Hindawi Publishing Corporation
Scientifica
Volume 2016, Article ID 8309253, 10 pages
http://dx.doi.org/10.1155/2016/8309253

*Research Article*

# Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest

**Zhijun Liao,[1,2] Ying Ju,[3] and Quan Zou[2,4]**

[1] *School of Basic Medical Sciences, Fujian Medical University, Fuzhou, Fujian 350108, China*
[2] *School of Computer Science and Technology, Tianjin University, Tianjin 300350, China*
[3] *School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China*
[4] *State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, China*

Correspondence should be addressed to Quan Zou; zouquan@tju.edu.cn

Received 26 May 2016; Revised 26 June 2016; Accepted 30 June 2016

Academic Editor: Wei Chen

G protein-coupled receptors (GPCRs) are the largest receptor superfamily. In this paper, we try to employ physical-chemical properties, which come from SVM-Prot, to represent GPCR. Random Forest was utilized as classifier for distinguishing them from other protein sequences. MEME suite was used to detect the most significant 10 conserved motifs of human GPCRs. In the testing datasets, the average accuracy was 91.61%, and the average AUC was 0.9282. MEME discovery analysis showed that many motifs aggregated in the seven hydrophobic helices transmembrane regions adapt to the characteristic of GPCRs. All of the above indicate that our machine-learning method can successfully distinguish GPCRs from non-GPCRs.

## 1. Introduction

The G protein-coupled receptors (GPCRs) are only discovered in eukaryotes, which constitute a vast protein family and perform their various functions always through coupling with G proteins in the cell. GPCRs have many aliases such as heptahelical receptors, serpentine receptor, G protein-linked receptors (GPLR), and seven-transmembrane (7TM) domain receptors; all the GPCRs contain a single polypeptide chain that pass through the cell membrane seven times [1]. There are roughly 1000 GPCRs in human genome (accounting for about 2% coding genes); thus, they form the largest receptor superfamily [2]; they are also involved in various diseases and constituted approximately 40% of drug targets. Because Robert J. Lefkowitz and Brian K. Kobilka revealed the biochemical mechanism of GPCRs for signaling pathways, they were awarded with 2012 Nobel Prize in chemistry [3].

Many different approaches have been utilized for GPCRs classification, such as protein motif-based systems, machine-learning methods [4], and other techniques. Based on the original sequence similarity and phylogenetic studies, GPCRs superfamily can be divided into five, six, or seven classes at different periods [5, 6]. According to GPCRdb (http://gpcrdb.org/) database developed by Kolakowski and updated by Horn et al. [7], which contains data, diagrams, and web tools involving collection of both GPCRs crystal structures and receptor mutants, GPCRs are classified into six main families: class A (Rhodopsin), class B1 (Secretin), class B2 (Adhesion), class C (Glutamate), class F (Frizzled), and other GPCRs. The former five classes are consistent with the Glutamate, Rhodopsin, Adhesion, Frizzled, and Secretin (GRAFS in short) classification system [8, 9]. Table 1 shows the protein number and composition for every class.

Class A rhodopsin-like receptors constitute the largest (more than 80%) of the human GPCR subtypes. They mediate numerous effects of endogenous peptides including neurotransmitters, hormones, and paracrine signals. For example, biogenic amines [10] such as norepinephrine, dopamine, and serotonin commonly play their role of drugs for pathological diseases through binding to GPCRs. Although the N-terminal extracellular domain is very short, class A receptors can form dimers, in homo/heterodimerization [11]. This class also includes approximately 60 orphan receptors which have no defined ligands or functions at all [12, 13].

Class B1 secretin-like receptors belong to one of hormone and neuropeptide receptor families; they consist of a large

Table 1: The number of proteins and composition for every class of GPCRs (from GPCRdb).

| GPCRdb family | Number of proteins (human) | Composition |
|---|---|---|
| Class A (rhodopsin) | 16526 (311) | Aminergic receptors, peptide receptors, protein receptors, lipid receptors, melatonin receptors, nucleotide receptors, steroid receptors, alicarboxylic receptors, sensory receptors, orphan receptors, and others |
| Class B1 (secretin) | 748 (15) | Peptide receptors |
| Class B2 (adhesion) | 381 (33) | Orphan receptors |
| Class C (glutamate) | 1038 (22) | Ion receptors, amino acid receptors, sensory receptors, and orphan receptors |
| Class F (frizzled) | 48 (11) | Peptide receptors |
| Other GPCRs | 37 (6) | Orphan receptors |

and versatile N-terminal extracellular domain (ECD) which functions as an affinity trap to hormone [14]. Moreover, they are of ancient origin and can bind with various peptides such as secretin, corticotrophin releasing factor, glucagon, parathyroid hormone, calcitonin, growth hormone releasing hormone, and calcitonin gene-related peptide [15].

Class B2 adhesion-like receptors are also known as the adhesion G protein-coupled receptors (ADGRs) with ancient origin; they make the function in various tissues include synapses of the brain [16]. Most ADGRs contain various domains in the N-terminus provided for binding site of other cells [17]; these domains have over sixteen types, including cadherin-like repeats, thrombospondin-like repeats, and calnexin domain. ADGRs have the characteristic of N-terminal adhesive domains [18]. For example, ADGR subfamily G4 (ADGRG4) has the sequence characteristics of a unique highly conserved motif and some functionally important motifs similar to class A, class B1, and combined elements [19].

Class C GPCRs mainly comprise metabotropic glutamate receptors (mGluRs), one type of L-glutamate binding receptors; another type is ionotropic glutamate receptors (iGluRs) which belong to a ligand-gated ion channels not the GPCR family. Class C GPCRs contain a large N-terminal domain for ligand-binding. There exist 8 isoforms of mGluRs to form signaling molecules via second messenger systems [20], which transfer extracellular signal through the mechanism of receptor dimer packing and allosteric regulation [21]. The activation of mGluRs is an indirect metabotropic process by the aid of binding to glutamate, a major excitatory neurotransmitter in the brain. The extracellular glutamate concentration (at micromolar range) is lower than the intracellular (at millimolar range) in neuron [22]. Human mGluRs are found in pre- and postsynaptic neurons, including the hippocampus, cerebellum, and other brain regions' synapses, and in peripheral tissues. mGluRs play an important role in regulating neuronal excitability and synaptic plasticity and in serving as mental disorders drug targets [23].

Class F frizzled/smoothened receptors are involved in Wnt binding whereas the smoothened receptor (belongs to GPCRs) reconciles hedgehog signaling via the required region cysteine-rich domain (CRD) in the N-terminus [24], because smoothened protein sequence is homologous to frizzled. The two proteins have the same 7TM structure and evolutionary relationship [25]. But the secreted frizzled-related proteins can exert its function by promoting or blocking Wnt3$\alpha/\beta$-catenin signaling in different concentration of secreted frizzled-related protein 1 and cellular context [26].

Other GPCRs include some orphan receptors except for the above classes; the characteristics of these receptors are that they have a similar structure to other identified receptors but lack endogenous ligand. They have altogether 37 proteins and 6 in human. Among them, Gpr175 (also called Tpra1) and GPR157 are well studied. Gpr175 is an orphan GPCR with positive regulation of the Hedgehog signaling pathway [27]; GPR157 couples with Gq protein and then activate $IP_3$-mediated $Ca^{2+}$ cascade, which is also a signaling molecule involved in positive regulation of neuronal differentiation of radial glial progenitors through the GPR157-Gq-$IP_3$ cascade pathway [28].

Generally, GPCRs interact with a varieties of ligands which can be classified as agonists, antagonists, or inverse agonists, three classes based on the receptor effect [29, 30]; these include different forms of "information," such as photons, taste, odorants [31], ions, pheromones, eicosanoids, nucleotides, nucleosides [9], neurotransmitters, amino acids [32], peptides, proteins, and hormones [33]. These ligands vary in size containing small molecules and large proteins.

GPCRs are transmembrane receptors that transduce extracellular stimuli into intracellular signals through activating intracellular heterotrimeric G protein complex, which comprise 15 G$\alpha$ subunits, 5 G$\beta$ subunits, and 12 G$\gamma$ subunits. Based on the sequence similarity and functional characteristics of G$\alpha$ subunits, G proteins are divided into four major classes: G$\alpha$s, G$\alpha$i/o, G$\alpha$q/11, and G$\alpha$12/13 [34]. G$\alpha$ activation or deactivation cycle controls the signal transduction, when cell is at resting mode, GDP binds to G$\alpha$ forming G$\alpha$-GDP and then joins G$\beta\gamma$ generating G$\alpha\beta\gamma$ complex, and G$\alpha$ is inactive at this stage; when stimulate signal is introduced from GPCR, G$\alpha$ raises a conformational change, GTP binds to G$\alpha$ forming G$\alpha$-GTP and destabilizing the G$\alpha\beta\gamma$ complex, G$\beta\gamma$ are disassociated and bound by G$\beta\gamma$ interacting proteins, and G$\alpha$ is active at this stage. When G$\alpha$ fulfilled signal transduction to the downstream pathway, G$\alpha$ hydrolyzes GTP to GDP through its intrinsic GTPase activity to form G$\alpha$-GDP and returns to the resting mode;

Table 2: The composition of 188D features of a protein.

| Physicochemical property | Dimensions |
|---|---|
| Amino acid composition | 20 |
| Hydrophobicity | 21 |
| Normalized Van der Waals volume | 21 |
| Polarity | 21 |
| Polarizability | 21 |
| Charge | 21 |
| Surface tension | 21 |
| Secondary structure | 21 |
| Solvent accessibility | 21 |
| Total | 188 |

this process constitutes a G protein cycle [35]. Activated G$\alpha$s catalyzes ATP to cAMP by adenylyl cyclase (AC) and results in the activation of protein kinase A (PKA) and phosphorylation of downstream effector. On the contrary, G$\alpha$i plays inhibition role of AC and suppresses cAMP production. G$\alpha$q/11 activates phospholipase C$\beta$ (PLC$\beta$) and produces inositol-1,4,5-trisphosphate ($IP_3$) and diacylglycerol (DAG) which can form PLC$\beta$-$IP_3$-DAG signaling pathway. G$\alpha$12/13 activates Rho GTPase families through RhoGEF to regulate cytoskeleton remodeling; these G protein families take the major effect in signal transduction [3]. Therefore, GPCR-G$\alpha$-AC-PKA and GPCR-G$\alpha$-PLC-$IP_3$ constitute two main signal transduction cascades within the cell.

In this paper, we performed an *in silico* analysis on the GPCRs amino acids information and other polypeptide physicochemical features and constructed 188D feature vectors (Table 2) of the proteins into an ensemble classifier [36–41]. The first 20D of 188D represents the 20 kinds of natural amino acids composition; the other 168D includes eight physical-chemical properties each deriving from the so-call CTD mode [42], where C stands for amino acid contents for each type of hydrophobic amino acids, T stands for the frequency of bivalent peptide, and D stands for amino acid distribution from five positions of a sequence. These 188D feature vectors have been integrated into software BinMemPredict which performed well in membrane protein prediction [42]. Moreover, we also performed motif analysis by MEME Suite (http://meme-suite.org/) because a motif may directly accord with the active site of an enzyme or a domain of the protein. MEME have been not only used to predict conserved motif regions but also employed for primers design with low quality sequence similarity patterns in multiple global alignments [43].

## 2. Materials and Methods

*2.1. Data Retrieval and Pretreatment.* GPCR sequences with fasta format were retrieved from the UniProt database (http://www.uniprot.org/); we obtained initial 5027 sequences altogether. To improve analysis performance, the raw dataset was preprocessed by the protein-clustering program CD-HIT (http://cd-hit.org/) for reducing the sequence homology bias of prediction; the sequence identity threshold was set at 0.80 and other parameters as default; thus, the highly homology sequences were removed, and finally 2495 GPCR protein sequences were gained as positive dataset, and the negative examples were from all the protein sequences but removing the positive ones, and 10386 entries (non-GPCRs) were acquired as negative dataset.

*2.2. Extracting the Discriminative Feature Vector for Classifying and Testing by Random Forest Classifier.* Protein features were extracted from the primary sequences according to their compositions of 20 kinds of amino acids and their eight types of physical-chemical properties; based on these characteristics, Cai et al. [44] and Zou et al. [42] had raised 188D feature vectors of SVM-Prot. The workflow was as follows:

(1) All distinct positive protein samples were employed to extract their corresponding protein families for Pfam number from the "Family and Domains" section of uniprot website and excluded the same and redundant Pfam number; the unique Pfam number set for positive dataset (in fasta format) was acquired.

(2) All the protein sequences were integrated into a Pfam number file; the same Pfam sequences were combined to the same file named with Pfam number; then, the positive Pfam number files were removed; the rest of Pfam number files were extracted only in the longest sequence for each Pfam as the negative dataset (in fasta format).

(3) Because the protein sequences possessed different length, each sequence needed to transform into fixed-size vectors for classification, both the positive and negative datasets were input to the 188D SVM-Prot programme for their feature vectors, the positive samples were given the label "1" at the end of vectors, the negative samples were given the label "−1" at the end of vectors, and the positive and negative files combined into a file with the filename format ended in .arff.

(4) The above file on positive and negative vector datasets was randomly divided into five parts, respectively, among which, every four parts were served as training examples and the remaining one part as test ones, every part contained both positive and negative samples (Table 3), and fivefold cross-validation was used.

(5) The training and test datasets were successively imported into weka data mining package (http://www.cs.waikato.ac.nz/ml/weka/), a machine-learning workbench. In weka, the training datasets were filtered with the synthetic minority oversampling technique (SMOTE) [45, 46] and changed the positive samples from 100 percent into 300 percent to overcome the highly imbalanced property of positive and negative cases; after preprocessing with SMOTE technique the two-group data kept an amount equilibrium, and the vector data were classified automatically via visualization analysis [47]. Based on the optimal features with some preliminary trials, we finally chose a Random Forest (RF) [48] module and "use training set" item on test options as classifier for training dataset, while for test dataset we chose "supplied test set" item on test options to predict the samples as GPCRs or non-GPCRs: that is, the prediction module

Table 3: The distribution of positive and negative sample numbers for training and test dataset.

| Performance | Part | Number of GPCRs | Number of non-GPCRs | Total number |
|---|---|---|---|---|
| 1st | Training | 1996 | 8309 | 10305 |
| 1st | Test | 499 | 2077 | 2576 |
| 2nd | Training | 1996 | 8309 | 10305 |
| 2nd | Test | 499 | 2077 | 2576 |
| 3rd | Training | 1996 | 8309 | 10305 |
| 3rd | Test | 499 | 2077 | 2576 |
| 4th | Training | 1996 | 8309 | 10305 |
| 4th | Test | 499 | 2077 | 2576 |
| 5th | Training | 1996 | 8308 | 10304 |
| 5th | Test | 499 | 2078 | 2577 |

Table 4: Performance measures for random forest from SVM-Prot feature.

| Measure | Formula | Meaning |
|---|---|---|
| Sensitivity | $\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$ | Measure to avoid type II error |
| Specificity | $\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$ | Measure to avoid type I error |
| Accuracy | $\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$ | Measure of correctness |
| Matthew's correlation coefficient | $\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$ | Correlation coefficient |

TP (true positive) stands for the number of true GPCRs that are predicted correctly, TN (true negative) stands for the number of true non-GPCRs that are predicted correctly, FP (false positive) is the number of true non-GPCRs that are incorrectly predicted to be GPCRs, and FN (false negative) is the number of true GPCRs that are incorrectly predicted to be non-GPCRs.

using the results of the just training set to distinguish the two classes.

To measure the performance quality of the statistical classification more intuitively in the field of machine learning, we adopted 5-fold cross-validation for test dataset and calculated four common parameters [49, 50]: sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC) to adopt for evaluating the SVM-Prot features and classifier, which are formulated as Table 4.

*2.3. Conserved Motif Analyses of Human GPCR Proteins.* Online MEME Suite 4.11.0 (http://meme-suite.org/) was used to analyze conserved motif analyses. MEME was a powerful, comprehensive web-based tool for mining sequence motifs in proteins, DNA, and RNA [51]. Currently, the MEME Suite has added 6 new tools since the *Nucleic Acids Research Web Server Issue* in 2009, and the web-based version tools reached 13. The maximum motif width, the minimal motif width, and the maximum number of motifs were set to 50, 6, and 10, respectively.

## 3. Results

*3.1. Reclassification of Positive and Negative Proteins on Five Test Datasets.* We obtained the 188D feature vectors containing positive and negative samples and divided them into training and test datasets as input to the Weka explorer, respectively, the results showed exactly classifying for all the five training datasets; therefore, the trained classifier could be utilized to verify the predication effect, and the test dataset was used to predict its class label directly. The correctly classified rates for five testing datasets were 90.64%, 90.37%, 88.04%, 93.28%, and 95.73%, respectively (mean ± SD: 91.61% ± 2.96%); the other indices were shown in Table 5.

*3.2. Conserved Motifs Analysis for Human GPCRs.* For the purpose of disclosing the evolutionary relationship of the conserved motifs of GPCRs, we randomly selected six classes of human GPCRs and gained 66 protein sequences which were analyzed by MEME software. The multiple local alignments were performed by MEME to generate the most significant 10 conserved motifs for the sequences (Figure 1 and Table 6).

## 4. Discussion

In this study we show that the novel SVM-Prot features based binary classifier can well discriminate GPCRs from non-GPCRs; we obtain exact classification model from the five training datasets and the AUC equals 1, and on the five testing datasets we get the average correctly classified rates of 91.61% and the average AUC of 0.9282; these indicate that predicted GPCRs and true GPCRs have a good overall consistency. AUC is a plot with $x$-axis representing false positives (equal to

Table 5: Performance qualities measure for test dataset by using the models from the corresponding training dataset.

| Test dataset | Sn | Sp | Acc | MCC | AUC* |
|---|---|---|---|---|---|
| 1st | 0.5952 | 0.9812 | 0.7882 | 0.6248 | 0.930 |
| 2nd | 0.5832 | 0.9807 | 0.7820 | 0.6146 | 0.909 |
| 3rd | 0.6013 | 0.9620 | 0.7817 | 0.5763 | 0.879 |
| 4th | 0.7675 | 0.9726 | 0.8700 | 0.7562 | 0.943 |
| 5th | 0.9238 | 0.9654 | 0.9446 | 0.8900 | 0.980 |
| Mean ± SD | 0.6942 ± 0.1491 | 0.9724 ± 0.0087 | 0.8333 ± 0.0726 | 0.6924 ± 0.1296 | 0.928 ± 0.038 |

*AUC, also called receiver operating characteristic (ROC) area, means the area under the receiver operating characteristic curve which is a measure of the accuracy of a classification model.

Table 6: Human top 10 conserved motifs of GPCR sequences found by the MEME system.

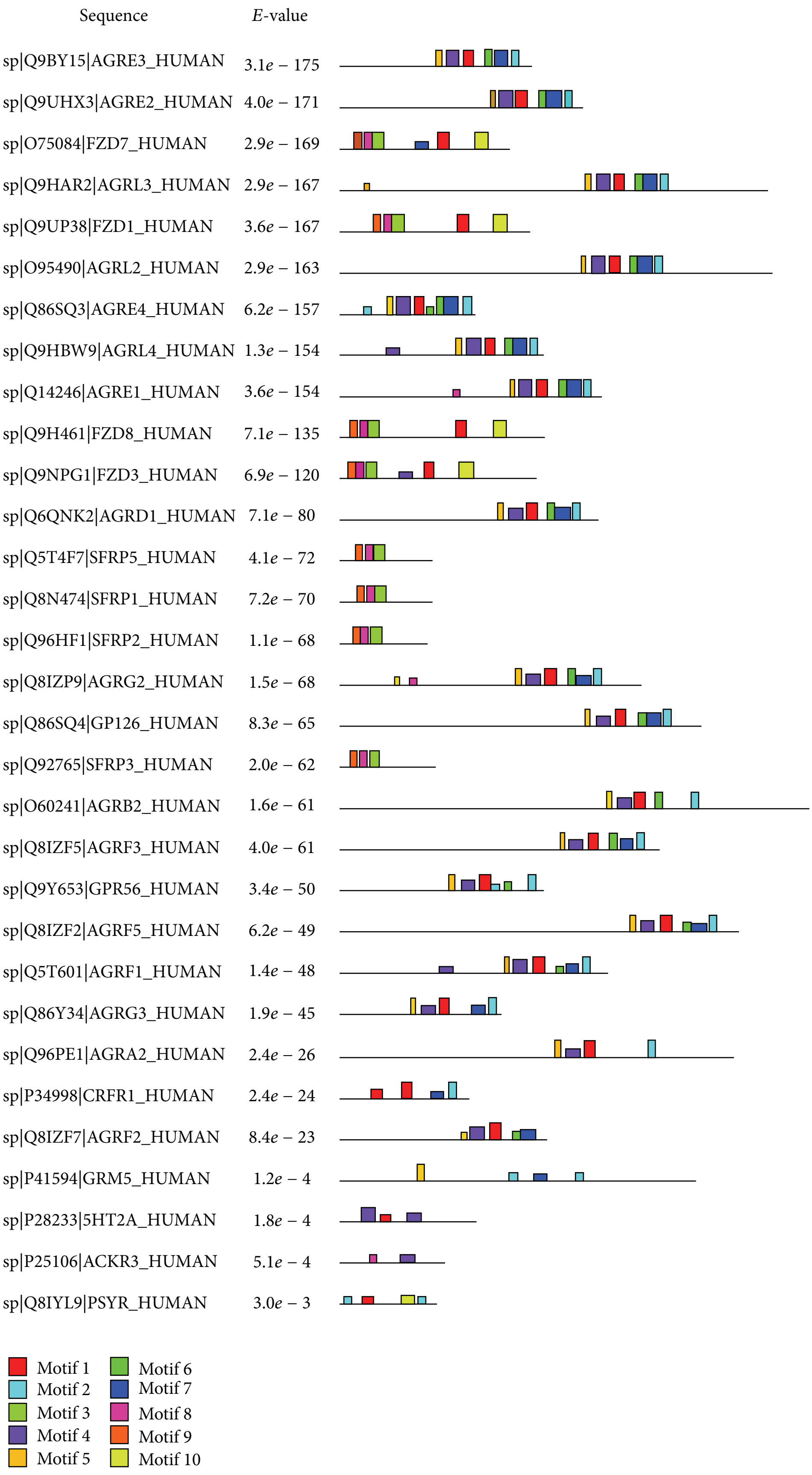
| Motif | Width | *E*-value | Best possible match |
|---|---|---|---|
| 1 | 40 | $4.3e - 239$ | KMACTIMAMFLHYFYLAAFFWMLIEGLHLYLMAVMVWHHE |
| 2 | 29 | $1.5e - 168$ | VMHYLFTIFNSFQGFFIFIFHCLLNRQVR |
| 3 | 41 | $4.4e - 105$ | CLDRPIPPCRSLCERARQGCEPLMNKFGFPWPEMMKCDKFP |
| 4 | 50 | $5.3e - 098$ | VITWVGIIISLVCLLICIFTFLFCRAIQNTRTSIHKNLCICLFLAHLLFL |
| 5 | 21 | $3.8e - 088$ | NKTHTTCRCNHLTNFAVLMAH |
| 6 | 29 | $1.0e - 076$ | GTDKRCWLHLDKGFIWSFIGPVCVIILVN |
| 7 | 50 | $3.9e - 063$ | IFFIITLWIMKRHLSSLNPEVSTLQNTRMWAFKAFAQLFILGCTWCFGIL |
| 8 | 29 | $1.8e - 054$ | LQVHQWYPLVKKQCHPDLKFFLCSMYAPV |
| 9 | 29 | $1.6e - 052$ | CQPIDIPLCHDIGYNQMIMPNLLNHETQE |
| 10 | 50 | $2.0e - 052$ | MKHDGTKTEKLEKLMIRIGVFSVLYTVPATIVIACYFYEQAFRDHWERTW |

1 – specificity) and $y$-axis representing true positives (equal to sensitivity), which is based on different cutoff values of a score from a binary classifier [52, 53]. AUC of 1 represents a perfect model; the more AUC is close to 1, the better prediction model we can develop, but if the value is reduced to 0.5, the model becomes no predictive ability at all. On our binary classification model we acquired high specificity and accuracy for testing datasets, but the values of sensitivity and Matthew's correlation coefficient were relatively low at about 0.7; this might be due to the problem of imbalance dataset where the size of positive was less than negative with the proportion of about 1 : 4; thus the false negative rate was relatively higher. This defect may also come from the intrinsic restriction of supervised learning algorithm, because the classification model built from training dataset can only have a good predictive effect on the test dataset having the same probability distribution as the training dataset [54].

The top ten human GPCR motifs show the feature of some motifs aggregation that appeared from the block diagram; this reflected in the structure characteristic of 7TM helices regions of GPCRs. Motifs 1,4,6,7, and 10 belonged to these 7TM domains; among them, the former 4 motifs displayed containing the region highly homologous to the class B1 secretin family, and motif 10 was a Fz domain in the membrane spanning region which is located near to the intracellular C-terminal region of GPCRs, which contained an alpha-helical Cys-rich domain (CRD) of Frizzled that was essential for Wnt binding [55, 56]. Motifs 3, 8, and 9 were CRD Frizzled-1 like domains involved in Wnt signal as well [57]. Motif 5 was latrophilin/CL-1-like G protein–coupled receptor proteolysis site motif (GPS) which was first identified in a neuronal $Ca^{2+}$-independent receptor of alpha-latrotoxin (CIRL)/latrophilin, an orphan GPCR [58]. GPS was a part of GPCR autoproteolysis-inducing (GAIN) domain which held a formative feature of adhesion GPCRs, and GPS cleavage process played an important role in renal organ physiology [59]. Take the first sequence Q9BY15, for instance, there listed 3 kinds of conserved domains start from the N-terminus: calcium-binding EGF domain (not shown), GPS domain, and 7TM domain of secretin family. The latter two domains appeared with concentration on the block diagram.

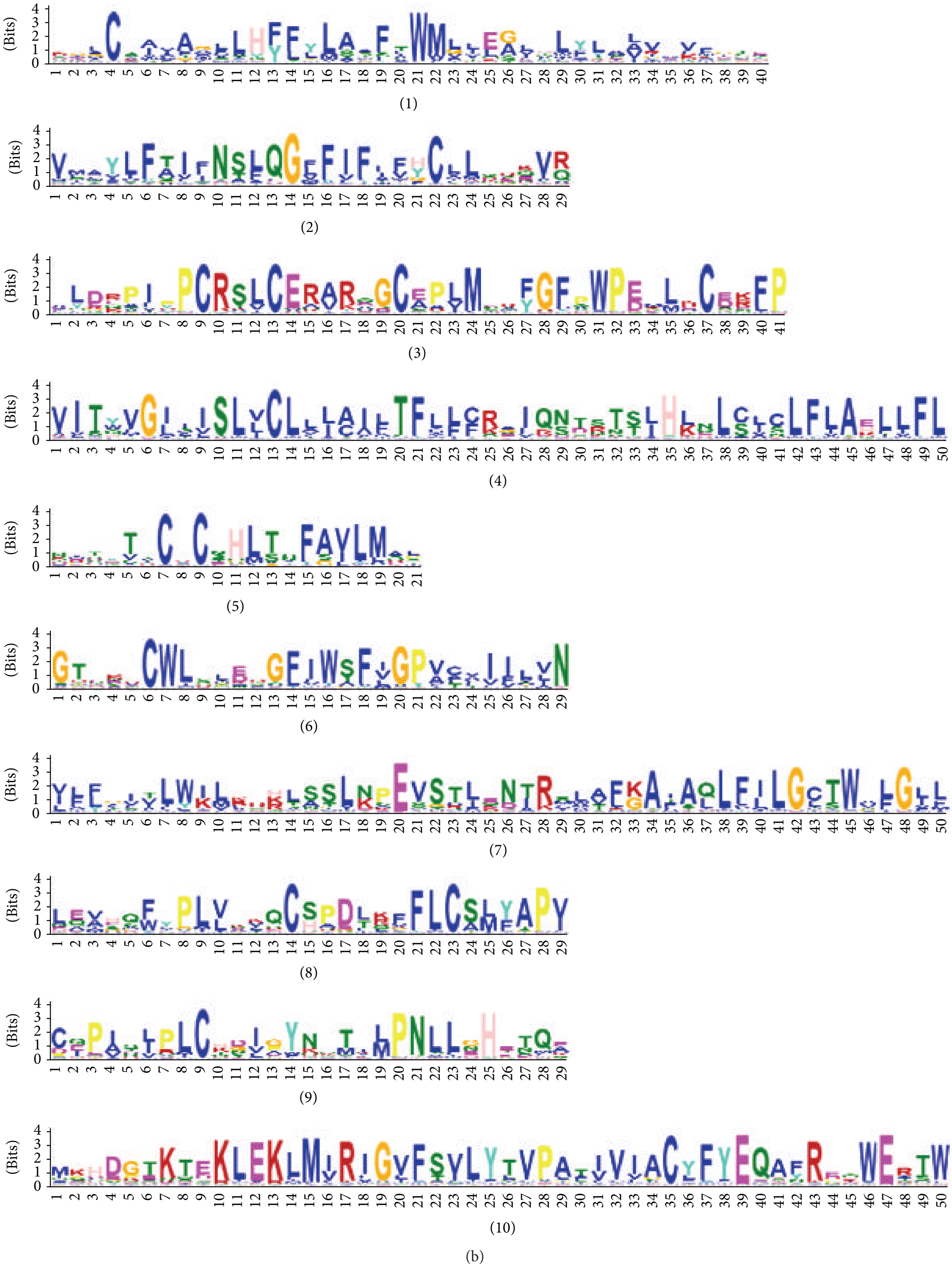
Support Vector Machine (SVM) is a supervised machine-learning algorithm on the basis of statistical learning theory [53, 60–65]. Due to the robustness, rapidness, and repeatability, machine-learning method is regarded as one of the best ways to efficiently classify numerous protein molecules. In two-class problems, our SVM classifier mapped the input 188D feature vectors into a higher dimensional feature space and then founded the optimal separation hyperplane [66] for GPCRs and non-GPCRs, while avoiding overfitting and underfitting problems. This approach belongs to linear classification model [67].

All the GPCR superfamily contains seven highly conserved 7TM regions with the feature of hydrophobicity; these 7TM can be identified by Hidden Markov Models (HMMs) and machine-learning methods [68]. The GPCRs structure researchers revealed that the classical sequence contained the following: the seven-transmembrane segments [TM1–7], three extracellular loops [EL1–3], three



(a)

Figure 1: Continued.



(1)

(2)

(3)

(4)

(5)

(6)

(7)

(8)

(9)

(10)

(b)

Figure 1: The discovered motifs of human GPCRs from the MEME system (for details see Table 6). (a) MEME run showing combined block diagram for top ten motifs distribution with corresponding sequence ID and *E*-value (*E*-value threshold: 0.01, showing 31 GPCR sequences). (b) The ten motif logos found by MEME.

intracellular loops [IL1–3], and the protein termini. Therefore, GPCR can be sequentially distributed into the following regions: N-terminus-TM1-IL1-TM2-EL1-TM3-IL2-TM4-EL2-TM5-IL3-TM6-EL3-TM7-C terminus. In summary, we have successfully developed a SVM-Prot features based Random Forest for identifying GPCRs from non-GPCRs based on the protein sequence information and their physicochemical properties. Nevertheless, this prediction model needs to be further explored so as to discriminate the subfamily and sub-subfamily of GPCRs.

## Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The work was supported by the Natural Science Foundation of Fujian Province of China (no. 2016J01152), the Natural Science Foundation of China (no. 61370010), and the State Key Laboratory of Medicinal Chemical Biology (no. 201601013).

## References

[1] B. Trzaskowski, D. Latek, S. Yuan, U. Ghoshdastider, A. Debinski, and S. Filipek, "Action of molecular switches in GPCRs—theoretical and experimental studies," *Current Medicinal Chemistry*, vol. 19, no. 8, pp. 1090–1109, 2012.

[2] D. M. Shore and P. H. Reggio, "The therapeutic potential of orphan GPCRs, GPR35 and GPR55," *Frontiers in Pharmacology*, vol. 6, article 69, 2015.

[3] H.-H. Lin, "G-protein-coupled receptors and their (Bio) Chemical significance win 2012 nobel prize in chemistry," *Biomedical Journal*, vol. 36, no. 3, pp. 118–124, 2013.

[4] Q. Zou, "Machine learning techniques for protein structure, genomics function analysis and disease prediction," *Current Proteomics*, vol. 13, no. 2, pp. 77–78, 2016.

[5] Y. Que, L. Xu, Q. Wu et al., "Genome sequencing of Sporisorium scitamineum provides insights into the pathogenic mechanisms of sugarcane smut," *BMC Genomics*, vol. 15, no. 1, article 996, 2014.

[6] D. M. Rosenbaum, S. G. F. Rasmussen, and B. K. Kobilka, "The structure and function of G-protein-coupled receptors," *Nature*, vol. 459, no. 7245, pp. 356–363, 2009.

[7] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen, and G. Vriend, "GPCRDB information system for G protein-coupled receptors," *Nucleic Acids Research*, vol. 31, no. 1, pp. 294–297, 2003.

[8] A. Krishnan, M. S. Almén, R. Fredriksson, and H. B. Schiöth, "The origin of GPCRs: identification of mammalian like rhodopsin, adhesion, glutamate and frizzled GPCRs in fungi," *PLoS ONE*, vol. 7, no. 1, Article ID e29817, 2012.

[9] K. Kochman, "Superfamily of G-protein coupled receptors (GPCRs)—extraordinary and outstanding success of evolution," *Postepy Higieny i Medycyny Doswiadczalnej*, vol. 68, pp. 1225–1237, 2014.

[10] S. Balfanz, N. Jordan, T. Langenstück, J. Breuer, V. Bergmeier, and A. Baumann, "Molecular, pharmacological, and signaling properties of octopamine receptors from honeybee (*Apis mellifera*) brain," *Journal of Neurochemistry*, vol. 129, no. 2, pp. 284–296, 2014.

[11] R. Franco, E. Martínez-Pinilla, J. L. Lanciego, and G. Navarro, "Basic pharmacological and structural evidence for Class A G-protein-coupled receptor heteromerization," *Frontiers in Pharmacology*, vol. 7, article 76, 2016.

[12] B. D. Shepard, N. Natarajan, R. J. Protzko, O. W. Acres, and J. L. Pluznick, "A cleavable N-terminal signal peptide promotes widespread olfactory receptor surface expression in HEK293T cells," *PLoS ONE*, vol. 8, no. 7, Article ID e68758, 2013.

[13] S. Sreedharan, M. S. Almén, V. P. Carlini et al., "The G protein coupled receptor Gpr153 shares common evolutionary origin with Gpr162 and is highly expressed in central regions including the thalamus, cerebellum and the arcuate nucleus," *FEBS Journal*, vol. 278, no. 24, pp. 4881–4894, 2011.

[14] L.-H. Zhao, Y. Yin, D. Yang et al., "Differential requirement of the extracellular domain in activation of class B G protein-coupled receptors," *The Journal of Biological Chemistry*, vol. 291, no. 29, pp. 15119–15130, 2016.

[15] J. C. R. Cardoso, V. C. Pinto, F. A. Vieira, M. S. Clark, and D. M. Power, "Evolution of secretin family GPCR members in the metazoa," *BMC Evolutionary Biology*, vol. 6, article 108, 2006.

[16] J. G. Duman, Y.-K. Tu, and K. F. Tolias, "Emerging roles of BAI adhesion-GPCRs in synapse development and plasticity," *Neural Plasticity*, vol. 2016, Article ID 8301737, 9 pages, 2016.

[17] J. Hamann, G. Aust, D. Araç et al., "International union of basic and clinical pharmacology. XCIV. Adhesion G protein-coupled receptors," *Pharmacological Reviews*, vol. 67, no. 2, pp. 338–367, 2015.

[18] T. Langenhan, G. Aust, and J. Hamann, "Sticky signaling—adhesion class g protein-coupled receptors take the stage," *Science Signaling*, vol. 6, no. 276, article re3, 2013.

[19] M. C. Peeters, I. Mos, E. B. Lenselink, M. Lucchesi, A. P. IJzerman, and T. W. Schwartz, "Getting from A to B-exploring the activation motifs of the class B adhesion G protein-coupled receptor subfamily G member 4/GPR112," *The FASEB Journal*, vol. 30, no. 5, pp. 1836–1848, 2016.

[20] W. Spooren, A. Lesage, H. Lavreysen, F. Gasparini, and T. Steckler, "Metabotropic glutamate receptors: their therapeutic potential in anxiety," *Current Topics in Behavioral Neurosciences*, vol. 2, pp. 391–413, 2010.

[21] Q. Bai and X. Yao, "Investigation of allosteric modulation mechanism of metabotropic glutamate receptor 1 by molecular dynamics simulations, free energy and weak interaction analysis," *Scientific Reports*, vol. 6, article 21763, 2016.

[22] J. Lewerenz and P. Maher, "Chronic glutamate toxicity in neurodegenerative diseases-what is the evidence?" *Frontiers in Neuroscience*, vol. 9, article 469, 2015.

[23] R. Vafabakhsh, J. Levitz, and E. Y. Isacoff, "Conformational dynamics of a class C G-protein-coupled receptor," *Nature*, vol. 524, no. 7566, pp. 497–501, 2015.

[24] S. Nachtergaele, D. M. Whalen, L. K. Mydock et al., "Structure and function of the Smoothened extracellular domain in vertebrate Hedgehog signaling," *eLife*, vol. 2, Article ID e01340, 2013.

[25] J. Pei and N. V. Grishin, "Cysteine-rich domains related to Frizzled receptors and Hedgehog-interacting proteins," *Protein Science*, vol. 21, no. 8, pp. 1172–1184, 2012.

[26] C. P. Xavier, M. Melikova, Y. Chuman, A. Üren, B. Baljinnyam, and J. S. Rubin, "Secreted Frizzled-related protein potentiation versus inhibition of Wnt3a/$\beta$-catenin signaling," *Cellular Signalling*, vol. 26, no. 1, pp. 94–101, 2014.

[27] J. Singh, X. Wen, and S. J. Scales, "The orphan G protein-coupled receptor Gpr175 (Tpra40) enhances Hedgehog signaling by modulating cAMP levels," *The Journal of Biological Chemistry*, vol. 290, no. 49, pp. 29663–29675, 2015.

[28] Y. Takeo, N. Kurabayashi, M. D. Nguyen, and K. Sanada, "The G protein-coupled receptor GPR157 regulates neuronal differentiation of radial glial progenitors through the Gq-IP$_3$ pathway," *Scientific Reports*, vol. 6, Article ID 25180, 2016.

[29] E. Mathew, A. Bajaj, S. M. Connelly et al., "Differential interactions of fluorescent agonists and antagonists with the yeast G protein coupled receptor ste2p," *Journal of Molecular Biology*, vol. 409, no. 4, pp. 513–528, 2011.

[30] W. Kuohung, M. Burnett, D. Mukhtyar et al., "A high-throughput small-molecule ligand screen targeted to agonists and antagonists of the g-protein-coupled receptor GPR54," *Journal of Biomolecular Screening*, vol. 15, no. 5, pp. 508–517, 2010.

[31] D. C. Gonzalez-Kristeller, J. B. P. do Nascimento, P. A. F. Galante, and B. Malnic, "Identification of agonists for a group of human odorant receptors," *Frontiers in Pharmacology*, vol. 6, article 35, 2015.

[32] R. L. Thurmond, "The histamine H4 receptor: from orphan to the clinic," *Frontiers in Pharmacology*, vol. 6, article 65, 2015.

[33] X. Lv, J. Liu, Q. Shi et al., "In vitro expression and analysis of the 826 human G protein-coupled receptors," *Protein Cell*, vol. 7, no. 5, pp. 325–337, 2016.

[34] J. D. Hildebrandt, "Role of subunit diversity in signaling by heterotrimeric G proteins," *Biochemical Pharmacology*, vol. 54, no. 3, pp. 325–339, 1997.

[35] M. Sato, "Roles of accessory proteins for heterotrimeric G-protein in the development of cardiovascular diseases," *Circulation Journal*, vol. 77, no. 10, pp. 2455–2461, 2013.

[36] Z. Yu, L. Li, J. Liu, and G. Han, "Hybrid adaptive classifier ensemble," *IEEE Transactions on Cybernetics*, vol. 45, no. 2, pp. 177–190, 2015.

[37] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, article e56499, 2013.

[38] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, and Z. Hong, "Improving tRNAscan-SE annotation results via ensemble classifiers," *Molecular Informatics*, vol. 34, no. 11-12, pp. 761–770, 2015.

[39] Z. Yu, H. Chen, J. Liu et al., "Hybrid $\kappa$—nearest neighbor classifier," *IEEE Transactions on Cybernetics*, vol. 46, no. 6, pp. 1263–1275, 2016.

[40] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive noise immune cluster ensemble using affinity propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 12, pp. 3176–3189, 2015.

[41] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[42] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, "BinMemPredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[43] M. Sahu, J. Sahu, S. Sahoo et al., "An approach to delineate primers for a group of poorly conserved sequences incorporating the common motif region," *Bioinformation*, vol. 8, no. 4, pp. 181–184, 2012.

[44] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3692–3697, 2003.

[45] K. H. Chen, K. Wang, A. M. Adrian, and N. Teng, "Diagnosis of brain metastases from lung cancer using a modified electromagnetism like mechanism algorithm," *Journal of Medical Systems*, vol. 40, no. 1, p. 35, 2016.

[46] W. Wiharto, H. Kusnanto, and H. Herianto, "Intelligence system for diagnosis level of coronary heart disease with K-star algorithm," *Healthcare Informatics Research*, vol. 22, no. 1, pp. 30–38, 2016.

[47] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[48] B. Liu, R. Long, and K. Chou, "iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework," *Bioinformatics*, 2016.

[49] R. Wang, Y. Xu, and B. Liu, "Recombination spot identification based on gapped k-mers," *Scientific Reports*, vol. 6, Article ID 23934, 2016.

[50] B. Liu, S. Wang, Q. Dong, S. Li, and X. Liu, "Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning," *IEEE Transactions on NanoBioscience*, 2016.

[51] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME suite," *Nucleic Acids Research*, vol. 43, no. 1, pp. W39–W49, 2015.

[52] K. Uno, K. Yoshizaki, M. Iwahashi et al., "Pretreatment prediction of individual rheumatoid arthritis patients' response to anti-cytokine therapy using serum cytokine/chemokine/soluble receptor biomarkers," *PLoS ONE*, vol. 10, no. 7, Article ID e0132055, 2015.

[53] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

[54] M. N. Davies, D. E. Gloriam, A. Secker et al., "Proteomic applications of automated GPCR classification," *Proteomics*, vol. 7, no. 16, pp. 2800–2814, 2007.

[55] T. Tsukiyama, A. Fukui, S. Terai et al., "Molecular role of RNF43 in canonical and noncanonical Wnt signaling," *Molecular and Cellular Biology*, vol. 35, no. 11, pp. 2007–2023, 2015.

[56] E. Brinkmann, B. Mattes, R. Kumar et al., "Secreted frizzled-related protein 2 (sFRP2) redirects non-canonical Wnt signaling from Fz7 to Ror2 during vertebrate gastrulation," *The Journal of Biological Chemistry*, vol. 291, no. 26, pp. 13730–13742, 2016.

[57] S. Thysen, F. Cailotto, and R. Lories, "Osteogenesis induced by frizzled-related protein (FRZB) is linked to the netrin-like domain," *Laboratory Investigation*, vol. 96, no. 5, pp. 570–580, 2016.

[58] V. G. Krasnoperov, M. A. Bittner, R. Beavis et al., "$\alpha$-Latrotoxin stimulates exocytosis by the interaction with a neuronal G-protein-coupled receptor," *Neuron*, vol. 18, no. 6, pp. 925–937, 1997.

[59] M. Trudel, Q. Yao, and F. Qian, "The role of G-protein-coupled receptor proteolysis site cleavage of polycystin-1 in renal physiology and polycystic kidney disease," *Cells*, vol. 5, no. 1, 2016.

[60] H. Ma, W. Chang, and G. Cui, "Ecological footprint model using the support vector machine technique," *PLoS ONE*, vol. 7, no. 1, article e30396, 2012.

[61] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection

and analysis," *Molecular Biosystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[62] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[63] D. Li, Y. Ju, and Q. Zou, "Protein folds prediction with hierarchical structured SVM," *Current Proteomics*, vol. 13, no. 2, pp. 79–85, 2016.

[64] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[65] L.-F. Yuan, C. Ding, S.-H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[66] B. Haasdonk, "Feature space interpretation of SVMs with indefinite kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 482–492, 2005.

[67] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, C. Ostermann, and A. Zell, "Large-scale learning of structure-activity relationships using a linear support vector machine and problem-specific metrics," *Journal of Chemical Information and Modeling*, vol. 51, no. 2, pp. 203–213, 2011.

[68] H. Bouziane, B. Messabih, and A. Chouarfia, "Profiles and majority voting-based ensemble method for protein secondary structure prediction," *Evolutionary Bioinformatics*, vol. 2011, no. 7, pp. 171–189, 2011.