W148–W153 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki495

GRIFFIN: a system for predicting GPCR–G-protein coupling selectivity using a support vector machine and a hidden Markov model

Yukimitsu Yabuki[1,2], Takahiko Muramatsu[1,3], Takatsugu Hirokawa[1], Hidehito Mukai[4] and Makiko Suwa[1,3,*]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [2]Information and Mathematical Science Laboratory (IMS) Inc., Meikei Building, 1-5-21 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan, [3]Nara Institute of Science and Technology, Graduate School of Information Science, 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan and [4]Mitsubishi Kagaku Institute of Life Sciences, 11 Minamiooya, Machida, Tokyo 194-8511, Japan

Received February 14, 2005; Revised April 15, 2005; Accepted April 26, 2005

ABSTRACT

We describe a novel system, GRIFFIN (G-protein and Receptor Interaction Feature Finding INstrument), that predicts G-protein coupled receptor (GPCR) and G-protein coupling selectivity based on a support vector machine (SVM) and a hidden Markov model (HMM) with high sensitivity and specificity. Based on our assumption that whole structural segments of ligands, GPCRs and G-proteins are essential to determine GPCR and G-protein coupling, various quantitative features were selected for ligands, GPCRs and G-protein complex structures, and those parameters that are the most effective in selecting G-protein type were used as feature vectors in the SVM. The main part of GRIFFIN includes a hierarchical SVM classifier using the feature vectors, which is useful for Class A GPCRs, the major family. For the opsins and olfactory subfamilies of Class A and other minor families (Classes B, C, frizzled and smoothened), the binding G-protein is predicted with high accuracy using the HMM. Applying this system to known GPCR sequences, each binding G-protein is predicted with high sensitivity and specificity (>85% on average). GRIFFIN (http://griffin.cbrc.jp/) is freely available and allows users to easily execute this reliable prediction of G-proteins.

INTRODUCTION

G-protein coupled receptors (GPCRs) with seven transmembrane helices are the major membrane proteins that play the important interface role for signaling to the inner cell. An external ligand stimulus to a GPCR induces the coupling with G-proteins ($G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$) followed by different kinds of signal transduction. Since about half (1) of all drugs distributed throughout the world are designed to control these mechanisms, GPCRs are important targets in the development of effective drugs.

From the viewpoint of drug design, it will be of utmost importance to screen a drug for its ability to effectively control the activation of a specific G-protein, by monitoring the stimulation by different ligands. In general, it is quite difficult to develop such a high-throughput experimental system; however, G-protein activity prediction made using bioinformatics techniques contributes to the design of an effective experimental system. Therefore, our purpose is to develop a program to predict GPCR–G-protein binding selectivity when both the GPCR sequence and ligand information are submitted.

The established way to predict protein function is to classify proteins into functional groups whose members are linked by sequence similarity using a conventional sequence search method such as BLAST (2) and FASTA (3). However, in the case of GPCRs, the function-similarity relationship is unclear. For example, (i) some homologous GPCR pairs with the same ligands bind to different kinds of G-protein; (ii) those pairs that bind to the same type of G-protein bind to a different ligand; and furthermore (iii) some GPCR pairs bind to both the same ligand and the same G-protein even though they show sequence similarity of <25% (4). Given this situation, various computational methods have been developed to understand the GPCR signaling mechanisms using not simple sequence searches but more powerful methods such as hidden Markov models (HMMs), support vector machines (SVMs) and statistical analysis. These methods are divided into two

*To whom correspondence should be addressed. Tel: +81 3 3599 8051; Fax: +81 3 3599 8081; Email: m-suwa@aist.go.jp

main branches: classification of GPCRs by ligand type (5–10) and classification of GPCRs by G-protein type (11–13). As a result, classification cannot be determined by the relationship between the external ligand and the G-protein type.

Compared with previous work, our work is unique because we intend to develop a program for predicting GPCR–G-protein coupling specificity from the ligand information as well as the GPCR sequence. To develop this program, we assume that the ligand, GPCR and G-protein form a complex, and therefore that structural information about the ligand, extracellular loops, intracellular loops and the transmembrane domain of GPCRs is essential for describing the binding of G-proteins. Since the SVM algorithm has been verified to be a high-performance classifier, especially for discriminating multidimensional parameters (5), we have used the SVM method in this work. We collected the combination of existing ligand, GPCR and G-protein, picked up various characteristic quantitative features as feature vector elements in the SVM from their structural information and, from these quantities, determined G-protein type. The predicting system includes a hierarchical SVM classifier using the feature vectors, which is useful for the Class A GPCR group. For opsins and olfactory receptors belonging to Classes A, B, C, frizzled and smoothened families, we apply an HMM classification, since these subfamilies can be directly assigned to a G-protein type. Thus we constructed the hierarchical system including the HMM and SVM components. Applying this system to known GPCR sequences, each binding G-protein is predicted with high sensitivity and specificity (>85% on average). Based on this study, we developed a GRIFFIN web server (http://griffin.cbrc.jp/) that can predict G-protein coupling specificity using the SVM and HMM methods.

METHODS

In order to predict GPCR–G-protein coupling selectivity, GRIFFIN implements two processes, SVM and HMM. The SVM process is suitable for predicting G-protein coupling selectivity for the Class A GPCR family. It is well known that the Class A GPCR family is huge, as it is the major family, and its large-scale diversity makes it difficult to predict its coupling G-proteins using only sequence similarity information. Therefore, we first applied the SVM method using characteristic quantities extracted from the ligand information and GPCR structure.

The HMM method is suitable for predicting the coupling selectivity of G-proteins with GPCRs belonging to opsins and olfactory receptors (in Class A), Class B, C, frizzled and smoothened families. Although it is still unclear what kind of G-protein binds to frizzled and smoothened GPCRs, these two families can be used as a filter to classify other GPCRs. For these families, the G-protein prediction is easier because sequence similarity information (described in terms of the HMM) directly correlates with functional annotation of the binding G-protein type.

The SVM and HMM calculations were performed using the LIBSVM (14) and HMMER (15) software packages, respectively. The detailed parameters and thresholds used are described below.

Training dataset

For Class A GPCRs, amino acid sequences were obtained from the SwissProt and TrEMBL databases. These GPCR sequences include both ligand and G-protein information written in TiPS (16) and GPCRDB (17). The number of Class A GPCR sequences selected as training data for SVM classification is 132 ($G_{i/o}$ binding type: 61 sequences; $G_{q/11}$ binding type: 47 sequences; G_s binding type: 24 sequences). And in this work, GPCRs which are coupled with multiple G-proteins and the $G_{12/13}$ G-protein family are not considered because there are not sufficient data to construct a prediction system.

The redundancy of these sequences was evaluated by analyzing clusters formed under sequence similarity set to decrease from 100% to 30% in steps of 1% using BLASTCLUST from the BLAST software package (2). One cluster consisted of two GPCRs (SwissProt IDs PKR1_HUMAN and PKR2_HUMAN) and appeared at 87% sequence identity, and the other clusters were not detected until the sequence identity reached 68%. This result shows that most GPCRs do not have strong similarities with each other. Though PKR1_HUMAN and PKR2_HUMAN show strong sequence similarity, as described above, they bind to different ligands and, therefore, both sequences should be used as training datasets. For this reason, 132 sequences are used in this work without a process of elimination of redundancy.

For opsins and olfactory receptors, Classes B, C, frizzled and smoothened families, sequences were obtained from the SwissProt and TrEMBL databases as well as the above-mentioned 132 Class A GPCRs. Class C GPCR sequences can be classified into two types, $G_{i/o}$ binding type and $G_{q/11}$ binding type; therefore, this family is separately collected into two groups. The numbers of GPCRs are 170, 394, 34, 20, 9, 40 and 5 for opsins, olfactory receptors, Class B, Class C for $G_{i/o}$ specific, Class C for $G_{q/11}$ specific, frizzled and smoothened families, respectively.

Determination of characteristic quantities used in SVM

To develop the program, we assumed that the ligand, GPCR and G-protein form a complex, and that therefore the interactions among this complex are all essential factors for activating G-protein bindings. From this viewpoint, structural characteristics should be extracted comprehensively from the ligand, extracellular loops, intracellular loops and transmembrane domain of GPCRs, although the tertiary positions of some characteristics are distant from the G-protein binding site.

To calculate these parameters, the boundaries of the transmembrane helix and loop regions of GPCR sequences were determined from multiple alignments of known Class A families with bovine rhodopsin as a three-dimensional structure template (PDB ID: 1f88) using CLUSTAL W (18).

In addition to the above parameters, two bit scores are calculated from GPCR sequences. One is obtained when a query GPCR sequence is searched against the HMM profile (peptide profile) constructed from multiple alignments of GPCR groups binding to a peptide ligand. The other is calculated by the HMM profile (amine profile) of a GPCR group which is bound to a small amine ligand. Each GPCR in these two groups was obtained from SwissProt: 439 and 243 for the

peptide ligand type and the small amine ligand type, respectively. Detailed parameter information is listed in Table 1. A Class A GPCR can be plotted to multidimensional space using a vector composed of these multiple parameters.

SVM classifies these vector representations (feature vectors) of GPCRs using a multidimensional hyperplane called the kernel function. Since the SVM is a classifier used to divide data into two groups, classifications such as (G_s binding type and others), ($G_{q/11}$ binding type and others) and ($G_{i/o}$ binding type and others) are performed.

In order to calculate the accuracy in discriminating each G-protein type ($G_{i/o}$, $G_{q/11}$ and G_s for the training dataset containing $G_{i/o}$, $G_{q/11}$ and G_s binding GPCRs), SVM training was performed by changing the combination of the feature vector elements, kernel functions (linear, polynomial, RBF and sigmoid formula) with parameters C and γ, which determine the shape of kernel function. A cross-validation test is performed for each combination of parameter sets. The variable ranges of the parameters C, γ and cross-validation fold are from 2^{-5} to 2^{15}, from 2^{-13} to 2^3, and from 2 to 5, respectively.

The best combination of feature vector elements and kernel functions is determined when the product of sensitivity and specificity shows the highest value of accuracy for evaluating G-protein coupling prediction. As indicated in Table 2, the discrimination of the G_s binding type is the most successful, with the following five feature vector elements: (i) the third intracellular loop length, (ii) the C-terminal loop length, (iii) the total number of arginines and lysines in the C-terminal region of the intracellular loop, (iv) the existence of proline at the position corresponding to the 170th residue on rhodopsin, and (v) the bit score of the amine profile. However, under the same condition, $G_{i/o}$ and $G_{q/11}$ types cannot be classified with high accuracy. Thus, in order to predict $G_{i/o}$ or $G_{q/11}$ from the two proteins with high accuracy, SVM training was performed again. The best performance results for $G_{i/o}$ and $G_{q/11}$ classifications are shown in Table 3. The highest sensitivity and specificity for classifying as $G_{i/o}$ type or $G_{q/11}$ type were achieved when five parameters [(i), (ii), (v) and two additional parameters: (vi) the bit score of the peptide profile and (vii) the ligand molecular weight] were used.

Table 1. Feature quantities used in SVM training as feature vector elements

Feature quantities from structural information of ligands and GPCRs
1. Length of N-terminal loop
2. Length of the first intracellular loop between TMH1 and TMH2
3. Length of the first extracellular loop between TMH2 and TMH3
4. Length of the second intracellular loop between TMH3 and TMH4
5. Length of the second extracellular loop between TMH4 and TMH5
6. Length of the third intracellular loop between TMH5 and TMH6
7. Length of the third extracellular loop between TMH6 and TMH7
8. Length of the C-terminal loop
9. Averaged hydrophobicity of TMH1
10. Averaged hydrophobicity of TMH2
11. Averaged hydrophobicity of TMH3
12. Averaged hydrophobicity of TMH4
13. Averaged hydrophobicity of TMH5
14. Averaged hydrophobicity of TMH6
15. Averaged hydrophobicity of TMH7
16. Bit score calculated when a query is searched against a profile which is made from sequences of amine binding GPCRs
17. Bit score calculated when a query is searched against a profile which is made from sequences of peptide binding GPCRs
18. Existence of Pro on the position corresponding to the 170th residue on BOVIN rhodopsin
19. Existence of Lys or Arg on the position corresponding to 148th residue on BOVIN rhodopsin
20. Molecular weight of the ligand
21. Number of Lys or Arg corresponding to the 244th, 247th, 248th and 251st residues on the third intracellular loop of BOVIN rhodopsin
22. Number of Lys or Arg corresponding to the 243rd, 244th, 247th, 248th and 251st residues on the third intracellular loop of BOVIN rhodopsin
23. Number of Phe, His, Tyr or Trp that exist in the C-terminal residue of the third intracellular loop to the 9th residue in the N-terminal residue of this loop
24. Number of Asp or Glu that exist in the third intracellular loop

TMH: transmembrane helix.

Making HMM profiles

HMM profiles were made from each member of the opsins, olfactory receptors, Classes B, C, frizzled and smoothened families using the HMMER program (15) (Class C HMM profiles were made separately from two groups which bind to $G_{i/o}$ or $G_{q/11}$). To verify the reliability of each profile used for prediction, all GPCRs were first picked up from GPCRDB to add as false data for each family. Each family was divided into four subgroups and 4-fold cross-validation tests were executed to verify the reliability of HMM profiles (that is, three-fourths of the subgroups are used as training datasets and the remaining fourth are used as test data). As a result, for each HMM profile, we determined the safe threshold score to discriminate subfamilies with the highest sensitivity and specificity (Table 4). As shown in Table 4, most families can be predicted with 100% sensitivity and specificity at each threshold. These results suggest that for each family, sequence information (described in terms of the HMM) corresponds to

Table 2. The prediction accuracy of SVM part performed with 132 GPCRs

G-protein type	*n*	Sensitivity (%)	Specificity (%)	Number of cross-validations	Best kernel function
$G_{i/o}$	61	77.0	78.3	4	RBF
$G_{q/11}$	47	68.1	72.7	4	RBF
G_s	24	83.3	95.2	4	RBF

Table 3. The prediction accuracy of SVM part performed with 108 GPCRs

G-protein type	*n*	Sensitivity (%)	Specificity (%)	Fold number of cross-validation	Best kernel function
$G_{i/o}$	61	91.8	94.9	4	Polynomial
$G_{q/11}$	47	93.6	89.8	4	Polynomial

Table 4. The prediction accuracy of HMM part performed with 4-fold cross-validation

Family	G-protein type	Sensitivity (%)	Specificity (%)	Threshold of bit score
Opsin	G_t	99.7	100.0	153.9
Olfactory	G_{olf}	100.0	100.0	151.2
Class B	G_s	100.0	100.0	68.0
Class C	$G_{i/o}$	93.5	100.0	1054.6
	$G_{q/11}$	100.0	100.0	1325.3
Frizzled	Unclear	100.0	100.0	168.7
Smoothened	Unclear	100.0	100.0	627.6

specific G-protein type: opsins bind to G_t, olfactory receptors bind to G_{olf}, most of the Class B family binds to G_s, and the Class C family binds to $G_{i/o}$ or $G_{q/11}$. The type of G-protein binding to the frizzled and smoothened families is still unclear; therefore, these GPCRs can be classified as 'unknown G-protein type'. Thus, an HMM profile search can directly link the G-protein information, and these profiles are useful filters in the classification of the Class A GPCRs and others.

The integrated system for predicting GPCR–G-protein coupling selectivity

The integrated system for predicting GPCR–G-protein coupling selectivity is shown as the flowchart in Figure 1. As input data, this system requires the sequence of query GPCR and ligand molecular weight, which are converted to feature vector elements. At the first stage, a query sequence is searched against the HMM profiles of the opsins and olfactory receptor subfamilies, Classes B, C, frizzled and smoothened families by the HMMER program (15) with high accuracy, as shown in Table 4. If the computed HMM profile score is larger than the threshold of a certain subfamily (see Table 4), the query sequence can immediately link to the G-protein information, and GRIFFIN stops the prediction process.

However, if the query does not meet the above conditions (i.e. all profile scores are less than each corresponding family threshold), GRIFFIN continues the processing to the second stage, which uses the SVM with feature vectors which are converted from sequence and ligand molecular weight. Since the prediction of G_s from other G-proteins and $G_{i/o}$ or $G_{q/11}$ in these two proteins requires different parameter sets and conditions to achieve the best performance, we constructed the following hierarchical system. First, this system determines whether the query sequence is binding to G_s by using five parameters and the RBF function. If the query is predicted to be of the G_s binding type (with 95.2% specificity and 83.3% sensitivity, Table 2), this result is displayed and GRIFFIN stops the prediction process. If it is predicted not to

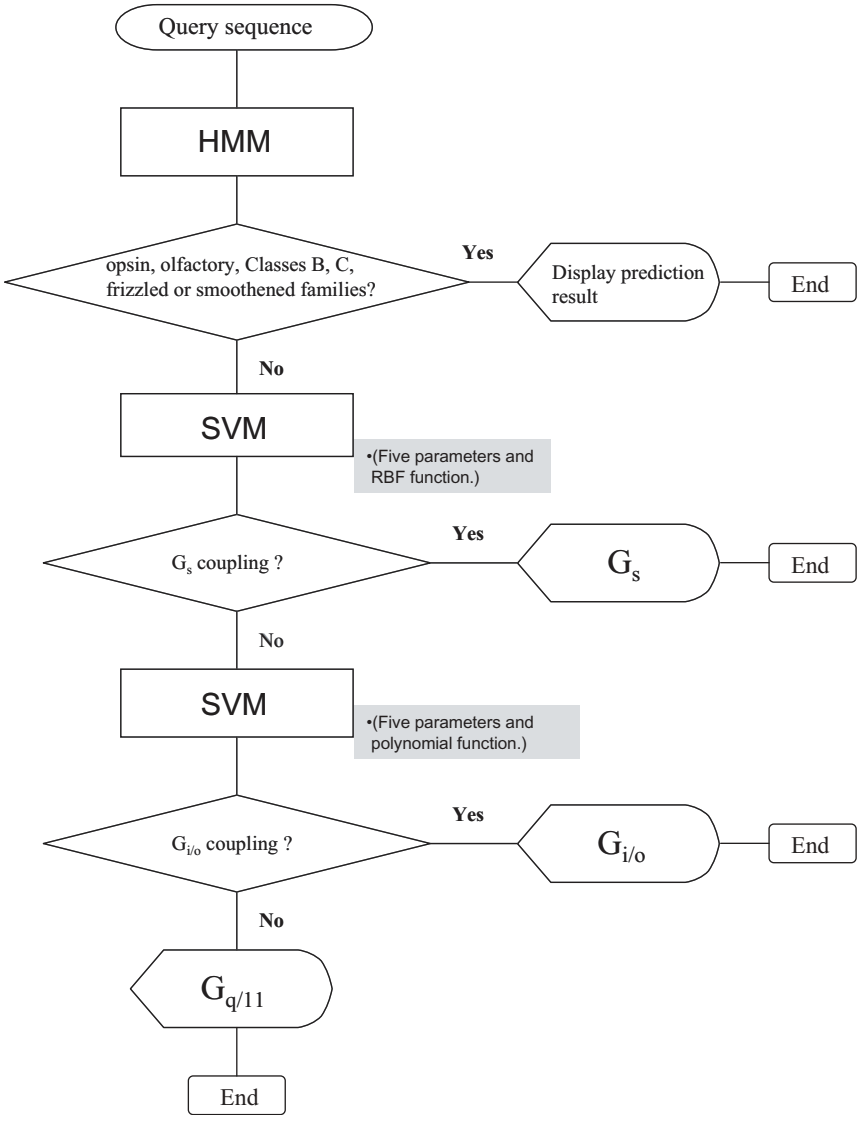

Figure 1. A flowchart of the integrated system for predicting GPCR–G-protein coupling selectivity.

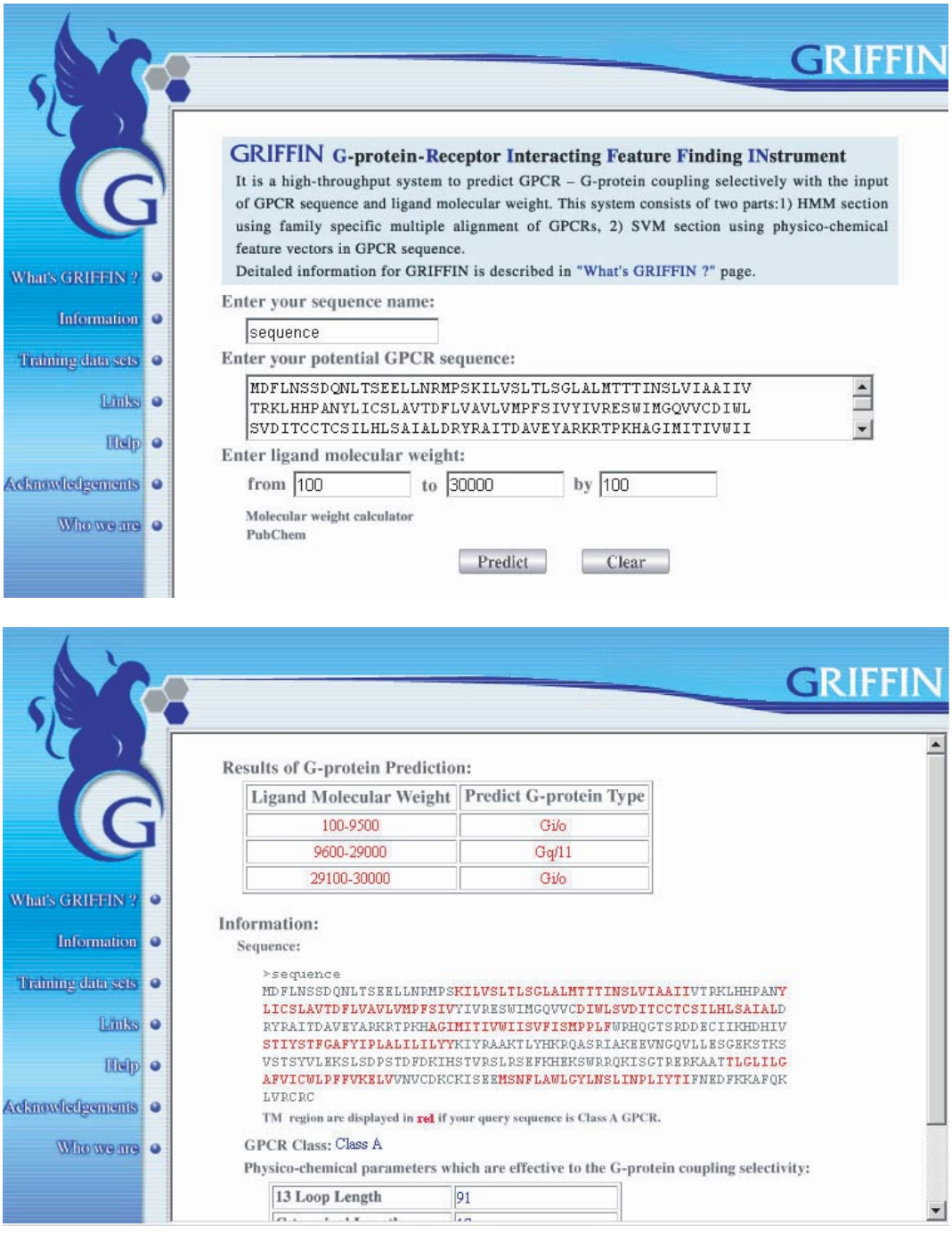

Figure 2. (**a**) The top of the GRIFFIN website, where the GPCR sequence and ligand molecular weight can be entered. (**b**) The result page of a GRIFFIN calculation, where the predicted G-proteins of the user-defined sequence are indicated together with physicochemical parameters used in the SVM or HMM calculation.

be of the G_s binding type, GRIFFIN changes the process to predict $G_{i/o}$ or $G_{q/11}$ coupling selectivity by using the other five parameters and the polynomial function, with high sensitivity and specificity as shown in Table 3. After applying this hierarchical system to known sequences through 10 000 rounds of 4-fold cross-validation, the average discrimination sensitivities and specificities were 87% and 88% for $G_{i/o}$, 85% and 84% for $G_{q/11}$, and 85% and 89% for G_s, respectively. In previous studies, three methods (11–13) were developed in order to predict G-protein binding selectivity. The method of Cao *et al*. is based on the naive Bayes model and it predicted the G-protein with 72% sensitivity from 55 GPCRs (11). Möller *et al*. indicated >90% specificity with 30–40% sensitivity using pattern extraction (12). Sreekumar *et al*. succeeded in reducing the error rate of prediction to <1% (13) using HMM classification. Our method shows better accuracy than Cao *et al*. and Möller *et al*. since it can predict G-proteins with both high sensitivity and specificity of >85%. HMM profiles by Sreekumar *et al*. indicated higher performance of prediction compared with our method.

However, it is difficult to compare their performance with our method because the GPCR sequences used in their dataset and their evaluation methodology are different from ours. Most importantly, our method is the first one to predict G-protein types by inputting both ligand molecular weight and GPCR sequence, and this prediction processing is available on the useful web server GRIFFIN.

THE WEB SERVER: GRIFFIN

Figure 2a shows the home page of the GRIFFIN website (http://griffin.cbrc.jp/). On this home page, there are small and large text boxes for entering a sequence name and an amino acid sequence, respectively. The three small text boxes at the bottom of the page are for entering the range of ligand molecular weight (onset, termination and differential values) from left to right. With this function, by changing the range of ligand molecular weight, the user can perform the computational experiment to monitor G-protein binding for orphan receptors whose ligands are still unknown. Of course, the user can also predict the type of G-protein by entering only one value for ligand molecular weight. If 'Molecular weight calculator' is clicked, it navigates to a page where the user can calculate the molecular weight of a chemical compound when the chemical equation is entered in the text box and the 'submit' button is clicked. To calculate the molecular weight of the peptide ligand, this page is linked to the PeptideMass website (19). The PubChem website, which contains chemical compound information, is also available via a link at the top of the page.

When the 'Predict' button is clicked, the GRIFFIN system navigates to the results page (Figure 2b). When the user enters a range of ligand molecular weight, and if this range matches a certain G-protein type, the results are displayed with each line representing a predicted G-protein type. For example, Figure 2b shows the result when a wide molecular range (from 100 to 30 000 in steps of 100) is entered; this query sequence changes between the coupling G-proteins $G_{i/o}$ and $G_{q/11}$. The query sequence and user-defined name are displayed in the FASTA format, with transmembrane regions colored in red, when the query GPCR belongs to the Class A family. In addition, feature vector elements and their scores, which are calculated in the process of prediction, are displayed in a table.

We believe that GRIFFIN will contribute to the research into functional assignment of orphan GPCRs and to the design of experimental systems for screening effective drugs.

ACKNOWLEDGEMENTS

We would like to thank Dr Taisin Kin for useful discussions pertaining to SVMs. Funding to pay the Open Access publication charges for this article was provided by the grant-in-aid for AIST and Mitsubishi Kagaku collaboration.

Conflict of interest statement. None declared.

REFERENCES

1. Drews,J. (1996) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.*, **14**, 1516–1518.
2. Altschul,S.F., Adden,T.L., Schaffer,A.A., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped Blast and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
4. Gaulton,A. and Attwood,T.K. (2003) Bioinformatics approaches for the classification of G-protein-coupled receptors. *Curr. Opin. Pharmacol.*, **3**, 114–120.
5. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
6. Bhasin,M. and Raghava,G.P.S. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.*, **32**, 383–389.
7. Lapnish,M., Gutcaits,A., Prusis,P., Post,C., Lundstedt,T. and Wikberg,J.E.S. (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principle chemical properties of primary amino acid sequences. *Protein Sci.*, **11**, 795–805.
8. Huang,Y., Cai,J., Ji,L. and Li,Y. (2004) Classifying G-protein coupled receptors with bagging classification tree. *Comput. Biol. Chem.*, **28**, 275–280.
9. Qian,B., Soyer,O.S., Neubig,R.R. and Goldstein,R.A. (2003) Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett.*, **554**, 95–99.
10. Attwood,T.K., Croning,M.D.R. and Gaulton,A. (2001) Deriving structural and functional insights from a ligand-based hierarchical classification of G-protein-coupled receptors. *Protein Eng.*, **15**, 7–12.
11. Cao,J., Panetta,R., Yue,S., Steyaert,A., Young-Bellido,M. and Ahmad,S. (2003) A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics*, **19**, 234–240.
12. Möller,S., Vilo,J. and Croning,M.D.R. (2001) Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics*, **17**, 174–181.
13. Sreekumar,K.R., Huang,Y., Pausch,M.H. and Gulukota,K. (2004) Predicting GPCR–G-protein coupling using hidden Markov models. *Bioinformatics*, **20**, 3490–3499.
14. Chang,C.-C. and Lin,C.-J. (2001) Training nu-support vector classifiers: theory and algorithms. *Neural Comput.*, **9**, 1443–1471.
15. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
16. Alexander,S., Peters,J., Mead,A. and Lewis,S. (Eds) (1999) TiPS receptor and ion channel nomenclature supplement. *Trends Pharmacol. Sci.*, **19**, 5–85.
17. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G-protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
18. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
19. Wilkins,M.R., Lindskog,I., Gasteiger,E., Bairoch,A., Sanchez,J.C., Hochstrasser,D.F. and Appel,R.D. (1997) Detailed peptide characterisation using PEPTIDEMASS—a World-Wide-Web-accessible tool. *Electrophoresis*, **18**, 403–408.