# Development of Prediction Method for GPCR–G-protein Coupling Selectivity Using Amino Acid Properties

Yukimitsu Yabuki[*,1,2], Masami Ikeda[3,4], Yuri Mukai-Ikeda[5] and Yoshihisa Ishida[1,5]

[1]*Department of Electrical Engineering, Graduate School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki-shi, Kanagawa 214-8751, Japan*

[2]*Information and Mathematical Science Laboratory (IMS) Inc. Meikei Building, 1-5-21 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan*

[3]*Consolidated Research Institute for Advanced Science and Medical Care (ASMeW), Waseda University, 513 Wasedatsurumaki-cho, Shinjuku-ku, Tokyo 162-0041, Japan*

[4]*Research Institute of Information Technological Biology, Information Technology Research Organization, Waseda University, AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan*

[5]*Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki-shi, Kanagawa 214-8751, Japan*

**Abstract:** We describe a novel method for predicting G-protein coupled receptor (GPCR) - G-protein coupling selectivity using amino acid properties of specific residues in GPCR sequences. We have evaluated various amino acid properties obtained with experimental or theoretical studies. The GPCRs having reliable G-protein binding information were collected from Guide to Receptors and Channels (GRAC) and gpDB databases, and these sequences were aligned with the amino acid sequence of bovine rhodopsin, whose structure is known, to identify positions of each amino acid residue and its secondary structure. The collected properties were used as feature values to calculate Fisher's ratio (FR) for each residue in GPCRs related with rhodopsin residue numbers. Some amino acid properties with high FR value were picked up as the effective characteristics for selecting G-protein type, and they were used as feature vectors in support vector machine (SVM) to predict GPCR - G-protein coupling selectivity. Applying this method to known GPCR sequences, each binding G-protein is predicted with high sensitivity and specificity of more than 96%. This result strongly suggests that the amino acid properties of specific residues are appreciably important for GPCR - G-protein coupling to determine G-protein binding selectivity and our method could be an effective tool to investigate the mechanism of GPCR - G-protein coupling through site directed mutagenesis experiments.

**Keywords:** G-protein coupled receptor, G-protein, G-protein coupling selectivity, amino acid property, support vector machine.

## INTRODUCTION

G-protein coupled receptors (GPCRs) are the major integral membrane proteins that exist in cellular membrane of eukaryotic cells and constitute the largest membrane protein family. On the basis of sequence similarities, GPCRs are generally classified into three classes, i.e., Class A (rhodopsin like receptors), Class B (secretin like receptors) and Class C (metabotropic glutamate/pheromone receptors), and members of these families have a common behavior in the point of containing characteristic seven transmembrane domains with highly conserved structural conformation in each class. GPCRs are known to play an important role as interface for transmitting signal to the inner cell and are involved in the system that are necessary to preserve life, such as neurotransmission system or endocrine system. Hence, GPCRs are regarded as important targets in the development of effective drugs. In fact, nearly half of all therapeutic agents distributed throughout the world are designed to control the mechanism involved in GPCRs [1]. As a system of the function of GPCRs, a specific external ligand, such as a neurotransmitter, a hormone or a small compound, stimulate a certain GPCR and the GPCR is coupled with one or more G-proteins with conformational changes, followed by various kinds of signals that are transmitted to the inner cell. G-proteins are soluble proteins that exist as a form of heterotrimeric complexes that are composed of a $G_\alpha$ subunit interacting with a $G_\beta$ and $G_\gamma$ complex combined tightly, and they are mainly classified into four families based on $G_\alpha$ subunit type, that is, $G_{i/o}$, $G_{q/11}$, $G_s$ and $G_{12/13}$. They have an essential role of determining first step of acting as trigger for subsequent

*Address correspondence to this author at the Information and Mathematical Science Laboratory (IMS), Inc. Meikei Building, 1-5-21 Otsuka, Bunkyo-ku, Tokyo, 112-0012, Japan; Tel: +81-3-5319-2011; Fax: +81-3-5319-2020; E-mail: yabuki.yukimitsu@imslab.co.jp

1874-1991/09 

various signals. $G_s$ and $G_{i/o}$ stimulate and inhibit adenylyl cyclase, respectively, and $G_{q/11}$ activates phospholipase C, and $G_{12/13}$ is involved in Rho family GTPase signaling [2]. In this way, first step of signaling depends on the type of G-protein and studying the mechanism of GPCR - G-protein coupling selectivity is of considerable significance to achieve new insights not only for transducing mechanism into inner cell but also developing effective therapeutic agents. Various researchers have succeeded to determine grave amino acid residues for G-protein coupling by site directed mutagenesis experiments [3-5], however, most of these studies are narrowed down to specific GPCRs, and features common to each of G-protein binding types are still unclear. This is due to the fact that no amino acid pattern is found commonly in each of the specific G-protein binding types.

Recently, various methods have been developed to predict GPCR - G-protein coupling selectivity using machine learning procedures, hidden Markov Models (HMM) [6], HMM and neural networks (NN) [7], and autocross-covariance (ACC) transform based support vector machines (SVM) [8]. These methods predicted G-protein binding types with an accuracy of over 90% and are effective for GPCRs that are still unknown for G-protein binding types. However, they are inadequate in the perspective for detecting the amino acid residues and the properties, which contribute to determine G-protein binding types. The prediction of G-protein binding types with high accuracy using amino acid properties of specific residues is useful for investigating the mechanism of GPCR - G-protein coupling through site directed mutagenesis experiments.

In this paper, we introduce a procedure for predicting GPCR - G-protein coupling selectivity using amino acid properties of specific residues in GPCR sequence. Compared with previous work, our work is a novel one because this method offers information on amino acid properties of specific residues for discriminating G-protein binding types with high accuracy. First we made Class A GPCRs dataset with reliable G-protein binding type information, and these GPCRs were aligned with rhodopsin sequence, whose structure is known, to identify positions of amino acid residues in structural regions that comprise of extracellular loops, transmembrane regions, and intracellular loops. The amino acid properties were collected from experimental or theoretical studies reported in literature. Using these properties, Fisher's ratios (FR) were calculated to detect differences of feature quantities in every G-protein binding type for each amino acid residue in accordance with rhodopsin residue numbers. As results of FR calculations, we selected the amino acid properties of specific residues with high FR value and these selected characteristics were applied as feature vector elements in the SVM to predict G-protein binding types. The SVM algorithm has been verified to be a high-performance classifier, especially for discriminating multidimensional parameters. Applying this method to known GPCR sequences, each binding G-protein was predicted with high accuracy of over 96%. This study is not only to predict G-protein binding types with high sensitivity/specificity but also to make contribution for understanding the mechanism of GPCR - G-protein coupling through site directed mutagenesis experiments.

## METHODS

### Development of Datasets

Functionally non-redundant human Class A GPCRs, which have experimentally evidences of G-protein coupling, were collected from Guide to Receptors and Channels (GRAC) 3rd edition [9] and gpDB [10] (Table **1**). The numbers of human Class A GPCRs registered in the GRAC and gpDB are 180 and 174, respectively. Comparing these two databases, there are some differences for the kinds of GPCRs registered and G-protein coupling information of individual GPCR (Fig. **1**). Consequently, 127 GPCRs, which have the same G-protein coupling information and registered in both databases, were pick up as highly-reliable data (Table **1**). In this study, a dataset was prepared by using a total of 112 GPCRs, which consist of $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types to detect amino acid residues as well as the physicochemical and structural characteristics that are thought to be important in GPCR - G-protein coupling selectivity. These amino acid sequences were downloaded from SWISS-PROT database (http://www.expasy.org). The remaining 15 GPCRs, which coupled to $G_{12/13}$ and other three G-protein binding types were neglected for the reason that there are too few GPCRs

**Table 1. The Number of Human Class A GPCRs Registered in the GRAC and gpDB**

| G-protein Type | GRAC | gpDB | GRAC ∩ gpDB [a] |
|---|---|---|---|
| $G_{i/o}$ | 73 | 71 | 64 |
| $G_{q/11}$ | 51 | 41 | 29 |
| $G_s$ | 25 | 26 | 19 |
| $G_{i/o}$ $G_{q/11}$ | 17 | 20 | 12 |
| $G_{i/o}$ $G_s$ | 4 | 6 | 0 |
| $G_{q/11}$ $G_s$ | 6 | 6 | 1 |
| $G_{i/o}$ $G_{q/11}$ $G_s$ | 4 | 4 | 2 |
| $G_{12/13}$ [b] | 8 | 25 | 2 |
| Total | 180 | 174 | 127 |

[a] The number of human Class A GPCRs that are registered with the same G-protein coupling information in the GRAC and gpDB.

[b] $G_{12/13}$ binding type GPCRs are also coupled with one or more other three G-proteins, and thus this type GPCRs are not included in the total number of GPCRs.

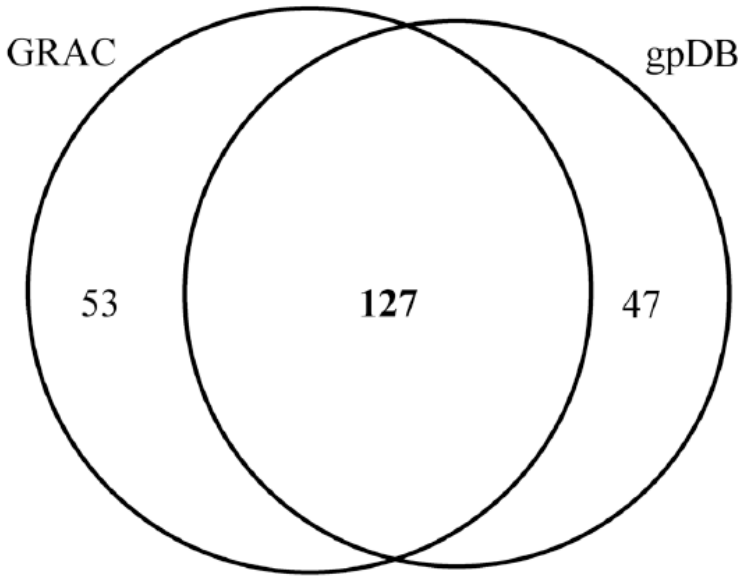


**Fig. (1). Venn diagram of GPCR datasets.** Left and right circles represent GPCRs registered in GRAC and gpDB, respectively. The intersection area represents GPCRs that have completely-consistent information for G-protein binding types in both GRAC and gpDB. The numeric character reflects the number of GPCRs.

of these types, which cannot be adequately used as training data to detect characteristics for GPCR - G-protein coupling selectivity.

The SWISS-PROT IDs and ACs for 127 GPCRs that are registered with the same G-protein binding types in above both databases are given in supplemental Table **S1**.

### Collecting Structural or Physicochemical Characteristics Corresponding to each of Amino Acids

In order to detect structural or physicochemical characteristics of amino acid residues that are thought to play an important role of selecting G-protein types in amino acid residues on each GPCR, various structural or physicochemical characteristics corresponding to 20 amino acids, that is, amino acid properties, were collected from literature that were reported by using experimental or theoretical analysis. These collected properties were used to calculate differences in each amino acid residue among $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types. We have used a set of 200 amino acid properties for the present work and the finally selected ones are listed in Table **2**.

### Alignment of GPCR with Bovine Rhodopsin

Generally, there is no conserved amino acid motif that discriminates G-protein binding types. However, the tertiary structures of GPCRs are comparatively conserved within each GPCR class, although there are some differences in extracellular or intracellular loop structures. This tendency implies that adoption of information based on tertiary structure for each GPCR is important for elucidating the interaction mechanism to G-protein and its grave amino acid residues. At present structural information is available only for few GPCRs that include bovine rhodopsin [11], squid rhodopsin [12], β-1 adrenergic [13] and β-2 adrenergic receptors [14]. In almost all GPCRs the exact positions of transmembrane, extracellular or intracellular loops are not clearly known. Hence, mapping GPCR sequences to those regions was attempted by using known bovine rhodopsin structure (PDB code: 1L9H) in which whole structure was solved at high resolution. The boundaries of the transmembrane helix and loop regions of GPCR sequences were determined by aligning them with bovine rhodopsin sequence using CLUSTAL W [15] and the alignments were adjusted with manual inspection.

### Calculation of Fisher's Ratio in each Amino Acid Residues

Fisher's Ratio (FR) is a useful method for discriminating variables into two classes. It is measured as the ratio of "between-class distance" and "within-class variance". FR has been effectively used in several applications of bioinformatics [16-17]. In this study, the FR was applied to detect the difference between the structural information or physicochemical characteristics of each amino acid residue in query sequence and rhodopsin sequence to distinguish one G-protein binding type (positive dataset) and others (negative dataset). In this instance, combinations of positive and negative datasets are treated as three ways: (i) $G_{i/o}$ binding type as positive and others as negative, (ii) $G_{q/11}$ binding type as positive and others as negative, and (iii) $G_s$ binding type as positive and others as negative datasets. The FR equation is as follows.

$$FR_i = \frac{(\mu_{p,i} - \mu_{n,i})^2}{\sigma_{p,i}^2 + \sigma_{n,i}^2} \quad (1)$$

Here, $\mu_{p,i}$ and $\sigma^2_{p,i}$ being the means and variance of the *i*-th amino acid residues of positive dataset in accordance with rhodopsin residue numbers and $\mu_{n,i}$ and $\sigma^2_{n,i}$ being the means and variance of negative datasets in the same way. The FR calculation was not applied for positions that have gaps in alignment. However, the gap regions have been considered with a central moving average method and the

**Table 2. The Selected Amino Acid Properties with high FR Values Among G-protein Binding Types**

| Domain | Res. Num.[a] | WS[b] | FR | Amino Acid Property |
|---|---|---|---|---|
| **(A) $G_{i/o}$** | | | | |
| TM2 | 94 | 3 | 0.89 | Transfer free energy (AcWI-X-LL peptides from bilayer interface to water) [20] |
| EL1 | 109 | 9 | 0.73 | Normalized frequency of C terminal helix [21] |
| IL2 | 138 | 5 | 0.75 | Slopes tripeptide [22] |
| IL2 | 144 | 9 | 0.98 | Refractivity [23] |
| IL2 | 145 | 11 | 0.99 | Hydrophobicity coefficient (RP-HPLC, C18 with 0.1%TFA/MeCN/$H_2O$) [24] |
| TM5 | 211 | No | 0.80 | Normalized frequency of extended structure [25] |
| TM6 | 259 | 7 | 0.70 | Transfer free energy (CHP/water) [26] |
| TM6 | 263 | 7 | 0.73 | Normalized average hydrophobicity scale [27] |
| **(B) $G_{q/11}$** | | | | |
| EL1 | 98 | 3 | 0.64 | van der Waals parameter epsilon [31] |
| EL1 | 101 | 11 | 0.61 | Retention coefficient in $NaH_2PO_4$ [32] |
| IL2 | 134 | 3 | 0.62 | Optimized side chain interaction [33] |
| IL2 | 143 | No | 0.63 | Average side chain orientation angle [34] |
| TM6 | 273 | 15 | 0.75 | Relative preference value at C1 [35] |
| TM7 | 307 | No | 0.92 | Size [36] |
| TM7 | 307 | No | 0.85 | Number of atoms in the side chain labelled 2+1 [37] |
| **(C) $G_s$** | | | | |
| IL2 | 149 | 7 | 1.77 | Hydrophilicity [39] |
| TM4 | 170 | No | 2.27 | Hydropathy scale based on self-information values in the two-state model (36% accessibility) [40] |
| TM4 | 170 | No | 2.14 | Optimized average non-bonded energy per atom [33] |
| TM4 | 171 | 3 | 2.32 | Free energy change of epsilon (i) to alpha (Rh) [41] |
| IL3 | 230 | No | 1.69 | Normalized positional residue frequency at helix termini N1 [42] |
| IL3 | 233 | No | 1.51 | Average relative fractional occurrence in AL (i) [43] |

TM: transmembrane, IL: intracellular loop, EL: extracellular loop

[a] Res. num.: Sequential number of residues in rhodopsin which were used as template for structural alignment to each GPCR.

[b] WS: Residue size of the central moving average method. No: The value of FR without using the central moving average method.

amino acid properties for each of amino acid residues can be computed as follows:

$$\overline{P}_i = \frac{1}{w}\sum_{k=i-m}^{i+m} P(k), \quad \left(m = \frac{w-1}{2}\right) \tag{2}$$

Here, $P(k)$ is the value of amino acid properties corresponding to each 20 amino acid at a sequence position $i$, and $w$ is the sliding window size for average calculation and is set to every odd numbers from 3 to 15. Among these results, some amino acid properties with high FR values were picked up as parameters that are used to train in a machine learning method to predict G-protein binding selectivity of each GPCR.

## Training and Testing for Predicting GPCR - G-protein Coupling Selectivity by using Support Vector Machines

Support Vector Machine (SVM) was adapted to discriminate G-protein binding types by utilizing amino acid properties selected from the results of FR calculation as vector representations (feature vectors) in SVM. The SVM calculation was performed by using LIBSVM 2.6 package [18]. SVM classifies these feature vectors of GPCRs using a multidimensional hyperplane called the kernel function. The SVM is a classifier used to divide data into two classes. In this study, classifications were performed under the condition that the targeted G-protein type and others as positive and negative datasets, respectively. In order to calculate the accuracy of discriminating each G-protein type ($G_{i/o}$, $G_{q/11}$ and $G_s$), SVM training was performed by RBF kernel with parameters C and G which determine the shape of kernel function, and by changing the combination of the feature vector elements. The variables C and G range from $2^{-5}$ to $2^{15}$, and $2^{-13}$ to $2^3$, respectively. Furthermore, five-fold cross validation tests were performed for each combination of parameter sets to achieve more reliable prediction. In case of 5-fold cross-validation test, four-fifths of the dataset are randomly picked up and are used as training datasets and the remaining fifth is used as test data.

The best combination of feature vector elements was determined when the product of sensitivity and specificity showed the highest value for evaluating G-protein coupling prediction. Furthermore, the Matthews Correlation Coefficient (MCC) [19] was also calculated to take into account different sizes between positive and negative datasets. The sensitivity, specificity and MCC are defined as follows,

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TP}{TP + FP} \quad (4)$$

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

Here, TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively.

## RESULTS AND DISCUSSION

### Detection of Residue Positions and its Amino Acid Properties Specific for Discriminating G-protein Binding Types

The amino acid properties of specific residues with high FR values were picked up from results of FR calculation for discriminating G-protein binding types. The detailed information for specific amino acid residues and the assignment of secondary structures based on the mapping with rhodopsin structure are presented in Table **2** and Fig. (**2**), respectively. In Table **2** and Fig. (**2**), there are eight, seven and six characteristics for $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types, respectively. The number of features selected are reasonable to discriminate each G-protein binding type. As shown in Table **2A**, the typical characteristics for the $G_{i/o}$ binding type are the eight amino acid properties with specific residue positions: transfer free energy (AcWI-X-LL peptides from bilayer interface to water) [20] at position 94 in 2nd transmembrane (TM2), normalized frequency of C terminal helix [21] at position 109 in 1st extracellular loop (EL1), slopes tripeptides [22], refractivity [23] and hydrophobicity

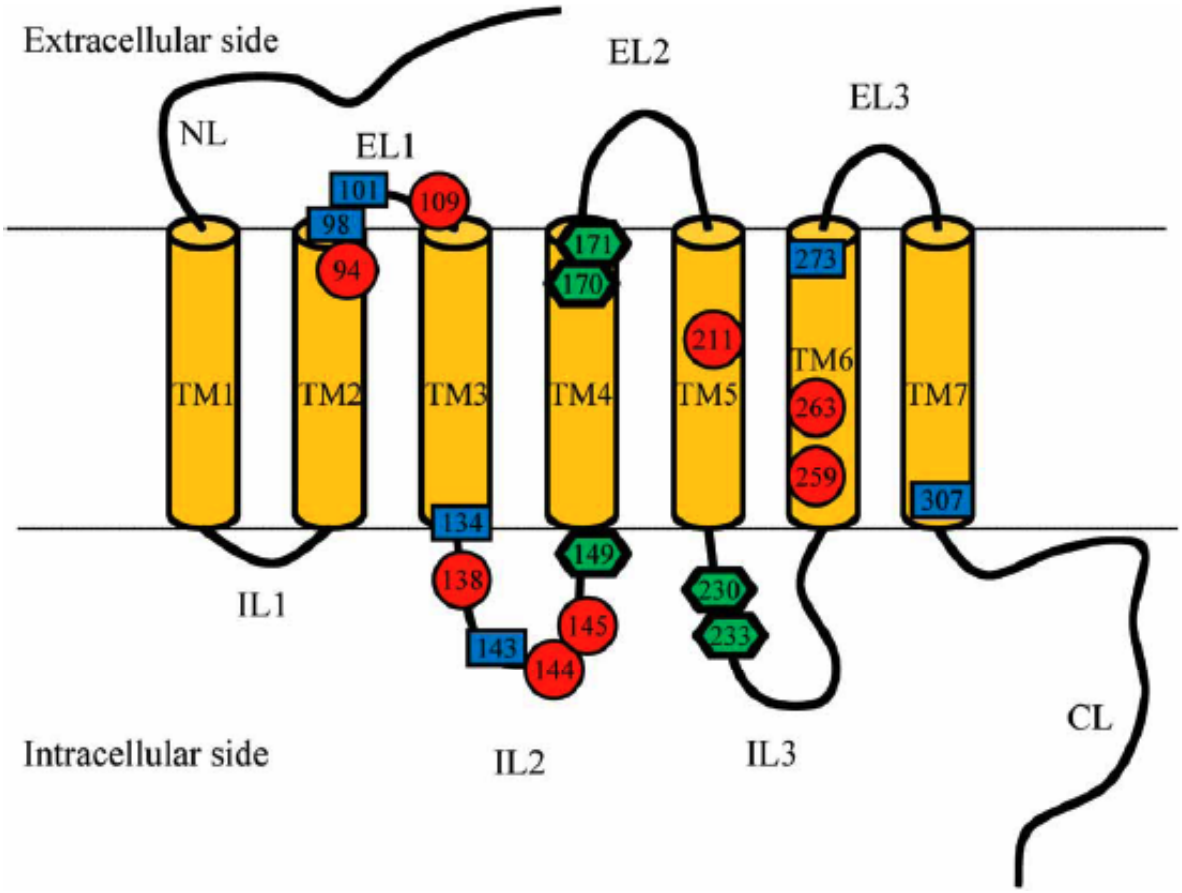


**Fig. (2). View showing a frame format of secondary structure of GPCR.** The two horizontal lines, circular cylinders filled in yellow, and black curving lines represent cellular membrane, transmembrane (TM) regions, and loop regions (N terminal loop (NL), extracellular loop (EL), intracellular loop (IL) and C terminal loop (CL)), respectively. The circle filled in red, square filled in blue and hexagonal shape filled in green show positions of amino acid residue that are effective for discriminating $G_{i/o}$ and other G-protein binding types, $G_{q/11}$ and other G protein binding types, and $G_s$ and other G-protein binding types, respectively. The numbers within symbols are residue numbers of bovine rhodopsin used as template for structure based sequence alignment with GPCRs. In addition, ranges of residue numbers in each region are as follows: NL:1-36, TM1:37-63, IL1:64-73, TM2:74-96, EL1:97-110, TM3:111-133, IL2:134-152, TM4:153-173, EL2:174-202, TM5:203-224, IL3:225-252, TM6:253-274, EL3:275-286, TM7:287-308, CL:309. The residue numbers are based on bovine rhodopsin.

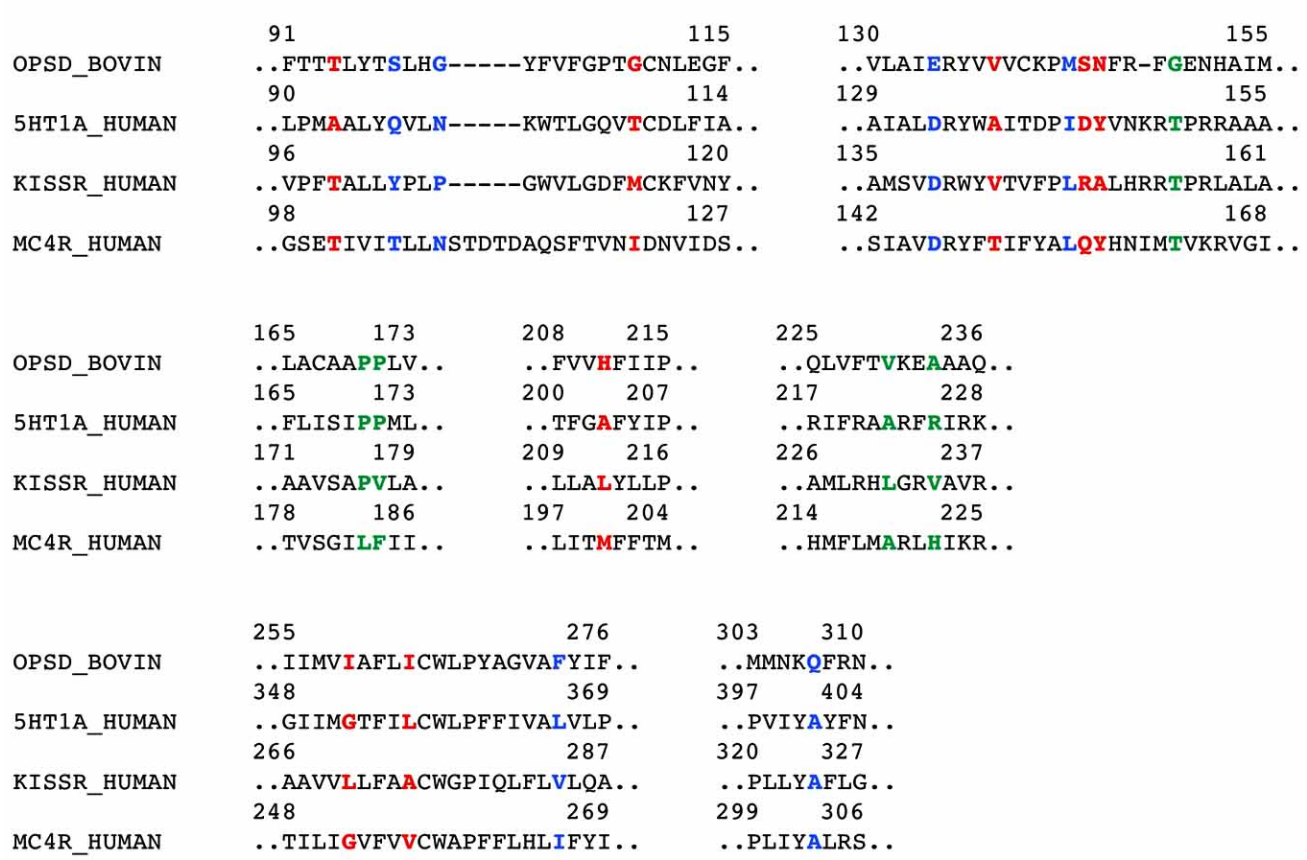


**Fig. (3). Alignments of GPCRs with bovine rhodopsin sequence (SWISS-PROT ID: OPSD_BOVIN).** 5HT1A_HUMAN, KISSR_HUMAN and MC4R_HUMAN (SWISS-PROT ID) are 5-hydroxytryptamine-1A, KISS1 and melanocortin 4 receptors, respectively. Numeric characters above each sequence line indicate residue numbers. Red, blue and green characters indicate amino acid residues that are effective for discriminating $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types, respectively. Hyphen symbol shows gaps between GPCR sequences.

coefficient (RP-HPLC, C18 with 0.1%TFA/MeCN/$H_2O$) [24] at positions 138, 144 and 145 in 2nd intracellular loop (IL2), respectively, normalized frequency of extended structure [25] at position 211 in TM5, transfer free energy (CHP/water) [26] and normalized average hydrophobicity scale [27] at positions 259 and 263, respectively. The position numbers correspond to the residue numbers in bovine rhodopsin used as template to align each GPCR. For above detected characteristics, the values of FR ranged from 0.70 to 0.98, and in particular, refractivity [23] and hydrophobicity coefficient (RP-HPLC, C18 with 0.1%TFA/MeCN/$H_2O$) [24] at positions 144 and 145 in IL2 showed high FR values, respectively. Site directed mutagenesis experiments demonstrated that the position 144 is critical for G-protein binding in 5-hydroxytryptamine-1A receptor, which is $G_{i/o}$ binding type [3]. Intriguingly, the position 144 in 5-hydroxytryptamine-1A receptor accords with the position 145 and the property at 145 discriminates $G_{i/o}$ binding type (Fig. **3**), which indicates its important role for coupling with $G_{i/o}$. The property at position 145 in IL2 was calculated as moving average of nine residues, which was influenced with neighboring four residues on both sides. Hence, the confidence level of the features at this position depends on the number of GPCR data available with reliable G-protein coupling information. The observation of three out of eight selected features located in IL2 was supported by previous reports that IL2 region plays an important role for GPCR - G-protein coupling [28]. On the subject of two features located on cytoplasmic side of TM6, previous studies have demonstrated that the opening of this cytoplasmic face of the receptor structure and G-protein coupling is caused by repulsive separation of the juxtacytoplasmic TM6 and TM3 in TSH receptor [29]. This result implies that the properties at positions 259 and 263 reflect the composition of any properties required in movement of TM6. Furthermore, as is obvious from Fig. (**2**), positions 94, 109 and 211 in TM2, EL1 and TM5 are located in extracellular side, respectively. These positions are clearly distant from G-protein binding site because G-protein interacts with intracellular loops. However, recent study has demonstrated that several different ligands lead to GPCR conformational changes [30]. Selectivity of G-protein binding type may affect with conformational changes caused by binding ligands, and those positions have possibilities in relation with $G_{i/o}$ binding.

The importance of amino acid properties of specific residues detected with high FR for $G_{i/o}$ binding type have been illustrated with a scatter plot of $G_{i/o}$ and other G-protein binding types (Fig. **4A**). Feature values of transfer free energy (AcWI-X-LL peptides from bilayer interface to water) [20] at position 94 in TM2 and normalized frequency of C terminal helix [21] at position 145 in IL2 are represented on horizontal and vertical axes, respectively. As

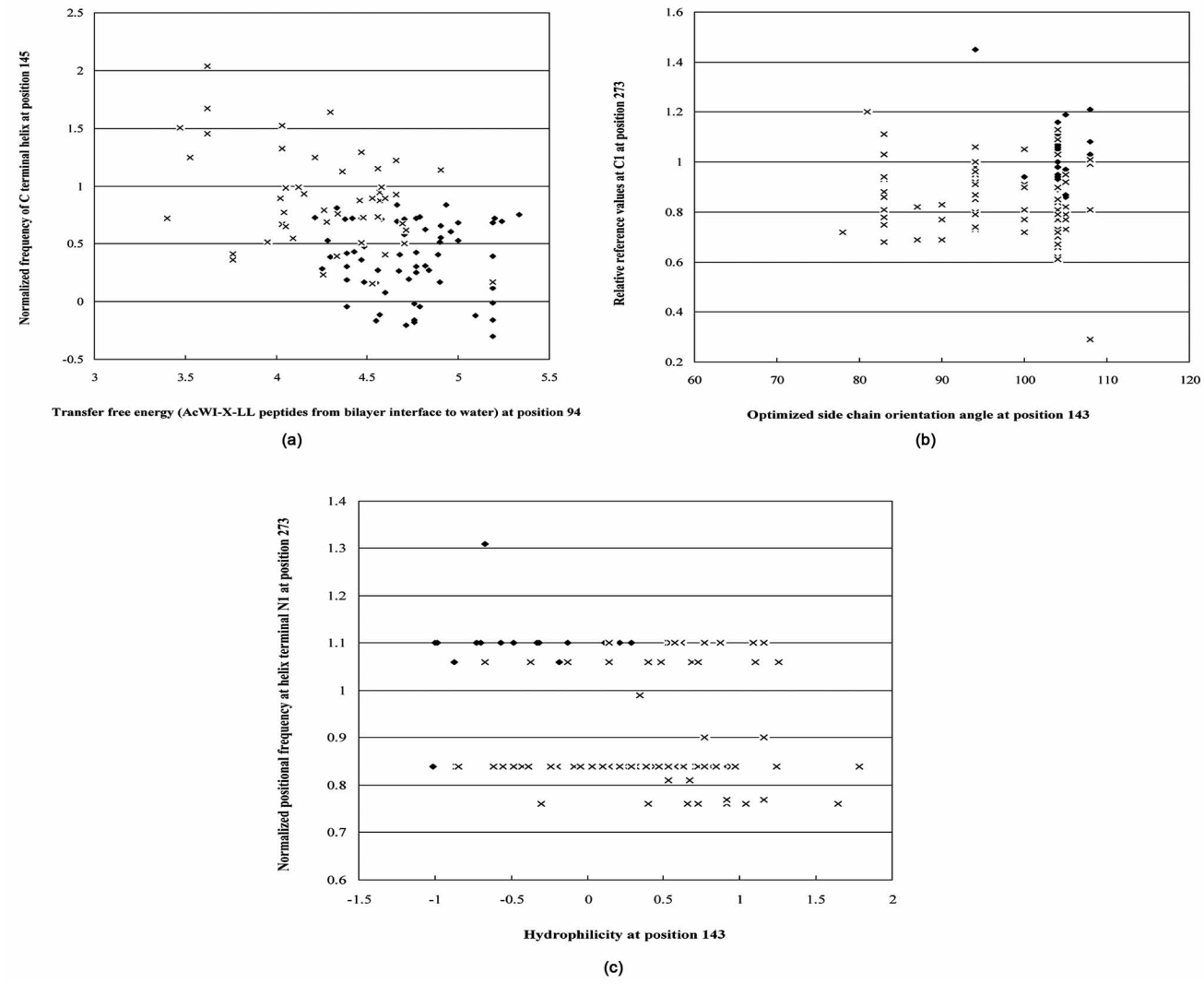


**Fig. (4). Scatter plots of feature values of specific amino acid residues in GPCRs.** The diamond and cross shapes represent positive and negative data, respectively. (**A**) Scatter plot of $G_{i/o}$ and other G-protein binding types. Feature values of transfer free energy (AcWI-X-LL peptides from bilayer interface to water) [20] at position 94 in TM2 and normalized frequency of C terminal helix [21] at position 143 in IL2 are represented on horizontal and vertical axes, respectively. (**B**) Scatter plot of $G_{q/11}$ and other G-protein binding types. Feature values of optimized side chain orientation angle [34] at position 143 in IL2 and relative reference value at C1 [35] at position 273 in TM6 are represented on horizontal and vertical axes, respectively. (**C**) Scatter plot of $G_s$ and other G-protein binding types. Feature values of hydrophilicity [39] at position 149 in IL2 and normalized positional residue frequency at helix terminal N1 [42] at position 230 in IL3 are represented on horizontal and vertical axes, respectively.

evidenced by the scatter plot, each distribution of $G_{i/o}$ and other G-protein binding types has a tendency of separation while there is intersection at the boundary of these distributions. This result shows that only those two parameters are inadequate to divide distributions with high accuracy and the additional other features are required for decreasing its intersection area.

The typical characteristics for the $G_{q/11}$ binding type are the following seven amino acid properties at specific residue positions as shown in Table **2B**: van der Waals parameter epsilon [31] and retention coefficient in $NaH_2PO_4$ [32] at positions 98 and 101 in EL1, respectively, optimized side chain interaction [33] and average side chain orientation angle [34] at positions 134 and 143 in IL2, respectively, relative preference value at C1 [35] at position 273 in TM6, and size [36] and number of atoms in the side chain labeled 2+1 [37] at position 307. Comparing to $G_{i/o}$ binding type, FR values of amino acid properties detected as characteristics of

$G_{q/11}$ binding type were slightly low. These values ranged from 0.61 to 0.92 and especially two amino acid properties, size [36] and number of atoms in the side chain labeled 2+1 [37], at positions 307 in TM7 showed high FR values. Han *et al.* [38] has reported that in M3 muscarinic acetylcholine receptor which is $G_{q/11}$ binding type, distance between the cytoplasmic ends of TM1 and TM7 is increased by agonist activation. This result is interesting because position 307 is located in C terminal of TM7 (Fig. **2**), which implies the possibility that above mentioned properties at position 307 have the tendency to change the conformation of GPCR to interact with $G_{q/11}$. Furthermore, site directed mutagenesis experiments demonstrated that the mutation of L148S in KISS1 receptor inhibits the catalytic activation of $G_q$ and L132S mutation of $\alpha_{1A}$-adrenergic receptor also recapitulates the effects as observed with L148S of KISS1, and suggested the critical importance of these residues for GPCR functional coupling [5]. Interestingly, positions 148 and 132 in IL2 in KISS1 and $\alpha_{1A}$-adrenergic receptor, respectively, correspond to the position 143 in IL2 of rhodopsin (Table **2**) which acted as properties for discriminating $G_{q/11}$ binding type (Fig. **3**). This result implies that the residue plays an important role for binding with $G_{q/11}$ and also supports the previous study that the IL2 region plays an essential role of coupling with G-protein [28]. Along with $G_{i/o}$ binding type, positions 98, 101 and 273 in EL1, EL1 and TM6 are located in extracellular side. These positions are distant from G-protein binding site and they have some sort of influence for $G_{q/11}$ binding.

Scatter plot of $G_{q/11}$ and other G-protein binding types is shown in Fig. (**4B**). Feature values of average side chain orientation angle [34] on position 143 in IL2 and relative reference value at C1 [35] on position 273 in TM6 are represented on horizontal and vertical axes, respectively. As observed for $G_{i/o}$ binding type, the distributions are comparatively divided although some of them are located at the intersection. This result also shows that additional features are required for discriminating the distributions with high accuracy.

The typical characteristics for the $G_s$ binding type are the following six amino acid properties at specific residue positions as shown in Table **2C**: hydrophilicity [39] at position 149 in IL2, hydropathy scale based on self-information values in the two-state model (36% accessibility) [40] and optimized average non-bonded energy per atom [33] at position 170 in TM4, free energy change of epsilon (i) to alpha (Rh) [41] at position 171 in TM4, and normalized positional residue frequency at helix termini N1 [42] and average relative fractional occurrence in AL (i) [43] at positions 230 and 233 in IL3, respectively. The FR values were remarkably very high when comparing other two G-protein types and these values ranged from 1.51 to 2.32. Kim *et al.* [4] have demonstrated that position 219 in the IL3 of melanocortin 4 receptor may favor interaction with the $G_s$ in the basal state. Interestingly, this position corresponds to the position 230 in IL3 shown in Table **2** as properties for discriminating $G_s$ binding type (Fig. **3**). This result implies that the position 230 in IL3 have important role for binding with $G_s$. In addition, Sugimoto *et al*. [44] have revealed that IL3 of prostaglandin EP2 receptor plays an essential role in $G_s$ activation using chimera experiment. Thus, positions 230 and 233 located in IL3 are reasonable for detecting as features to discriminate $G_s$ binding type. Furthermore, along with $G_{i/o}$ and $G_{q/11}$ binding types, a discrimination parameter was detected in IL2 (position 149) and this result bears out the possibility that position 149 in IL2 performs a crucial function in $G_s$ coupling process. In all detected features for discriminating $G_s$ binding type, three amino acid properties on positions 170 and 171 in extracellular side of TM4 were detected with particular high FR values. Although these positions are distant from intracellular side coupled with G-proteins, the high values of FR on the positions imply the possibilities that there are some kinds of relevance for coupling with $G_s$.

Scatter plot of $G_s$ and other G-protein binding types is shown in Fig. (**4C**). Feature values of hydrophilicity [39] on position 149 in IL2 and normalized positional residue frequency at helix termini N1 [42] on position 230 in IL3 are represented on horizontal and vertical axes, respectively. Comparing with scatter plots of $G_{i/o}$ and $G_{q/11}$ binding types, the distributions in $G_s$ binding type are clearly divided and there is a little intersection area. This result reflects the fact that FR values of $G_s$ binding type calculated in this study are high in comparison with other G-protein binding types.

The numerical values for the 20 amino acid residues for the amino acid properties selected in this work are given in supplemental Table **S2**.

## The Results of Discriminating G-protein Binding Type Using Support Vector Machines

The results obtained for discriminating $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types are shown in Table **3** and we noticed that all G-protein binding types were discriminated with high accuracy. The average specificity, sensitivity and MCC are respectively, 99%, 98% and 0.97. In particular, the $G_s$ binding type was the most successful and was achieved to completely classify without false positives and negatives, and the $G_{q/11}$ binding type was also predicted with no false positives. The scatter plots (Fig. **4**) showed that there are some mixed areas of positive and negative data around the boundary of distributions even as trending to separation. Eventually these mixed areas were discriminated with high accuracy by using SVM, which is appropriate method for high dimensional vectors, with feature vectors of eight, seven and six for $G_{i/o}$, $G_{q/11}$ and $G_s$ binding types, respectively. This result implies that the differences of amino acid residues and its properties involved in determining selectivity of GPCR - G-protein coupling are affected by delicate balances of structural information or physicochemical characteristics appropriate to some margins of structures in GPCRs. As for these results, amino acid properties of specific residues indicated in Table **2** are proved to be effective for discriminating G-protein binding types with high accuracy. Furthermore, some of them correspond to the residues that were verified to play crucial roles for binding G protein with experiments of amino acid alterations (mutation or deletion or insertion) [3-5]. This fact implies that those amino acid residues and the properties are important for investigating the mechanism of GPCR - G-protein coupling by experiments. However, the size of dataset used in this study is small because of pursuit of the reliability of G-protein binding information for each GPCR. Although this indication implies that accuracies of dis-

**Table 3. The Prediction Accuracy of SVM with 112 GPCRs**

| G-protein Binding Type | Number[a] | Specificity (%) | Sensitivity (%) | MCC |
|---|---|---|---|---|
| $G_{i/o}$ | 64 | 96.9 | 96.9 | 0.93 |
| $G_{q/11}$ | 29 | 100.0 | 96.6 | 0.98 |
| $G_s$ | 19 | 100.0 | 100.0 | 1.00 |

[a] Number: Number of GPCRs.

criminating G-protein binding types have the possibility to decrease with increasing numbers of GPCR with more reliable G-protein binding information, by using our method, increasing of numbers of GPCR will lead to extraction of characteristics that are more effective for further reliable prediction of G-protein binding types.

Recently, several methods have been reported for predicting GPCR - G-protein coupling selectivity using machine learning algorithms such as HMM [6], HMM and NN [7] and ACC transform based SVM [8]. These methods can predict GPCR - G-protein coupling selectivity with the sensitivity and specificity of over 90%. As a whole machine learning procedures are effective for discriminating G-protein binding types with high accuracy. Guo and coworkers [8] used various physicochemical properties to predict G-protein coupling types, however, it is unclear in the perspective that distinguish effective amino acid residues and its properties in terms of discriminating G-protein binding types. As far as Sgourakis and coworkers [7] and Sreekumar and coworkers [6] are concerned, intracellular loop regions of GPCRs are used as target regions for discriminating of G-protein binding types, however, it seems hard for experimentalists to know specific amino acid residues and the properties to be important for various experiments. When our method is compared with other previous methods, as shown in Table **3**, it is clear that our method is one of the most effective procedures for discriminating G-protein binding type. It is not appropriate to compare directly with other methods due to the differences of numbers of GPCR, its G-protein binding types and species of GPCRs. The advantage of our method is not only the prediction with high accuracy but also the identification of residue numbers and its amino acid properties that effectively discriminate G-protein binding types. In this regard, our method draws a line of demarcation from other methods. Our purpose is to come through concrete amino acid properties and the positions that are used as navigation marks for experiments with mutation or deletion of amino acid residues in GPCRs to investigate the mechanism of GPCR - G-protein coupling.

## CONCLUSIONS

In this study, we developed a novel method for predicting GPCR - G-protein coupling selectivity with high accuracy of more than 96% using the amino acid properties of specific residues by SVM. The novel point of our method is to detect amino acid properties of specific residues that are thought to play an important role for coupling with each G-protein binding type, and furthermore some of these residues have been verified by experiments of amino acid alteration such as mutation or chimera [3-5]. In addition, in all binding types ($G_{i/o}$, $G_{q/11}$ and $G_s$) our finding is interesting that some residues detected in IL2 are important for discriminating each G protein binding type. Experiments also support that IL2 of GPCR plays an important role for coupling with G-proteins [28]. Furthermore, in extracellular side, some properties were detected in all of G-protein binding types. The extracellular side is opposite to intracellular side and is distant from sites coupled with G-protein. This result implies the possibilities that there are some kinds of relation about coupling with G-protein because recent study has verified that several different ligands lead to GPCR conformational changes [30], and some properties of extracellular side detected in this study may be reflected to the conformational changes of GPCRs. It is necessary to point out that the size of dataset used in this study is small because of the pursuit of reliability of G-protein binding information, which may slightly alter the accuracy and selected amino acid properties of specific residues for discriminating G-protein binding types with increasing number of GPCRs and more reliable G-protein binding information. However, increasing of numbers of GPCR will lead to extraction of characteristics that are more effective for further reliable discrimination of G-protein binding types. Our study will make effective contribution for investigating the mechanism of coupling GPCR with G-protein by site directed mutagenesis experiments.

## ACKNOWLEDGEMENT

We are grateful to Dr. Shinsuke Yamada (CBRC, AIST) for helpful discussions.

## SUPPLEMENTARY MATERIAL

Supplementary material can be viewed at www.bentham.org/open/tosbj

## REFERENCES

[1] Bridges TM, Lindsley CW. G-protein-coupled receptors: from classical modes of modulation to allosteric mechanisms. ACS Chem Biol 2008; 3: 530-41.

[2] Mao J, Xie W, Yuan H, Simon MI, Mano H, Wu D. Tec/Bmx non-receptor tyrosine kinases are involved in regulation of Rho and serum response factor by Galpha12/13. EMBO J 1998; 17: 5638-46.

[3] Kushwaha N, Harwood SC, Wilson AM, *et al*. Molecular determinants in the second intracellular loop of the 5-hydroxy-tryptamine-1A receptor for G-protein coupling. Mol Pharmacol 2006; 69: 1518-26.

[4] Kim DH, Shin SW, Baik JH. Role of third intracellular loop of the melanocortin 4 receptor in the regulation of constitutive activity. Biochem Biophys Res Commun 2008; 365: 439-45.

[5] Wacker JL, Feller DB, Tang XB, *et al*. Disease-causing mutation in GPR54 reveals the importance of the second intracellular loop for class A G-protein-coupled receptor function. J Biol Chem 2008; 283: 31068-78.

[6] Sreekumar KR, Huang Y, Pausch MH, Gulukota K. Predicting GPCR–G-protein coupling using hidden Markov models. Bioinformatics 2004; 20: 3490-9.
[7] Sgourakis NG, Bagos PG, Hamodrakas SJ. Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. Bioinformatics 2005; 21: 4101-6.
[8] Guo Y, Li M, Lu M, Wen Z, Huang Z. Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins 2006; 65: 55-60.
[9] Alexander SPH, Mathie A, Peters JA. Guide to Receptors and Channels (GRAC), 3rd edition. Br J Pharmacol 2008; 153 (Suppl 2): S1-S209.
[10] Theodoropoulou MC, Bagos PG, Spyropoulos IC, Hamodrakas SJ. gpDB: a database of GPCRs, G-proteins, effectors and their interactions. Bioinformatics 2008; 24: 1471-2.
[11] Palczewski K, Kumasaka T, Hori T, *et al*. Crystal structure of rhodopsin: a G protein-coupled receptor. Science 2000; 289: 739-45.
[12] Shimamura T, Hiraki K, Takahashi N, *et al*. Crystal structure of squid rhodopsin with intracellularly extended cytoplasmic region. J Biol Chem 2008; 283: 17753-6.
[13] Warne A, Serrano-Vega MJ, Baker JG, *et al*. Structure of the beta1-adrenergic G protein-coupled receptor. Nature 2008; 454: 486-91.
[14] Cherezov V, Rosenbaum DM, Hanson MA, *et al*. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 2008; 318: 1258-65.
[15] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22: 4673-80.
[16] Liu Q, Zhu Y, Wang B, Li Y. Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. Comput Biol Chem 2003; 27: 355-61.
[17] Park KJ, Gromiha MM, Horton P, Suwa M. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 2005; 21: 4223-9.
[18] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. 2001. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[19] Matthews BW. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405: 442-51.
[20] Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 1996; 10: 842-8.
[21] Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978; 47: 45-148.
[22] Avbelj F. Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins. J Mol Biol 2000; 300: 1335-59.
[23] McMeekin TL, Groves ML, Hipp NJ. Refractive Indices of Amino Acids, Proteins and Related Substances. In: Stekol JA, Ed. Amino Acids and Serum Proteins. American Chemical Society: Washington DC 1964; pp. 54-66.
[24] Wilce MC, Aguilar MI, Hearn MT. Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides. Anal Chem 1995; 67: 1210-9.
[25] Tanaka S, Scheraga HA. Statistical mechanical treatment of protein conformation. 5. A multistate model for specific-sequence copolymers of amino acids. Macromolecules 1977; 10: 9-20.
[26] Lawson EQ, Sadler AJ, Harmatz D, *et al*. A simple experimental model for hydrophobic interactions in proteins. J Biol Chem 1984; 259: 2910-2.
[27] Cid H, Bunster M, Canales M, Gazitua F. Hydrophobicity and structural classes in proteins. Protein Eng 1992; 5: 373-5.
[28] Wess J. Molecular basis of receptor/G-protein-coupling selectivity. Pharmacol Ther 1998; 80: 231-64.
[29] Ringkananont U, Van Durme J, Montanelli L, *et al*. Repulsive separation of the cytoplasmic ends of transmembrane helices 3 and 6 is linked to receptor activation in a novel thyrotropin receptor mutant (M626I). Mol Endocrinol 2006; 20: 893-903.
[30] Hoffmann C, Zürn A, Bünemann M, Lohse MJ. Conformational changes in G-protein-coupled receptor--the quest for functionally selective conformations is open. Br J Pharmacol 2008; 153: 5358-66.
[31] Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976; 104: 59-107.
[32] Meek JL, Rossetti ZL. Factors affecting retention and resolution of peptides in high-performance liquid chromatography. J Chromatogr 1981; 211: 15-28.
[33] Oobatake M, Kubota Y, Ooi T. Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteins. Bull Inst Chem Res Kyoto Univ 1985; 63: 82-94.
[34] Meirovitch H, Rackovsky S, Scheraga HA. Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. Macromolecules 1980; 13: 1398-1405.
[35] Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science 1988; 240: 1648-52.
[36] Dawson DM. In: Brock DJH, Mayo O. (eds). The Biochemical Genetics of Man. Academic Press, New York, pp. 1-38, 1972.
[37] Charton M, Charton BI. The dependence of the Chou-Fasman parameters on amino acid side chain structure. J Theor Biol 1983; 102: 121-34.
[38] Han SJ, Hamdan FF, Kim SK, *et al*. Identification of an agonist-induced conformational change occurring adjacent to the ligand-binding pocket of the M (3) muscarinic acetylcholine receptor. J Biol Chem 2005; 280: 34849-58.
[39] Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981; 78: 3824-8.
[40] Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001; 42: 452-9.
[41] Wertz DH, Scheraga HA. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules 1978; 11: 9-15.
[42] Aurora R, Rose GD. Helix capping. Protein Sci 1998; 7: 21-38.
[43] Rackovsky S, Scheraga HA. Differential geometry and polymer conformation 4. conformational and nucleation properties of individual amino acids. Macromolecules 1982; 15: 1340-6.
[44] Sugimoto Y, Nakato T, Kita A, *et al*. Functional domains essential for Gs activity in prostaglandin EP2 and EP3 receptors. Life Sci 2003; 74: 135-41.

Received: February 04, 2009 Revised: February 23, 2009 Accepted: February 23, 2009