Kernel based machine learning algorithm for the efficient prediction of type III polyketide synthase family of proteins

V Mallika[1], KC Sivakumar[2], S Jaichand[2], EV Soniya[1*]

[1] Plant Molecular Biology Division, 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

{mallika,sivakumar,evsoniya}@rgcb.res.in

tojaichand@gmail.com

Summary

Type III Polyketide synthases (PKS) are family of proteins considered to have significant role in the biosynthesis of various polyketides in plants, fungi and bacteria. As these proteins show positive effects to human health, more researches are going on regarding this particular protein. Developing a tool to identify the probability of sequence, being a type III polyketide synthase will minimize the time consumption and manpower efforts. In this approach, we have designed and implemented PKSIIIpred, a high performance prediction server for type III PKS where the classifier is Support Vector Machine (SVM). Based on the limited training dataset, the tool efficiently predicts the type III PKS superfamily of proteins with high sensitivity and specificity. PKSIIIpred is available at http://type3pks.in/prediction/. We expect that this tool may serve as a useful resource for type III PKS researchers. Currently work is being progressed for further betterment of prediction accuracy by including more sequence features in the training dataset.

1 Background

Polyketide synthases, also known as PKS, are family of enzyme complexes that produce large class of secondary metabolites known as polyketides which possess pharmacologically important properties including antibiotic, antifungal, antitumor and immunosuppressive activities. Three types of PKS are known ([1], [2]) to date, and they are quite different from each other in their structures and functions.

Among the types of PKS, type III PKS is responsible for the synthesis of polyketides including chalcone and stilbene. The well studied chalcone synthase play a significant role in the biosynthesis of various polyketides, including substances required for flower colour, defence against pathogen, protection from ultraviolet light, interaction with microorganism and fertility [3]. In many of the plants, chalcone synthase exist as multigene families ([4], [5]). In addition to the CHS, the super family also includes functionally divergent non-chalcone forming members like 2-pyrone synthase, phloroisovalerophenone synthase, trihydroxybenzophenone synthase, acridone synthase, stilbene synthase, pentaketide chromone synthase, octaketide synthase, resveratrol synthase, benzalacetone synthase, aloesone synthase and stilbene carboxylate synthase [1]. Extensive gene duplication followed

* Corresponding author

doi:10.2390/biecoll-jib-2010-143

by the functional divergence is believed to have played an important role in generating the biochemical diversity of PKS superfamily.

Structurally type III PKS proteins are small homodimeric proteins of approximately 40-45kDa, with a catalytic triad of Cys164-His303-Asn336 at the active site. Structural insight into the PKS reaction mechanism elucidated for the CHS from the *Medicago sativa* indicates the presence of three inter connected cavities: the CoA binding tunnel, a coumaroyl binding pocket and a cyclisation pocket [6] at the active site. Type III PKS monomer utilizes the triad [7] within an internal active site cavity that is connected to the surrounding aqueous phase by a narrow CoA binding tunnel.

Studies on various type III polyketide derivatives propose positive effect on human health and it is found that their products play therapeutically important roles [8] in disease treatment. For example, resveratrol, a stilbene synthase derivative from grapes shows cancer chemopreventive activity in murine models ([9], [10]). Currently there are more studies happening on type III PKS family of proteins and the volume of the polyketide data is increasing day by day. With this regard identification of proteins by laboratory experiments are difficult and also time consuming. In this approach, we attempted to predict type III PKS superfamily of proteins by using support vector machine.

2 Methods

2.1 Dataset

Two datasets were considered for the development of the prediction tool. Positive (+) dataset comprised of 70 selected type III PKS protein sequences from plants, fungi and bacteria. Similarly negative (-) dataset was created by using same numbers of non-type III PKS protein sequences. The sequences were retrieved from Swiss-Prot in FASTA format (http://www.expasy.org/sprot/) and used to train and model SVM for predicting the type III PKS. To test the reliability of the prediction server we also prepared a test set of 1000 type III PKS and non type III PKS proteins which were not the part of training set. In the test set of non type III PKS proteins, we included ketosynthase proteins as they adopt similar structural fold and can often show sequence similarity to type III PKS proteins. The test set is available online as supplementary data at http://type3pks.in/prediction/faq.php.

2.2 Support Vector Machine

The support vector machines (SVM) are group of fast optimization machine learning algorithms with strong theoretical foundation developed by Vapnik and his team at AT&T Bell Labs, which have been used for many kinds of pattern recognition ([11], [12], [13]). SVM can be trained by using various sequence features for the successful prediction of type III PKS. In the given work, SVM has been implemented by using SVM^{light} package (http://svmlight.joachims.org/) which possesses two modules: SVM_learn and SVM_classify. The first module is concerned for preparing models learned from the training dataset (+ve and -ve) and the latter one classifies the data by using the models prepared by SVM_learn. Here we have trained the SVM by using a set of positive and negative datasets, produces a model (classifier) that can be used to identify the potential type III PKS. This package allows users to select various parameters and various kernel functions (linear, polynomial, radial basis, sigmoid or any other user defined kernel). The selection of optimal kernel function is very important in SVMs. Here in the creation of SVM models, we have used four types of kernel functions: linear, polynomial, sigmoid and radial. The performance of SVM based methods has been optimized by tuning SVM parameters, in order to achieve maximum accuracy.

2.3 Numerical properties for SVMs

The SVM models were trained by using dipeptide and multiplet frequencies ([14], [15]) of amino acid composition. The total number of amino acids is 20 and therefore the theoretical number of possible dipeptides is 400. For each protein, a matrix of 400 dipeptides was generated and fed as an input to SVM. The frequency of each dipeptide is calculated by the formula: $DF_{ij}=N_{ij}/N$ where '$N_{ij,}$' count of the ij^{th} dipeptide; N, total number of possible dipeptides; i, j = 1-20. The repetitiveness of the amino acid sequences were also analyzed by means of multiplet which comprises homopolymeric stretches of any length (homodipeptides XX, homotripeptides XXX, etc.) $(X)_n$ where 'X' denotes any specific amino acid and $n\geq 2$. This measure of homopeptide density can be evaluated by using the formula: MFi = (counts of the i^{th} amino acid occurring as multiplet)/I, where i=1-20 and 'I' is the length of the protein.

2.4 Implementation

We have implemented the prediction tool in a web server which is available at: http://type3pks.in/prediction/. The program is written entirely in PHP and JAVA, and is hosted by Apache on a linux server. The home page serve as the platform for submitting data where users can either paste or upload sequence which should be in standard FASTA format. The pages also outline the program's features and introduces user about the protein type III PKS.

2.5 Performance assessment

To evaluate the reliability of the prediction tool, we used four parameters: sensitivity, specificity, accuracy and Matthews correlation coefficient. The sensitivity gives the fraction of positive events predicted by the tool and the specificity indicates how many false subjects are incorrectly recognized as positives. Both sensitivity and specificity ranges within '0' and '+1', the latter value represents accurate prediction. Accuracy is the proportion of correctly predicted proteins.

$$\text{Specificity (SP)} = TP/(TP+FP)$$

$$\text{Sensitivity (SN)} = TP/(TP+FN)$$

$$\text{Accuracy (AC)} = (TP+TN)/(TP+FP+TN+FN)$$

Matthews correlation coefficient or MCC ([16], [17]) is a statistical parameter which also used to estimate the accuracy of prediction. MCC may range from -1 to +1 and the highest MCC value indicates better prediction [18].

The MCC is given as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Where TP, number of true positives; TN, number of true negatives; FP, number of false positives; FN, number of false negatives. In this work type III PKS are true positives and non-type III PKS are true negatives.

3 Results

We have designed and implemented a high performance prediction server towards developing a robust protocol for identification of type III PKS proteins where the classifier is Support Vector Machines (SVM). Web interface and overall architecture of the prediction server is depicted in the figure 1 and 2 respectively. Based on the very limited training dataset, the tool efficiently predicts the type III PKS family of proteins with high sensitivity and specificity.

Table 1: Amino acid composition of type III PKS and non-Type III PKS.

Amino acid	No of residues in type III PKS	No of residues in Non-Type III PKS
Ala (A)	35 (9.0%)	29 (12.8%)
Arg (R)	16 (4.1%)	15 (6.6%)
Asn (N)	10 (2.6%)	10 (4.4%)
Asp (D)	19 (4.9%)	13 (5.7%)
Cys (C)	7 (1.8%)	6 (2.6%)
Gln (Q)	13 (3.3%)	10 (4.4%)
Glu (E)	28 (7.2%)	8 (3.5%)
Gly (G)	31 (8.0%)	23 (10.1%)
His (H)	8 (2.1%)	4 (1.8%)
Ile (I)	25 (6.4%)	6 (2.6%)
Leu (L)	35 (9.0%)	24 (10.6%)
Lys (K)	27 (6.9%)	4 (1.8%)
Met (M)	14 (3.6%)	6 (2.6%)
Phe (F)	14 (3.6%)	5 (2.2%)
Pro (P)	21 (5.4%)	8 (3.5%)
Ser (S)	20 (5.1%)	17 (7.5%)
Thr (T)	21 (5.4%)	11 (4.8%)
Trp (W)	4 (1.0%)	3 (1.3%)
Tyr (Y)	11 (2.8%)	7 (3.1%)
Val (V)	30 (7.7%)	18 (7.9%)
Pyl (O)	0 (0.0%)	0 (0.0%)
Sec (U)	0 (0.0%)	0 (0.0%)

It has been shown in the past that amino acid composition can be used to classify proteins ([19], [20], [21]). It was observed that amino acid compositions of type III PKS proteins were significantly different from that of non-PKS proteins. Difference in amino acid composition of type III PKS and non-PKS calculated by using ProtParam [22] (http://au.expasy.org/tools/protparam.html) is represented in the table 1. Here chalcone synthase (Swiss-Prot ID: P30073) and type I polyketide synthase ketosynthase domain

(Swiss-Prot ID: Q32YX0) were taken as representatives. Thus it is possible to discriminate type III PKS proteins from other proteins.

The SVM based classifier was developed using the numerical properties obtained by running the dataset using dipeptide frequency. The composition of dipeptides differs significantly in the type III PKS and non-PKS proteins. Different kernel and parameter functions were also tried to get the best performance in the given training dataset. The final output score for a given sequence was obtained using a combination of all the available models.

To access the performance of PKSIIIpred, we preferred a test set of 1000 type III PKS sequence and 1000 non type III PKS sequences that were not the part of the training dataset. It is interesting to find that PKSIIIpred could identify 99% of the known type III PKS sequences with 99.5% accuracy with MCC observed to be highest (0.99) where the sensitivity and specificity were nearly the same (99.3% and 99.6%). We could observe that the combination of models provided the best results.

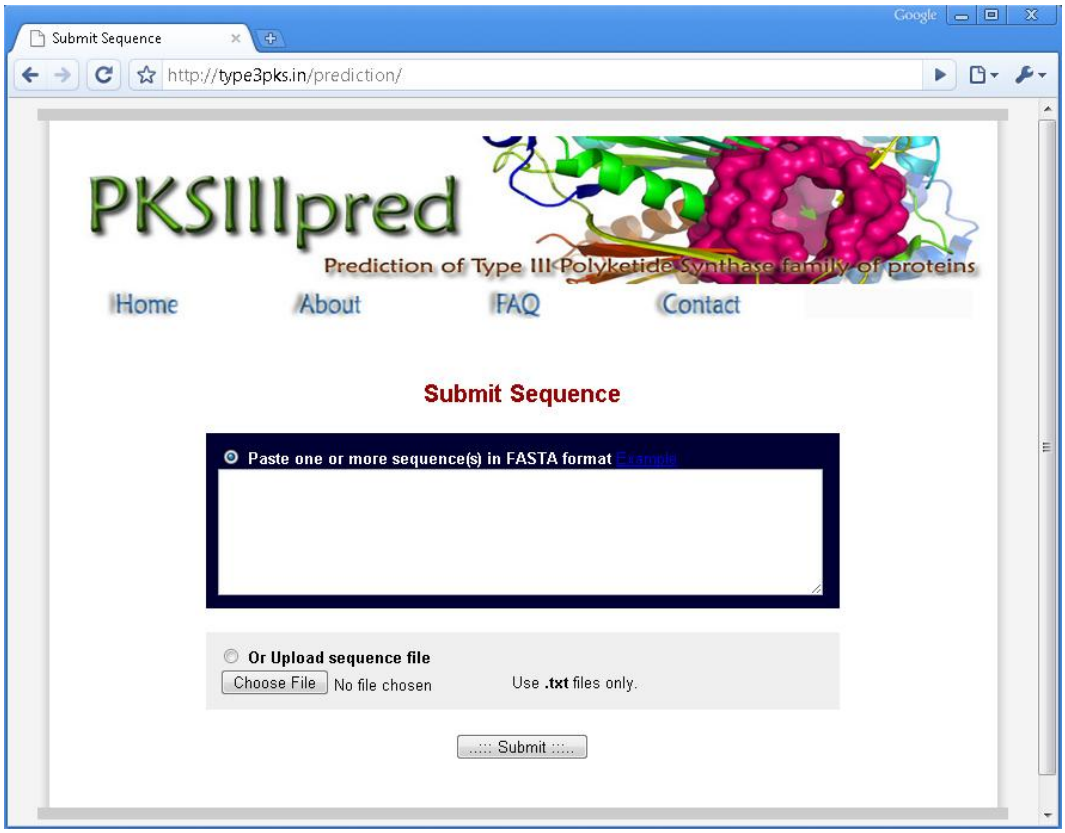

Figure 1: Web interface of PKSIIIpred.

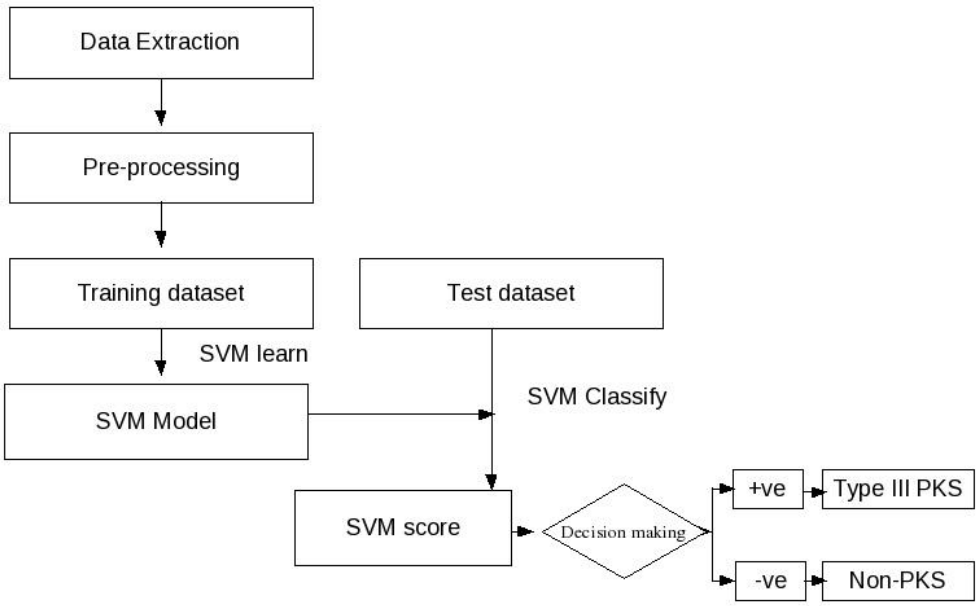

Figure 2: The Architecture of the PKSIIIpred server.

Web server: The method presented here is available on the World Wide Web in the form of a server, "PKSIIIpred". The Uniform Resource Link is http://type3pks.in/prediction/. The user can enter a protein sequence or upload a file consisting of multiple sequences in the standard FASTA format. If the user wishes to upload more sequence, it must be in .txt format.

4 Discussion

With the rapidly increasing amount of data generated by large-scale sequencing and intensifying attention on the study of type III PKS, methods that can discriminate type III PKS with high reliability and fast speed are important. NRPS-PKS is a reported web based server for PKS proteins developed by National Institute of Immunology [23] which can identify the potential PKS protein domains and also substrate specificity. Here in our approach, we are using a different methodology of machine learning to discriminate type III PKS and non-PKS proteins. We have provided type III PKS amino acid sequences from bacteria, fungi and bryophytes in the training dataset, so they can be perfectly predicted during user investigation.

In this approach, by giving a training set with known class labels (+1 for type III PKS, -1 for non-PKS), the SVM in training learns a hyperplane that optimally separates the items of the two classes. To calculate the decision factor, query is compared with the sequence composition of training set using the kernel function. If query is more 'similar' to the type III PKSs from the training set a positive score is obtained, otherwise score is negative. Depending on whether score is larger than or smaller than 0, proteins are usually classified into one of the two classes by the SVM.

The server uses far fewer features but achieved better performance in the evaluation. The values of statistical parameters (sensitivity, specificity, accuracy and MCC) are found to be very modest. These values rapidly enhance if further subjects are integrated in the learning dataset. It is noted that the server efficiently predicts type I polyketide synthase, ketosynthase domain as negative which adopts similar structural fold and can often shows sequence similarity to type III PKS proteins. These results demonstrated that the sequence features used by PKSIIIpred have powerful discriminating power.

It is encouraging to note that the ability of Support Vector Machines to learn and classify the PKS sequences accurately using small sized training dataset. Thus by using fewer, sequence-based features and also by significantly reduced computing cost, a web server to be developed.

5 Conclusions

In this work, we have described SVM based approach for the prediction of type III PKS proteins based on amino acid composition. Based on this method we have developed and implemented an efficient and easy to use online user-friendly prediction server called PKSIIIpred. The sensitivity and specificity reaches nearly 99% for prediction of type III PKS. We expect that the tool may serve as a useful resource for researchers as it is freely available. Currently work is being progressed for further betterment in prediction accuracy by including more sequence features for the training dataset.

Acknowledgements

The authors wish to thank BTISNet, Department of Biotechnology, Government of India for the Bioinformatics facility. This work was supported by a grant from Department of Information Technology, Government of India.

References

[1] M.B. Austin and J.P. Noel, The chalcone synthase superfamily of type III polyketide synthases, *Nat Prod Rep,* vol. 20, pp. 79–110, 2003.

[2] B. Shen, Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms, *Curr Opin Chem Biol*, vol. 7, no. 2, pp. 285-295, 2003.

[3] J. Schroder, A family of plant-specific polyketide synthases: facts and predictions, *Trends Plant Sci*, vol. 2, pp. 373–378, 1997.

[4] R.E. Koes, C.E. Spelt, J.N.M Mol, and A.G.M. Geratas, The chalcone synthase multigene family of Petunia hybrida (V30) Sequence homology, chromosomal location and evolutionary aspects, *Plant Mol Biol*, vol. 10, pp. 375-385, 1987.

[5] T.B. Ryder, S.A. Hedrick, J.N. Bell, X. Liang, S.D. Clouse, and C.J. Lamb, Organisation and differential activation of a gene family encoding the plant defense enzyme CHS in Phaseolus vulgaris, *Mol Gen Genet*, vol. 210, pp. 219-233, 1987.

[6] J.L. Ferrer, J.M. Jez, M.E. Bowman, R.A. Dixon, and J.P. Noel, Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis, *Nat Struct Biol*, vol. 6, pp. 775-784,1999.

[7] M.B. Austin, M. Izumikawa, M.E. Bowman, D.W. Udwary, J.L. Ferrer, B.S. Moore and J.P. Noel, Crystal Structure of a Bacterial Type III Polyketide Synthase and Enzymatic Control of Reactive Polyketide Intermediates, *J Biol Chem*, vol. 279, pp. 45162–45174, 2004.

[8] F.E. Koehn and G.T. Carter, The evolving role of natural products in drug discovery, *Nat Rev Drug Discov*, vol. 4, no. 3, pp. 206-220, 2005.

[9] M. Jang, L. Cai, G.O. Udeani, K.V. Slowing, C.F. Thomas, C.W. Beecher, H.H. Fong, N.R. Farnsworth, A.D. Kinghorn, R.G. Mehta, R.C. Moon, and J.M. Pezzuto, Cancer

chemopreventive activity of resveratrol, a natural product derived from grapes, *Science*, vol. 275, no. 5297, pp. 218-220, 1997.

[10] M.V. Clement, J.L. Hirpara, S.H. Chawdhury, and S. Pervaiz, Chemopreventive Agent Resveratrol, a Natural Product Derived From Grapes, Triggers CD95 Signaling-Dependent Apoptosis in Human Tumor Cells, *Blood*, vol. 92, no. 3, pp. 996-1002, 1998.

[11] V.N. Vapnik, An overview of statistical learning theory, *Neural Networks, IEEE Transactions*, vol. 10, pp. 988-999, 1999.

[12] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning*, vol. 20, pp. 273-297, 1995.

[13] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[14] V. Brendel, P. Buchert, I.R. Nourbakhsh, B.E. Blaisdell, and S. Karlin, Methods and algorithms for statistical analysis of protein sequences, *Proc Natl Acad Sci USA*, vol. 89, pp. 2002-2006, 1992.

[15] S. Karlin, V, Brendel, and P. Bucher, Significant similarity and dissimilarity in homologous proteins, *Mol Biol Evol*, vol. 9, no. 1, pp. 152-167, 1992.

[16] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim Biophys Acta*, vol. 405, pp. 442-451, 1975.

[17] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics*, vol. 16, pp. 412-424, 2000.

[18] O. Carugo, Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots, *BMC Bioinformatics*, vol. 8, pp. 380, 2007.

[19] A. Garg, M. Bhasin, G.P. Raghava. Support Vector Machine-based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search. *J. Biol. Chem.* Vol. 280, pp. 14427-14432, 2005.

[20] M. Bhasin and G.P. Raghava. ELSpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acid Res*, Vol. 32, W414–W419, 2004.

[21] S. Mei and W. Fei. Amino acid classification based spectrum kernel fusion for protein subnuclear localization. *BMC Bioinformatics*, 11 (Suppl 1): S17 doi:10.1186/1471-2105-11-S1-S17, 2010.

[22] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, Protein Identification and Analysis Tools on the ExPASy Server, (In) John M. Walker (ed), The Proteomics Protocols Handbook, Humana Press. pp. 571-607, 2005.

[23] M. Z. Ansari, G, Yadav, R. S. Gokhale, and D. Mohanty, NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases, *Nucleic Acids Res*, vol. 32, pp. W405-W413, 2004.