Hindawi Publishing Corporation
ISRN Computational Biology
Volume 2014, Article ID 581245, 11 pages
http://dx.doi.org/10.1155/2014/581245

Research Article

Application of Hybrid Functional Groups to Predict ATP Binding Proteins

Andreas N. Mbah

Center for Bioinformatics & Computational Biology, Department of Biology, Jackson State University, Jackson, MS 39217, USA

Correspondence should be addressed to Andreas N. Mbah; nji41@yahoo.com

Received 2 September 2013; Accepted 29 October 2013; Published 8 January 2014

Academic Editors: S.-A. Marashi and B. Oliva

Copyright © 2014 Andreas N. Mbah. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The ATP binding proteins exist as a hybrid of proteins with Walker A motif and universal stress proteins (USPs) having an alternative motif for binding ATP. There is an urgent need to find a reliable and comprehensive hybrid predictor for ATP binding proteins using whole sequence information. In this paper the open source LIBSVM toolbox was used to build a classifier at 10-fold cross-validation. The best hybrid model was the combination of amino acid and dipeptide composition with an accuracy of 84.57% and Mathews Correlation Coefficient (MCC) value of 0.693. This classifier proves to be better than many classical ATP binding protein predictors. The general trend observed is that combinations of descriptors performed better and improved the overall performances of individual descriptors, particularly when combined with amino acid composition. The work developed a comprehensive model for predicting ATP binding proteins irrespective of their functional motifs. This model provides a high probability of success for molecular biologists in predicting and selecting diverse groups of ATP binding proteins irrespective of their functional motifs.

1. Introduction

Recent advances in the next generation sequencing and human genome projects have resulted in rapid increase of protein sequences, thus widening the protein sequence-structure gap [1, 2], leading to diverse protein functions from common family. Computation prediction tools for predicting protein structure and function are highly needed to narrow the widening gap [3]. The ATP binding proteins (ATP-BPs) are a diverse family of proteins in terms of amino acid sequences, function, and their three-dimensional structures. These proteins hydrolyze ATP to provide the energy necessary to drive biochemical reactions in the cell [4]. There are two distinct functional groups of ATP binding proteins.

The first functional group has the Walker A motif [GXXXXGK (T/S) or G-4X-GK (T/S)] in their sequences for ATP binding [5]. Many members are transmembrane proteins and are responsible for transporting a wide variety of substrates across extra- and intracellular membranes [6]. The biochemical functions of ATP binding proteins are well exhibited within the ABC transporters group. In bacteria cell, ABC transporters pump substances such as sugars, vitamins, and metal ions into the cell, while in eukaryotes they transport molecules out of the cell [7]. They are also known to transport lipids and play a protective role to the developing fetus against xenobiotics [7]. ABC transporters are crucial in the development of multidrug resistance, with the ATP binding sites exploitable as targets for chemotherapeutic agents [8]. The mechanism of action in multidrug transportation is unclear. However, one model called hydrophobic vacuum cleaner states that, in P-glycoprotein, the drugs are bound indiscriminately from the lipid phase based on their hydrophobicity [9].

The second evolutionary diverse functional class of ATP binding proteins is called universal stress proteins (USPs). The universal stress proteins (USPs) are found in diverse group of organisms like archaea, eubacteria, yeast, fungi, and plants; their expressions are triggered by variety of environmental stressors [10]. These stressors might include but are not limited to starvation of nutrients such as carbon, nitrogen, phosphate, sulfate and the required amino acid and variety of toxicants and other agents such as heavy metals, oxidants, acids, heat shock, DNA damage, phosphate, uncouplers of the electron transport chain, and ethanol

[11, 12]. The USPs bind to ATP through the ATP binding motif [G-2X-G-9X-G(S/T)] [13]. Members of the USPs will segregate into two groups based on whether or not they bind to ATP [13].

Experimental efforts are underway to determine the function of newly discovered proteins [14], but these experimental methods are costly and time consuming and at times are unsuccessful, due to the complexity involved in protein crystallization process. Several methods had been studied based on predicting ATP binding residues from their known structural features but with low accuracies [15, 16]. Some predictors of ATP binding proteins have been developed with promising results such as those in [17, 18], including Green et al. [19] article on an effective method to recognize ATP binding proteins by testing parallel cascade identification and KNN. Unfortunately these methods were adapted to ATP binding proteins containing only the classical Walker A motif [G-4X-GK (T/S)] in their sequences. The objective of this research reported here was to introduce a classifier built from a pool of protein sequences containing both ATP binding motifs of G-4X-GK (T/S) and G-2X-G-9X-G(S/T). To achieve the objective, support vector machine (SVM) approach is proposed which predicts protein functions based on the discriminative features that map protein sequences to biological functions [20–23] using the sequence pool ATP hybrid motifs.

There is aneed to develop an automated predictor for ATP binding USP encoded proteins to speed experimental designs and study how these proteins function under diverse environmental stressors. This research has developed hybrid ATP binding protein predictor using the open source LIBSVM toolbox classification. The best model was the combination of amino acid and dipeptide composition of the sequences with an accuracy of 84.57% and Mathews correlation coefficient (MCC) value of 0.693%. This model shows a striking overall performance in sensitivity (82.46%), specificity (87.00%), and precision (87.85%) with area under the ROC curve (AUC) value of 0.849219. The general trend shows that combinations of descriptors perform better and improved the overall performances of individual descriptors, particularly when combined with amino acid composition. This model provides a high probability of success for molecular biologists in predicting and selecting diverse motif groups of ATP binding proteins.

2. Materials and Method

2.1. Datasets. Balanced datasets of ATP and non-ATP binding proteins were constructed from the UniProt protein database (UniProt release 2011_11) (http://www.uniprot.org/), Protein Data Bank (http://www.rcsb.org/pdb/home/home.do), IMG/M database (http://img.jgi.doe.gov/cgi-bin/m/main.cgi), and published literatures [24–26] which contain diverse universal stress proteins.

2.1.1. Extraction of Walker A Motif Dataset. A total of 2000 protein sequences which belong to Walker A motif positive dataset were retrieved. Redundancy due to homologous sequences was removed using CD-HIT [27] and PISCES [28] servers at a threshold of 25%. This threshold statistically retains adequate number of protein sequences for analysis as well as avoids bias that might result from high homology. Dataset obtained was manually reviewed through literature search and information from the protein data bank [2] to ensure they represent ATP binding proteins. A total of 100 sequences were randomly selected from the original dataset and retained for training and testing to represent Walker A motif positive (ATP binding) dataset. The Walker A motif negative dataset (non-ATP binding) was taken from Yu et al. 2006 [29]. This was the "negative" dataset used for nucleic acid binding proteins. This is because ATP binding proteins are members of nucleotide binding protein family; hence the negative dataset used in [29] for predicting nucleotide binding protein family was considered useful. Redundancy was also maintained at 25% threshold and each protein was verified to be non-ATP binding using both the literature and protein data bank information. A total of 100 sequences were also randomly selected from [29] and retained for training and testing to represent Walker A motif negative (non-ATP binding) dataset.

2.1.2. Extraction of USP Protein Dataset. The extracted USP sequences were tested for the presence or absence of the G-2X-G-9X-G(S/T) motif in their sequences using the NCBI conserved domain search tool [30]. The USP sequences were divided into two groups based on the presence or absence of ATP binding motif [13]. The redundancy was also maintained at 25% threshold and 100 sequences were selected for each class of proteins (200 sequences in total).

The overall summary of the data prepared for analysis was as follows: (i) 100 ATP binding proteins with Walker A motif; (ii) 100 without ATP binding proteins without Walker A motif, (iii) 100 USP sequences with ATP binding motif [G-2X-G-9X-G(S/T)], and (iv) 100 USP sequences without ATP binding motif [G-2X-G-9X-G(S/T)]. The 400 sequences were separated into two hybrid groups as follows: 200 ATP binding sequences and 200 sequences without ATP binding motifs and were used to generate the feature vector. The feature vector was generated from the entire sequences of the proteins (not only the ATP-binding domains) via PROFEAT server using 1497 descriptor set [31]. Physicochemical and sequence attributes of biologically informative were prioritized for investigation. The attributes were incorporated into LIBSVM classifier to find the best hybrid model for predicting ATP binding proteins.

2.2. LIBSVM Classifier. Support vector machines (SVM) recognized objects to be classified as points in a high-dimensional space needing a hyperplane to separate them [32]. The biological molecules are represented with descriptor set. With a proper mapping furnished by a kernel function, SVM classifiers separate transformed data with a hyperplane in a high-dimensional space to predict the correct classification of protein functional classes. SVMs have been widely used in supervised classification problems in bioinformatics, such as [33–36]. The LIBSVM package which is freely

downloadable at (http://www.csie.ntu.edu.tw/~cjlin/libsvm) was adopted and used to evaluate the attributes and build the final classifier, using the radial basis function (RBF) as the kernel function [37–39].

A "grid-search" was employed to select the proper values of the parameter of RBF and the penalty parameter (C) of the soft margin SVM. C was set to $2^{-5}, 2^{-3}, \ldots, 2^{15}$ and γ to $2^{-15}, 2^{-13}, \ldots, 2^{3}$. All the combinations of C and γ were tested and the pair with the best cross-validation accuracy for each feature set or combination of feature sets was selected. A smaller γ value makes the decision boundary smoother. The SVM training parameter C is the regularization factor, which controls the tradeoff between low training error and large margin [37, 40]. Throughout this work, the parameter C was maintained at $C = 4$ after trial and error assessment as the best value. The optimal value of γ was obtained for each descriptor set for best results. The entire sets of attributes were evaluated in terms of their association with ATP binding protein and a final subset with good predictive power was selected. In this research a 10-fold cross validation (10CV) was implemented. The objective of training is to maximize the ability of the SVM predictor to discriminate between classes while avoiding overfitting.

2.3. Tenfold Cross-Validation Analysis. The technique to evaluate any newly developed method has become a major challenge to investigators. The jack-knifing leave-one-out cross-validation (LOOCV) [41–43] is the popular technique for evaluating models. During this procedure one sequence is used for testing and the left over sequences are used for training. This process is repeated many times and each sequence is used once for testing. Even though this method is popular, it is computer intensive with considerable labor time.

In this work, 10-fold cross-validation was used to train and test the dataset with sequences randomly partitioned into ten sets. This cross-validation ensures that the dataset was split at the protein level in addition to the stratified partition, thus ensuring a more rigorous evaluation. During the procedure, the positive and negative data samples are distributed randomly into 10 sets or the so-called fold. In each of the 10 round steps, 9 of the 10 sets are used to construct a classifier (training), and then the classifier is evaluated using the remaining set (testing). This procedure was repeated ten times in a manner where each set was used for testing [44, 45]. The overall performance was the average of the performances of all the 10 sets.

2.4. The LIBSVM Performance Evaluation. The standard parameters used in evaluating the performance of the LIBSVM are indicated below. The overall accuracy (Acc) is the intuitive measurement of the performance on a balance dataset whereas Matthew's correlation coefficient (MCC) [46] is more realistic than Acc in measuring performance when using an unbalanced dataset [47, 48]. When both MCC and Acc values are high, the overall performance of the predicted model is better. In addition to Acc and MCC, the following parameters below were also calculated. Sensitivity is the percentage of correctly predicted binding proteins to the total binding proteins.

True positive (TP).

True negative (TN).

False positive (FP) (false alarm).

False negative (FN).

False positive rate (FPR).

Sensitivity/recall or True positive rate (TPR) TPR = TP/P = TP/(TP + FN).

Precision = TP/(TP + FP).

Accuracy (Acc) = (TP + TN)/(P + N) = (TP + TN)/(TP + TN + FP + FN).

Specificity (SPC) SPC = TN/N = TN/(FP + TN) = 1 – FPR.

Matthew's correlation coefficient (MCC).

((TP × TN) − (FP × FN))/[sqrt ((TN + FN) × (TN + FP) × (TP + FN) × (TP + FP))] OR

$$\mathrm{MCC} = \frac{(\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN})}{\sqrt{\mathrm{PNP'N'}}}. \quad (1)$$

Here TP is the number of true positives (ATP-BPs), TN is the number of true negatives (non ATP-BPs), FP is the number of false positives, and FN is the number of false negatives.

2.5. Area under the ROC Curve (AUC) for LIBSVM. It is a plot between true positive proportion (TP/TP + FN) and false positive proportion (FP/FP + TN). The StatsDirect was used package to plot ROC and calculates the area under the ROC curve directly by an extended trapezoidal rule [49]. The confidence interval was constructed using DeLong's variance estimate [50] embedded in the statistic package.

3. Results and Discussion

The ATP binding proteins are known to play key roles in the biochemical functioning of the cell. In signaling pathways ATP molecules are substrates for protein kinase phosphorylation. It is difficult to identify ATP binding proteins due to lack of experimentally determined protein structures [51–53]. This is because the growth of protein sequences from various genomic projects exceeds the capacity of experimental techniques in determining protein structures and their binding reactions which are time consuming and at times unsuccessful. Therefore there is an urgent need to develop automated expert methods for determining the functional class of proteins such ATP binding proteins from their primary sequence information.

The general assumption here is that every protein that binds to ATP molecule either USPs or those having Walker A motif will have some common features embedded in their sequences. In both the USP (G-2X-G-9X-G(S/T)) and Walker A (G-4X-GK (T/S)) motifs, the G, K, T, and S denote glycine, lysine, threonine, and serine, respectively, and X denotes any

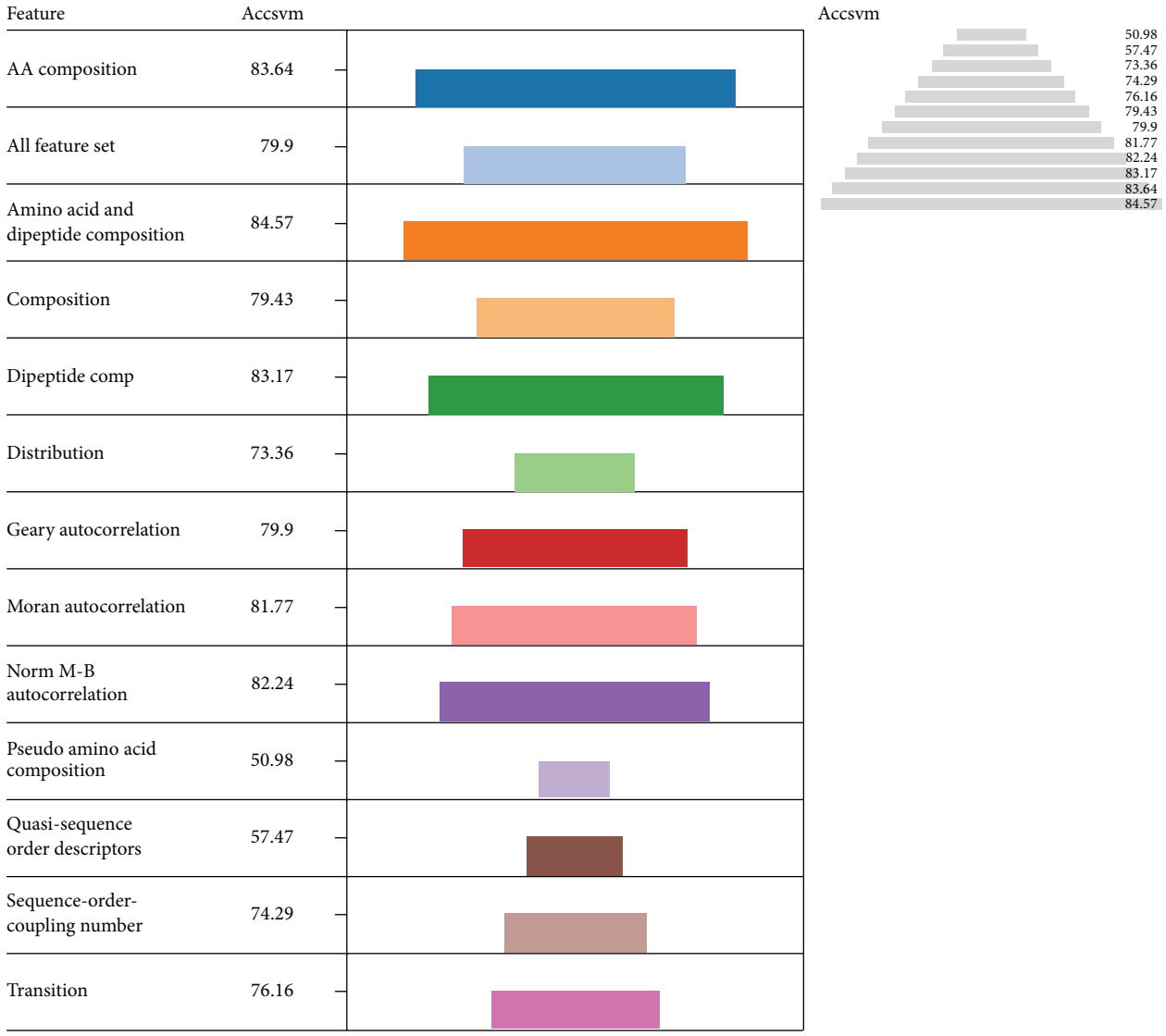


Figure 1: The performances of descriptors with LIBSVM in terms of accuracy. The length of each color coded descriptor and the pyramidal view are a measure of their performances in terms of accuracy (Accsvm). In terms of accuracy the best descriptor was combination of amino acid and dipeptide composition (84.57%), followed by amino acid composition (83.64%), dipeptide composition (83.17%), and Norm M-B autocorrelation in that order. The pseudo amino acids and Quasi sequence order descriptors perform poorly.

amino acid residue. The lysine (K) residue in the Walker A motif is crucial for nucleotide binding [54] in this class of proteins. It interacts with the phosphate groups of the nucleotide and with the magnesium ion, which coordinates the β- and γ-phosphates of the ATP molecule [55, 56].

The universal stress proteins bind to ATP through the ATP binding motif G-2X-G-9X-G(S/T), with the -G(S)/T as essential residues for ATP binding and phosphorylation [13]. Therefore, members of this class of proteins will segregate into two groups, based on whether or not they bind to ATP [13, 57]. Thus, it is important to identify ATP binding USPs and other ATP binding proteins. Several methods have been studied based on predicting ATP interacting residues if the protein structures are known, with some results showing very low accuracies [15, 16, 58, 59]. This work has predicted ATP binding proteins in general with high accuracy irrespective of their structural information using SVM classifier. The training and prediction statistics for each of the descriptor sets used were visualized and discussed below. The visualizations were constructed using Tableau Public Software (http://www.tableausoftware.com/public).

The objective in this report was to find the best descriptor set which can be use to build a predictive model for a reliable and effective server for predicting ATP-BPs in general, irrespective of their subfunctional classes. Throughout this work, the parameter C was maintained at $C = 4$, while the optimal value of γ for each descriptor was obtained and used in evaluating their performances. Their performances

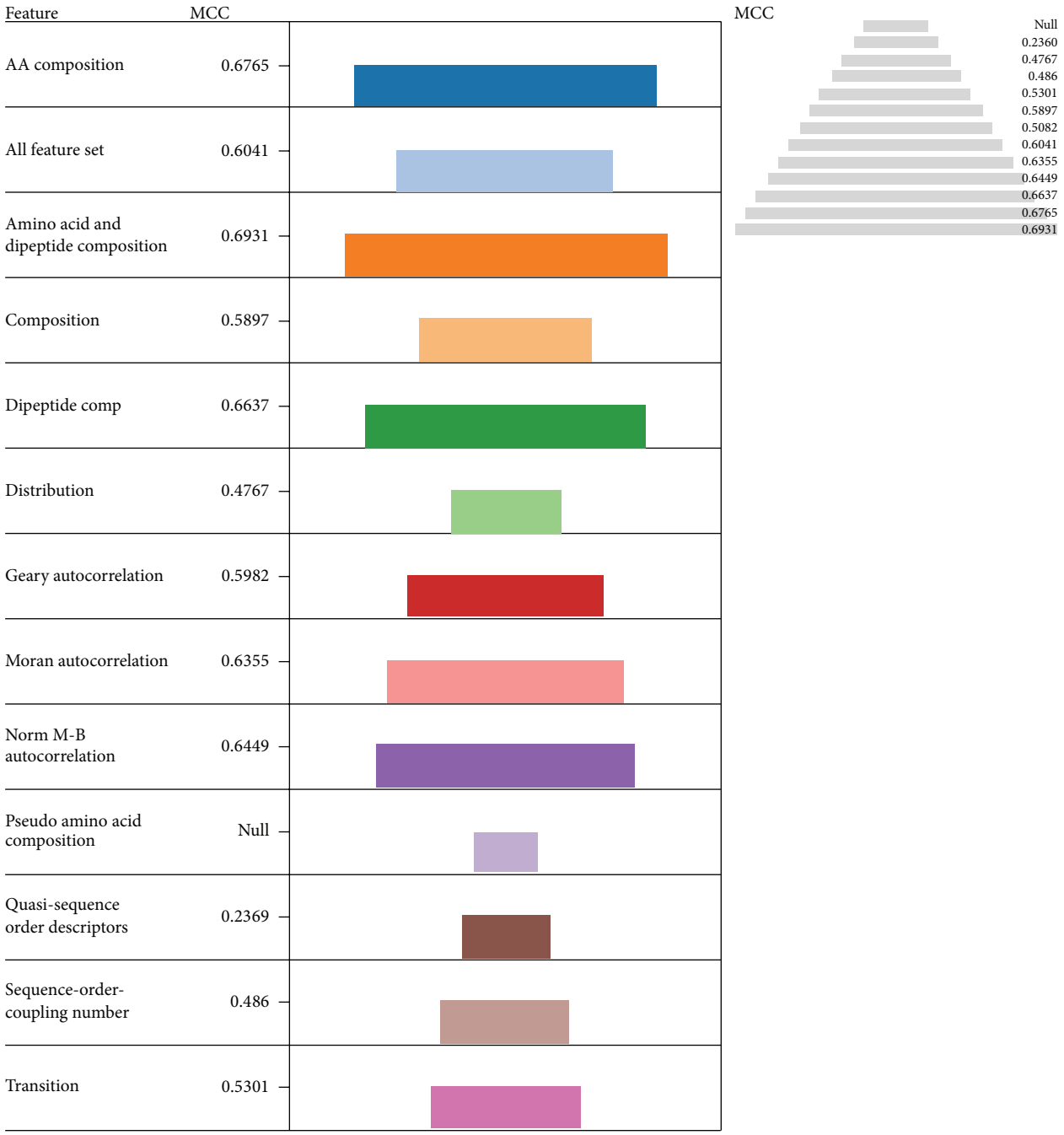


Figure 2: The performances of descriptors with LIBSVM in terms of Mathew's correlation coefficient (MCC). The length of each color coded descriptor and the pyramidal view are a measure of their performances in terms of MCC. The best performer was amino acid and dipeptide composition in combination (0.6931) followed by amino acid composition (0.6765), dipeptide composition (0.6637), and Norm M-B autocorrelation (0.6449) in that order.

were evaluated based on five computed parameters consisting of their accuracies, sensitivities, specificities, precisions, and MCC, after a 10-fold cross validation (CV10).

The performance of pseudo amino acid composition was evaluated with only accuracy due to lack of sufficient sequence information. The lengths of the color coded descriptors were used as a measure of their performances. In terms of accuracy the best descriptor was the combination of amino acid with dipeptide composition (84.57%), followed by amino acid composition alone (83.64%), dipeptide composition (83.17%), and Norm M-B autocorrelation in that order (Figure 1). The pseudo amino acids and Quasi sequence order descriptors performed poorly compared to the other descriptors. However, the overall performances of the other

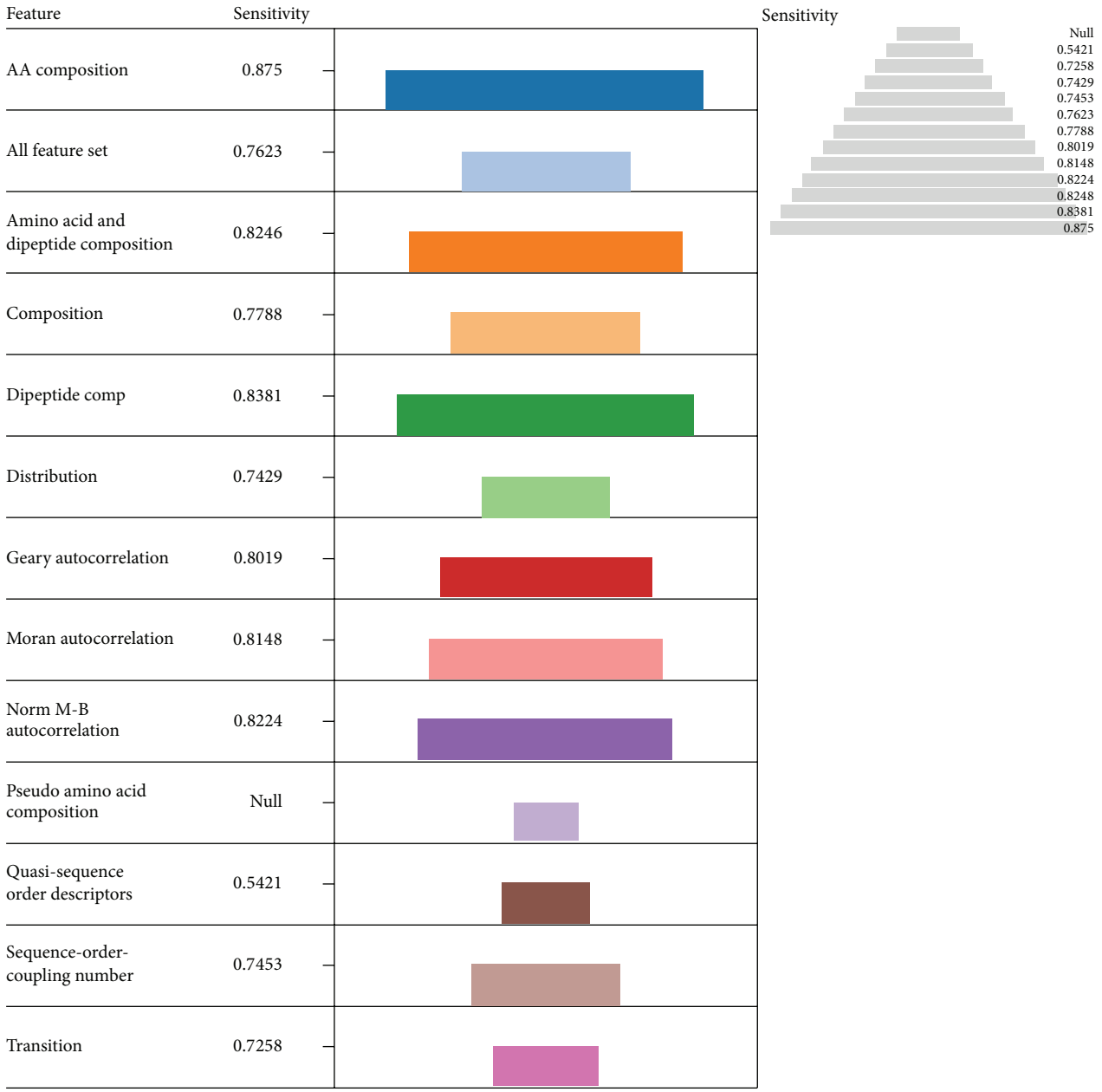


Figure 3: The performances of descriptors with LIBSVM in terms of sensitivity. The length of each color coded descriptor and the pyramidal view are a measure of their performances in terms of sensitivity. The most sensitive descriptor was amino acid composition (0.875) followed by dipeptide composition (0.8381), amino acid and dipeptide composition in combination (0.8246), and Norm M-B autocorrelation (0.8224) in that order.

descriptors were better as most of them registered accuracy values greater than 70.00%. These high performers might be due to the rigorous refinement of protein sequences. Thus protein function classification with SVM classifiers can be improved drastically using rigorously refined protein sequences.

The individual performances of amino acid composition (83.64%) and dipeptide composition (83.17%) were increased to 84.57% when both descriptors were combined together. This indicates that the combination of descriptors can enhance the individual performance of other descriptors, particularly those combining with amino acid composition. This is a binary classification problem involving a balance dataset and accuracy (Acc) is the best parameter for evaluating performance based on balance dataset whereas Matthew's correlation coefficient (MCC) is more realistic than Acc when using an unbalanced dataset [47, 48]. But when both MCC and Acc values are high, the overall performance of the predicted model is better.

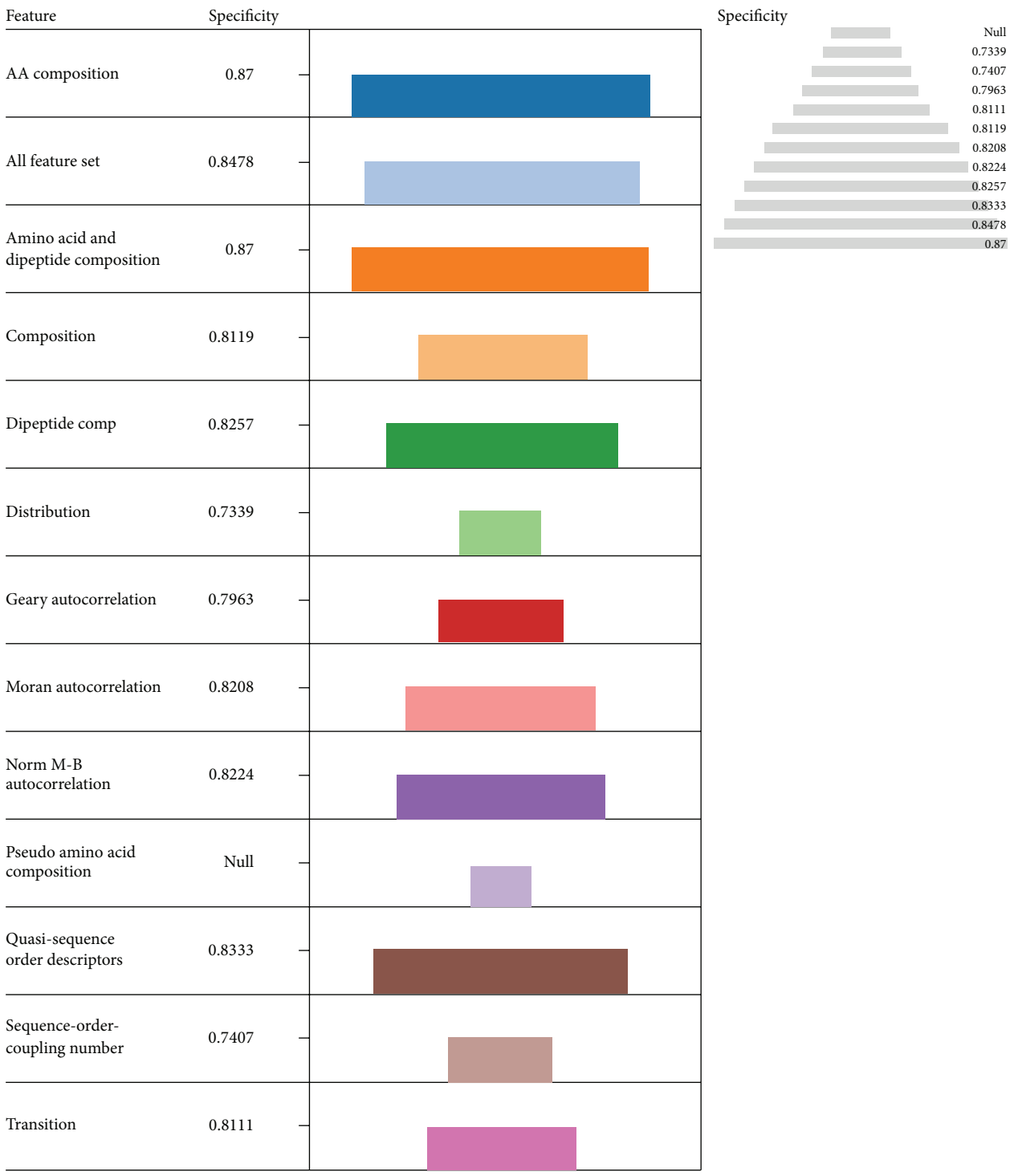


Figure 4: The performances of descriptors with LIBSVM in terms of Specificity. The length of each color coded descriptor and the pyramidal view are a measure of their performances in terms of specificity. The most specific descriptor was amino acid composition and amino acid/dipeptide composition (0.87) followed by all using all the feature set (0.8478), Quasi sequence order descriptors (0.8333), and dipeptide composition (0.8257) in that order.

The performances of the models were evaluated based on MCC (Figure 2). The pyramidal view and the length of the color coded descriptors were used for performance visualization. The best performer was amino acid and dipeptide composition in combination (0.6931) followed by amino acid composition (0.6765), dipeptide composition (0.6637), and Norm M-B autocorrelation (0.6449) in that order. This order is in line with their performances measured using accuracy

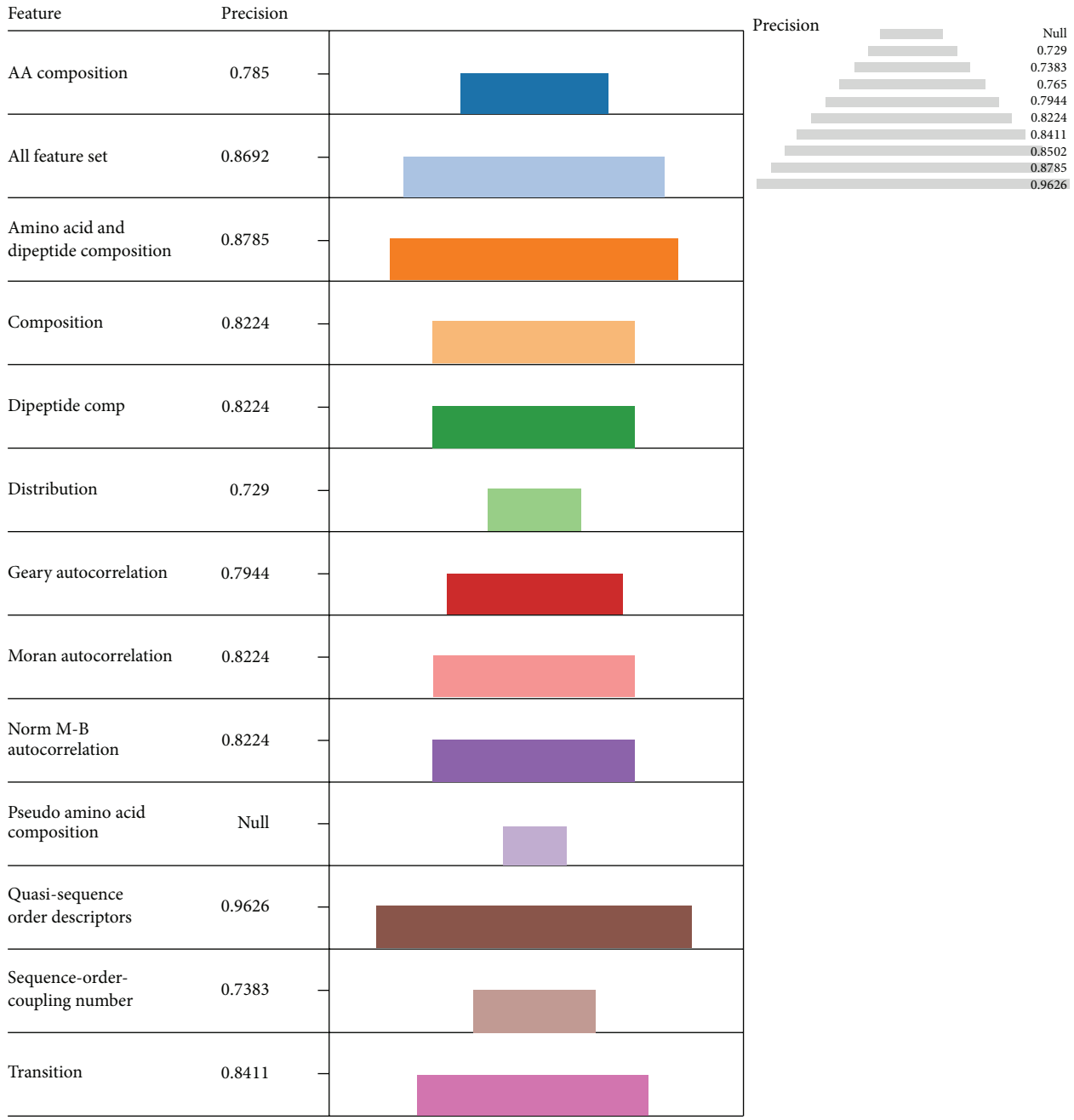


Figure 5: The performances of descriptors with LIBSVM in terms of Precision. The length of each color coded descriptor and the pyramidal view are a measure of their performances in terms of precision. The most precise descriptor was Quasi sequence order descriptors (0.9626) followed by amino acid and dipeptide composition in combination (0.8785), all feature set (0.8692) and Transition (0.8411) in that order.

as the parameter. This result justifies the performance of the overall model. In general the combination of descriptor sets performs better than individual descriptors, particularly when combined with amino acid composition.

Therefore from the statistical point of view the use of combination sets particularly with amino acid composition tend to give better prediction performance than individual-sets [53]. The amino acid composition generally increases the overall accuracies of other descriptors in combination. One of the shortcoming of amino acid composition as a descriptor is that the same amino acid composition may correspond to diverse sequences due to the loss of sequence order [28, 60]. This sequence order information can be partially covered by combination with dipeptide composition, but dipeptide composition itself lacks information on the fraction of the individual residue in the sequence, as such a combination set is expected to give a better prediction result [27, 61] as shown above due to masking effect.

The models were further investigated based on their sensitivity to predict ATP-BPs and the results displayed in

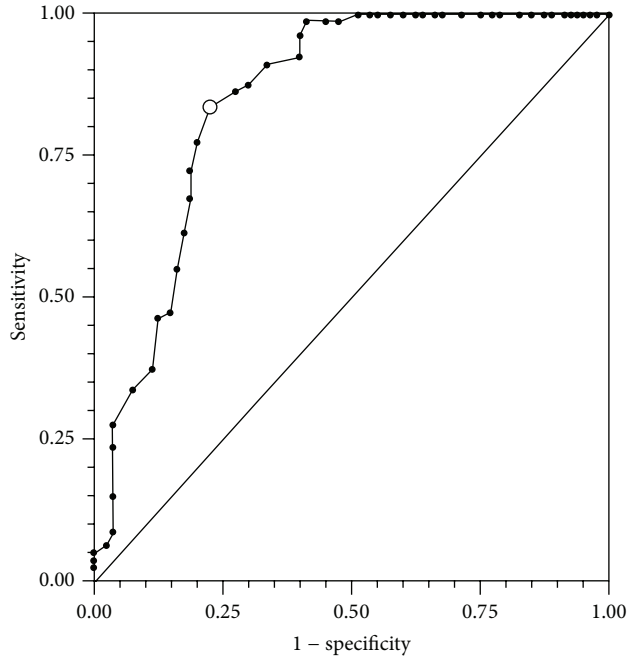


Figure 6: The ROC plot: the plot shows the performance of the LIBSVM model generated with StatsDirect package using an extended trapezoidal rule and a nonparametric method analogous to the Wilcoxon/Mann-Whitney test to calculate the area under the ROC curve. The calculated AUA was 0.849219.

pyramidal view (Figure 3). The most sensitive descriptor was amino acid composition (0.875) followed by dipeptide composition (0.8381), amino acid/dipeptide composition in combination (0.8246), and Norm M-B autocorrelation (0.8224) in that order.

These descriptors were among the best four performers in terms of Acc and MCC. Evaluation based on specificity indicates that amino acid composition (0.87) was more specific followed by using the entire feature set (0.8478), Quasi sequence order descriptors (0.8333), and dipeptide composition (0.8257) in that order (Figure 4). This information highlights the vital role played by amino acid composition in protein function predictions in general. Interestingly the Quasi sequence order descriptors (0.9626) had the highest precision followed by amino acid and dipeptide composition in combination (0.8785), entire feature set (0.8692), and Transition (0.8411) in that order (Figure 5).

The overall model evaluation shows that the amino acids and dipeptide composition was the best model for predicting ATP-BPs from diverse functional classes using whole sequence information. The use of "all the descriptor" set did not generally result in a better model in classification. The "all features" descriptor accuracy was 79.9% against 84.57% for amino acids/dipeptide in combination. This finding is in accordance with [62, 63], on their work on molecular descriptors for predicting compounds of specific properties using "all features" set. The reduction in accuracy might be due to noise generated by the use of many overlapping and redundant descriptors. Hence the accuracy of the classifier algorithms can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance in solving the classification problem in question. The performance of the SVM model using ROC plot (Figure 6) has a value of AUC of 0.849219. This highlights a better model based on whole sequence analysis.

4. Conclusions

The prediction of ATP-binding proteins has been exploited using a battery of descriptor sets and a hybrid functional group. Also for the first time the prediction of ATP binding in universal stress proteins had been investigated using the support vector machine. The best hybrid model was the combination of amino acid and dipeptide composition of the sequences with an accuracy of 84.57% and Mathews correlation coefficient (MCC) value of 0.693. The general trend is that combination of descriptors will perform better and improve the overall performances of individual descriptors, particularly when combined with amino acid composition. This model provides a high probability of success for molecular biologists in predicting and selecting diverse groups of ATP binding proteins.

Conflict of Interests

The author reports no conflict of interests in this work including the mentioned trademarks.

Acknowledgments

The research reported was supported by the National Institutes of Health (NIH-NIGMS-1T36GM095335) and the National Science Foundation (EPS-0903787; EPS-1006883). The content is solely the responsibility of the author and does not necessarily represent the official views of the funding agencies.

References

[1] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 45–48, 2000.

[2] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[3] J. Guo, H. Chen, Z. Sun, and Y. Lin, "A novel method for protein secondary structure prediction using dual-layer SVM and profiles," *Proteins*, vol. 54, no. 4, pp. 738–743, 2004.

[4] C. Bustamante, Y. R. Chemla, N. R. Forde, and D. Izhaky, "Mechanical processes in biochemistry," *Annual Review of Biochemistry*, vol. 73, pp. 705–748, 2004.

[5] J. E. Walker, M. Saraste, M. J. Runswick, and N. J. Gay, "Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold," *The EMBO Journal*, vol. 1, no. 8, pp. 945–951, 1982.

[6] N. Hirokawa and R. Takemura, "Biochemical and molecular characterization of diseases linked to motor proteins," *Trends in Biochemical Sciences*, vol. 28, no. 10, pp. 558–565, 2003.

[7] C. Gedeon, J. Behravan, G. Koren, and M. Piquette-Miller, "Transport of glyburide by placental ABC transporters: implications in fetal drug exposure," *Placenta*, vol. 27, no. 11-12, pp. 1096–1102, 2006.

[8] A. Maxwell and D. M. Lawson, "The ATP-binding site of type II topoisomerases as a target for antibacterial drugs," *Current Topics in Medicinal Chemistry*, vol. 3, no. 3, pp. 283–303, 2003.

[9] H. Ashida, T. Oonishi, and N. Uyesaka, "Kinetic analysis of the mechanism of action of the multidrug transporter," *Journal of Theoretical Biology*, vol. 195, no. 2, pp. 219–232, 1998.

[10] K. Kvint, L. Nachin, A. Diez, and T. Nystrom, "The bacterial universal stress protein: function and regulation," *Current Opinion in Microbiology*, vol. 6, no. 2, pp. 140–145, 2003.

[11] T. Nystrom and F. C. Neidhardt, "Cloning, mapping and nucleotide sequencing of a gene encoding a universal stress protein in *Escherichia coli*," *Molecular Microbiology*, vol. 6, no. 21, pp. 3187–3198, 1992.

[12] A. Diez, N. Gustavsson, and T. Nystrom, "The universal stress protein a of *Escherichia coli* is required for resistance to DNA damaging agents and is regulated by a RecA/FtsK-dependent regulatory pathway," *Molecular Microbiology*, vol. 36, no. 6, pp. 1494–1503, 2000.

[13] M. C. Sousa and D. B. Mckay, "Structure of the universal stress protein of *Haemophilus influenzae*," *Structure*, vol. 9, no. 12, pp. 1135–1141, 2001.

[14] V. J. Promponas, C. A. Ouzounis, and I. Iliopoulos, "Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey," *Briefings in Bioinformatics*, 2012.

[15] J. S. Chauhan, N. K. Mishra, and G. P. Raghava, "Identification of ATP binding residues of a protein from its primary sequence," *BMC Bioinformatics*, vol. 10, article 434, 2009.

[16] T. Guo, Y. Shi, and Z. Sun, "A novel statistical ligand-binding site predictor: application to ATP-binding sites," *Protein Engineering, Design and Selection*, vol. 18, no. 2, pp. 65–70, 2005.

[17] K. Chen, M. J. Mizianty, and L. Kurgan, "ATPsite: sequence-based prediction of ATP-binding residues," *Proteome Science*, vol. 9, article S4, supplement 1, 2011.

[18] Y. N. Zhang, D. J. Yu, S. S. Li, Y. X. Fan, Y. Huang, and H. B. Shen, "Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features," *BMC Bioinformatics*, vol. 13, article 118, 2012.

[19] J. R. Green, M. J. Korenberg, R. David, and I. W. Hunter, "Recognition of adenosine triphosphate binding sites using parallel cascade system identification," *Annals of Biomedical Engineering*, vol. 31, no. 4, pp. 462–470, 2003.

[20] A. Garg, M. Bhasin, and G. P. S. Raghava, "Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search," *The Journal of Biological Chemistry*, vol. 280, no. 15, pp. 14427–14432, 2005.

[21] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, no. 4, pp. 477–486, 2004.

[22] X. Xiao, P. Wang, and K.-C. Chou, "GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes," *Journal of Computational Chemistry*, vol. 30, no. 9, pp. 1414–1423, 2009.

[23] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "Prediction of RNA binding sites in a protein using SVM and PSSM profile," *Proteins*, vol. 71, no. 1, pp. 189–194, 2008.

[24] B. S. Williams, R. D. Isokpehi, A. N. Mbah et al., "Functional annotation analytics of bacillus genomes reveals stress responsive acetate utilization and sulfate uptake in the biotechnologically relevant bacillus megaterium," *Bioinformatics and Biology Insights*, vol. 6, pp. 275–286, 2012.

[25] R. D. Isokpehi, O. Mahmud, A. N. Mbah et al., "Developmental regulation of genes encoding universal stress proteins in *Schistosoma mansoni*," *Gene Regulation and Systems Biology*, vol. 5, pp. 61–74, 2011.

[26] A. N. Mbah, O. Mahmud, O. R. Awofolu, and R. D. Isokpehi, "Inferences on the biochemical and environmental regulation of universal stress proteins from Schistosomiasis parasites," *Advances and Applications in Bioinformatics and Chemistry*, vol. 6, pp. 15–27, 2013.

[27] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, no. 3, pp. 282–283, 2001.

[28] G. Wang and R. L. Dunbrack Jr., "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[29] X. Yu, J. Cao, Y. Cai, T. Shi, and Y. Li, "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines," *Journal of Theoretical Biology*, vol. 240, no. 2, pp. 175–184, 2006.

[30] A. Marchler-Bauer, C. Zheng, F. Chitsaz et al., "CDD: conserved domains and protein three-dimensional structure," *Nucleic Acids Research*, vol. 41, pp. D348–D352, 2013.

[31] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen, "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Research*, vol. 34, pp. W32–W37, 2006.

[32] Z. Bikadi, I. Hazai, D. Malik et al., "Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein," *PLoS ONE*, vol. 6, no. 10, Article ID e25815, 2011.

[33] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005.

[34] M. P. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[35] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[36] K.-C. Chou and Y.-D. Cai, "Predicting protein-protein interactions from sequences in a hybridization space," *Journal of Proteome Research*, vol. 5, no. 2, pp. 316–322, 2006.

[37] M. E. Matheny, F. S. Resnic, N. Arora, and L. Ohno-Machado, "Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality," *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 688–697, 2007.

[38] F. Javed, G. S. Chan, A. V. Savkin et al., "RBF kernel based support vector regression to estimate the blood volume and heart rate responses during hemodialysis," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '09)*, pp. 4352–4355, 2009.

[39] C.-C. Chang and C.-J. Lin, "Training nu-support vector classifiers: theory and algorithms," *Neural Computation*, vol. 13, no. 9, pp. 2119–2147, 2001.

[40] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113–126, 2004.

[41] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, pp. 275–349, 1995.

[42] C. Chen, L. Chen, X. Zou, and P. Cai, "Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine," *Protein and Peptide Letters*, vol. 16, no. 1, pp. 27–31, 2009.

[43] H. Ding, L. Luo, and H. Lin, "Prediction of cell wall lytic enzymes using chou's amphiphilic pseudo amino acid composition," *Protein and Peptide Letters*, vol. 16, no. 4, pp. 351–355, 2009.

[44] J. Bondia, C. Tarin, W. Garcia-Gabin et al., "Using support vector machines to detect therapeutically incorrect measurements by the MiniMed CGMS," *Journal of Diabetes Science and Technology*, vol. 2, pp. 622–629, 2008.

[45] S. Chen, S. Zhou, F.-F. Yin, L. B. Marks, and S. K. Das, "Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis," *Medical Physics*, vol. 34, no. 10, pp. 3808–3814, 2007.

[46] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[47] L. Bao and Y. Cui, "Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information," *Bioinformatics*, vol. 21, no. 10, pp. 2185–2190, 2005.

[48] R. J. Dobson, P. B. Munroe, M. J. Caulfield, and M. A. S. Saqi, "Predicting deleterious nsSNPs: an analysis of sequence and structural attributes," *BMC Bioinformatics*, vol. 7, article 217, 2006.

[49] J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[50] E. R. Delong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[51] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins," *The EMBO Journal*, vol. 5, no. 4, pp. 823–826, 1986.

[52] A. M. Lesk and C. Chothia, "How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins," *Journal of Molecular Biology*, vol. 136, no. 3, pp. 225–270, 1980.

[53] M. Hilbert, G. Bohm, and R. Jaenicke, "Structural relationships of homologous proteins as a fundamental principle in homology modeling," *Proteins*, vol. 17, no. 2, pp. 138–151, 1993.

[54] P. I. Hanson and S. W. Whiteheart, "AAA+ proteins: have engine, will work," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 7, pp. 519–529, 2005.

[55] K. M. Ferguson, T. Higashijima, M. D. Smigel, and A. G. Gilman, "The influence of bound GDP on the kinetics of guanine nucleotide binding to G proteins," *The Journal of Biological Chemistry*, vol. 261, no. 16, pp. 7393–7399, 1986.

[56] F. Jurnak, A. Mcpherson, A. H. J. Wang, and A. Rich, "Biochemical and structural studies of the tetragonal crystalline modification of the *Escherichia coli* elongation factor Tu," *The Journal of Biological Chemistry*, vol. 255, no. 14, pp. 6751–6757, 1980.

[57] T. I. Zarembinski, L. I.-W. Hung, H.-J. Mueller-Dieckmann et al., "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 26, pp. 15189–15193, 1998.

[58] M. Saito, M. Go, and T. Shirai, "An empirical approach for detecting nucleotide-binding sites on proteins," *Protein Engineering, Design and Selection*, vol. 19, no. 2, pp. 67–75, 2006.

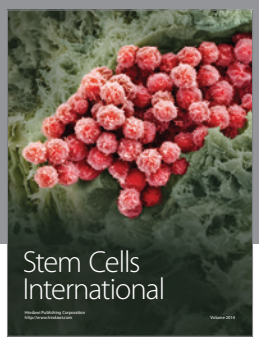
[59] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman, "Automated analysis of interatomic contacts in proteins," *Bioinformatics*, vol. 15, no. 4, pp. 327–332, 1999.

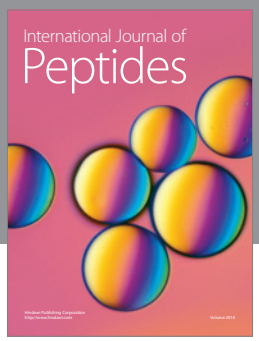
[60] R. E. Schapire and Y. Singer, "Boostexter: a boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

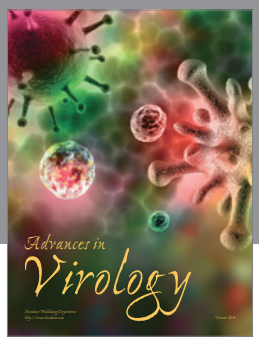
[61] S. A. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, "Efficacy of different protein descriptors in predicting protein functional families," *BMC Bioinformatics*, vol. 8, article 300, 2007.

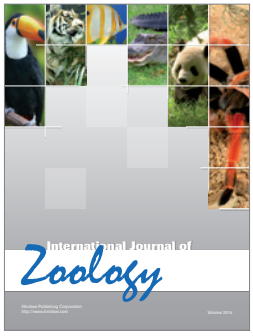
[62] L. Xue and J. Bajorath, "Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening," *Combinatorial Chemistry and High Throughput Screening*, vol. 3, no. 5, pp. 363–372, 2000.

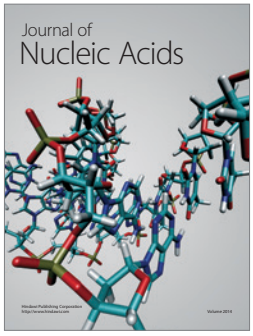
[63] L. Xue, J. W. Godden, and J. Bajorath, "Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1227–1234, 2000.

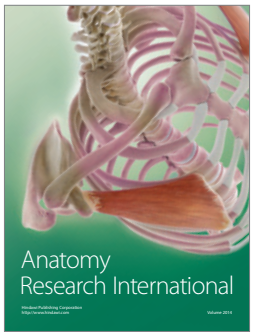


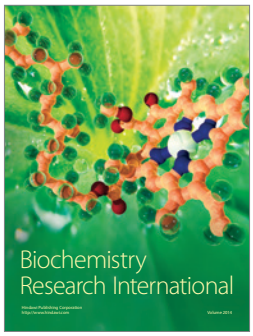


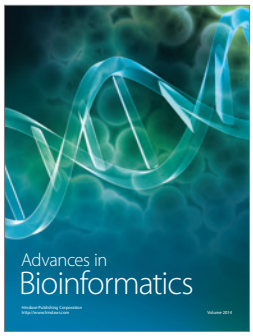


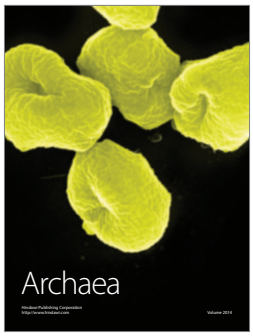


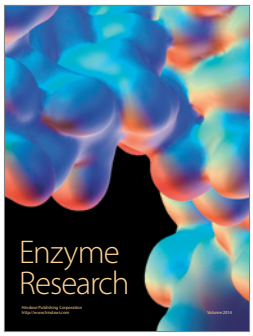


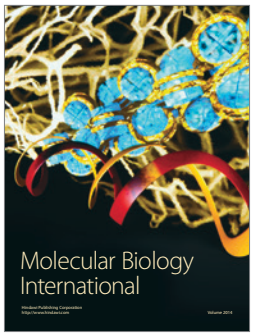












Hindawi

Submit your manuscripts at
http://www.hindawi.com