International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 5 (2013): 239-244
Research Article
Open Access

ISSN: 2320-0634

# Protein Function Prediction Using Support Vector Machine

Arvind Kumar Tiwari[1]* and R B Mishra[2]

*Department of Computer Engineering, IIT BHU, Varanasi, India*

*Corresponding author: Arvind Kumar Tiwari; e-mail: arvind.rs.cse12@itbhu.ac.in

Received: 15 August 2013 Accepted: 25 August 2013 Online: 01 September 2013

## ABSTRACT

Protein function prediction is very important and challenging task in Bioinformatics. In this paper we have used proteins represented by a set of enzymes i.e. Oxidoreductase, Transferase, Hydrolase, Isomerase, Ligase, and Lyase, extracted from the Enzyme Commission (EC) classification to build the models. In this paper we have used Support vector machine to predict protein function which is more efficient for resolving linear and non linear classification problems. We have used protein dataset available at PDB using features such as primary structures, secondary structures , molecular weight, structural molecular weight, chain length, atom count, ligand molecular weight and residue count as training parameters and EC number as corresponding output . Here we used expert model of support vector machine, with RBF kernel function where width is 0.10 and parameter C is 10. The result in this paper using these parameters shows that the overall average accuracy is 84.07%.

**Keywords:** Proteins, Function prediction, Support vector machine, classification, Enzyme

## INTRODUCTION

Proteins are formed from a set of 20 amino acids and the function of a protein is closely related to the structure. There are various function of protein such as catalysis, transport and Information. Enzyme behaves like a catalyst which speed up the rate of reaction without becoming the part of reaction. The primary structure of a protein is the sequence of amino acids, secondary structure is the formation of alpha helixes ,beta sheets and loops and the tertiary structure is responsible for the spatial arrangement of the protein and the quaternary structure refers to the proteins that have more than one chain of amino acids . In this paper we used the proteins that are classified according to EC number. Finding protein function is an important task which supports the research for novel drug design. In this paper we used six classes of enzymes Oxidoreductase, Transferase, Hydrolase, Isomerase, Ligase, and Lyase. In this paper we used eight features primary structures, secondary structures, molecular weight, structural molecular weight, chain length, atom count, ligand molecular weight and residue count to predict the protein function. Using these features we construct the expert model of support vector machine to predict the protein function.

Here we describe the previous research work carried out for the protein function prediction or classification and we also discuss about the various classifiers models that are used in this study. L.Y. Han et al [4] proposed a method to predict functional family of protein that is useful for protein function prediction. Every protein sequence is represented by a set of amino acid composition by using these composition he used SVM, supervised machine learning and the result of this model is compared with the Naïve Bayes and C 4.5. C.Z. Cai et al [2] used the SVM for protein function classification. He used a various protein classes such as RNA-binding, homodimer, drug absorption, drug excretion etc. He found the testing accuracy between 84-96%. Paul D. Dobson et al [7] proposed a method that can assign the function from the structure of protein by using EC number. He used one-class versus one-class SVM to predict the protein function. He found

the accuracy between 35-60%. Luiz C. Borro et al [6] proposed a method for predicting EC number. He used various features of the protein structure find from STRING_DB and used Bayesian classifier to predict the protein function. He found the accuracy 45.3%. Yong-Cui Wang et al [10] proposed a method to predict enzyme functions using amino acid composition, their neighborhood relationship to each other, and the hierarchical structure of the class. He compared the results from the attributes considered and concludes that the information from all three together offers better results. Using the SVM classifier, they obtain a prediction rate of between 81% and 98%.

## MATERIALS AND METHODS

### Data source

The coding regions of mitochondrial genomes of the five species of phylum Chlorophyta were retrieved from the GenBank (http://www.ncbi.nlm.nih.gov/GenBank) database of NCBI. The species used in the analysis and their accession numbers are listed below (Table 1).

### Support Vector Machine

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data The SVM can be characterized as a machine learning algorithm capable of resolving linear and non-linear classification problems. The principal idea of classification by support vector is to separate examples with a linear decision surface and maximize the margin of separation between the classes to be classified. The SVM is more useful for analyzing large number of datasets, that is, those with a large number of predictor fields .SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, and then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong. After the transformation, the boundary between the two categories can be defined by a hyperplane.The mathematical function used for the transformation is known as the kernel function. SVM supports the Linear, Polynomial, Radial basis function (RBF) and Sigmoid kernel types. When there is a straightforward linear separation of then linear function is used otherwise we used Polynomial, Radial basis function (RBF) and Sigmoid kernel function. Besides the separating line between the categories, a classification SVM model also finds marginal lines that define the space between the two categories. The data points that lie on the margins are known as the support vectors. The margin will be wider between the two categories for better model and prediction of new record. When the margin is not wider then the model is called overfitted. A small amount of misclassification can be accepted in order to widen the margin. To find the optimum balance between a wide margin and a small number of misclassified data points. The kernel function has a regularization parameter (C), a highly sensitive parameter which determines the flexibility of the margin of the hyperplanes and therefore controls the trade-off between these two values that is error and margin. These kernel functions are capable of mapping the data set in different spaces, making it possible to use a hyper plane to do the separation. This directly influences the results obtained by the classifier.

### Linear Classification

Consider the training sample ($x_i$, $y_i$) for i=1, 2 ….n where $x_{i \in R^N}$ is the input pattern and $y_i \in \{-1, +1\}$ is the desired output. Given a weight vector w and bias b. The two classes can be separated by two margins parallel to hyper plane.

$$w^T x_{i+} b \geq 1 \quad \text{For } y_i=1 \text{ and } x \in P \qquad (1)$$

$$w^T x_{i+} b \leq -1 \quad \text{For } y_i=-1 \text{ and } x \in N \qquad (2)$$

where w= $(w_1, w_2, w_3 \ldots\ldots w_n)^T$ is a vector of n element. Equation (1) and (2) are the equation of a decision surface in the form of a hyperplane. It is given by

$$f_{w,b}(x) = sign(w^T x + b) \qquad (3)$$

Combined the inequalities (1) and (2) we get

$$y_i(w^T x_i + b) \geq 1 \quad \text{i=1,2, 3……..n} \qquad (4)$$

So there exist a hyperplane for each group of training data. The goal of SVM is to determine an optimal weight $w_0$ and optimal bias $b_0$ such that the selected hyperplane separates the training data with maximum margin. The hyperplane determined by $w_0$ and $b_0$ is called optimal separating hyperplane (OSH).

So the equation of optimal hyperplane is given by

$$w_o^T x_i + b_0 = 0 \qquad (5)$$

For any particular data point ($x_i$, $y_i$) for which equation (5) is satisfied are called support vectors.

To find optimal hyperplane

$$y_i(w^T x_i + b) \geq 1 \quad \text{i=1, 2, 3……..n}$$

With maximum margin $2\eta$, where $\eta = 1/\|w\| = 2/\|w\|$ (6)

It is equivalent to minimize the cost function

$$\phi(w) = \frac{1}{2} w^T w = \frac{1}{2}\|w\|^2 \; where \|w\| = w^T w \qquad (7)$$

Subject to $y_i(w^T x_i + b) \geq 1$ i=1,2, 3……..n

To solve this optimization problem for a given training sample ($x_i$, $y_i$) for i=1, 2 ….n. find the Lagrange multiplier $\alpha_i$ for i =1, 2……n. that maximize the objective function.

$$L(w,b,\alpha)=\frac{1}{2}w^Tw-\sum_{i=1}^{n}\alpha_i[y_i(w^Tx_i+b)-1] \quad (8)$$

Here $\alpha_i \geq 0$ are the Lagrange multipliers.

The solution of this quadratic problem is given by maximizing L with respect to $\alpha \geq 0$ and minimizing L with respect to w, b.

Differentiating equation (8) with respect to w and b and setting the derivatives to 0 we get .

$$\frac{\delta L(w,b,\alpha)}{\delta w}=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0$$

$$w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

So (9)

$$\frac{\delta L(w,b,\alpha)}{\delta b}=0-\sum_{i=1}^{n}\alpha_i y_i=0$$

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

So (10)

By putting the value of equation (9) and (10) in equation (8) the QP becomes the maximization of the equation.

$$L(\alpha)=\frac{1}{2}(\sum\alpha_i y_i x_i)^T.(\alpha_j y_j x_i)-\sum_{i=1}^{n}\alpha_i y_i(w^Tx_i+b)+\sum_{i=1}^{n}\alpha_i$$

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

here

$$L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i^T.x_j)$$

So (11)

So maximize equation (11)

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

Subject to and $\alpha_i \geq 0$

Find the optimal weight vector $w_0$

$$w_0=\sum_{i=0}^{n}\alpha_i y_i x_i \quad (12)$$

It is also called the support vector.

The optimal bias is given by

$$b_0=y_i-w_0^Tx \quad (13)$$

After getting support vector and bias the decision function that separates two classes is given by

$$f(x)=sign[\sum_{SV}^{n} y_i\alpha_i x_i^T x_i+b_0] \quad (14)$$

**Non Linear Classification**

In non linear classification the original training data x in the input space X is projected in to a high dimensional feature space F via a Mercer Kernel Function K. The optimal separating hyperplane is constructed in the feature space.

Here the mapping of input space X to feature space F is given by $X \to F$

$$x \to \phi(x) \quad (15)$$

So the classifier is translated into the form

$$f(x)=sign(\phi(x)^Tw_0+b_0) \quad (16)$$

$$w_0=\sum_{i=1}^{n}\alpha_i y_i\phi(x_i)$$

Since

$$f(x)=sign(\sum_{i=1}^{n}y_i\alpha_i\phi(x)^T\phi(x_i)+b_0)$$

So (17)

$$f(x)=sign(\sum_{i=1}^{n}y_i\alpha_i K(x_i,x)+b_0)$$

Here (18)

Here K is a Symmetric positive definite function which satisfy the Mercers condition.

Kernel Function $K(x,z)=\phi(x)^T\phi(z)$ (19)

$$K(x,y)=\sum_{i=1}^{\infty}\alpha_i\,\phi(x)^T\phi(y) \quad , \quad \alpha_i \geq 0 \quad (20)$$

The kernel represent the inner product in input space

$$K(x,y)=\phi(x)^T\phi(y)$$

So the F space in dual Lagrangian is given by

$$L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i.x_j)-\lambda\sum_{i=1}^{n}\alpha_i y_i$$

$$\sum_{i=1}^{n}\alpha_i y_i=0$$

Subject to and $\alpha_i \geq 0$ , i= 1, 2, .....n.

So the decision function is given as

$$f(x)=sign[\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b_0] \quad (21)$$

So here the optimal bias $b_o$ is given by

$$b_o=y_i-w_0^T\phi(x_i)$$

Hence for any support vector $x_i$

$$b_o=y_i-\sum_{j=1}^{n}y_i\alpha_i(x_j,x_i) \quad (22)$$

There are number of Kernel function used in support vector machine including

Polynomial $K(x,y)=(1+x.y)^d$

Gaussian RBF $K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$

Exponential RBF $K(x, y) = \exp(-\|x - y\| / 2\sigma^2)$

**Data Set Preparation**

The protein raw data set used in this paper is obtained from PDB. In the data set 3831 protein or enzymes taken from PDB are classified according to EC Number and Enzyme name. Eight features, primary protein structures (Sequence), secondary protein structures (PSS), molecular weight (MW), structural molecular weight (SMW), chain length, atom count, and ligand molecular weight and residue count are extracted from PDB. Table I shows the description of the data set and table II shows the proteins according to class and the total set of each class taken for training, and validation. Data preparation and all manipulations have been done using Microsoft Excel.

**Table 1.** Data set description

| S. No | Fields | Description |
|---|---|---|
| 1 | Sequence | The linear amino acid sequence of a protein |
| 2 | Sequence and Secondary Structure | The linear amino acid sequence of a protein and chain of α-helices, β-sheets, turns and loops |
| 3 | Molecular Weight | Protein Molecular Weight |
| 4 | Ligand Molecular Weight | Ligand (Ion or Functional Molecule) Molecular Weight |
| 5 | Structure Molecular Weight | Structure Molecular Weight |
| 6 | Residue Count | Residue count of protein |
| 7 | Chain Length | Polymer Chain Length |
| 8 | Atom Count | Atom count of protein |

**Table 2.** Database description of six enzymes

| EC No. | Class (Enzyme) | Function | Total Set |
|---|---|---|---|
| 1 | Oxidoreductases | Catalyze the reduction-oxidation reactions. | **327** |
| 2 | Transferases | Transfer a functional grouping and a donor group to a receptor. | **1765** |
| 3 | Hydrolases | Catalyze hydrolysis, the breaking of links and structures by the action of water. | **773** |
| 4 | Lyases | Enzymes which catalyze the cleavage of C-C, C-O and C-N links. | **430** |
| 5 | Isomerases | Catalyze the isomerization reactions of simple molecules. | **282** |
| 6 | Ligases | Formation of links by condensation of substances. | **254** |
| | | **Total** | **3831** |

**Performance Evaluation**

The performance of support vector machine is measured by the quantity of True positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Where TP (True Positive) is the number of positive instances that are classified as positive, FP (False Positive) is the number of Negative instances that are classified as positive, TN(True Negative) is the number of Negative instances that are classified as Negative and FN(False Negative) is the number of positive instances that are classified as Negative. By using these quantities standard Accuracy, Sensitivity, Specificity and Precision performance measure is defined by

**Accuracy**= (TP+TN)/ (P+N) -- The proportion of instances that are correctly classified.

**Sensitivity**=TP/P--The proportion of positive instances that are correctly classified as positive.

**Specificity**=TN/N--The proportion of negative instances that are correctly classified as negative.

**Precision**=TP/ (TP+FP)--The proportion of instances classified as positive that are really positive.

It is described as the following table

| | **Predicted Class** | | |
|---|---|---|---|
| | **Positive** | **Negative** | **Total** |
| **Positive** | **TP** | **FN** | **P** |
| **negative** | **FP** | **TN** | **N** |

## RESULTS AND DISCUSSION

In this paper we use 5 fold cross validation method to measure the performance of the support vector machine classifier.

**(A) Performance Evaluation Oxidoreductases (EC-1)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| **K=1** | 37 | 7 | 44 | 0 | 12 | 12 |
| **K=2** | 32 | 21 | 53 | 0 | 22 | 22 |
| **K=3** | 22 | 11 | 33 | 0 | 19 | 19 |
| **K=4** | 40 | 10 | 50 | 0 | 16 | 16 |
| **K=5** | 24 | 11 | 35 | 0 | 16 | 16 |
| **Total** | 155 | 60 | 215 | 0 | 85 | 85 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 155 | 60 | 215 |
| Negative | 0 | 85 | 85 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 80 | 72.09 | 100 | 100 |

**(B) Performance Evaluation Transferases (EC-2)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| **K=1** | 273 | 6 | 279 | 13 | 14 | 27 |
| **K=2** | 258 | 4 | 262 | 2 | 18 | 20 |
| **K=3** | 210 | 2 | 212 | 5 | 19 | 24 |
| **K=4** | 213 | 4 | 217 | 8 | 6 | 14 |
| **K=5** | 235 | 5 | 240 | 8 | 19 | 27 |
| **Total** | 1189 | 21 | 1210 | 36 | 76 | 112 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 1189 | 21 | 1210 |
| negative | 36 | 76 | 112 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 95.68 | 98.26 | 67.85 | 97.06 |

**(C) Performance Evaluation Hydrolases (EC-3)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| K=1 | 74 | 19 | 93 | 0 | 133 | 133 |
| K=2 | 95 | 24 | 119 | 0 | 97 | 97 |
| K=3 | 85 | 32 | 117 | 1 | 113 | 114 |
| K=4 | 78 | 28 | 106 | 0 | 117 | 117 |
| K=5 | 78 | 29 | 107 | 1 | 93 | 94 |
| Total | 410 | 132 | 542 | 2 | 553 | 555 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 410 | 132 | 542 |
| negative | 2 | 553 | 555 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 87.78 | 75.64 | 99.63 | 99.51 |

**(D) Performance Evaluation Lyases (EC-4)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| K=1 | 44 | 20 | 64 | 0 | 44 | 44 |
| K=2 | 37 | 14 | 51 | 0 | 58 | 58 |
| K=3 | 39 | 18 | 57 | 1 | 42 | 43 |
| K=4 | 42 | 20 | 62 | 1 | 44 | 45 |
| K=5 | 42 | 16 | 58 | 0 | 40 | 40 |
| Total | 204 | 88 | 292 | 2 | 228 | 230 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 204 | 88 | 292 |
| negative | 2 | 228 | 230 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 82.75 | 69.86 | 99.13 | 99.02 |

**(E) Performance Evaluation Isomerases (EC-5)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| K=1 | 14 | 38 | 52 | 0 | 23 | 23 |
| K=2 | 10 | 24 | 34 | 0 | 20 | 20 |
| K=3 | 17 | 25 | 42 | 0 | 35 | 35 |
| K=4 | 6 | 25 | 31 | 0 | 29 | 29 |
| K=5 | 13 | 28 | 41 | 0 | 40 | 40 |
| Total | 60 | 140 | 200 | 0 | 147 | 147 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 60 | 140 | 200 |
| negative | 0 | 147 | 147 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 59.65 | 30 | 100 | 100 |

**(F) Performance Evaluation Ligases (EC-6)**

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| K=1 | 10 | 33 | 43 | 0 | 12 | 12 |
| K=2 | 10 | 21 | 31 | 0 | 18 | 18 |
| K=3 | 12 | 25 | 37 | 0 | 18 | 18 |
| K=4 | 5 | 32 | 37 | 0 | 14 | 14 |
| K=5 | 3 | 18 | 21 | 0 | 20 | 20 |
| Total | 40 | 129 | 169 | 0 | 82 | 82 |

**Average Performance Evaluation**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 40 | 129 | 169 |
| negative | 0 | 82 | 82 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 44.62 | 23.66 | 100 | 100 |

Here it is observed that of total six classes four classes presented the better result, one give the considerable result but one class Ligases give the low value of accuracy as well as sensitivity.

**(G) Table 3.** Result obtained for the six classes of protein analyzed

| Class | Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|---|
| Oxidoreductases (EC-1) | 80 | 72.09 | 100 | 100 |
| Transferases (EC-2) | 95.68 | 98.26 | 67.85 | 97.06 |
| Hydrolases (EC-3) | 87.78 | 75.64 | 99.63 | 99.51 |
| Lyases (EC-4) | 82.75 | 69.86 | 99.13 | 99.02 |
| Isomerases (EC-5) | 59.65 | 30 | 100 | 100 |
| Ligases (EC-6) | 44.62 | 23.66 | 100 | 100 |

**(H)** The Overall Performance Evaluation of Six Classes

| K Fold cross validation | TP | FN | Total | FP | TN | Total |
|---|---|---|---|---|---|---|
| K=1 | 452 | 123 | 575 | 13 | 239 | 252 |
| K=2 | 442 | 108 | 550 | 2 | 234 | 236 |
| K=3 | 385 | 113 | 498 | 7 | 246 | 253 |
| K=4 | 384 | 119 | 503 | 9 | 226 | 235 |
| K=5 | 395 | 107 | 502 | 9 | 218 | 227 |
| Total | 2058 | 570 | 2628 | 40 | 1163 | 1203 |

**Average Performance Evaluation of Six classes**

| | Positive | Negative | Total |
|---|---|---|---|
| Positive | 2058 | 570 | 2628 |
| negative | 40 | 1163 | 1203 |

| Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|
| 84.07 | 78.31 | 96.67 | 98.09 |

Here we observed that the overall accuracy of support vector machine classifier by using expert model with RBF kernel where width is 0.10 and parameter C is 10 is 84.07%.

## CONCLUSION

In this paper we proposed support vector machine based method for the classification of Enzyme which has great potential for linear and non linear classification. The result shoes that it is capable for classification of different enzyme functions. The expert model of SVM is useful tool for protein function prediction. Here we found the overall accuracy of

expert model of SVM with RBF kernel where width is 0.10 and parameter C is 10 is 84.07% which is better than the previous proposed approaches. Here we used only eight features for the prediction of protein function so in future the performance of the classifier may be increased by using more features and different features selection algorithms for selecting more relevant features for protein function prediction.

## REFERENCES


1. Burbidge, R., Buxton, B.: An introduction to support vector machines for data mining. In M. Sheppee (Ed.), Keynote Papers, Young OR12, University of Nottingham, páginas 3–15,.Operational Research Society: Operational Research Society, March (2001).
2. C.Z. Cai, W.L. Wang, L.Z. Sun, and Y.Z. Chen, "Protein function classification via support vector machine approach," Mathematical Biosciences, vol. 185, pp. 111-122, 2003.
3. Cortes, C. Vapnik, V. Support-vector networks. Machine Learning, 3(20):273–297 (1995).
4. L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, and Y.Z. Chen, "Predicting functional family of novel enzymes irrespective of sequence similarity," Nucleic Acids Research, vol. 32, pp. 6437- 6444, 2004.
5. Lu, L., Qian, Z., Cai, Y. D., Li, Y. (2007). ECS: An automatic enzyme classifier based on functional domain composition,Comput *Biol Chem*, 31 (3), 226-232.
6. Luiz C. Borro, Stanley R.M. Oliveira, Michel E.B. Yamagishi, AdaultoL. Mancini, Jose G. Jardine, Ivan Mazoni, Edgard H. dos Santos,Roberto H. Higa, Paula R. Kuser, and GoranNeshich, "Predictingenzyme class from protein structure using Bayesian classification,"Genetics and Molecular Research, vol. 5, pp. 193- 202, 2006.
7. Paul D. Dobson and Andrew J. Doig, "Predicting Enzyme Class fromProtein Structure without Alignments," JMB, vol. 345, pp. 187-199,2005.
8. Qiu, J. D., Huang, J. H., Shi, S. P., Liang, R. P. (2010). Using the Concept of Chou's Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine Based on Discrete Wavelet Transform. *Protein Pept Lett.*, 17(6), 715-22.
9. Wang, Y. C., Wang, X. B., Yang, Z. X., Deng, N. Y. (2010). Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept. Lett.*, 17, pp. 1441–1449.
10. Yong-Cui Wang, Yong Wang, Zhi-Xia Yang, Nai-Yang Deng. Support vector machine prediction of enzyme function with conjoint triadfeature and hierarchical context . BMC Systems Biology, 5(Suppl 1):S6(2011).




*****