ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

# Prediction of Post-Synaptic Activity in Proteins Using Recursive Feature Elimination

Bernardo Penna Resende de Carvalho[1], Ricardo de Souza Ribeiro[1] and Talles Henrique de Medeiros[1]

1- Ottimah Process Improvements - Dep. of Research and Development
Padre Marinho St. 37, 13th floor, Santa Efigênia, BH (MG), 30.140-040 - Brazil
E-mail: {bernardo.penna,ricardo.souza}@ottimah.com

2- Federal University of Minas Gerais - Computational Intelligence Lab.
Antônio Carlos Av. 6627, Pampulha, Belo Horizonte (MG), 31.270-901 - Brazil
E-mail: talles@cpdee.ufmg.br

**Abstract.**
This work presents a new approach to predict post-synaptic activities in proteins. It uses a feature selection technique, called Recursive Feature Elimination, in order to select only the relevant features from the complete database. Once the reduced subset is found, Least Squares Support Vector Machine, a SVM based classifier, is used to predict its classes. The experiments were performed on a database that was harvested from Swiss Prot/Uniprot, a public domain database with a rich source of information for a very large number of proteins. The obtained results show that the proposed approach led to a reduced representation to the database, using only 6% of the original information, and yielded an improvement into the classification when compared to another two prediction techniques applied to the complete database, Decision Tree and Least Squares Support Vector Machine.

## 1 Introduction

Bioinformatics is an emerging field that deals with the greatest challenge of modern biology: understanding vital processes in molecular level. Many Computational Intelligence techniques are used in order to help the development of new drugs and treatments, as well as reducing the market time entrance of a new medicine. This can be possible because *in silico* tests are much faster to be carried through than *in vitro* ones.

Support Vector Machine (SVM) [1] became a very popular machine learning in the last decade. Its success is mainly due to a solid formal basis and the elegant approach for margin maximization based on support vectors. Least Squares Support Vector Machine (LS-SVM) [2] is a modified version of SVM, which can be easily implemented and used for prediction tasks. The optimization problem generated by LS-SVM can be solved with a system of linear equations, which is less complex than the quadratic programming used in SVM.

Recursive Feature Elimination (RFE) [3] is a feature selection technique that uses SVM to detect the most relevant features of a given database. The great advantage of RFE is the minimization of the original search space of the problem

in a significantly simpler space. This fact is very usefull for bioinformatics, because it implies in the economy of resources for *in vitro* experiments.

This work presents a new approach to predict post-synaptic activities in proteins. Many proteins with post-synaptic activity have been functionally characterized by biochemical, immunological and proteomic exercises, and are now extensively catalogued and annotated [4]. The proposed approach uses RFE technique in order to reduce a protein database of post-synaptic activities and, in a second step, it applies LS-SVM to classify the reduced subset. This database was first predicted with a Decision Tree technique [5], whose results are presented in the experiments.

More interesting than only determining if a given protein possess post-synaptic activity or not, it is important to discover which are the most relevant features related to this task, which is done by RFE. The usage of LS-SVM classifier without a first step of RFE was also compared, in order to verify the efficiency of our proposition. The results show that the proposed approach led to a reduced representation of the database, using only 6% of the original information, and yielded an improvement into the classification when compared with other prediction techniques.

The remainder of this paper is organized as follows. In section 2, it will be presented the bioinformatics context of this work. In Section 3.1, there is a description of the Computational Intelligence techniques used in the experiments, LS-SVM and RFE. In Section 4, it's presented the results and discussion. And finally, in the last section it's presented the conclusions of this work.

## 2 Bioinformatics

Bioinformatics is a scientific area based on computing and molecular biology, which aims to analyze large amounts of biologic data, predicting the genes functions or demonstrating the relations between genes and proteins.

### 2.1 Proteins

The proteins play an essential role in some biological processes, as sustentation mechanics, enzymatic catalysis, transport and storage, immunities protection, generation and transmission of nervous impulses, control of metabolism and growth, among others [4]. The protein synthesis takes place in ribosomes, in a two-stage process: synthesis of a mRNA molecule from the DNA information (transcription) and translation, made by tRNAs, from mRNA nucleotide sequences to amino acid sequences. Since mRNAs are originated from DNA information, the proteins are considered as the expression of DNA.

### 2.2 Post-synaptic activities

Synapses are spaces filled with a fluid, between the neurons, which separate the emitting and receiving cells. The nervous signal is transmitted from the pre-synaptic to the post-synaptic neuron, through the synapse. Some types

of proteins can be found in the post-synaptic neurons, being important to the transmition of information. Many proteins with post-synaptic activity already had its functions characterized in world-wide databases [5].

The database used in the experiments was harvested from Swiss Prot/Uniprot, a public domain database with a rich source of information for a very large number of proteins [5]. It contains 4303 proteins, being 260 with post-synaptic activity and the remain without this activity. For each protein, there is 443 attributes indicating existence (goal attribute equal 1) or absence (null attribute goal) of determined amino acid standard, called motif. Motifs represent specific standards that can exist in some proteins, each one indicating amino acid sequences or gaps between them, that can be used to indicate functional protein behaviors.

## 3 Computational Intelligence applied to Bioinformatics

The computational analysis are used to discover unknown details and arrangements in the organization of genomics and proteomics data, helping to clarify the structure and the function of the genes and proteins. Many computational intelligence techniques can be used in bioinformatics [6]. In this section it will be presented LS-SVM [7] and RFE [3], both used in our experiments.

### 3.1 Least Squares Support Vector Machines

Given the training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, with input $\mathbf{x}_i \in \Re^n$ and corresponding binary output $y_i \in \{-1, +1\}$, LS-SVM maps the input data in a high dimensional space, called feature space, to build a linear separation hyperplane $f(\mathbf{x}) = 0$, such that

$$\omega^T \varphi(\mathbf{x}) + b = 0 \tag{1}$$

where $\omega$ is the weight vector, $b$ is the bias term and $\varphi(.)$ is the mapping function applied to data to represent them at the feature space.

The primal problem of the LS-SVM is defined as

$$min_{\omega,b,\mathbf{e}} \quad J_P(\omega, b, \mathbf{e}) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{2}$$

subject to

$$y_i[\omega^T\varphi(\mathbf{x}_i) + b] = 1 - e_i, \quad i = 1..., N$$

where $\gamma$ controls the two terms of $J_P(\omega, b, \mathbf{e})$ and $e_i$ is the error of pattern $\mathbf{x}_i$.

We apply the Lagrangean, in order to incorporate the equality constraints of the primal problem in the dual cost function, using the Lagrange multipliers $\alpha$. Deriving the Lagrangean problem with respect to the primal and dual variables and setting the result to zero, which is needed to find its saddle point, gives

$$\begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i y_j \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) + \frac{1}{\gamma}) + yb = 1. \end{cases} \tag{3}$$

We can describe (3) like a linear system $\mathbf{AX} = \mathbf{B}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & -\mathbf{Y}^T \\ \mathbf{Y} & \mathbf{H} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} b \\ \alpha \end{bmatrix} \; and \; \mathbf{B} = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}. \tag{4}$$

The matrix $\mathbf{H}$ in (4) obeys the Mercer Theorem, that deals with the conditions that a given function $K(\mathbf{x}_i, \mathbf{x}_j)$ must have to be a kernel function. Therefore $\mathbf{H}$ is positive definite, the matrix $\mathbf{A}$ is not. The solution of the system of linear equations (4) is the same of the primal problem (2).

Once we have got the values of the Lagrange multipliers, the output of the LS-SVM can be calculated, like the SVM, as

$$f(\mathbf{x}) = sign[\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b]. \tag{5}$$

### 3.2 Recursive Feature Elimination

Given the training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, with input $\mathbf{x}_i \in \Re^n$ and corresponding binary output $y_i \in \{-1, +1\}$, the weight vector gotten after SVM training is [1]

$$\omega = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \tag{6}$$

when using the linear kernel.

According to the output $f(\mathbf{x})$ in (5), we can consider that the relevance of feature $j$ is $\omega_\mathbf{j}$. In order to compare the relevances of negative and positive classes, ${\omega_\mathbf{j}}^2$ is used as a relevance criterion to the features.

The RFE technique is described as following [3]:

1. The training process of SVM is evaluated using the linear kernel function.
2. The Lagrange multipliers $\alpha_i$ associated to the input points $\mathbf{x}_i$ are obtained.
3. The Lagrange multipliers are used to compute the weight vectors $\omega$, as described in (6).
4. The relevance criterion ${\omega_\mathbf{j}}^2$ for each feature $j$ is calculated.
5. The feature with smaller ${\omega_\mathbf{j}}^2$ values are eliminated from the dataset.
6. If the number of eliminated features is smaller than a given parameter, go to step 1. Otherwise, the recursive feature elimination is finished.

## 4 Results and discussion

In this section we present the results of the experiments with the database Swiss Prot/Uniprot, in Tables 1 and 2. These results are compared with Decision Tree [5] through the generation of decision rules, using the same fraction of training

Table 1: Result of Swiss Prot/Uniprot database with different approaches

| Method | Features used | True positive ratio (TPR) | True negative ratio (TNR) |
|---|---|---|---|
| $DecisionTree$ | 445 | 0.85 | 0.98 |
| $LS-SVM$ | 445 | 0.87 | 0.95 |
| $RFE+LS-SVM$ | 27 | 0.89 | 0.97 |

and testing sets used on [8]. One of the main contributions of this work is the minimization of the complex space of original problem search ($2^{445}$) for a significantly simpler space ($2^{27}$).

In Table 1 we can note $RFE+LS-SVM$ approach selected only 27 features as relevant ones, while other approaches used the complete database. True positive ratio (TPR), the ratio between predicted positive examples and all positive ones (with post-synaptic activity), is presented for all prediction strategies. $RFE+LS-SVM$ got the best TPR result, followed by $LS-SVM$. True negative ratio (TNR), the ratio using negative examples, is also presented in this table. It's observed that $RFE+LS-SVM$ has a result 1% smaller than $DecisionTree$ in this aspect.

Although TPR is more important than TNR, since the discovery of the proteins that posses post-synaptic activity is the main task of the dataset, the best way to compare different approaches is the product of both ratios. This product is the metric used to evaluate the methods in this work, because for unbalanced data the accuracy is not appropriate to reflect the sensitivity of a classifier [5]. In Table 2 we observe that $RFE+LS-SVM$ got the higher $TPRxTNR$, 0.86, while $LS-SVM$ got 0.83 and $DecisionTree$ 0.84. The accuracy of $DecisionTree$ was the highest due to its $TNR$ of 0.98, followed by an accuracy of 96.35 of $RFE+LS-SVM$.

The 27 selected motifs found by the proposed approach, considered the most relevance ones for this task, are PS01113, PS00018, PS00022, PS00479, PS00484, PS00518, PS00687, PS00904, PS00107, PS00108, PS00589, PS00856, PS00120, PS00941, PS00867, PS00713, PS00714, PS00405, PS00410, PS00319, PS00320, PS00236, PS00237, PS00979, PS00980, PS00981, PS00678. The results show that using only these motifs is not only enough, but also better than using the complete dataset, to predict the post-synaptic acivity in the proteins in Swiss Prot/Uniprot database.

Table 2: Accuracy and Product of Ratios on Swiss Prot/Uniprot database

| Method | Accuracy (%) | TPR x TNR |
|---|---|---|
| $DecisionTree$ | 97.85 | 0.84 |
| $LS-SVM$ | 95.30 | 0.83 |
| $RFE+LS-SVM$ | 96.35 | 0.86 |

The same algorithms of rules generation used in [5] can be applied into the reduced dataset of 27 selected motifs and analyzed by specialists, in order to validate new rules. These new rules will be simpler than $DecisionTree$ ones, because the search space was reduced by RFE, as pointed in this work.

## 5 Conclusion

This work presents a new approach to predict post-synaptic activities in proteins. The proposed approach uses RFE technique in order to reduce a protein database of post-synaptic activities and, in a second step, it applies LS-SVM to classify the reduced subset. More interesting than only determining if a given protein possess post-synaptic activity or not, it is important to discover which are the most relevant features related to this task, which is done by RFE.

The obtained results show that the proposed approach led to a reduced representation for the database, using only 6% of the original information, and yielded an improvement into the classification when compared with another two prediction techniques applied to the complete database, Decision Tree and Least Squares Support Vector Machine. The $RFE + LS - SVM$ strategy guaranteed an improvement in the ratio of the minority class, represented by TPR, while the general precision of the classifier, measured for the product $TPRxTNR$, also got sensible improvement, when comparing with other classifiers.

Future works can evaluate traditional feature selection techniques, as PCA or ICA, in order to compare with the RFE utilization proposed in this work. Another interesting idea is applying the same algorithms of rules generation used in [5] and analyze the results, in order to validate new rules. These new rules will be simpler than $DecisionTree$ ones, because the search space was reduced by RFE.

## References

[1] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[2] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[3] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3), 2002.

[4] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 2004.

[5] G. L. Pappa, A. J. Baines, and A. A. Freitas. Predicting post-synaptic activity in proteins with data mining. *Bioinformatics*, 21(Suppl. 2):ii19–ii25, September 2005.

[6] M. C. P. Souto, Ana Carolina Lorena, Alexandre Claúdio Botazzo Delbem, and André C. P. L. F. de Carvalho. Técnicas de aprendizado de máquinas para problemas em biologia molecular. In *II Jornadas de Atualização em Inteligência Artificial*, Unicamp, SP, 2003.

[7] B. P. R. Carvalho. New strategies for support vectors' automatic detection in least squares support vector machines. *Master thesis*, 2005. URL: www.cpdee.ufmg.br/~bpenna/bpenna_dissertacao_2005.pdf.

[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.