PROTEIN STRUCTURE PREDICTION USING SUPPORT VECTOR MACHINE

Anil Kumar Mandle[1], Pranita Jain[2], and Shailendra Kumar Shrivastava[3]

[1]Research Scholar, Information Technology Department Samrat Ashok Technological InstituteVidisha, (M. P.) INDIA
akmandle@gmail.com
[2]Astt.Prof. Information Technology Department Samrat Ashok Technological Institute Vidisha, (M. P.) INDIA
Pranita.jain@gmail.com
[3]HOD, Information Technology Department Samrat Ashok Technological Institute Vidisha, (M. P.) INDIA
shailendrashrivastava@rediffmail.com

ABSTRACT

Support Vector Machine (SVM) is used for predict the protein structural. Bioinformatics method use to protein structure prediction mostly depends on the amino acid sequence. In this paper, work predicted of 1-D, 2-D, and 3-D protein structure prediction. Protein structure prediction is one of the most important problems in modern computation biology. Support Vector Machine haves shown strong generalization ability protein structure prediction. Binary classification techniques of Support Vector Machine are implemented and RBF kernel function is used in SVM. This Radial Basic Function (RBF) of SVM produces better accuracy in terms of classification and the learning results.

KEYWORDS

Bioinformatics, Support Vector Machine, protein folding, protein structure prediction.

1. INTRODUCTION

A protein is a polymeric macromolecule made of amino acid building blocks arranged in a linear chain and joined together by peptide bonds. The primary structure is typically represented by a sequence of letters over a 20-letter alphabet associated with the 20 naturally occurring amino acids. Proteins are the main building blocks and functional molecules of the cell, taking up almost 20% of a eukaryotic cell's weight, the largest contribution after water (70%). Protein structure prediction is one of the most important problems in modern computational biology. It is therefore becoming increasingly important to predict protein structure from its amino acid sequence, using insight obtained from already known structures.The secondary structure is specified by a sequence classifying each amino acid into the corresponding secondary structure element (e.g., alpha, beta, or gamma).

DOI : 10.5121/ijsc.2012.3106

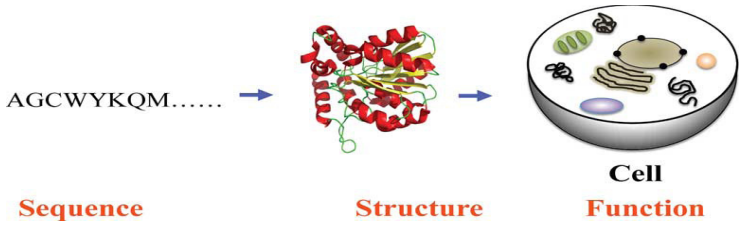

Fig 1: Protein Sequence-Structure-function.

Proteins are probably the most important class of biochemical molecules, although of course lipids and carbohydrates are also essential for life. Proteins are the basis for the major structural components of animal and human tissue. Extensive biochemical experiments [5], [6], [9], [10] have shown that a protein's function is determined by its structure. Experimental approaches such as X-ray crystallography [11], [12] and nuclear magnetic resonance (NMR) spectroscopy [13], [14] are the main techniques for determining protein structures. Since the determination of the first two protein structures (myoglobin and hemoglobin) using X-ray crystallography [5], [6], the number of proteins with solved structures has increased rapidly. Currently, there are about 40 000 proteins with empirically known structures deposited in the Protein Data Bank (PDB) [15].

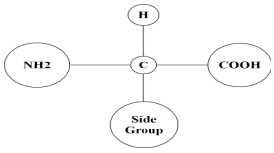

Fig 2: Structure of Amino Acid

2. PROTEIN STRUCTURE

2.1 Primary Structure

The primary structure refers to amino acid sequence is called primary structure. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code. Post-translational modifications such as disulfide formation, phosphorylations and glycosylations are usually also considered a part of the primary structure, and cannot be read from the gene [16].

2.2 Secondary Structure

Alpha helix may be considered the default state for secondary structure. Although the potential energy is not as low as for beta sheet, H-bond formation is intra-strand, so there is an entropic advantage over beta sheet, where H-bonds must form from strand to strand, with strand segments that may be quite distant in the polypeptide sequence. Two main types of secondary structure, the alpha helix and the beta strand, were suggested in 1951 by Linus Pauling and coworkers. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles

ψ and φ on the Ramachandran plot. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone [17].

2.3 Tertiary structure

The folding is driven by the non-specific hydrophobic interactions (the burial of hydrophobic residues from water), but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment [18].

2.4 Quaternary structure

Protein quaternary structure can be determined using a variety of experimental techniques that require a sample of protein in a variety of experimental conditions. The experiments often provide an estimate of the mass of the native protein and, together with knowledge of the masses and/or stoichiometry of the subunits, allow the quaternary structure to be predicted with a given accuracy. It is not always possible to obtain a precise determination of the subunit composition for a variety of reasons. The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer. Multimers made up of identical subunits are referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits are referred to with a prefix of "hetero-" (e.g. a heterotetramer, such as the two alpha and two beta chains of hemoglobin) [19].

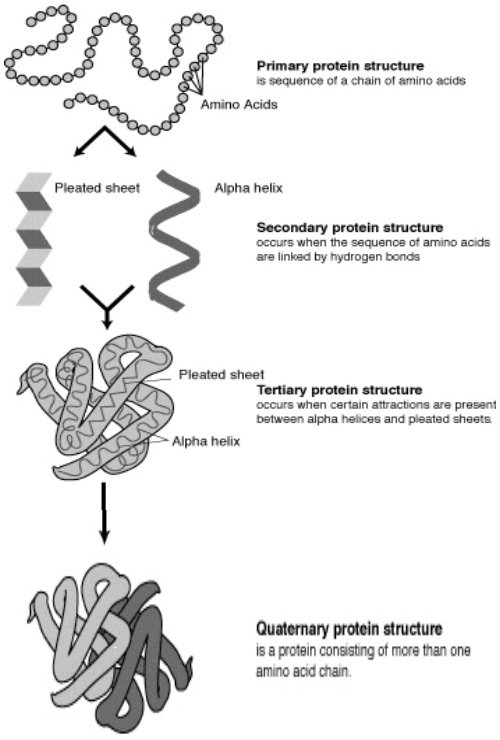

Fig 3: Four levels of protein structure.

3. METHOD AND MATERIAL

3.1. Database of homology-derived structure (HSSP)

3.1.1. Content of database

More than 300 files were produced, one for each PDB protein from the fall 1989 release of PDB with release 12 of EMBL/Swissprot (12305 sequences). This corresponds to derived structures for 3512 proteins or protein fragments; 1854 of these are homologous over a length of at least 80 residues. Some of these proteins are very similar to their PDB cousin, differing by as little as one residue out of several hundred.

3.1.2. Size of database

The increase in total information content in HSSP over PDB is as difficult to quantify as the increase in information when a homologous protein is solved by crystallography. A rough conservative estimate can be made as follows. The average number of aligned sequences is 103 per PDB entry. Of the 3512 aligned sequences (counting each protein exactly once) 1831 are more than 50% different (sequence identity) from any PDB cousin; after filtering out short fragments and potential unexpected positives by requiring an alignment length of at least 80 residues [29].

3.1.3. Limited database

Any empirical investigation is limited by the size of the database. Deviations from the principles observed here are possible as more and perhaps new classes of protein structures become known.

3.2. Dictionaries Secondary Structure Protein (DSSP)

The DSSP classifies residues into eight different secondary structure classes: H (o-helix), G (3_{10} - helix), I (21-helix), E (strand), B (isolated n-bridge), T (turn), S (bend), and- (rest). In this study, these eight classes are reduced into three regular classes based on the following Table 1. There are other ways of class reduction as well but the one applied in this study is considered to be more effective [21].

DSSP Class	8-state symbol	3-state symbol	Class Name
3_{10}-helix α-helix π-helix	G H I	H	Helix
B-strand	E	E	Sheet
Isolated β-bridge Bend Turn Rest(connection region)	B S T -	C	Loop

Table: 1 8-To-3 state reduction method in secondary structure

The RS 126 data set is proposed by Rost & Sander and according to their definition, it is non-homologous set.

3.3. Data Coding

Feature extraction is a form of pre-processing in which the original variables are transformed into new inputs for classification. This initial process is important in protein structure prediction as the primary sequences of the data are presented as single letter code. It is therefore important to transform them into numbers. Different procedures can be adopted for this purpose, however, for the purpose of the present study, orthogonal coding will be used to convert the letters into numbers.

Input and coding system

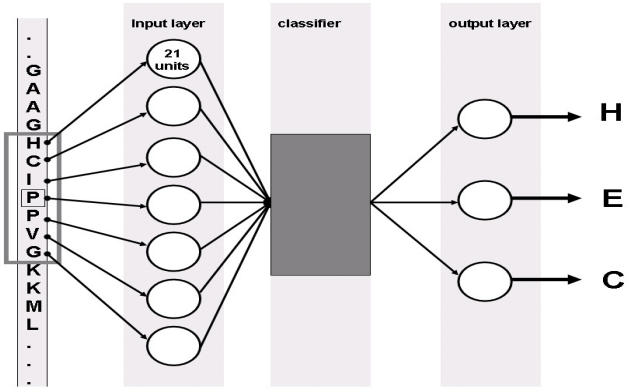

Fig 4: Input and output coding for protein secondary structure prediction

Fig: 3 give a network structure for a general classifier. The primary sequences are used as inputs to the network. To determine these inputs, a similar coding scheme as used by Holley and Karplus [28] has been adopted. To read the inputs into the network, a network encodes a moving window through the primary sequences.

4. STRUCTURE PREDICTION

4.1. Structure Prediction 1-D

Protein Sequence: Input 1D

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPMVLSEGEWQLVLHVWAKV

CCCCCHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHCCCCCHHHEEEEEEHHHHH

Protein Structure: Output 1D

4.2. Structure Prediction 2-D

Example depicts a predicted 2-D contact map with an 8 Angstrom cutoff. The protein sequence is aligned along the sides of the contact map both horizontally and vertically showing fig.4.

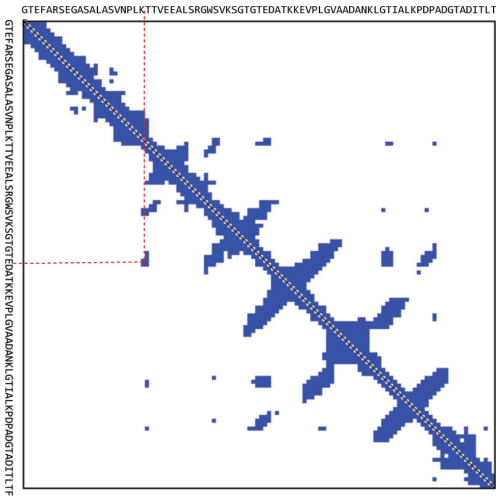

Fig 5: Two-dimensional protein structure prediction.

4.3. Structure Prediction 3-D

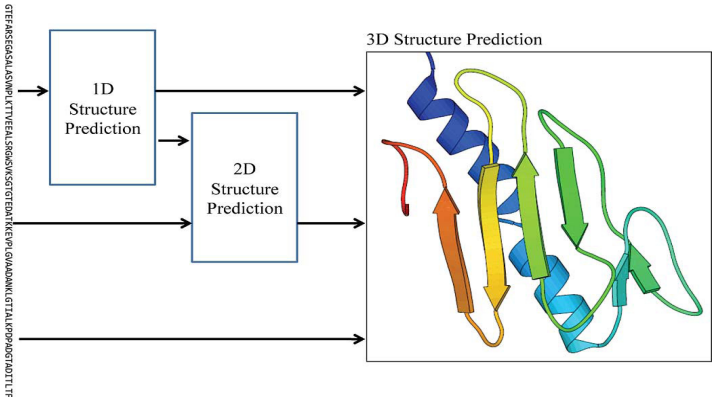

Fig 6: Three-dimensional protein structure prediction

There are 20 different amino acids that can occur in proteins. Their names are abbreviated in a three letter code or a one letter code. The amino acids and their letter codes are given in Table.

Glycine	Gly	G	Tyrosine	Try	Y
Alanine	Ala	A	Methionine	Mer	M
Serine	ser	S	Tryptophan	Trp	T
Threonine	Thr	T	Asparagine	Asn	A
Cysteine	Cys	C	Glutamine	Gln	G
Valine	Val	V	Histidine	His	H
Isoleucine	Ile	I	Aspartic Acid	Asp	A
Leucine	Leu	L	Glutamic Acid	Glu	G
Proline	Pro	P	Lysine	Lys	L
Phenylalanine	Phe	P	Arginine	Arg	A

Table: 2 Amino acids

5. INTRODUCTION SUPPORT VECTOR MACHINE

Support Vector Machine is supervised Machine Learning technique. The existence of SVM is shown in figure 6. Computer Vision is the broad area whereas Machine Learning is one of the application domains of Artificial Intelligence along with pattern recognition, Robotics, Natural Language Processing [18]. Supervised learning, Un-supervised learning, Semi-supervised learning and reinforcement learning are various types of Machine Learning.

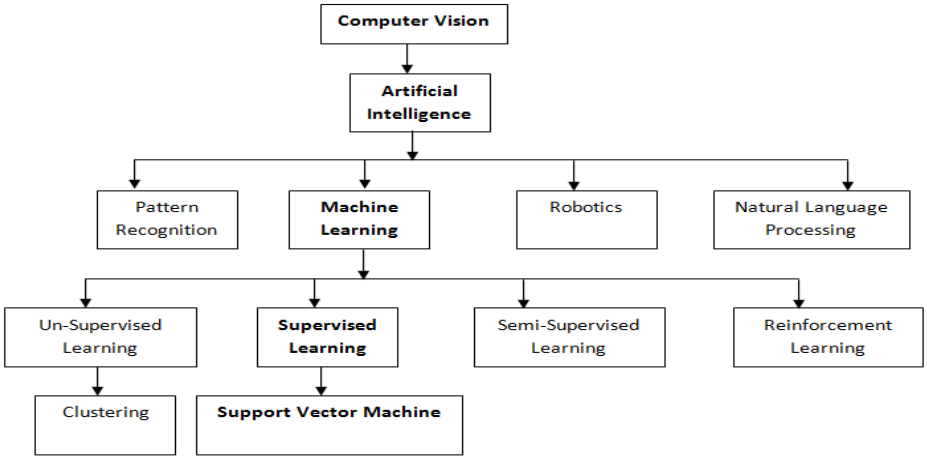

Fig 7: Existence of Support Vector Machine

5.1 Support Vector Machine

In this section, we give a brief review of SVM classification. SVM is a novel-learning machine first developed by Vapnik[16]. We consider a binary classification task with input variables X_i (i 1,...,l) = having corresponding labels Y_i= {-1,+1} . SVM finds the hyperplane to separate these two classes with a maximum margin. This is equivalent to solving the following optimization problem:

$$\text{Min } ½ \, w^T w \qquad \text{..... (1)}$$

$$\text{Subject to: } y_i(w.x_i+b) \qquad \text{.....(2)}$$

In Fig. 8, is a sample linearly separable case, soli d points and circle points represent two kinds of sample e separately. H is the separating hyperplane. H1 and H2 are two hyperplane through the closest points (the Support Vectors, SVs). The margin is the perpendicular distance between the separating hyperplane H1 and H2.

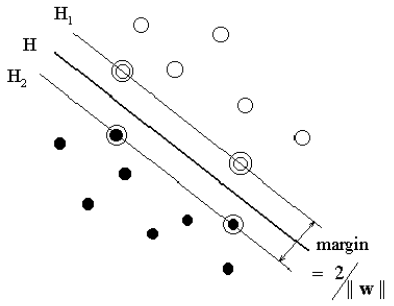

Fig 8: Optimal separation hyperplane

To allow some training errors for generalization, slack variables ξ_i and penalty parameter C are introduced. The optimization problem is re-formulated as:

$$\text{Min } ½\ w^T.w + C\sum_{i=1}^{l} \xi_i \qquad \text{..... (3)}$$

$$\text{Subject to: } \begin{matrix} y_i\,(w.x_i+b) \geq 1-\xi_i \\ i = 1.....l; \quad \xi_i \geq 0 \end{matrix} \qquad \text{..... (4)}$$

The purpose of $C\sum_{i=1}^{l} \xi_i$ is to control the number of

Misclassified samples. The users choose parameter C so that a large C corresponds to assigning a higher penal ty to errors [20].

By introducing Lagrange multiples α_i for the constraints in the (3) (4), the problem can be transformed into its dual form

$$\min_{\alpha} ½\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i . x) - \sum\alpha_i \qquad \text{..... (5)}$$

$$\text{Subject to: } \sum_{i=1}^{l} y_i \alpha_i , 0 \leq \alpha_i \leq C, i=1....,l \qquad \text{.....(6)}$$

Decision function as:

$$f(x) = \text{sign } (w.x + b)$$

$$= \text{sign } (\sum_{i=1}^{l} \alpha_i y_i (x_i .x) + b) \qquad \text{.....(7)}$$

For nonlinear case, we map the input space into high dimension feature space by a nonlinear mapping.

However, here we only need to select a kernel function and regularization parameter C to train the SVM. Our substantial tests show that the RBF (radial basis function) kernel , defined as,

$$K (x_i , x_j) = \exp (-\Upsilon \| x_i – x_j \|^2) \qquad \text{.....(8)}$$

With a suitable choice of kernel RBF (Radial Basis Function) the data can become separable in feature space despite being non-separable in the original input space.

6. RESEARCH AND METHODOLOGY

The proposed method used to RBF (Radial Basis Function) of SVM . Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its secondary, and tertiary from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Bioinformatics method to used protein secondary structure prediction mostly depends on the information available in amino acid sequence. SVM represents a new approach to supervised pattern classification which has been successfully applied to a wide range of pattern recognition problems, including object recognition, speaker identification, gene function prediction with microarray expression profile, etc. It is a good method of protein structure prediction which is based on the theory of SVM [26].

6.1. Protein Structure Prediction Based SVM

Protein structure prediction has performed by machine learning techniques such as support vector machines (SVM's).Using SVM to construct classifiers to distinguish parallel and antipararallel beta sheets. Sequences are encoded as psiblast profiles. With seven-cross validation carried on a non-homologous protein dataset, the obtain result shows that this two categories are separable by sequence profile [27]. β-turns play an important role in protein structures not only because of their sheer abundance, which is estimated to be approximately 25% of all protein residues, but also because of their significance in high order structures of proteins. Hua-Sheng Chiu introduces a new method of β-turn prediction based SVM that uses a two-stage classification scheme and an integrated framework for input features. The experimental results demonstrate that it achieves substantial improvements over Beta turn, the current best method [28].

7. RESULT ANALYSIS

$$\textbf{Accuracy rate} = \frac{\sum \textbf{Correct Predicted Instance}}{\sum \textbf{No. of Instance}} * 100$$

Protein Name	Protein ID	Accuracy Structure Prediction (1D)	Accuracy Structure Prediction (2D)	Accuracy Structure Prediction (3D)	Average Accuracy Prediction	Execute Time
BACTERIOPHYTOCHROME	1ztu	99.75	100.00	99.72	99.82	2m5 Sec
RIBONUCLEASE	2bir	98.19	100.00	98.28	98.82	15 Sec
GLOBIN	2w31	98.17	100.00	99.72	99.31	20 Sec
MYOGLOBIN	101m	98.74	100.00	99.60	99.45	32 Sec

Table: 3 Protein Prediction 1D, 2D, and 3D Accuracy

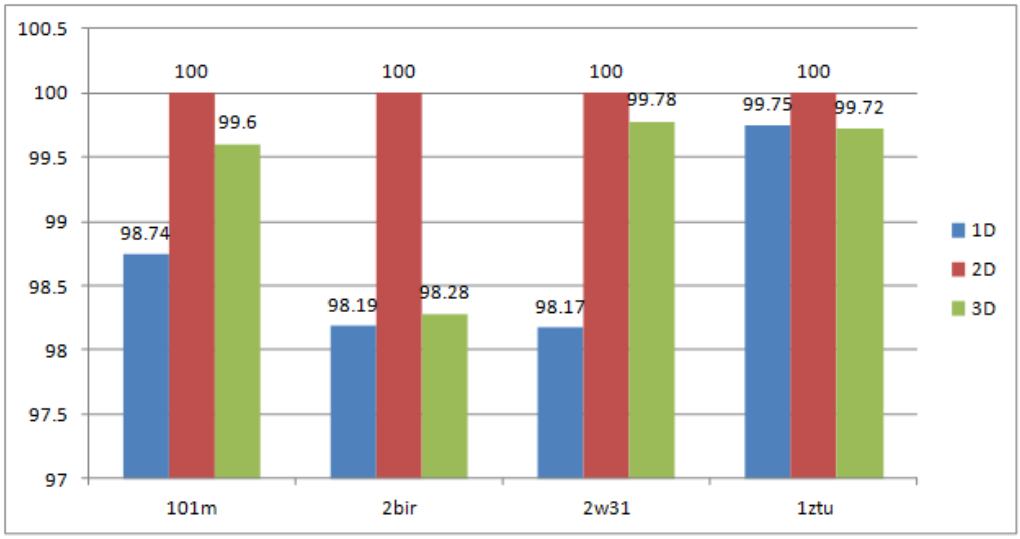

Fig 9: Prediction Accuracy of 1-D, 2-D, and 3-D Structure

Average Accuracy rate = ∑ [1D+2D+3D] /3 *100

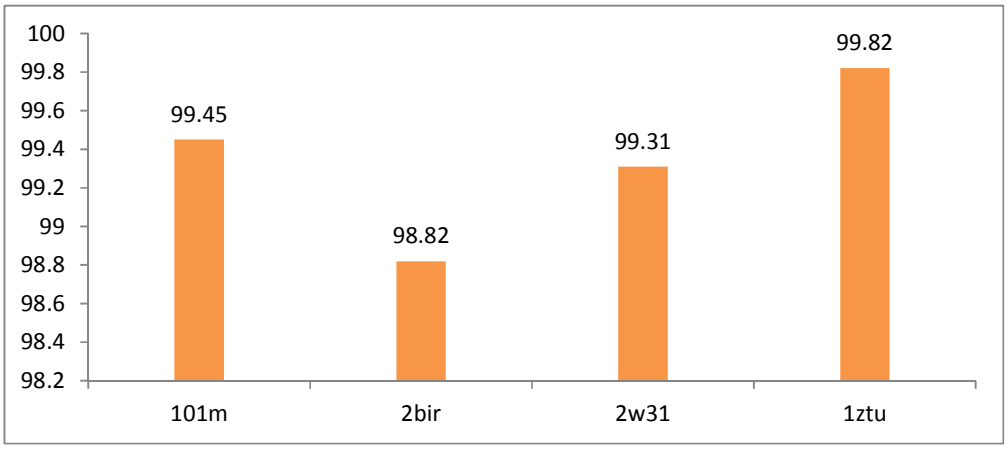

Fig 10: Protein Prediction 1-D, 2-D, and 3-D Average Accuracy

8. CONCLUSIONS

Support Vector Machine is learning system that uses a high dimensional feature space, trained with a learning algorithm from optimization theory. Since SVM has many advantageous features including effective avoidance of over-fitting, the ability to manage large feature spaces, and information condensing of the given data, it has been gradually applied to pattern classification problem in biology.

As a part of future work, accuracy rate need to be tested by increasing datasets. This paper implements RBF kernel function of SVM. By application of other kernel functions such as Linear Function, Polynomial Kernel function, Sigmoidal function, this accuracy of Protein Structure Prediction can be further increased.

9. REFERENCES

[1] Rost, B. and Sander, C. "Improved prediction of protein secondary structure by use of sequence profile and neural networks.", Proc Natl Acad Sci U S A 90, pp. 7558-62, 1993

[2] Blaise Gassend et al. Secondary Structure Prediction of All-Helical Proteins Using HiddenMarkov Support Vector Machines. Technical Report MIT-CSAIL-TR-2005-060, MIT, October 2005..

[3] Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Research, 2006, Vol. 34.

[4] Yann Guermeur, Gianluca Pollastri et al.Combining protein secondary structure prediction models with ensemble methods of optimal complexity. Neurocomputing, 2004(56):305-327.

[5] Jayavardhaha Rame G.L. etal,"Disulphide Bridge Prediction using Fuzzy Support Vector Machines",Intelligent Sensing and Information Processing, pp.49-54 2005.

[6] Long-Hui Wang, Juan Liu. "Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme", Genome Informatics, 2004, 15(2): pp. 181–190.

[7] K. A. Dill, "Dominant forces in protein folding," Biochemistry, vol. 31, pp. 7134–7155, 1990.

[8] R. A. Laskowski, J. D. Watson, and J. M. Thornton, "From protein structure to biochemical function?," J. Struct. Funct. Genomics, vol. 4, pp. 167–177, 2003.

[9] A. Travers, "DNA conformation and protein binding," Ann. Rev. Biochem., vol. 58, pp. 427–452, 1989.

[10] P. J. Bjorkman and P. Parham. "Structure, function and diversity of class I major histocompatibility complex molecules," Ann. Rev. Biochem., vol. 59, pp. 253–288, 1990.

[11] L. Bragg, "The Development of X-Ray Analysis", London, U.K.: G. Bell, 1975.

[12] T. L. Blundell and L. H. Johnson, Protein Crystallography. New York: Academic, 1976.

[13] K. Wuthrich, NMR of Proteins and Nucleic Acids. New York: Wiley, 1986.

[14] E. N. Baldwin, I. T. Weber, R. S. Charles, J. Xuan, E. Appella, M. Yamada, K. Matsushima, B. F. P. Edwards, G. M. Clore, A. M. Gronenborn, and A. Wlodawar, "Crystal structure of interleukin 8: Symbiosis of NMR and crystallography," Proc. Nat. Acad. Sci., vol. 88, pp. 502–506, 1991.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," Nucl. Acids Res., vol. 28, pp. 235–242, 2000.

[16] C Cortes and V Vapnik. Support-vector networks. Machine Learning, 1995, 20(3): 273-297.

[17] Osuna E., Freund R., Girosi F.: "Support vector machines: Training and applications", Massachusetts Institute of Technology, AI Memo No. 1602. 1997.

[18] Sujun Hua and Zhirong Sun. "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach",J. Mol. Biol.(2001)308, 397-407.

[19] Jian Guo, Hu Chen, Zhirong Sun. "A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles.PROTEINS: Structure, Function, and Bioinformatics", 54:738–743 (2004).

[20] Shing-Hwang Doong,Chi-yuan Yeh.A Hybrid Method for Protein Secondary Structure Prediction. Computer Symposium,Dec.12-17,2004,Taipei,Taiwan

[21] Ian H. Witten, Eibe Frank, "Data Mining-Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, Second Edition, 2005, pp. 7-9.

[22] Sujun Hua and Zhirong Sun. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach .J. Mol. Biol. (2001)308, 397-407.

[23] Anjum Reyaz-Ahmed and Yan-Qing Zhang" Protein Secondary Structure Prediction Using Genetic Neural Support Vector Machines"

[24] Lipontseng Cecilia Tsilo Prediction secondary structure prediction using neural network and SVM.33

[25] Jianlin Cheng, Allison N. Tegge, Member, IEEE, and Pierre Baldi, Senior Member,IEEE REVIEWS IN BIOMEDICAL ENGINEERING,VOL.1,2008."Machine Learning Method for Protein Structure Prediction.

[26] Olav Zimmermann and Ulrich H.E. Hansmann. Support Vector Machines for Prediction of Dihedral Angle Regions.Bioinformatics Advance Access. September 27, 2006

[27] Longhui Wang, Olav Zimmermann, and Ulrich H.E. Hansmann. "Prediction of Parallel and Antiparallel Beta Sheets Based on Sequence Profiles Using Support Vector Machines", Neumann Institute for Computing Workshop, 2006.

[28] Hua-Sheng Chiu, Hsin-Nan Lin, Allan Lo, Ting-Yi Sung et al, "A Two-stage Classifier for Protein β-turn Prediction Using Support Vector Machines" IEEE International Conference on Granular Computing,2006.

[29] Chris Sander and Reinhard Schneider "Database of homology derived protein structures and the structural meaning of sequence alignment." Proteins, 9, 56-69, (1991).

Authors

Anil Kumar Mandle: Has completed his B.E. in Information Technology from Guru Ghasidas University Bilaspur (C.G.) in the year 2007 and is pursuing M.Tech from SATI Vidisha (M.P.)

Pranita Jain: Astt.Prof. Information Technology Department Samrat Ashok Technological Institute Vidisha, (M. P.) INDIA

Prof. Shailendra Kumar Srivastava: HOD, Information Technology Department Samrat Ashok Technological Institute Vidisha, (M. P.) INDIA