Computer Abetted Drug Design: Application of drug-target like protein prediction using learning algorithms of support vector machine

Mickey Sahu[1] and Ashish Shrivastava[2]
[1]*Department of Computer Science, Aisect University, Bhopal, 464993 (M.P)*
[2]*Department of Biotechnology, C.S.A. Govt. P. G. Nodal College, Sehore, 466001 (M.P.)*

Abstract: *In this paper, we explore the use of statistical learning approaches to predict drug-target like proteins from their primary sequences in order to facilitate the rapid discovery of new potential therapeutic targets from the large quantity of sequences in human genome. It was found that the Support Vector Machine (SVM) algorithm with a fine-tuned Gaussian kernel was able to make reasonably accurate prediction, which showed its potential to be used in the genome scale rapid drug target discovery, as a novel in silico approach supplementary to the conventional experimental approaches.*

Keyword: *SVM, Therapeutic, PCA, ICA*

I. Introduction :

Target discovery constitutes one of the main components of today's early stage pharmaceutical research [1,2]. The aim of target discovery is to identify and validate suitable drug targets (i.e. proteins or nucleotides to which drug binding produces therapeutic effects) for therapeutic interventions. Only a small fraction of proteins are actually targeted by today's drugs. The discovery of targets that are sufficiently robust to yield marketable therapeutics is an enormous challenge. Through the years, many approaches have been used with varying degrees of success. These are mainly wet-lab based approaches which require the consumption of large amount of money and time. Statistical learning approaches have been applied to find the relationship between protein sequences and their functions [3-7], which lead to the hypothesis that the statistical learning methods may be equally applicable in prediction of drug-target like proteins, which is an efficient approach to pick out candidate targets from the huge number of proteins in the human genome. The establishment of therapeutic target database has provided a useful resource for statistical model training.

II. Prediction of drug-target like proteins :

In order to evaluate different classification and pre-processing techniques, an efficient tool to implement different algorithms is needed. The matrix operation support provided by MatLab [8] makes the representation of numerical data and implementation of the different algorithms much easier. Therefore, we choose MatLab as our platform of computation. With the help of standard MatLab matrix functions and standard toolboxes, i.e. statistics toolbox and optimization toolbox, we implemented the algorithms for scaling, PCA, decision tree, k-nearest neighbor, and support vector machine. An ICA package for MatLab, FastICA 2.1, was used for ICA analysis, which is developed by Jarmo Hurri et.al. [9]. The support vector machine algorithm was implemented with a Gaussian kernel, $K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ . .This is because the Gaussian kernel always performs better than others in our previous study of protein function classification [10, 11].

III. Support vector machine prediction :

On original data sets, SVMs are trained with the kernel parameter σ scanned in the range of [1...75] with an interval of 1, which is the range that empirically gives optimal classification results in protein function classification [10]. The measurements concerned, *A* , *F* and *G* , are plotted against σ in Figure 3.1. The best *A* , best *F* , and best *G* , as summarized in Table 3.1, are 87.28%, 56.72%, and 72.47% respectively. On scaled data sets, the kernel parameter σ is scanned in the range of [0.04..3] with an interval of 0.04. The *A* , *F* , and *G* obtained with different σ are plotted in Figure 3.1. The best *A* , best *F* , and best *G* are found in a single SVM model with σ = 1.28, which are 89.91%, 68.49%, and 75.63% respectively. These results are better than those of the original data sets.

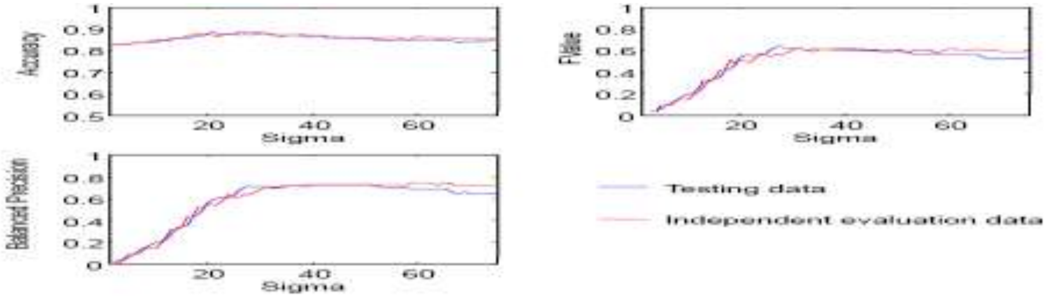

Figure 3.1: Support vector machine prediction of drug-target like proteins on original data sets.

Table 3.1: Summary of the SVM performance on original data sets.

Measure Optimized	Kernel Parameter σ	Overall Accuracy	F Value	Balanced Precision
Overall Accuracy	27	0.8728	0.5672	0.6350
F Value	27	0.8728	0.5672	0.6350
Balanced Precision	50	0.8596	0.6000	0.7247

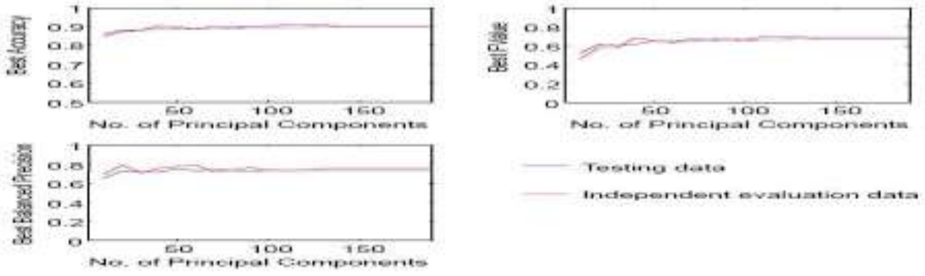

Figure 3.2: Support vector machine prediction of drug-target like proteins after PCA dimensionality reduction.

Table 3.2: Summary of the SVM performance on PCA pre-processed data sets.

Measure Optimized	No. of Principal Components	Overall Accuracy	F Value	Balanced Precision
Best Overall Accuracy	130	0.8991	0.6849	0.7563
Best F Value	130	0.8991	0.6849	0.7563
Best Balanced Precision	50	0.8991	0.6512	0.7730

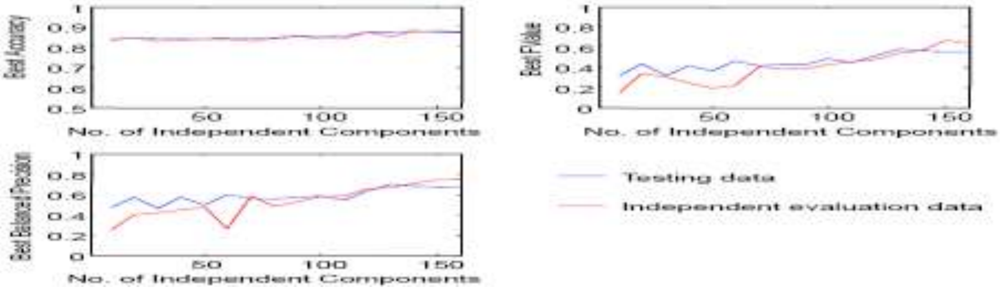

Figure 3.3: Support vector machine prediction of drug-target like proteins after ICA dimensionality reduction.

Table 3.3: Summary of the SVM performance on ICA pre-processed data sets.

Measure Optimized	No. of Independent Components	Overall Accuracy	F Value	Balanced Precision
Best Overall Accuracy	140	0.8772	0.5634	0.7046
Best F Value	130	0.8509	0.5385	0.6657
Best Balanced Precision	130	0.8509	0.5385	0.6657

In summary, SVM performance can be improved by the pre-processing of scaling and PCA. Among all the three classification techniques explored, SVM classifies our data best. The best *A* , best *F* , and best *G* , if optimized individually, can reach 89.91%, 68.49% and 77.30% respectively.

IV. Prediction results and analysis :

Table 4.1 summarizes the comparison between different statistical learning methods evaluated in this work. Overall, SVM gives the best results with the best *A* , best *F* , and best *G* reaching 89.91%, 68.49% and 77.30% in different SVM models. The accuracy of SVM prediction, if successfully generalized in real-world application, is reasonably good to provide valuable information for genome scale target discovery.

Table 4.1: Performance comparison between different statistical methods

Measurement Optimized	Decision Tree	K-nearest Neighbor	Support Vector Machine
Best Overall Accuracy	85.09%	83.77%	89.91%
Best F Value	54.05%	56.84%	68.49%
Best Balanced Precision	68.40%	75.30%	77.30%

Errors in statistical learning arise for a number of reasons. It is not expected that exhaustive experiments have been done to verify whether each known protein is a target or not. Also, the therapeutic targets collected in TTD are not complete. This may result in that, with a small possibility, some drug targets are included in the negative examples. Although, most of the statistical learning methods are able to deal with a certain level of noise, these approaches are generally based on a large number of observations (examples).

V. Conclusion :

A number of statistical learning methods and pre-processing techniques are investigated for the application of drug-target like protein prediction, which includes the learning algorithms of decision tree, k-nearest neighbor and support vector machine and the pre-processing techniques of scaling, PCA and ICA dimensionality reduction. The support vector machine approach gives the best classification results. Performance and applicability of the statistical learning methods may be further improved by incorporation of new information. Efficiency and accuracy of statistical learning methods in prediction of drug-target like proteins can also be enhanced from new progress in learning algorithms, descriptors, and pre-processing techniques.

References:

[1] Dannhardt, G. and Laufer, S., Structural approaches to explain the selectivity of COX-2 inhibitors: is there a common pharmacophore?, C*urr Med Chem*, 7 (2000) 1101-12.

[2] Kurogi, Y. and Guner, O.F., Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, C*urr Med Chem,* 8 (2001) 1035-55.

[3] Klein, P., Kanehisa, M. and DeLisi, C., Prediction of protein function from sequence properties. Discriminant analysis of a data base, Biochi*m Biophys Acta, 787* (1984) 221-6.

[4] Nakai, K., Kidera, A. and Kanehisa, M., Cluster analysis of amino acid indices for prediction of protein structure and function, Protei*n Eng, 2 (1*988) 93-100.

[5] Fetrow, J.S. and Skolnick, J., Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases, J Mol *Biol, 281* (1998) 949-68.

[6] Edwards, Y.J. and Cottage, A., Prediction of protein structure and function by using bioinformatics, Method*s Mol Biol, 175* (2001) 341-75.

[7] Baxter, S.M. and Fetrow, J.S., Sequence- and structure-based protein function prediction from genomic information, Curr O*pin Drug Discov Devel, 4 (2*001) 291-5.

[8] Hanselman, D.C. and Littlefield, B., Master*ing MATLAB 6 : a comprehensive tutorial and reference, Pren*tice Hall, Upper Saddle River, N.J., 2001, xviii, 814 p. pp.

[9] Hyvarinen, A., Fast and robust fixed-point algorithms for independent component analysis, Ieee T*ransactions on Neural Networks, 10 (*1999) 626-634.

[10] Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. and Chen, Y.Z., SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nuclei*c Acids Res, 31 (*2003) 3692-7.

[11] Cai, C.Z., Wang, W.L. and Chen, Y.Z., Support Vector Machine Calssification of Physical and Biological Datasets., Inter. *J. Mod. Phys. C, Acce*pted (2003).

[12] Li, A.P., Screening for human ADME/Tox drug properties in drug discovery, Drug D*iscov Today, 6 (*2001) 357-366.