BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 6(3)

PKSIIIexplorer: TSVM approach for predicting Type III polyketide synthase proteins

Mallika Vijayan[1], Sivakumar Krishnankutty Chandrika[2], Soniya Eppurathu Vasudevan[1]*

[1]Plant Molecular Biology, Rajiv Gandhi Centre for Biotechnology, Thycaud P O, Poojappura, Thiruvananthapuram - 695 014, Kerala, India; [2]Bioinformatics Facility, Rajiv Gandhi Centre for Biotechnology, Thycaud P O, Poojappura, Thiruvananthapuram - 695 014, Kerala, India; Soniya EppurathuVasudevan - Email: evsoniya@rgcb.res.in; Phone: 91- 471-2529454; Fax: 91-471-2348096; *Corresponding author

Received March 14, 2011; Accepted March 23, 2011; Published April 22, 2011

Abstract:
PKSIIIexplorer, a web server based on 'transductive Support Vector Machine' allows fast and reliable prediction of Type III polyketide synthase proteins. It provides a simple unique platform to identify the probability of a particular sequence, being a type III polyketide synthases or not with moderately high accuracy. We hope that our method could serve as a useful program for the type III polyketide researchers. The tool is available at "http://type3pks.in/tsvm/pks3".

Keywords: Type III polyketide synthase, PKSIIIexplorer, TSVM, Chalcone synthase.

Abbreviations: PKS, Polyketide synthase; CHS, Chalcone synthase; SVM, Support vector machine; MCC, Matthews Correlation Coefficient.

Background:
Type III polyketide synthases (type III PKSs) are large superfamily of proteins that produce wide variety of secondary metabolites which possess antibiotic, antifungal, antitumor and immunosuppressive activities **[1]**. For example, resveratrol, a stilbene synthase derivative from grapes shows cancer chemopreventive activity in murine models **[2]**. To discover more of these novel proteins, Support Vector Machines (SVMs) have been used successfully for the purpose of classification. Earlier we have developed SVM based "PKSIIIpred", in which only labelled data were used for training set **[3]**. But in the improved version, we used an innovative variant of SVM, the so-called 'Transductive SVM' (TSVM) that not only take into account the labeled training data but also integrate unlabeled data.

Methodology:
Dataset:
Positive (type III PKSs) and negative (non-type III PKSs) datasets were prepared (1000 each). Sequences were retrieved in FASTA format from Swiss-Prot. Unlabelled dataset (2000) was generated by profile hidden Markov models (HMMs) using the positive dataset to extract certain proteins from Swiss-Prot. In the case of unlabeled dataset, we are not sure whether they are type III PKS or not. BLASTCLUST was used to verify the non-redundancy of datasets **[4]**.

TSVM- implementation:
SVMs are group of fast optimization machine learning algorithms which have been used for many kinds of pattern recognition **[5]**. The performance of SVM based methods has been optimized by tuning SVM parameters (linear, polynomial, radial or sigmoid). In classical SVMs, the training data that are used to build the model ideally cover the whole problem space; the model is then used to predict the labeling of new data points. But in most of the biological datasets the number of labeled data points is rather small, but a large number of unlabeled data points are available. To take advantage of these unlabeled data, the so called 'TSVMs' have been developed **[6]**. Here, TSVM was implemented using SVMlight package which posess two modules: SVM_learn (preparing models) and SVM_classify (classifying samples). For each cluster of composite specificity, we prepared a feature file with the sequences belonging to this specificity labeled +1, all other sequences with different but known specificity labeled –1, and the uncharacterized sequences labeled 0. TSVM was trained as described above, to obtain a model for composite specificity. During several rounds of evaluation, many parameters produced poorly performing models with poor MCC values. Therefore, selected a set of consistently performing parameters for identifying the optimally performing models. After training the SVM models, it is necessary to combine the predictions of all models to one single prediction. Here the SVM that outputs the largest score is used to assign the specificity to the unknown sequence.

Numerical properties:
The models were trained by using dipeptide and multiplet frequencies **[7]** of amino acid composition. For each protein, a matrix of 400 dipeptides was generated and fed as an input to SVM. The repetitiveness of the amino acid sequences were analyzed by means of multiplet which comprise homopolymeric stretches of any length (XX, XXX, ... (X)n) where X denotes any specific amino acid and n≥2.

Webserver:
The server was prepared in Apache version 2.0 and the scripting was done in PHP version 5.3.2. The background running programs for dipeptide and multiplet frequencies were written in Perl 5.8.5.

Performance assessment:
Fivefold cross-validation technique was used to evaluate the performance of all the models. We computed the Error rate (err) specificity (SP), sensitivity (SN), and MCC **[8-9]** for assessing the performance of a method (given in Supplementary material). Sensitivity gives the fraction of positive events; specificity represents how many false subjects are incorrectly recognized as positives; the 'error rate' is the fraction of type III PKS data that is classified incorrectly **[9]**. MCC ranges from -1 to +1 and the highest value indicates better prediction. We identified the model with highest MCC value in each of the five subsets. In the second subset, three models with different parameters sets 47, 65 and 89 were equally good and therefore both of them were included **(Table 1 see Supplementary material).**

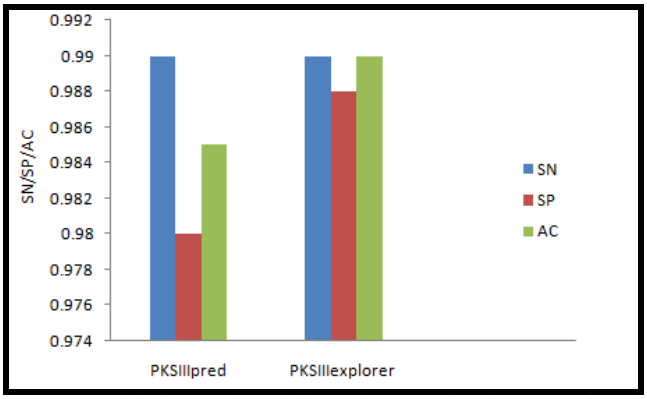

Figure 1: Statistical comparisons indicates that TSVM based PKSIIIexplorer is superior in the case of sensitivity (SN), specificity (SP) and accuracy (AC) than PKSIIIpred.

Discussion:
The web-interface of "PKSIIIexplorer" allows, one to 'upload' or 'paste-in' the sequences in fasta format. Here we describe the application of TSVMs to functionally predict the peptides, based on the chemical fingerprint of the residues. By using various kernal functions, we got the best results for polynomial and radial (RBFs), over linear and sigmoid **(Table 1)** and found that SVM models yield very good results (MCC = 0.84–0.97). In addition to the plant proteins, we also provided type III PKSs from bacteria, fungi and bryophytes in the training dataset, so they can be perfectly predicted during user investigation. It is noted that the server efficiently predicts type I PKS, ketosynthase domain as negative which adopts similar structural fold and shows sequence similarity to type III PKS. These results demonstrated that the sequence features used by PKSIIIexplorer have powerful discriminating power. The system also found to be superior **(Figure 1)** to the previous prediction server "PKSIIIpred" (http://type3pks.in/prediction/).

Conclusion:
Because of the diverse pharmacological functions, the volume of data on type III PKS is rapidly increasing. With this regard developing a highly sensitive method to identify the protein 'in silico' will accelarate the experimental research. Our results give high reliable predictions, even though the training data is relatively low, leaving a room for further improvement with a growing number of type III PKSs. BLAST could be helpful especially for rare specificities and therefore, we plan to integrate it in a future version of PKSIIIexplorer.

Acknowledgements:
Authors are thankful to Department of Information Technology, Government of India for financial support. The authors acknowledge BTISNet, Department of Biotechnology, Government of India for the Bioinformatics facility.

References:
[1] Austin MB & Noel JP. *Nat Prod Rep*. 2003 **20**(1):79 [PMID: 12636085]
[2] Jang M *et al*. *Science* 1997 **275**(5297): 218 [PMID: 8985016]
[3] Mallika V *et al*. *J Integr Bioinform*. 2010 **7**(1): 143 [PMID: 20625199]
[4] Altschul SF *et al*. *J Mol Biol*. 1990 **215**(3): 403 [PMID: 2231712]
[5] Vapnik VN. *IEEE Trans Neural Netw*. 1999 **10**(5): 988 [PMID: 18252602]
[6] http://svmlight.joachims.org/
[7] Brendel V *et al*. *Proc Natl Acad Sci U S A*. 1992 **89**(6): 2002 [PMID: 1549558]
[8] Matthews BW. *Biochim Biophys Acta*. 1975 **405**(2): 442 [PMID: 1180967]
[9] Baldi P *et al*. *Bioinformatics*. 2000 **16**(5): 412 [PMID: 10871264]

Edited by P Kangueane
Citation: Mallika *et al*. Bioinformation 6(3): 125-127 (2011)

BIOINFORMATION

open access

Supplementary material:

Computation of the Error rate (err) specificity (SP), sensitivity (SN), and MCC **[8-9]** for assessing the performance of a method (given in Supplementary material)

$$\text{Errorrate} = \text{err} = (\text{FP} + \text{FN})/(\text{FP} + \text{FN} + \text{TP} + \text{TN})$$

$$\text{Recall} = \text{sensitivity}\ S_n = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Precision} = \text{specificity}\ S_p = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Matthews correlation coefficient MCC} = \sqrt{\frac{(\text{TP}\cdot\text{TN}) - (\text{FN}\cdot\text{FP})}{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FN})(\text{TN}+\text{FP})}}.$$

Where, TP =True Positive; TN = True Negative; FP=False Positive; FN=False Negative.

Table 1: Parameter sets and performance of five selected models to identify type III PKS proteins are displayed in Table. In the test subset, the models with different parameters sets 47c, 65c and 89a were equally good and therefore included in the training set.

Selected best model (classifier)	Kernel type	Parameters	Error (err)	Sensitivity (Sn)	Specificity (Sp)	Mathew's correlation coefficient (MCC). Performance of the best model in the selected test subset
47b	Polynomial	d=2, C=0.001	0.7	0.94	0.94	0.93
47c	RBF	d=0.001, C=100	1.1	0.90	0.95	0.92
65c	RBF	d=0.01, C=10	3.3	0.81	0.90	0.84
89a	Polynomial	d=2, C=0.001	0.4	1.00	0.95	0.97
89c	RBF	d=0.001, C=100	0.7	0.88	0.88	0.87

Additional files

Supplementary data available online at http://type3pks.in/tsvm/pks3/faq.php