif is established.

Definition of protein

Proteins are large bimolecules, or macromolecules, consisting of one or more long chains of amino acid residues. Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in protein folding into a specific three-dimensional structure that determines its activity. Most proteins consist of linear polymers built from series of up to 20 different L-α-amino acids. The side chains of the standard amino acids, detailed in the list of standard amino acids, have a great variety of chemical structures and properties. Once linked in the protein chain, an individual amino acid is called a residue, and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone.

Amino acid

Amino acids are organic compounds containing amine (-NH2) and carboxyl (-COOH) functional groups, along with a side chain(R group) specific to each amino acid. The key elements of an amino acid are carbon, hydrogen, oxygen, and nitrogen, although other elements are found in the side chains of certain amino acids. About 500 amino acids are known (though only 20 appear in the genetic code) and can be classified in many ways. They can be classified according to the core structural functional groups locations as alpha(α), beta (β), gamma (γ) or delta (δ) amino acids; other categories relate to polarity, pH level, and side chain group type (aliphatic, acyclic, aromatic, containing hydroxyl or sulfur, etc.).

Tertiary structure with contact map generation

The tertiary structure is the 3 dimensional, native structure of a single polypeptide or protein. A protein normally is folded into a compact structure, usually referred to as 'globular protein', a term traditionally associated with water soluble proteins. The secondary structures are stabilized by the final, native fold. The native fold is defined as the active conformation.

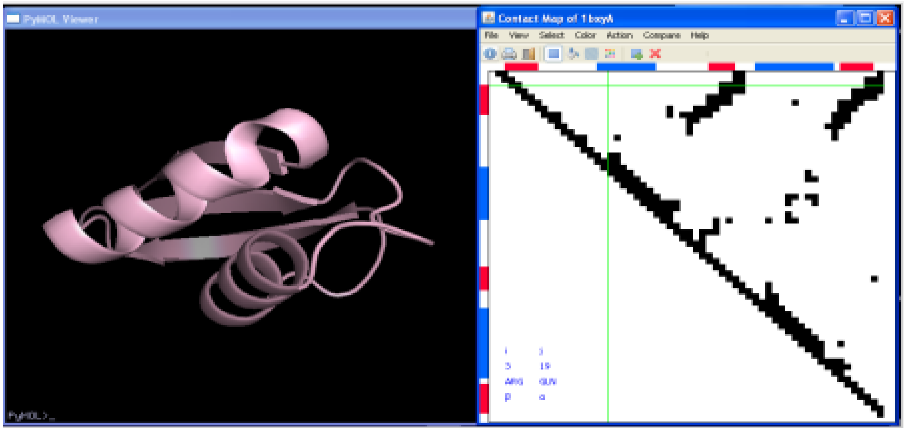

Figure 1: Tertiary Structure and contact map

Contact Map

Given a protein with N amino acids, its contact map C_{NxN} is a NxN binary symmetric matrix. While distance (D_{ij}) between the two amino acids is no longer than an arbitrary distance threshold (T), the values of elements of the contact map are assigned to 1. The values are set to 0. Various ways have been proposed to define the distance between two amino acids: the distance between their C atoms; the distance between the closest atoms belonging to each of the amino acids, respectively; the distance between the centers of mass of the two amino acids; and the distance between their C atoms (C for Glycine, since the side-chain of a Glycine only contains a hydrogen). The standard distance threshold is 6-12 angstroms (Å), which is widely used. The advantage is that contact maps are invariant to rotations and translations. It provides useful information about protein's structure.

Number of contacts

The number of contacts is proportional with the length of protein sequence, but has nothing with some parameters, such as the number of protein peptide chain, amount of each protein secondary structural types, and so on. Assuming that the sequence's length is N, then the number of contacts is D=k*(N-b), where k is a coefficient. Because the contact map has removed 2N elements in the principal diagonal, there is an offset b>0. For the given contact map, assuming that its number of contacts is D, and the sequence length is N, according to threshold we determine k1 k2 and b we could get y1=k1*(N-b) and y2=k2*(N-b). Design the punishment rule as follows:

$$\Delta pd=\begin{cases} y1\text{-}D & \text{if } (D>y1) \\ D\text{-}y2 & \text{if } (D<y2) \\ 0 & \text{otherwise} \end{cases} \quad \text{.............(1)}$$

2. Methodology

The methodology of the project, different steps for contact map prediction from protein sequences based on pdb ids and extracting features and structures.

Proposed System

The FASTA sequence/PDB Structures are considered as the input for which the proposed Contact Map Generation algorithm is applied to extract Feature extraction table. Different classification methods are performed. The methodology overview for prediction of contact maps are extracting from Protein Primary Sequences is given below.

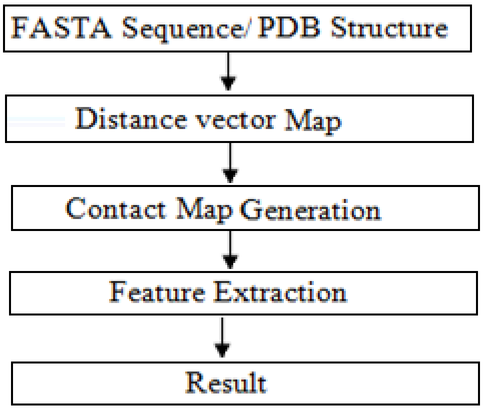

Figure 2: Methodology Overview for Contact Map Prediction

Dataset

The datasets are taken from RCSB PDB based on Breast Cancer and Cervical Cancer PDB id in the format of FASTA

Sequence consists of α-sheet, helix-sheet, ATOM . It consists of chains and describes minimum length of the sequence and residues of the related cancer proteins.

Table 1: Input Data Set

Name of the Cancer	PDB ID	Number of Proteins
Cervical	1C4Z,1JJ4, 2RSK, 5Y9C, 3F81, 3OFL, 4GIZ, 5J6R	8
Breast	1F1G, 1JBO, 1N50, 1R5K, 2A87,2DCN, 3A54, 3HY3, 3M47, 3S7S,5MOW, 5N31	12

Protein Primary Structure

Protein primary structure is the linear sequence of amino acids in a peptide or protein. By convention, the primary structure of a protein is reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end. Protein primary structures can be directly sequenced, or inferred from DNA sequences.

Secondary Structure

Protein secondary structure is the three dimensional form of local segments of proteins. The two most common secondary structural elements are alpha helices and beta sheets, though beta turns and omega loops occur as well. Structure prediction utilized statistical approaches, employing data collected from proteins with known secondary structure. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its three dimensional tertiary structure. Secondary structure is formally defined by the pattern of hydrogen bonds between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone.

Tertiary Structure

Tertiary structure refers to the three dimensional globular structure formed by bending and twisting of the secondary structural elements. The protein structure can be consider as a folding of secondary structure elements, such as α-helices and β-sheets, which together constitute the overall three-dimensional configuration of the protein chain. Protein Data Bank (PDB) gives entire information related to the structure from primary, secondary to tertiary structures of a protein.

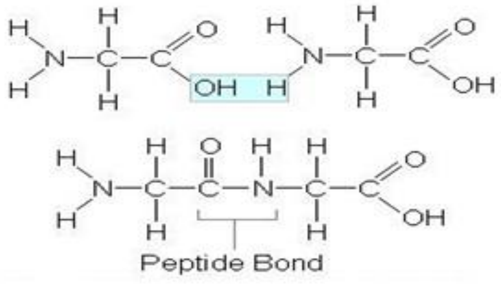

Figure 3: Tertiary Structure

FASTA Sequence/PDB Structure

FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.

Extraction of Helices

A helix/ helices is smooth space curve, i.e, a curve in three-dimensional space. It has the property that the tangent line at any point makes a constant angle with a fixed line called the axis. Helices are important in biology, as the DNA molecule is formed as two intertwined helices, and many proteins have helical substructures, known as alpha helices.

Algorithm

Input: Breast and Cervical CancerPDB files

Output: Prediction of Helices
Step-1: Reads sequence from PDB file.
Step-2: Identifies length of string starts with “TER”.
Step-3: Checks amino acids from the given sequence which starts with a three letter words.
Step-4: The final sequence prints in the sequence file.
Step-5: Reads the header sequence atoms related to A.
Step-6: It generates amino acids atoms related to the sequence of A.
Step-7: Generates individual helix count for generated amino acids.
Step-8: It gives the CA atoms from helix.

Contact-Map Generation

A concept map or conceptual diagram is a diagram that depicts suggested relationships between concepts. A concept map typically represents ideas and information as boxes or circles, which it connects with labeled arrows in a downward-branching hierarchical structure. The technique for visualizing these relationships among different concepts is called concept mapping.

Algorithm

Input: Cancer PDB files

Output: Prediction of Helices
Step-1: Read the PDB file.
Step-2: Define the matrix size.
Step-3: Read the length of the sequence from starting and ending codon.
Step-4: Increment the count until the length reach to end codon of a sequence.
Step-5: Display the sequence of atoms represents with A.

Distance vector

Distance-Vector is a table-driven routing scheme for networks based on the Bellman–Ford algorithm. The main contribution of the algorithm was to solve the routing loop problem. Each entry in the routing table contains a sequence number, the sequence numbers are even if a link is present; else, an odd number is used. The number is generated by the destination, and the emitter needs to send out the next update with this number.

Advantages

- The availability of paths to all destinations is always shows that less delay is required in the path set up process.
- The method of incremental update with sequence number labels, marks the existing wired network protocols.

Disadvantages

- It requires a regular updates of its routing tables, which uses up a small amount of bandwidth even when the network is idle.
- The topology of the network changes, a new sequence number is necessary before the network re-converges; thus, Distance Vector is unsuitable for highly dynamic or large scale networks.

Algorithm

Input: Cancer PDB files

Output: Identifying Distance-Vector values

```
Step-1: Read PDB file and length of the string.
Step-2: Identifies only CA atoms present in the PDB file.
Step-3: To find distance vector value.
m=Math.sqrt(Math.pow(xm[jm]-xm[im],2)+Math.pow(ym[jm]-ym[im],2)+Math.pow(zm[jm]-zm[im],2)
Step-4: Representing threshold value the matrix is formed with the coordinate values.
Step-5: Repeat the count until end of the sequence length to find distance vector values for protein sequence.
```

Accuracy calculation

Accuracy describing a combination of both types of observational error above (random and systematic), so high accuracy requires both high precision and high trueness. The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\text{......................(2)}$$

The specificity of a test is its ability to determine the healthy cases correctly. To estimate it, we should calculate the proportion of true negative in healthy cases.

Mathematically, this can be stated as:

$$\text{Specificity}=\frac{TN}{TN+FP}\text{........................(3)}$$

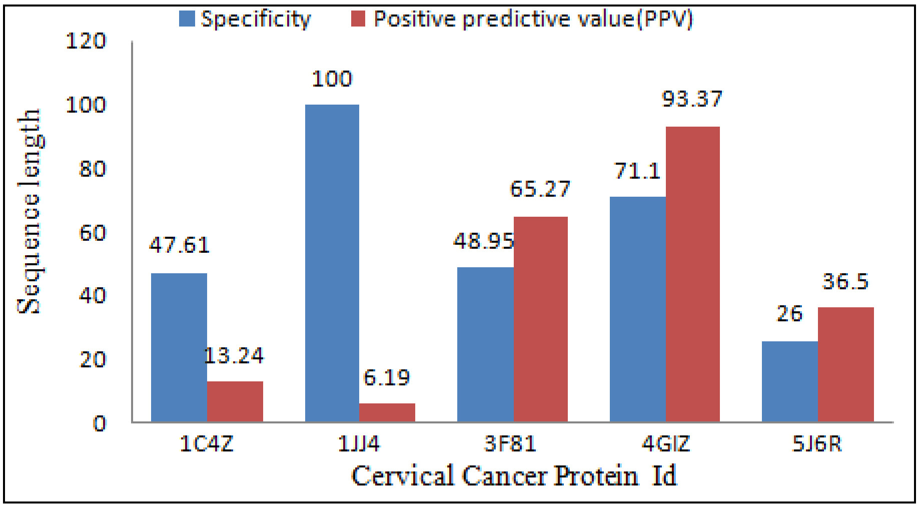

Figure 4: Accuracy calculation for Cervical Cancer

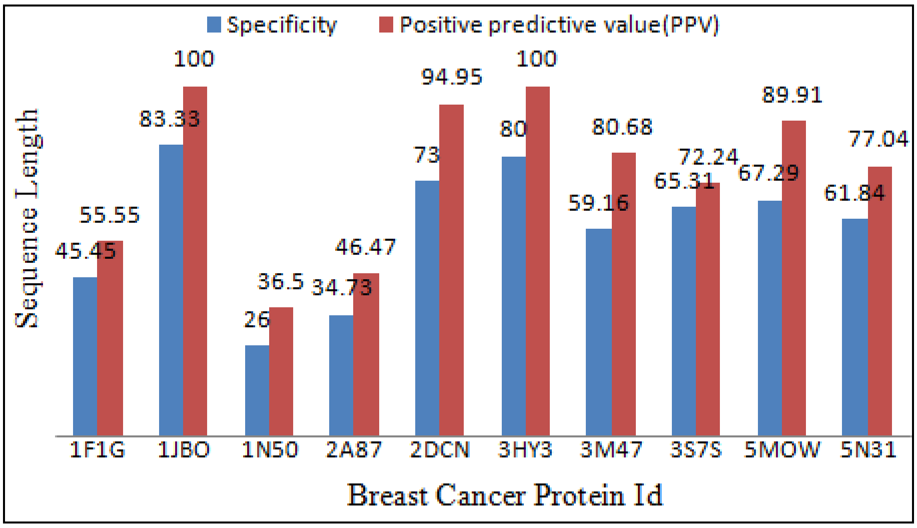

Figure 5: Accuracy calculation for Breast Cancer

Protein Contact Map

The problem of contact map prediction can be stated as a classification problem. Given a set of proteins with known structures, contact residues and non-contact residues are separated as positive instances and negative instances. For each instance, various features are collected to capture useful information of the pair of residues, including amino acid content, secondary structures, evolutionary correlation, and other information that can discriminate contacts from non-contacts. Then, these feature vectors of both positive instances and negative in-stances are used as the input to a classification tool to learn a classifier.

Features

Here used four types of features for each pair of positions in a protein sequence, that capture different aspects of the amino acids and the positions i and j:

- Sequence Length (SeqLen): is the number of residues in the target sequence.
- Sequence Separation (SeqSep): is the number of the residues in between i and defined by- j·
- Physicochemical and Biological Properties of Amino Acids: is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids.
- Predicted Secondary Structure (PSS): The predicted secondary structures for each protein, was obtained by PSIPRED. For each residue, three values are used to represent whether it belongs to an alpha-helix (H), beta strand (E) or coil (C).

Eight Neighbors Algorithm

Input: Contact-map

Output: Feature extraction

```
Step-1: K← number of nearest neighbor.
Step-2: For each object X in the test set do
Step-3: Calculate the distance D(X,Y) between X and every object Y in the training set.
Step-4: neighborhood← the k neighbors in the training set closed to X.
Step-5: X. class← Select Class (neighborhood).
Step-6: End for.
```

Diagonal values extraction

A diagonal matrix is a matrix in which the entries outside the main diagonal are all zero. The term usually refers to square matrices.

Baker-Bird Algorithm

Input: Mask-Map

Output: Cluster values

```
Step-1: Read Mask-Map values as input
Step-2: Preprocess the pattern as give each row of P a unique name using AC automaton of pattern rows.
Step-3: Represent P as a 1D vector and construct the 1D KMP automation.
Step-4: Row matching is label positions of T where suffix matches row of P using AC automaton.
Step-5: Column matching uses KMP on named columns of T to find pattern occurrences.
```

Evaluation Measures

Several evaluation scores are used to evaluate the performance of these methods. These evaluation scores take into account part or full out of four parameters of prediction quality, that is true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Although only two out of the four parameters are taken into account, accuracy and coverage are two widely accepted and used statistical indices.

The number of all predicted ones (sum of TP and FP), respectively. Coverage (also referred to as 'Sensitivity' or 'Recall') is another crucial performance criteria, defined as:

$$\text{Cov}=\frac{TP}{TP+FN} \qquad \text{.......................(4)}$$

and N_{obs} is the number of observed contacts (sum of TP and FN). While accuracy and coverage only take into account of contacts and ignore non-contacts (TN), F1-score and Matthews's correlation coefficient (MCC) are two more general evaluations. The F1 score is a weighted average of the accuracy and coverage. An F1 score reaches its best value at 1 and worst score at 0, defined as:

$$F_1=\frac{2(Acc*Cov)}{(Acc+Cov)} \qquad \text{.......................(5)}$$

Although the F1 score is a more comprehensive measure as it combines accuracy and coverage, true negative rate still is not considered.

The Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm.

$$\text{dist}((x_1y_1),(x_2,y_2),(x_{3,}y_3)= \sqrt{(x3-x2-x1)2+(y3-y2-y1)2} \text{ ....6}$$

Table 2: 1F1G Protein values for original and predicted Helix for Beta width-3

Protein	Original Helices		Predicted Helices	
			W=3	
1F1G	57	61	4	6
	133	138	9	12
	212	216	15	21
	288	293	23	30
	367	371	35	44
	443	448	41	70
	522	526	46	77
	598	603	72	95
	677	681	79	103
	751	758	99	114
	832	836	108	120
	906	913	116	129
			122	143
			131	

Table 3: 1C4Z Protein values for original and predicted Helix for Alpha width-3, 4, 5

Protein	Original Helices		Predicted Helices					
			W=3		W=4		W=5	
1C4Z	508	523	1	6	12	26	12	25
	524	530	8	43	28	31	49	61
	545	560	45	69	49	62	89	104
	561	565	71	108	89	104	116	121
	585	602	110	126	115	121	131	133
	611	619	128	158	128	134	136	145

	624	632	166	176	136	146	152	154
	632	645	178	185	152	155	190	204
	648	653	187	224	190	205	207	216
	686	700	226	230	207	220	233	238
	703	718	232	255	232	239	147	251
	729	738	258	263	247	252	265	274
	761	773	265	291	265	275	279	258
	775	787	296	298	279	289	335	346

3. Conclusion

The aim of this work is predicting cancer protein primary sequence is a fundamental problem in modern computational biology. It is the key for developing deeper artificial intelligence systems in biomedicine and bioinformatics. In this work, process is comparing with online web-server DISTILL and manual script by using machine learning algorithms to figure out the major difficulties of protein structure prediction, how to extract the most useful and suitable protein features in which the model parameters are jointly optimized for the target of protein structure prediction. DISTILL the result is based on Server response and given query. Future work, the process is to consider a bigger and deeper analysis to improve the performance of protein structure learning with different approaches.

References

[1] Distill http://distillf.ucd.ie/distill/.

[2] The Protein Data Bankhttps://www.rcsb.org/.

[3] Lin Bai, Lina Yang "A Unified Deep Learning Model for Protein Structure Prediction" IEEE 2017.

[4] Karina B. Santos, Gregorio K. Rocha, Fabio L. Custodio, Helio J. C. Barbosa "Improving De novo Protein Structure Prediction using Contact Maps Information" IEEE 2017.

[5] Gregorio K. Rocha, Jaqueline S. Angelo, Karina B. Santos, Fabio L. Custodio, Laurent E. Dardenne, and Helio J.C. Barbosa "Using an Aggregation Tree to Arrange Energy Function Terms for Protein Structure Prediction" doi: 10.1109/CIBCB.2017.8058533, IEEE 09 October 2017.

[6] Sheng Wang, Wei Li, Renyu Zhang, Shiwang Liu and Jinbo Xu "CoinFold: a web server for protein contact prediction and contact-assisted protein folding" Vol. 44, W361–W366, doi: 10.1093/nar/gkw307, Nucleic Acids Research, 2016.

[7] Jiang Xie, Wang Ding, Luonan Chen, Qiang Guoand Wu Zhang "Advances in Protein Contact Map Prediction Based on Machine Learning" Vol. 11, No. 3, Medicinal Chemistry, 2015.

[8] Cosme E. Santiesteban-Toca, Gerardo M. Casanola-Martinand Jesus S. Aguilar-Ruiz "A Divide-and-Conquer Strategy for the Prediction of Protein Contact Map" vol.12, Letters in Drug Design & Discovery 2015.

[9] Suvarna Vani K and M. Om Swaroopa and T.D. Sravani and K. Praveen Kumar "Frequent substructures and fold classification from protein contact maps" doi: 10.1109/CIBCB.2014.6845518, IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) May 2014.

[10] Matt Spencer, Jesse Eickholt, Jianlin Cheng " A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction" doi 10.1109/tcbb.2014.2343960, IEEE Transactions on Computational Biology and Bioinformatics 2013.

[11] Chan Weng Howe and Mohd Saberi Mohamad "Prediction of Protein Residue Contact Using Support Vector Machine" pp. 323–332 Springer-Verlag Berlin Heidelberg 2012.

[12] Pietro Di Lena, Ken Nagata and Pierre Baldi "Deep architectures for protein contact map prediction" Vol.28 no. 19 2012, pages 2449–2457 doi:10.1093/bioinformatics/bts475 Advance Access publication July 30, 2012.

[13] Shoshana Neuburger "Pattern Matching Algorithms: An Overview" September 15, 2009.

[14] Lan Huang, Guixia Liu, Rongxing Wang, Bin Yang, Chunguang Zhou "Improved Clonal Selection Algorithm for Protein Contact Map Prediction" Fifth International Joint Conference on INC, IMS and IDC 2009.

[15] Jianlin Cheng, Allison N. Teggeand Pierre Baldi "Machine Learning Methods for Protein Structure Prediction" vol. 1, IEEE 2008.

[16] Narjes K. Habibe, Kaveh Mahdaviane, Mohammad H. Saraee "Mining Protein Primary Structure Data Using Committee Machines Approach to Predict ProteinContact Map" International Conference on Emerging Technologies IEEE-ICET 2008.

[17] Amihood Amira, Oren Kapahb, Dekel Tsur "Faster two-dimensional pattern matching with rotations" doi:10.1016/j.tcs.2006.09.012, 196 – 204, Theoretical Computer Science 368 2006.

[18] Wei Chu, Zoubin Ghahramani, Alexei Podtelezhnikov, and David L. Wild "Bayesian Segmental Models with Multiple Sequence Alignment Profiles for Protein Secondary Structure and Contact Map Prediction" vol. 3, no. 2, IEEE APRIL-JUNE 2006.

[19] L. Gwenn Volkert and Deborah A.Staffer "A Comparison of Sequence Alignment Algorithms for Measuring Secondary Structure Similarity" IEEE 2004.

[20] Ying Zhao and George Karypis "Prediction of Contact Maps Using Support Vector Machines" IEEE 2003.