TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

A Machine Learning Approach for Prediction of Gibberellic Acid Metabolic Enzymes in Monocotyledonous Plants

Sreepriya P.[1], Naganeeswaran S. [1], Hemalatha N. [2], Sreejisha P. [1] and Rajesh M. K.[1,*]
[1]*Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod, Kerala, India*
[2]*AIMIT, St. Aloysius College, Mangalore, Karnataka, India*
sreepriya.pradeep@gmail.com, naganeeswaran@gmail.com, hemasree71@gmail.com, sreeji279@gmail.com, mkraju.cpcri@gmail.com
[*Corresponding author: Phone: +91-4994-232894-284; Fax: +91-4994-232322]

ABSTRACT

Gibberellins (GA) are one of the most important phytohormones that control different aspects of plant growth and influence various developments such as seed germination, stem elongation and floral induction. More than 130 GAs have been identified; however, only a small number of them are biologically active. In this study, five enzymes in GA metabolic pathway in monocots have been thoroughly researched namely, ent-copalyl-diphosphate synthase (CPS), ent-kaurene synthase (KS), ent-kaurene oxidase (KO), GA 20-oxidase (GA20ox), and GA 2-oxidase (GA2ox). We have designed and implemented a high performance prediction tool for these enzymes using machine learning algorithms. 'GAPred' is a web-based system to provide a comprehensive collection of enzymes in GA metabolic pathway and a systematic framework for the analysis of these enzymes for monocots. WEKA-based classifiers (Naïve-Bayes) and Support Vector Machine (SVM) based-modules were developed using dipeptide composition and high accuracies were obtained. In addition, BLAST and Hidden Markov Model (HMMER-based model) were also developed for searching sequence databases for homolog's of enzymes of GA metabolic pathway, and for making protein sequence alignments.

Keywords: GA, SVM, WEKA, BLAST, HMMER

1 INTRODUCTION

Gibberellic acids (GA) are naturally occurring phytohormones that regulate growth and influence various developmental processes, including stem elongation, germination, dormancy, flowering, enzyme induction, and leaf and fruit senescence [1]. They are also involved in the discernment of environmental stimuli, thus are significant not only for a plant's growth and development but also in awareness of its environment. Gibberellins are diterpenoid acids which

DOI: 10.14738/tmlai.24.375
Publication Date: 28th August 2014
URL: http://dx.doi.org/10.14738/tmlai.24.375

are formed by the terpenoid pathway in plastids and then modifying the endoplasmic reticulum (ER) and cytosol until they are biologically-active [2]. Gibberellins are derived by the *ent*-gibberellane skeleton, but are synthesized by *ent*-kaurene [3]. The GA biosynthetic pathway can be divided into three stages, each stage residing in a different cellular compartment *viz*. plastid, the endoplasmic reticulum, and the cytosol [4].

A number of experimental studies have explained thoroughly the biosynthetic functions of gibberellic acid. In this study, five enzymes involved in GA metabolic pathway in monocots have been thoroughly researched namely, *ent*-copalyl-diphosphate synthase (CPS), *ent*-kaurene synthase (KS), *ent*-kaurene oxidase (KO), GA 20-oxidase (GA20ox), and GA 2-oxidase (GA2ox) [5]. In this study, we have designed and implemented a high performance prediction tool based on kernel-based Machine Learning Algorithms *viz*., Support Vector Machine (SVM) and WEKA for prediction of enzymes in gibberellic acid metabolic pathway. In addition, standalone BLAST and Hidden Markov Model (HMMER-based model) were also developed for searching sequence databases for homolog's of enzymes of GA metabolic pathway, and for making protein sequence alignments. 'GAPred' was developed using the evolutionary and sequence features of a protein sequence and the performance of the each model was evaluated using cross-validation techniques. Based on our study, we have also created and hosted a web server for predicting enzymes involved in GA metabolism.

2 MATERIALS AND METHODS

2.1 Dataset

In the present study, two datasets were considered for the development of the prediction tool 'GAPred'. Positive (+ve) dataset comprised of 102 selected GA metabolic enzyme protein sequences from monocots *viz*., date palm (*Phoenix dactylifera*), coconut (*Cocos nucifera*), rice (*Oryza sativa*), barley (*Hordeum vulgare*), maize (*Zea mays*), banana (*Musa acuminata*), and brachypodium (*Brachypodium distachyon*), after redundancy elimination by using ClustalW. Similarly negative (-ve) dataset was created by using same numbers of non-GA metabolic enzymes sequences. The sequences were retrieved from NCBI in FASTA format (http://www.ncbi.nlm.nih.gov/). Domains of enzymes involved in GA metabolic pathway were identified using Pfam search and PRINTS search and most of the identified enzyme domains are known to be conserved in related species. To avoid the over estimation, we clustered the protein sequences from positive data (+ve) set with a threshold of 30% identity by CD-HIT (Cluster Database at High Identity with Tolerance). Out of 102 GA metabolic enzymes sequence, 62 proteins were randomly selected for the creation of training set. Similarly training set of non-GA metabolic enzymes sequence was created. To test the reliability of the prediction tool, we also prepared a test set of 40 GA metabolic enzymes sequences and non-GA metabolic enzymes sequences which were not the part of training set.

2.2 Support Vector Machine

Support Vector Machines (SVM) are a group of rapid optimization machine learning algorithms with strong theoretical foundation, which have been used for many kinds of pattern recognition [6-8]. SVMs are now extensively used for biological applications and methods such as classifying objects as diverse as protein and DNA sequences, mass spectra and microarray expression profiles [9]. In this work, SVM has been implemented by using $SVM^{multiclass}$ package [10] which possesses two modules: SVM_multiclass_learn and SVM_multiclass_classify. The first module (SVM_multiclass_learn) is concerned for preparing models learned from the training dataset (+ve and -ve) and the final one classifies the data by using the models prepared by SVM_multiclass_learn. Here, we have trained the $SVM^{multiclass}$ by using a set of positive and negative datasets, and produces a model (classifier) that can be used to identify the potential enzymes involved in gibberellic acid metabolic pathway. With the help of this package the user can select various kernel functions (linear, polynomial, radial basis, sigmoid or any other user defined kernel) for preparing models. In SVMs, the kernel function selected must be the most favorable one. Here in the creation of SVM models, we have used three types of kernel functions: linear, polynomial, and radial. The performance of SVM based methods has been optimized by regulating SVM parameters so that maximum accuracy could be obtained.

2.3 WEKA

WEKA stands for 'Waikato Environment for Knowledge Analysis' and is a free open source software developed by at the University of Waikato, New Zealand. This popular machine learning software contains a collection of algorithms and visualization tools for data analysis, analytical modeling and also graphical user interfaces for easy access to this functionality. In the given work, we used WEKA classification [11], where different attributes of a protein sequence are analyzed to classify the protein sequence into one of the predefined classes. Both train and test set was used to get the classification of the data set by using better algorithms. The performance of WEKA has been optimized by tuning evaluation parameters and visualization schemes, in order to analyze the accuracy of classifiers.

2.4 Sequence Features

Dipeptide composition gives comprehensive information about each protein sequence that possess sequence feature. Generally, the total number of amino acids is 20 and thus the theoretical number of possible dipeptides is 400. A matrix of these 400 dipeptides was generated for each protein and is then given as an input to both SVM and WEKA. Each dipeptide frequency is calculated by the formula: $DF_{ij}=N_{ij}/N$ where, N_{ij}=count of the ijth dipeptide; N=total number of possible dipeptides; i, j = 1-20 amino acid.

2.5 Performance Assessment of GAPred

With the help of statistical calculations, we generally examine the efficiency of a predictor either using single independent dataset test, cross-validation test or jackknife test. However, jackknife test method takes much longer time to examine a predictor based on SVM and WEKA [12] and therefore, in this work, we have adopted 10-fold cross-validation for WEKA and 5-fold cross-validation for SVM^{multiclass} and independent data set validation techniques were adopted for measuring performance. In 10-fold cross-validation test, the significant dataset was divided randomly into ten equally sized sets. The training and testing methods were carried out ten times with each individual set used for testing and for the nine sets left behind for training. Similarly in 5-fold cross validation, the dataset was partitioned randomly into five equally sized sets. In the independent dataset test, the training dataset used to train the predictor does not contain any data that is to be tested.

2.6 Evaluation Parameters

We had made use of five parameters to evaluate the reliability of the prediction tool, they are: Accuracy (Ac), Sensitivity (Sn), Specificity (Sp), Precision (Pr) and Matthew's Correlation Coefficient (MCC). Accuracy defines the proportion of correctly predicted proteins (Eq.1). The sensitivity (Sn) and specificity (Sp) represent the correct prediction ratios of positive (+ve) and negative data (-ve) sets of metabolic enzymes of gibberellic acid sequence respectively (Eq. 2 and 3). Precision is the proportion of the predicted positive cases that were correct (Eq.4). Matthew's correlation coefficient or MCC [13-14] is a statistical parameter which also used to estimate the accuracy of prediction (Eq.5). MCC may range from -1 to +1 and the highest MCC value indicates better prediction [15].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad (2)$$

$$Specificity = \frac{TN}{FP+TN} \times 100 \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (4)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Where TP=number of true positives; TN=number of true negatives; FP=number of false positives; FN=number of false negatives. In this work, metabolic enzymes of GA sequences are true positives and non- metabolic enzymes of GA sequences are true negatives.

2.6 Sequence Similarity search using standalone BLAST

The standalone BLAST programs are freely provided as open-source software by NCBI. With stand-alone BLAST we can made our own databases to search against. In this study, sequences were searched against the protein non-redundant (nr) database in association with standalone BLASTp and to detect homology of metabolic enzymes of gibberellic acid proteins and result was analysed.

2.7 Sequence Similarity Search using HMMER

Profile Hidden Markov models (profile HMMs) techniques are one of the most dominant methods for protein homology detection [16]. HMMER helps to find out protein sequences which are similar in sequence databases and to make protein sequence alignments. HMMER becomes particularly powerful when the query is a multiple sequence alignment of a sequence family rather than for single query sequences. It makes a profile of the query that assigns a position-specific scoring system for substitutions, insertions, and deletions [17]. HMMER profiles are probabilistic models called “profile Hidden Markov models” (profile HMMs). Because of the strength of its underlying probability models, HMMER aims to be much more accurate and more capable of finding out remote homolog’s rather than BLAST, FASTA or any of the other sequence alignment and database search tools based on older scoring methodology [18]. Hence, in this study, we have used HMMER to detect homology of metabolic enzymes of gibberellic acid proteins and a remarkable result was analyzed.

2.8 ROC Curves

By making use of ROC curves, a graph created by plotting the fraction of false positives (FPR) against true positives (TPR) at various threshold settings [19], we can explain the performance of multi class classifiers in SVM and WEKA more specifically. TPR is also known as sensitivity, and FPR is 1-specificity or true negative rate. ROC analysis is linked in a direct and natural method to benefit analysis of diagnostic decision making. ROC curves useful for the evaluation of machine learning techniques and data mining research.

2.9 Web-server

We have implemented the prediction tool “GAPred” in a web server. The program is written entirely in HTML, PHP and PERL program in a Linux platform. The tool page serve as the platform for submitting data where users can either paste or upload sequence which should be in standard FASTA format. It also provides a comprehensive collection of enzymes in GA metabolic pathway and introduces a user to gibberellic acid metabolic pathway.

3 GAPRED

3.1 Evaluation of Performance of GAPred

We have carried out 10-fold cross-validation for WEKA and 5-fold cross-validation for SVM and also independent data test validation to evaluate the performance of GAPred (Tables 1-4). Cross validation and independent data test results for SVM from the Tables 1 and 3 shows that cross validation has better result for dipeptide composition methods compared to independent data test. While in the case of WEKA, the independent data set has better result than cross validation method.

3.2 Comparison of GAPred with BLAST and HMMER

We have also used standalone BLAST and HMMER to detect homology of metabolic enzymes of gibberellic acid. This was used to compare an input protein sequence with a created database to generate the homology of the given sequence. A comparison of enzyme proteins was conducted with standalone BLAST and HMMER database and an accuracy of 99% and 93% were obtained (Tables 5-6). By making a comparison between SVM, WEKA, BLAST and HMMER from Table 7 and Figure 1, it can be seen that SVM has achieved 100% accuracy and MCC value equal to 1 that an ideal classification method should possess. Hence, SVM was selected to be the best model for GAPred.

Table 1 Validation of independent data test results of dipeptide composition of metabolic enzymes of gibberellic acid with SVM^{multiclass}

Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Linear	5	100	53	100	0.16
Polynomial	95	100	98	100	0.95
RBF	93	95	94	95	0.88

Table 2 Validation of independent data test results of dipeptide composition of metabolic enzymes of gibberellic acid with WEKA

Classifiers	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Naïve Bayes	98	100	99	100	0.98
Bayes Net	95	100	98	100	0.95
Decorate	83	100	91	100	0.84

Table 3 Comparison of the prediction performance of three kernels of SVMmulticlass with dipeptide composition technique using 5-fold cross validation

Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Linear	100	100	100	100	1
Polynomial	100	100	100	100	1
RBF	100	100	100	100	1

Table 4 Comparison of the prediction performance of three classifiers of WEKA with dipeptide composition technique using 10-fold cross validation

Classifiers	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Naïve Bayes	89	100	94	100	0.89
Bayes Net	95	97	96	97	0.92
Decorate	95	95	95	95	0.90

Table 5 Comparison of the prediction performance of standalone BLAST with created database of domains of metabolic enzymes of gibberellic acid

	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
BLAST	100	98	99	98	0.98

Table 6 Comparison of the prediction performance of HMMER with created database of domains of metabolic enzymes of gibberellic acid

	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
HMMER	90	97	93	96	0.87

Table 7 Comparison of the prediction performance of three methods with metabolic enzymes of gibberellic acid sequences

	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
SVM	100	100	100	100	1
WEKA	98	100	99	100	0.98
BLAST	100	98	99	98	0.98
HMMER	90	97	93	96	0.87

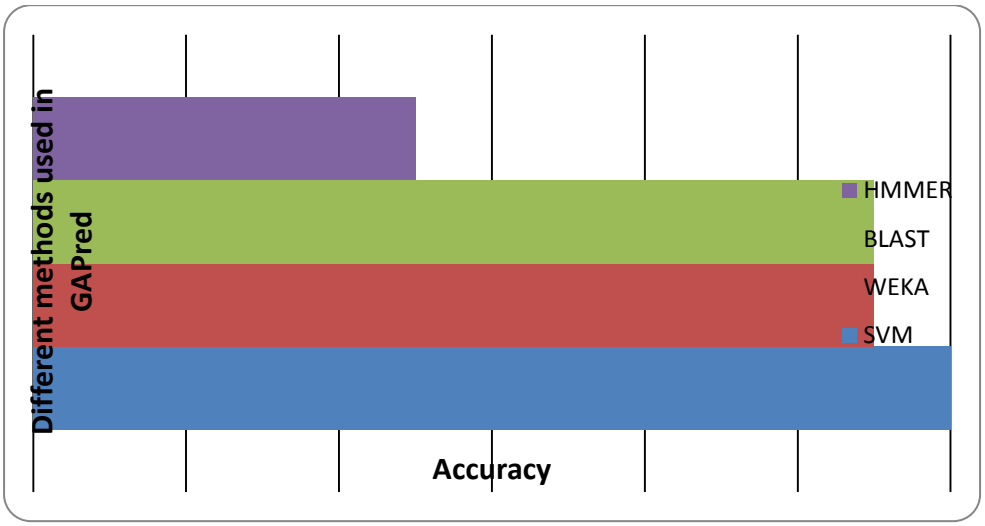

Figure 1: Comparison of performance validation of GAPred with different methods

3.3 ROC curve

We have plotted the ROC curves for SVM and WEKA based on the independent test performance of the dipeptide compositions. From the ROC curves (Figures 2-3), representing the relationship between sensitivity and (1-specificity) for a class, it is clear that the SVM composition module represents a perfect classifier since the curve obtained is an inverted 'L', which is a desirable characteristic of an ROC curve. Each point on the ROC curve was plotted based on different threshold scores.

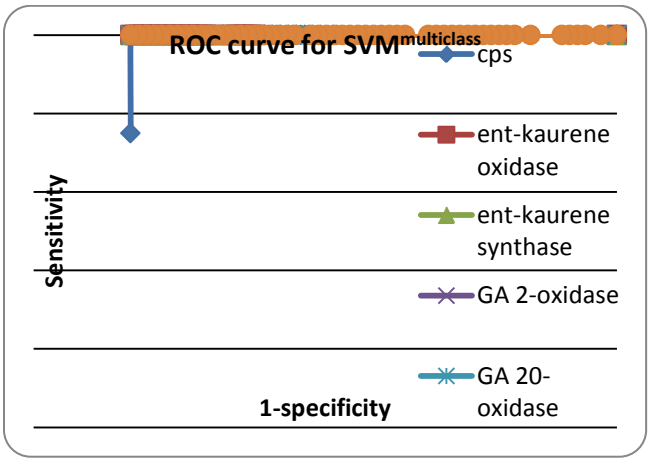

Figure 2: ROC curve for dipeptide composition in SVM^{multiclass} using independent test results

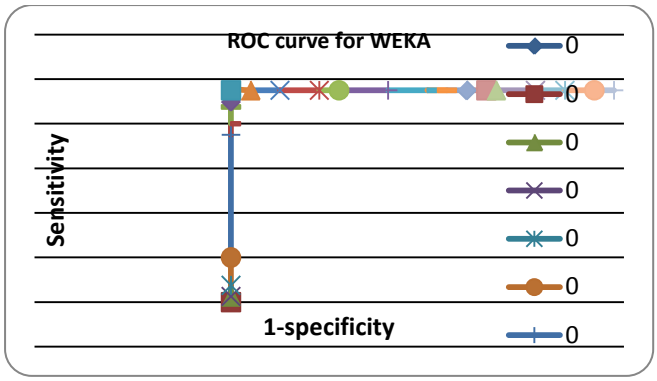

Figure 3: ROC curve for dipeptide composition in WEKA using independent test results

3.4 Description of Web Server

We have implemented the prediction tool "GAPred" in a web server. The tool was developed in PERL program and web interface in PHP and HTML to assess the user queries, in Linux platform. The tool page serve as the platform for submitting data where users can either paste or upload sequence which should be in standard FASTA format (Figure 4). It also provides a comprehensive collection of enzymes in GA metabolic pathway and introduces the user about gibberellic acid. The tool is freely available at http://gapred.cpcribioinformatics.in/gapred/

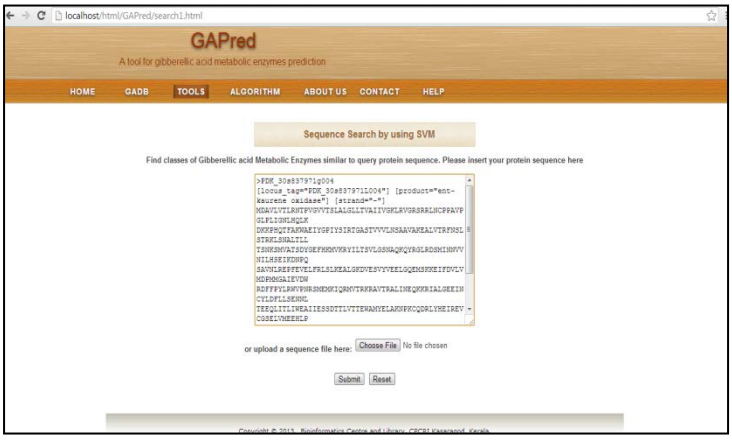

Figure 4: Web interface of GAPred

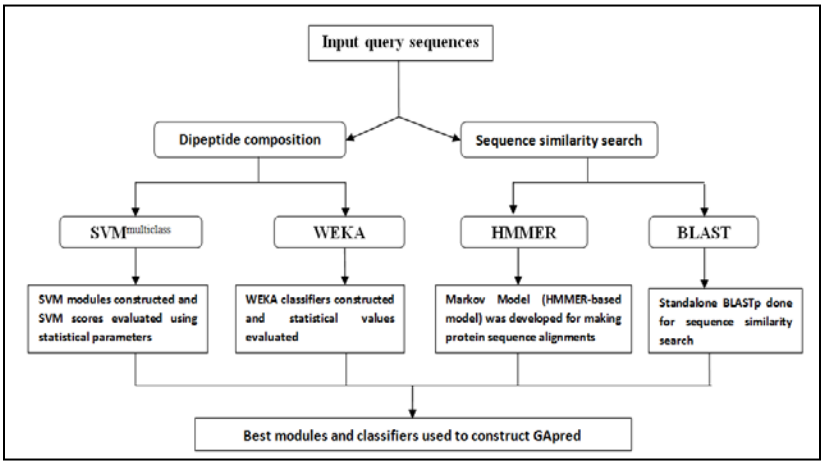

Figure 5: The architecture of the GAPred server.

4 CONCLUSION

In this work, we have described SVM and WEKA-based approaches for the prediction of enzymes in gibberellic acid metabolic pathway based on dipeptide composition. Comparison of standalone BLAST and HMMER- based homology searches with machine learning algorithms revealed that the latter performed better compared to homology-based tools. Based on kernel methods, we have developed and implemented an efficient and easy to use user-friendly prediction server called 'GAPred' for predicting five gibberellic acid metabolic enzymes. The sensitivity and specificity reaches 100% for prediction of gibberellic acid metabolic enzymes. We expect that the tool may be a useful resource for researchers as it is freely available.

REFERENCES

[1]. Phinney, B.O., The history of gibberellins. *In*: The Biochemistry and Physiology of Gibberellins, Crozier, A. (Ed.), Praeger Publishers, New York , USA , 1983. 1: p. 19–52.

[2]. Lichtenthaler, H.K., Rohmer, M. and Schwender, J., Two independent biochemical pathways for isopentenyl diphosphate biosynthesis in higher plants. Physiologia Plantarum, 1997. 101: p. 643–652.

[3]. Graebe, J.E., Gibberellin biosynthesis and control. Annual Review of Plant Physiology, 1987. 38: p. 419–465.

[4]. Chappell, J., Biochemistry and molecular biology of the isoprenoid biosynthetic pathway in plants. Annual Review of Plant Physiology and Plant Molecular Biology, 1995. 46: p. 521-547.

[5]. MacMillan, J., Biosynthesis of the gibberellin plant hormones. Natural Product Reports, 1997. 14: 221–244 .

[6]. Vapnik, V.N., An overview of statistical learning theory, Neural Networks. IEEE Transactions, 1999. 10: p. 988-999.

[7]. Cortes, C. and Vapnik, V. Support Vector Networks. Machine Learning, 1995. 20: p. 273-297.

[8]. Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998. 2: p. 121-167.

[9]. Noble, W.S., Support vector machine applications in computational biology. *In*: Kernel Methods in Computational Biology. Schoelkopf, B., Tsuda, K. and Vert, J. P. (Eds.), Cambridge, MA: MIT Press, 2004. p. 71–92.

[10]. Joachims, T., Making large-scale SVM learning practical. *In*: Advances in Kernel Methods: Support Vector Learning. Schoelkopf, B.,Burges, C. and Smola,A. (Eds.), Cambridge MA: MIT Press, 1999. p. 41–56.

[11]. Üney, F. and Türkay, M., A mixed-integer programming approach to multi-class data classification problem. European Journal of Operational Research, 2006. 173(3): p. 910-920.

[12]. Chou, K.C. and Zhang, C.T., Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology, 1995. 30: p. 275–349.

[13]. Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta, 1975. 405: p. 442-451.

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

[15]. Carugo, O., Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. BMC Bioinformatics, 2007. 8: p. 380.

[16]. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C., Sequence comparisons using multiple sequences detect three times as many remote homologues as pair-wise methods. Journal of Molecular Biology, 1998. 284: p. 1201-1210.

[17]. Hughey, R. and Krogh, A., Hidden Markov models for sequence analysis: extension and analysis of the basic method. Computer applications in the Biosciences, 1996. 12: p. 95-107.

[18]. Mitchison, G.J. and Durbin, R., Tree-based maximal likelihood substitution matrices and hidden Markov models. Journal of Molecular Evolution, 1995. 41: p. 1139-1151.

[19]. Swets, J.A., Measuring the accuracy of diagnostic systems. Science, 1998. 240: p. 1285–1293.