Protein Function Prediction using SVM Kernel Approach

Anjna Jayant Deen[1] Manasi Gayanchandani[2]

Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India
email: anjnadeen123@gmail.com[1], manasi_gyanchandani@yahoo.co.in[2]

***Abstract*— Bioinformatics are rapidly and fast growing area of proteomics and genome. Protein datasets and their vector space are large in size. This results in difficulty for SVM classifier to train large group of protein sequence datasets. Therefore, the basic classification algorithms cannot handled the problem to train large support vectors. The objective of this study is using kernel, optimize the classifier support vector and enhance the classification accuracy. In this work, k-fold cross validation is used on different type of SVM kernels ,experimental test accuracy of protein function classes is found to be in RBF kernel is 97.09%. This work suggests the usefulness of SVM kernel methods in the cataloguing of protein functional classes and its possible application in protein function prediction.**

Keywords*— *Machine Learning ,Support Vector Classifier, RBF, Linear kernel and Polynomial kernel and protein function.

I. INTRODUCTION

Machine learning classification algorithms are widely implemented in field of bioinformatics. Protein function prediction is major task to find the various functional and structural class of protein from sequence information alone[4]. Support vector machine (SVM) kernel methods are a class of algorithms for matching the patterns, it find the similarity and relation form heterogeneous and homogenous data sets[3]. Ordering this different pattern relation and similarity based on ranking, co-relation, classification, clusters and degree of similarity , kernel captures the inner information between all pair of sequence dataset in the feature space in the form of kernel trick. most kernel algorithm are based on optimization[2]. The kernel based training parameter tuning and feature selection parameter significant impact on the classification accuracy. Support vector classifier has an explore on a number of various protein classes in cell binding site and determining protein function[1]. These information utilize multiple data sources in a combined sequence pattern. Proteins function responsible in cell distribution of immune system, response to drug absorption, finding disease, and protein-protein interaction[5]. Kernel based learning methods problem formulate for dimension of the original vector space, due to this reasons high throughput protein data, the kernel replace the traditional Euclidean inner vector space. With the collected sequence information, main attention to the expansion of methods for the prediction of protein function [5,8,9,10] from the sequence. As a result, alternative classification methods to be developed in the study of protein function. In [17] have used Multiple Kernels methods for predicting protein function. Authors [16] have proposed a special multilable transductive classifier design to predict multiple functions form several unlabeled proteins sequence data. [15]The method is simpler and faster, and further composite networks with improved function prediction accuracy .In paper [18] have used ,an ensemble classifier to predict accuracies form subcellular location of protein function benchmark datasets using KNN and SVM algorithms. In papers [13,14,23] have ,the improved prediction accuracies and this reveals that gene ontology annotations and hydrophobicity of amino acids help to predict subcellular locations of eukaryotic proteins. In [12] have proposed combining heterogeneous sources of data is essential for accurate prediction of protein

function. SVM potentially have been used by various researcher for protein function prediction.

II *Support vector Machine (SVM)*

Support vector machine is the most popular classifier. It is supervisor learner based on statistical learning theory, widely used for proteins structure and function. Generally, SVM is used for both binary and multiclass separation. In the case of linear data, SVM draws a separating line called hyper-plane, which plays the role of decision surface to separate the data into two different classes.

A. Protein sample selection and kernel method

Machine learning tool is the knowledge learning based on decision in classification and prediction phase. Decision making parameters should be deployed classifier for fast identification of new patterns detection efficiently. Two important powerful tool in machine learning theory are neural network and kernel methods[5,23].Protein microarray data feature representation in vector space in a m×n data matrix form, its row and column vector depends on protein sequence analysis[5].

Feature representation in Vector Space: vector depends on sequence data analysis, classification of genes. The feature vectors will correspond to the rows (instead of columns), i.e. the feature vector of the j^{th} sample will be expressed as,

$y(j) = x(j)_1\ x(j)_2 \ldots x(j)_n$ for $j=1,\ldots,m$ (1), a feature vector will be represented by an m dimensional vector formed by m sequence data. Vector space Linear algebra provides the basic theory of manipulating the patterns in vector spaces. The basic vector algebra covers the ideas of linear independence, subspaces and their spans, norm and inner product, and linear transformations. The fundamental concept of the vector space (or linear space) plays a key role in most mathematical algorithms in machine learning, the notion of intrinsic space associated with a kernel function. The intrinsic space is so named because of independent of the training dataset. The dimension of the space is denoted by j and will be referred to as the intrinsic degree. This degree indicates the training efficiency and computational cost.

B Feature representation and dimension reduction

The kernel-based learning models may be based on the following representations.

Intrinsic-space representation:- This is conceptually simpler and involves a full process with explicit feature mapping to the intrinsic space, the learning model will treat the intrinsic dataset just as the original dataset.

In kernel based SVM classification transforms amino acid sequence into kernels and then integrated into a by following steps [4],[34]:

Step 1. Feature Mapping and then train a classifier.

Step 2. SVM kernel optimizes the weights of vectors.

Dimension of the intrinsic kernel vector space, both steps jointly determined by the kernel function and the training dataset as shown in fig.3.

III *Experimental Setup*

A. Data Set Description

In this study ,data collected from protein data bank Swiss-Prot (collected 560,459 amino acid sequences), namely Human(20,431), Mouse(17,019), and Rat(8,068). All these protein benchmarks amino acid sequence data are downloaded from the Swiss-Prot(Protein Data Bank) in FASTA format. fig 1 filtered the proteins in species class and their functional annotation, in this data preprocessing there is 45,518 protein sequence after data parsing as shown in fig. 2. Study by Mostafavi[19] filtered the proteins sequence,to include only those functions that had at least 30 proteins and at most 100 proteins. In paper [19], protein function annotated according to the biological process function categories in the gene ontology database. Thus, in all total of 3509 functions is annotated. The statistics of function annotated protein listed in Table 1.

TABLE1.Protein Function Prediction Benchmarks Statistics

Dataset	Proteins Sequence
Human	20,431
Mouse	17,019
Rat	8,068
Total Function	3509

\

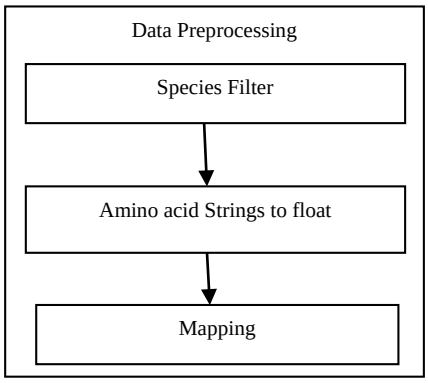


Fig.1 Data Pre-processing

Data parsing during pre-processing and feature selection are important in classification. For each amino acid sequence, feature vectors dimensions encoded to representations residue properties. In this study, we implement in Python (HP Z80 workstation).

IV Result and discussion

A. Accuracy

Training and testing accuracies: The training accuracy is a common metric used to learned classifier can differentiate the positive and negative data drawn from the training dataset.

In testing accuracy reflect the classification accuracy of the learned classifier on the testing dataset, which has no overlap with the training dataset. Testing and prediction accuracies It is commonly accepted that the testing accuracy serves as a reasonable indicator of the prediction performance. Therefore, all the learned kernel classifiers must undergo the evaluation process during the testing phase. The SVC based kernel receiving the best cross-validation will be deployed for protein function prediction fig 4. Prediction accuracy of results is commonly measured by the quantity of True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) [12],[21]. In additional quantity to measure these is sensitivity, specificity and overall accuracy (Q) performance measures defined by

$$\text{Sensitivity} = TP/(TP + FN), \quad (2)$$

$$\text{Specificity} = TN/(TN + FP), \quad (3)$$

$$Q(\text{overall accuracy}) = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

are also useful in assessing the prediction accuracy . All these quantities are used in the evaluation of SVM kernel classification of proteins in this work.

B. Parameter Tuning and Results

In this study K-fold cross validation is applied on protein sequence. The C and gamma **γ** functions are tuned to fitting data sample into train class and test class. as show in Table 2

Table 2. 5- fold cross validation parameter

Parameter Kernel	**C**	**γ**
RBF	2^{1}	2^{-7}
Linear	2^{0}	2^{-5}
Polynomial	2^{0}	2^{-6}

Prediction accuracy depends on various feature descriptor vector and diversity of protein sample. SVM kernel methods has been improved with the more protein data, shown in Table (3).

Table 3. Overall accuracy results based on support vector classifier with different kernels.

Protein datasets [Human, Mouse, Rat]	Overall Accuracy%
SVC with Linear Kernel	58.13
Linear SVC with Linear Kernel	63.03
SVC with RBF kernel	97.09
SVC with Polynomial Kernel	68.89

Accuracy range in this study is RBF kernel 97.09 % good perform as compare with other kernel. Accuracy of linear kernel is 58.13% and 63.03% and polynomial kernel accuracy is 68.89%. The sensitivity and specificity are the range of 51.10%-96.7% and 89.01-99.6%, respectively. Therefore in this experimental results observed RBF kernel is best in classification of proteins into specific functional class shown in fig. 4a and 4b.

Fig.3. SVM kernel based Protein function Prediction Model

V Conclusion

Testing results on protein data, Human, Mouse and Rat sequence functional classes suggests that SVM appears to be a potentially useful tool for protein function prediction by means of classification of proteins into specific functional classes. Further works on samples collection for every functional class, refined samples selection, and improvement of SVM kernel and feature vector selection will help in development of SVM into a practical protein function prediction tool. Machine learning methods may be further improved by choosing a more refined set of samples for each classes.

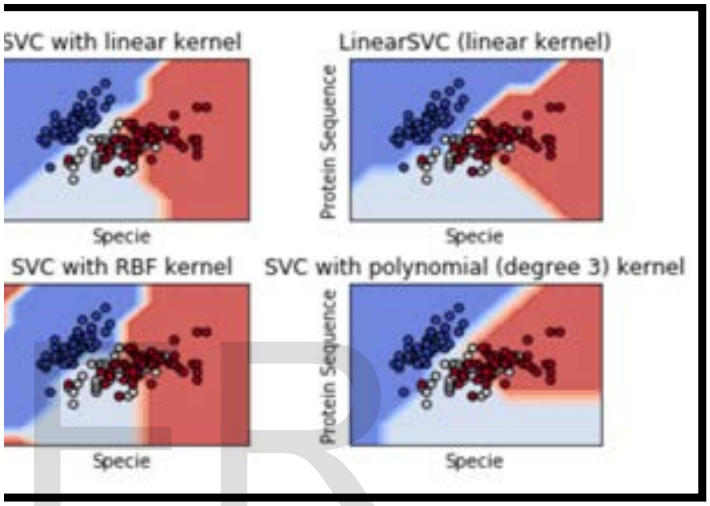


Fig. 4a. Classification graph of Support vector classifier

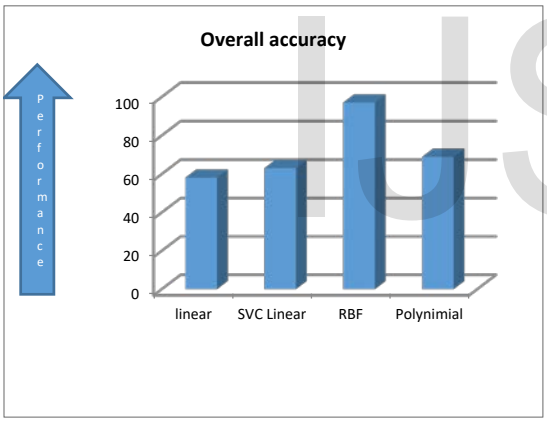


Fig. 4b. Classification Bar-chart of Support vector classifier

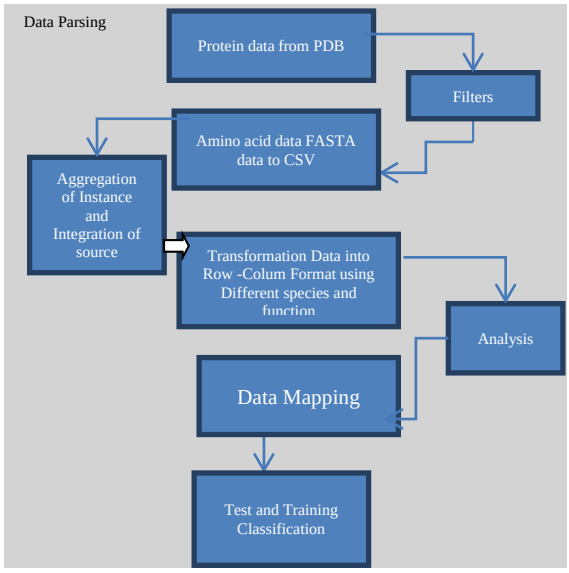
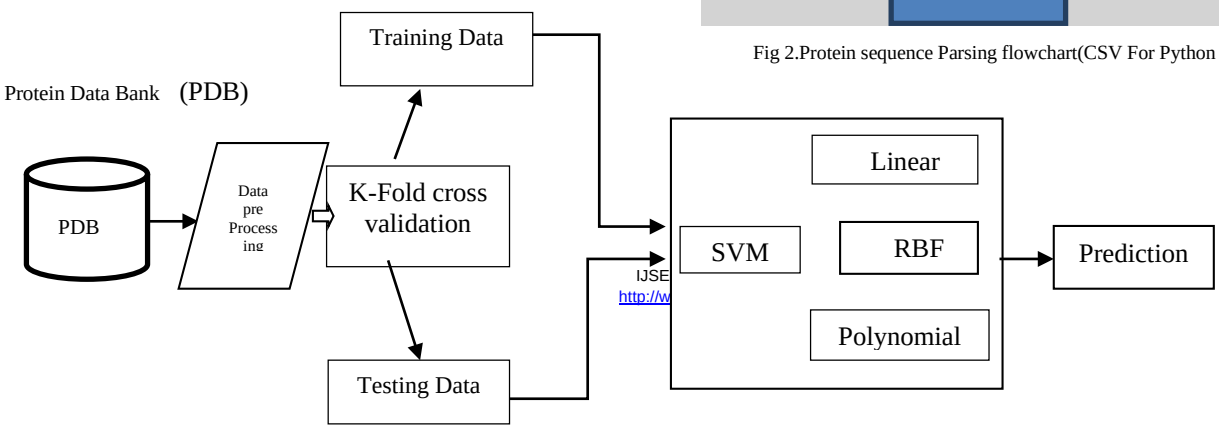


Fig 2.Protein sequence Parsing flowchart(CSV For Python Program)



References

[1] Cheng-lung Huang Jen Wang " GA- based feature selection and optimization support vector machine" https://doi.org/10.1016/j.eswa.2005.09.024
[2] Johar M. Ashfaque, Amer Iqbal "Introduction to Support Vector Machines and Kernel Methods" April 12, 2019 publication at https://www.researchgate.net/publication/332370436
[3] Mirko Polato, Ivano Lauriola and Fabio Aiolli" A Novel Boolean Kernels Family for Categorical Data" Entropy 20(6),June 2018 DOI: 10.3390/e20060444
[4] S. Y. Kung " Kernel Methods and Machine Learning" Cambridge University Press 978-1-107-02496-0
[5] S. Yaman ,J. Pelecanos, "Using Polynomial Kernel Support Vector Machines for Speaker Verification," in *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 901-904, Sept. 2013.
[6] J.W. Lengeler, Metabolic networks: a signal-oriented approach to cellular models, Biol. Chem. 381 (2000) 911.
[7] H. Siomi, G. Dreyfuss, RNA-binding proteins as regulators of gene expression, Curr. Opin. Genet. Dev. 7 (1997) 345.
[8] GiorgioValentini ,ReviewArticle Hierarchical Ensemble Methods for Protein Function Prediction; Hindawi Publishing Corporation ISRN Bioinformatics Volume 2014, Article ID 901419, 34 pages
[9] Hae-Jin Hu, Yi Pan, Senior Member, IEEE, Robert Harrison, and Phang C. Tai ;Improved Protein Secondary Structure Prediction Using Support Vector Machine With a New Encoding Scheme and an Advanced Tertiary Classifier; IEEE Transaction on Nanobioscience, Vol .3,no.4,dece 2004
[10] Robert E. Langlois, Alice Diec, Yang Dai, Hui Lu ;Kernel Based Approach for Protein Fold Prediction from Sequence; Department of Bioengineering, University of Illinois at Chicago, IL, USA
[11] A. Godzik, M. Jambon and I. Friedberg;Visions & Reflections (Minireview) Computational protein function prediction: Are we making progress? Cell.Mol.Life Sci. 64 (2007) DOI 10.1007/s00018-007-7211-y Birkhuser Verlag, Basel, 2007
[12]Karsten M. Borgwardt, HansPeter Kriegel ;Kernel Methods for Protein Function Prediction;
Institute for Informatics, LMU Munich, Oettingenstr. 67, 80538 Munic h, Germany ;Automatic Function Prediction - Special Interest Group (AFP-SIG), Detroit, USA, 2005
[13] Guoxian Yu, Guangyuan Fu, Jun Wang, and Hailong Zhu; Predicting Protein Function via Semantic Integration of Multiple Networks; IEEE/ACM transaction on computational biology and bioinformatics vol. 13, NO. 2, March 2016
[14] Jun Hu, Yang Li, Ming Zhang, Xibei Yang, Hong-Bin Shen, and Dong-Jun Yu Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-based Features and Boosting Multiple SVMs; IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1545-5963, 2016
[15] Zheng Rong Yang and Kuo-Chen Chou, Bio-support vector machines for computational proteomics; Vol. 20 no. 5 2004, pages 735–741 DOI: 10.1093/bioinformatics
[16] Guoxian Yu, Huzefa Rangwala, Carlotta Domeniconi, Guoji Zhang, and Zhiwen Yu ;Protein Function Prediction Using Multilabel Ensemble Classification; IEEE/ACM transaction on computational biology and bioinformatics vol. 10, no. 4, July 2013
[17] Guoxian Yu, Huzefa Rangwala, Carlotta Domeniconi, Guoji Zhang, and Zili Zhang ;Predicting Protein Function Using Multiple Kernels; IEEE/ACM transaction on computational biology and bioinformatics vol. 12, no. 1, Jan/Feb 2015
[18] Ali Al-Shahib, Rainer Breitling and David R Gilbert;Predicting Protein Function by Machine Learning on Amino Acid Sequence;BMC Genomics,2007
[19] Sara Mostafavi ,Qadic"Fast integration of heterogeneous data sources for predicting gene function with limited annotation" *Bioinformatics*, Volume 26, Issue 14, 15 July 2010, Pages 1759–1765.
[20] CJ.C. Burges, A tutorial on support vector machine for pattern recognition, Data Min. Knowl. Disc. 2 (1998) 121-123.
[21] C.H.Q. Ding, I, Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics 17 (2001) 349.
[23] Z. Yuan, J. Burrage, J.S. Mattick, Prediction of protein solvent accessibility using support vector machines, Proteins 48 (2002) 566.
[24] S.J. Hua, Z.R. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, J. Mol. Biol. 308 (2001) 397.
[25] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, Comput. Chem. 26 (2002) 293.
[26] H. Drucker, D.H. Wu, V.N. Vapnik, Support vector machine for spam categorization, IEEE T. Neur. Network 10 (1999) 1048.
[27] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal Mach. Learn. Res. 2 (2001) 45.
[28] Z.Y. Li, S.W. Tang, S.C. Yan, Multi-class SVM classifier based on pairwise coupling. Lect. Notes Comput. Sci. 2388 (2002) 321.
[30] N. Thubthong, B. Kijsirikul, Support vector machines for Thai phoneme recognition, International Journal Uncertain. Fuzz. 9 (2001) 803.
[31] M. Gordan, C. Kotropoulos, I. Pitas, A temporal network of support vector machine classifiers for the recognition of visual speech, Lect. Notes Anal. Intell. 2308 (2002) 355.
[32] V.V. Gavrishchaka, S.B. Ganguli, Support vector machine as an efficient tool for high-dimensional data processing: application to substorm forecasting, J. Geophys. Res. 106 (2001) 29911.
[33] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389.
[34]Karsten M. Borgwardt, HansPeter Kriegel ;Kernel Methods for Protein Function Prediction;
Institute for Informatics, LMU Munich, Oettingenstr. 67, 80538 Munic h, Germany ;Automatic Function Prediction - Special Interest Group (AFP-SIG), Detroit, USA, 2005
[35] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906.
[36] P. Pavlidis, J. Weston, J.S. Cai, W.S. Noble, Learning gene functional classifications from multiple data types, international Journal in Computer Biology. 9 (2002) 401.
[37] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Nat. Acad. Sci. USA 97 (2000) 262.
[38] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for predicting HIV protease cleavage sites in protein, J. Comput. Chem. 23 (2002) 267.
[39] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University, Cambridge, 2000.
[40] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Anderson, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412.
[41] Robert E. Langlois, Alice Diec, Yang Dai, Hui Lu ;Kernel Based Approach for Protein Fold Prediction from Sequence; Department of Bioengineering, University of Illinois at Chicago, IL, USA
[42] J.E. Roulston, Screening with tumor markers, Mol. Biotechnol. 20 (2002) 153.
[43] C.Z. Cai, W.L. Wang, Y.Z. Chen, Support vector machine classification of physical and biological datasets, Int. J. Mod. Phys. C 14 (5) (2003) in press.