Prediction of GPCR-G Protein Coupling Specificity Using Features of Sequences and Biological Functions

Toshihide Ono* and Haretsugu Hishigaki

Laboratory of Bioinformatics, Otsuka Pharmaceutical Co., Ltd., Kawauchi-cho, Tokushima 771-0192, Japan.

Understanding the coupling specificity between G protein-coupled receptors (GPCRs) and specific classes of G proteins is important for further elucidation of receptor functions within a cell. Increasing information on GPCR sequences and the G protein family would facilitate prediction of the coupling properties of GPCRs. In this study, we describe a novel approach for predicting the coupling specificity between GPCRs and G proteins. This method uses not only GPCR sequences but also the functional knowledge generated by natural language processing, and can achieve 92.2% prediction accuracy by using the C4.5 algorithm. Furthermore, rules related to GPCR-G protein coupling are generated. The combination of sequence analysis and text mining improves the prediction accuracy for GPCR-G protein coupling specificity, and also provides clues for understanding GPCR signaling.

Key words: GPCR, G protein, coupling specificity, NLP, C4.5, text mining

Introduction

Cell activity is regulated by extracellular signals that are transmitted into the cell interior through different classes of plasma membrane receptors. The vast majority of these receptors belong to the superfamily of G protein-coupled receptors (GPCRs), which is one of the largest protein families. Through their extracellular and transmembrane domains, GPCRs recognize a variety of ligands, resulting in the transmission of a range of signals across the cell membrane. G proteins are composed of α, β, and γ subunits; the intracellular signals relayed upon receptor activation are determined by specific classes of G proteins.

Heterotrimeric ($\alpha\beta\gamma$) G proteins are central components of the primary mechanism used by virtually all eukaryotic cells to receive, interpret, and respond to a wide range of structurally and chemically diverse extracellular stimuli (*1*, *2*). These G proteins assume different conformations and engage in distinct molecular interactions depending on the bound nucleotide. The interaction of an activated GPCR with a G protein catalyzes the exchange of guanosine triphosphate (GTP) with guanosine diphosphate (GDP) and results in the subsequent dissociation of the Gα-GTP complex from the $\beta\gamma$ complex; alternatively, it may result in the molecular rearrangement of G protein subunits (*3*). This enables both the Gα-GTP complex and the $\beta\gamma$ dimers to interact with a variety of downstream effectors.

Based on subunit sequence homology, the G protein α-subunit family can be divided into four subfamilies—G_s, $G_{i/o}$, $G_{q/11}$, and $G_{12/13}$ (*4*). The G_s subfamily, consisting of G_s and G_{olf} subtypes, is named for its stimulation of adenylyl cyclases (*5*). The $G_{i/o}$ subfamily is named for its inhibition of adenylyl cyclases (although not all $G_{i/o}$ isotypes share this property) (*6*). Members of the $G_{q/11}$ subfamily are involved in the stimulation of phospholipase Cβ isoforms (*7*). The activation of G_{12} and/or G_{13} proteins is associated with the stimulation of the low-molecular-weight G protein Rho and its downstream targets (*8*). The G protein α-subunit nomenclature is commonly used to classify GPCRs. Hence, these GPCRs are referred to as G_s-, G_i-, or G_q-coupled receptors that reflect their primary signal transduction pathway.

Characteristically, each GPCR subtype appears to only couple to a subset of G proteins that may be found in a particular cell. Elucidation of the mechanisms underlying this coupling specificity has been an important problem in the GPCR research. Chimeric receptors have been used to locate domains within receptor sequences that may define their coupling specificity (*4*). These studies reveal that the selec-

***Corresponding author.**
E-mail: ono@otsuka.gr.jp

tivity of G protein recognition is determined by multiple intracellular receptor regions. The most important regions appear to be in the second intracellular loop and in the start and end of the third intracellular loop, which are close to the cytoplasmic surface of the membrane (*9*).

It is important to determine GPCR-G protein coupling specificity in order to understand cell signaling. Firstly, this is essential for understanding the physiological mechanisms underlying the response mediated by the activation of a given GPCR. Secondly, from the viewpoint of drug development, such predictions would be very useful in devising experiments to screen orphan receptors for ligands, since these experiments monitor a specific intracellular response, which is determined by a receptor's coupling specificity (*10*). Therefore, the use of bioinformatic techniques for determining GPCR-G protein coupling specificity contributes to the development of effective experimental systems.

In recent years, several computational methods have been developed to understand GPCR coupling specificity, in which sequence features such as hidden Markov models (HMMs) and sequence motifs are used (*9, 11–15*). However, the accuracy of these methods tends to be biased by the degree of sequence similarity. For example, an orthologous relationship would lead to overestimation of the correct prediction accuracy of coupling specificity.

The natural language processing (NLP) technique, on the other hand, is a useful method for extracting biological knowledge from large text databases. To date, various approaches have been developed to comprehensively extract information by using NLP. The information is subsequently used to build up annotations within biological databases that describe relevant aspects of these proteins. Some approaches have been directly applied to derive important information on protein annotations (*16–18*).

In this study, we describe a novel method for predicting GPCR-G protein coupling specificity. This method uses the features of sequences and biological functions of GPCRs that are derived by sequence analysis and NLP text mining, respectively. The prediction accuracy is improved by adding the features of biological functions that contain information on diseases and molecular interactions as well as key words related to GPCRs. In addition, a machine learning algorithm, namely the C4.5 algorithm (*19*), is used to extract information on the characteristics of GPCR-G protein coupling.

Results

We evaluated our method for predicting coupling specificity by applying it to a dataset of 153 human GPCRs, including 84 G_i-, 33 G_q-, and 36 G_s-coupled sequences. To avoid the prediction bias caused by sequence similarity, that is, the orthologous relationship, we eliminated non-human sequences. The classification of this dataset was performed by using the C4.5 algorithm and then was evaluated by the leave-one-out cross-validation (LOOCV) test. The result of the classification sensitivity and specificity at the subtype level of G proteins is shown in Table 1. For the G_i- and G_q-coupled GPCRs, both sensitivity and specificity were more than 90%. In the case of G_s-coupled GPCRs, the specificity was more than 90% but the sensitivity was lower than 90%. This result indicates that G_s-coupling specificity should not be predicted first because its specificity and sensitivity are the lowest among the three coupling classes. We therefore investigated which prediction order could obtain the best result of the total predication accuracy (Table 2). As a result, the best accuracy of 92.2% was obtained with the prediction order of G_q, G_i, and G_s classes from first to last.

Next, we investigated the contribution of the features of biological functions to the prediction accuracy by predicting against three types of feature sets. The first set only contained the features of sequences; the

Table 1 Prediction Accuracy of GPCR-G Protein Coupling for Three G Protein Subfamilies

Subfamily	No. of GPCR sequences	Sensitivity (%)	Specificity (%)
G_i	84	96.4	96.4
G_q	33	90.9	98.3
G_s	36	83.3	93.2

Table 2 Total Prediction Accuracy for Different Prediction Orders

Prediction order			No. of correct predictions	Accuracy (%)
1st	2nd	3rd		
G_q	G_i	G_s	141	92.2
G_i	G_q	G_s	139	90.8
G_i	G_s	G_q	139	90.8
G_s	G_q	G_i	139	90.8
G_q	G_s	G_i	138	90.2
G_s	G_i	G_q	137	89.5

second set only contained the features of biological functions; and the last set contained both feature types. The result of their prediction accuracy is shown in Figure 1. The accuracy obtained by using only the first or the second set was 88.9% and 81.6%, respectively. However, the total accuracy of 92.2% was obtained when both feature types were used. This result suggests that the use of a combination of features derived from both sequences and biological functions improves the total prediction accuracy when compared with that obtained by using only a single type.

The advantage of using the C4.5 algorithm is that it builds decision trees and generates rules that can be used as predictors. Furthermore, the rules generated by the C4.5 algorithm can provide insights into the characteristics of GPCR-G protein coupling. To generate the rules, we applied the C4.5 algorithm to the dataset that included all the 153 GPCRs. Table 3 shows some examples of the generated rules. For instance, Rule 1 in Table 3 means "if a gene is not related to the calcium signaling pathway and has two G_i-specific motifs, then it is coupled with the G_i protein". The rules that contain the features of sequences had a low error rate (there was no discriminating error for Rules 1 and 2 in Table 3, while a few rules had minor errors), and their using frequencies were high. The other rules were constructed by using only the features of biological functions (such as Rules 3 and 4 in Table 3), and their using frequencies were relatively low. In addition, their discrimination error rates were higher than those of the rules using the features of sequences. However, the coupling specificity, which could not be predicted correctly with only the features of sequences, was predicted correctly using

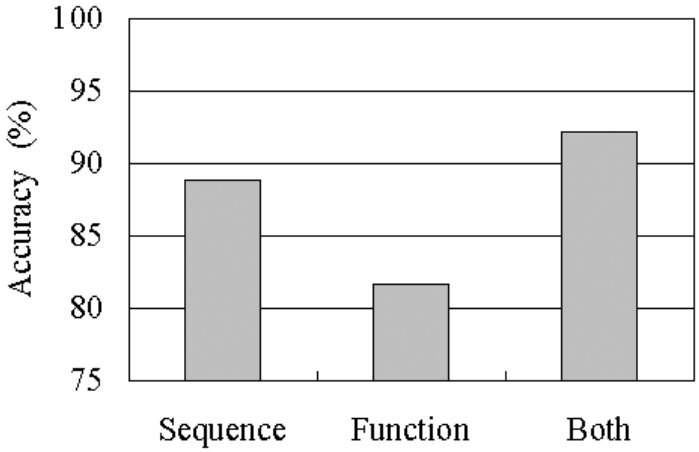

Fig. 1 The total prediction accuracy using different feature sets.

Table 3 Examples of the Rules Produced by the C4.5 Algorithm*

Rule	Antecedent		Consequence	Frequency	Error
	Feature of biological function	Feature of sequence	Coupled Gα type	of use	rate
1	calcium signaling pathway=0	pHMM_G_i_1 $\leq$ 0.0051; pHMM_G_i_7 $\leq$ 8.4e-5	G_i	66	0
2	potassium $\leq$ 2.46	pHMM_G_i_10 $>$ 0.00086; pHMM_G_q_5 $\leq$ 2.3e-5	G_q	17	0
3	cAMP $\leq$ 4.48; inositol phosphate >11.17		G_q	13	8%
4	5-hydroxytryptamine $\geq$ 8.19; cAMP $>$ 28.61		G_s	14	7%

*The features of biological functions contain the information and key words extracted from the literature that are related to the GPCR function. Their values are indicated by the scores according to our calculation (see Materials and Methods). The features of sequences contain HMM profiles with the format "pHMM_(Gα type)_(number of pHMM)". These values are indicated as the E-values obtained from hmmpfam. Frequency of use: the number of uses required to discriminate coupling specificity. Error rate: the error rate of discrimination.

these rules. This result indicates that for some GPCRs, the coupling specificity cannot be predicted only by the sequence clues; however, the integration of functional knowledge makes prediction feasible even in this case. Moreover, these rules represent the functional knowledge required for understanding GPCR-G protein coupling.

Discussion

Our primary aim was to develop a method that efficiently classifies GPCRs according to their coupling specificity in the case of three subfamilies of G proteins. The basis of our approach was as follows: (1) the use of features of sequences and biological functions to predict the non-biased sequence set; and (2) the generation of rules for understanding the specificity of GPCR-G protein coupling.

Our dataset excluded the GPCRs coupled to the $G_{12/13}$ subfamily for the main reason that the literature data on the coupling properties of this subfamily of G proteins were limited. Several similar methods that were reported previously also used restricted datasets for predicting the coupling specificity due to the same reason. The availability of a greater amount of data in the future will allow the prediction of $G_{12/13}$ coupling specificity.

The most important factor in our method is the inclusion of the features of biological functions derived by the NLP technique. Figure 1 shows the contribution of the features of biological functions to the total accuracy. Although it is possible to achieve high prediction accuracy by using only the features of sequences, the accuracy can be improved by adding the features of biological functions. Surprisingly, a prediction accuracy of more than 80% was obtained by using only the features of biological functions. It was thought that functional information regarding to GPCR signaling could be systematically accumulated by active research. Our method could then satisfactorily predict the coupling specificity using only the features of biological functions. These features are useful for determining the coupling specificity when the GPCR gene has no features of sequences similar to those of other genes. Furthermore, the result also suggests that our method is capable of handling problems such as sequence variances and sequence errors.

The importance of functional information is apparent as shown in Rule 1 of Table 3. The rule antecedents of Rule 1 consist of both features of sequences and biological functions. Using this rule, 66 GPCRs were predicted to be G_i-coupled receptors with the prediction accuracy of 100%. However, if the condition "calcium signaling pathway = 0" is removed from the rule, which is mainly related to G_q-coupled receptors (*7*), then three G_q-coupled GPCRs would be inaccuratly predicted as G_i-coupled ones because their sequences had a low E-value (high score) against the G_i-specific pHMM. On the contrary, by adding the condition "calcium signaling pathway = 0" pertaining to functional information, which implies that there is no relationship with the calcium signaling pathway, the three G_q-coupled receptors were classified correctly into the G_q class because their "calcium signaling pathway" scores were high.

In this study, the C4.5 alogrithm, which is a rule-based algorithm, has been applied for predicting the coupling specificity. The advantage of this algorithm is that examination of the rules generated by C4.5 can provide insights into the characteristics of GPCR-G protein coupling. An example is shown in Rule 3 in Table 3. This rule states that if the score of the key word "cAMP" is less than 4.48 and the score of the key word "inositol phosphate" is more than 11.17, then the receptor will couple with G_q proteins. Actually, G_q-coupled GPCRs stimulate inositol phosphate/Ca^{2+} intracellular signaling (*20*). Therefore, the high score of the key word "inositol phosphate" reflects this phenomenon. On the contrary, the low score of the key word "cAMP" suggests that cAMP is not related to G_q coupling. It is known that cAMP is related to the G_s- or G_i-class proteins, which activate or inhibit adenylyl cyclase. This rule is supported by obvious biological phenomena and proves the validity of the rule generated by our method. This result indicates that the advantage of our method lies in its ability to determine not only the coupling information but also the biological relationship. When more functional information is available, it would be possible to generate a rule that could identify more distant relationships, such as the relationship among GPCR-G protein coupling, indirectly interacting molecules, and certain diseases.

To our knowledge, the method reported here is the first approach to use the NLP technique for creating the features of biological functions. Although we only used simple NLP techniques in this study, such as part-of-speech and word frequency, we could obtain high prediction accuracy and related information on GPCR-G protein coupling specificity as the rules. With the progress in genome research and

the development of high-throughput techniques, large amounts of data have been generated on the mechanisms of gene expression, protein structure and interaction, interaction with small molecules, and related diseases. These ever increasing data are described in relevant literature. Increasing the availability of knowledge from biological literature and using more effective NLP techniques will be useful for clarifying the relationship between GPCR-G protein coupling and biological phenomenon along with improving the accuracy of GPCR-G protein coupling prediction.

Materials and Methods

Dataset

Our dataset contained 153 human GPCR sequences along with the information on their coupling specificity collected from the gpDB database (*21*). In this study, we excluded GPCRs of the $G_{12/13}$ coupling family because the number of their annotations was not sufficient for prediction. In addition, we also excluded the olfactory receptors because this class has high sequence homology that results in prediction bias in the case of G_{olf} proteins. As a result, the final dataset contained 84 G_i-, 33 G_q-, and 36 G_s-coupled sequences.

Construction of the features of GPCR sequences

The profile HMMs (pHMMs) were constructed from the motif models generated by the MEME (multiple EM for motif elicitation) (*22*) and MAST (motif alignment & search tool) programs. MEME uses the expectation maximization algorithm (*23*) to discover conserved regions or motifs in a dataset of protein sequences. The algorithm uses a heuristic criterion function based on a maximum likelihood ratio test to compare candidate motifs. MEME outputs models of conserved regions in a rank order; the strongest motif is represented by the first model. For the analysis reported here, we used MEME version 2.0 with the minimum width set at 12 amino acids and a Dirichlet mixture prior. The sequence alignment blocks of each type of coupled G proteins were constructed by MEME.

Based on the alignment blocks, we then constructed a pHMM using the hmmbuild program of the HMMER software package (*24*). The discriminative power of each pHMM was evaluated by a query against all GPCRs measuring the coverage (that is, the percentage of positives that scored an E-value lower than the lowest E-value scored by a negative example in the dataset). The result of this exhaustive search formed a library of 60 refined pHMMs (20 G_i-specific, 20 G_q-specific, and 20 G_s-specific pHMMs). A query GPCR sequence was searched against the pHMMs built for each G protein type, namely the G_i, G_q, and G_s classes, using the hmmpfam program. The E-values of each pHMM obtained from hmmpfam were used as the features for the C4.5 algorithm.

Construction of the features of biological functions

The features of biological functions consist of functional key words (biologically important terms) and information on diseases and molecular interactions. These features were extracted from biological literature. The extraction of key words is one of the main problems in text mining. However, since our aim was not to precisely extract biologically important terms from biological literature but to predict GPCR-G protein coupling specificity, the complicated NLP method was not used to extract the functional key words. Our extraction method was based on the frequency of biological terms because terms that are frequently used in a document and in a set of documents are considered to be "important" terms in this area (*25*). First, the abstracts from the literature on GPCRs were obtained from Entrez Gene (*26*), which includes related links to PubMed. Next, the text obtained from the abstracts was parsed with simple part-of-speech rules to exclude the noise terms by using the Brill POS tagger package (*27*). The excluded terms are as follows: slash, backslash, comma, semicolon, to, coordinating conjunction, preposition, subordinating conjunction, possessive ending, determiner, symbol, wh-determiner, wh-pronoun, and wh-adverb. The remaining words were defined as the functional key words. The score for the functional key words was calculated based on the frequency of their occurrence with the formula: $\text{Score}_{i,w} = F_{i,w}/L_i$, where $F_{i,w}$ is the frequency of the term w in literature related to GPCR gene i, and L_i is the number of literature sources related to GPCR gene i. This score was calculated for all of the functional key words and each GPCR. The obtained values were used as the features of biological functions for predicting the coupling specificity.

Decision tree algorithm and rule generator (C4.5)

The C4.5 algorithm was applied to predict GPCR-G protein coupling specificity. It is a rule induction approach derived from Quinlan's C4.5 decision tree (*19*). This decision tree was generated by an entropy-based selection measure to determine the feature that is most discriminatory. The rules generated by this approach are in the conjunctive form such as "if A and B, then C", where both A and B are the rule antecedents while C is the rule consequence.

Note that the comprehensibility of the rules generated by C4.5 is better than that of the decision tree. This is because the number of rules is usually less than the number of leaves in the tree, and the number of antecedents of a rule is usually less than the number of test conditions appearing in the corresponding path in the tree. Moreover, in some cases, the generalization ability of the rules may be even better than that of the tree.

Implementation

Firstly, we constructed the feature sets of the sequences and biological functions. Next, we predicted the coupling specificity by using C4.5. In order to calculate the accuracy of discriminating each G protein type, a LOOCV test was performed. At the subtype level of G proteins, the evaluation was performed based on sensitivity and specificity. Total accuracy was calculated as the rate that the relevant GPCRs were correctly predicted by applying our method.

Authors' contributions

TO conceived the study, performed the analysis, and drafted the manuscript. HH supervised the work. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Cabrera-Vesa, T.M., *et al.* 2003. Insights into G protein structure, function, and regulation. *Endocr. Rev.* 24: 765-781.
2. Kostenis, E., *et al.* 2005. Techniques: promiscuous Galpha proteins in basic research and drug discovery. *Trends Pharmacol. Sci.* 26: 595-602.
3. Bunemann, M., *et al.* 2003. Gi protein activation in intact cells involves subunit rearrangement rather than dissociation. *Proc. Natl. Acad. Sci. USA* 100: 16077-16082.
4. Wong, S.K. 2003. G protein selectivity is regulated by multiple intracellular regions of GPCRs. *Neurosignals* 12: 1-12.
5. Benjamin, D.R., *et al.* 1995. Solution structure of the GTPase activating domain of alpha s. *J. Mol. Biol.* 254: 681-691.
6. Johnston, C.A. and Watts, V.J. 2003. Sensitization of adenylate cyclase: a general mechanism of neuroadaptation to persistent activation of Galpha(i/o)-coupled receptors? *Life Sci.* 73: 2913-2925.
7. Exton, J.H. 1993. Role of G proteins in activation of phosphoinositide phospholipase C. *Adv. Second Messenger Phosphoprotein Res.* 28: 65-72.
8. Kurose, H. 2003. Galpha12 and Galpha13 as key regulatory mediator in signal transduction. *Life Sci.* 74: 155-161.
9. Moller, S., *et al.* 2001. Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics* 17: S174-181.
10. Minic, J., *et al.* 2005. Yeast system as a screening tool for pharmacological assessment of G protein coupled receptors. *Curr. Med. Chem.* 12: 961-969.
11. Cao, J., *et al.* 2003. A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics* 19: 234-240.
12. Sreekumar, K.R., *et al.* 2004. Predicting GPCR-G-protein coupling using hidden Markov models. *Bioinformatics* 20: 3490-3499.
13. Sgourakis, N.G., *et al.* 2005. A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile hidden Markov models. *BMC Bioinformatics* 6: 104.
14. Sgourakis, N.G., *et al.* 2005. Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. *Bioinformatics* 21: 4101-4106.
15. Yabuki, Y., *et al.* 2005. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Res.* 33: W148-153.
16. Andrade, M.A. and Valencia, A. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14: 600-607.
17. Xie, H., *et al.* 2002. Large-scale protein annotation through gene ontology. *Genome Res.* 12: 785-794.
18. Raychaudhuri, S., *et al.* 2002. Associating genes with gene ontology codes using a maximum entropy analy-

sis of biomedical literature. *Genome Res.* 12: 203-214.

19. Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, USA.
20. Mishra, J. and Bhalla, U.S. 2002. Simulations of inositol phosphate metabolism and its interaction with $InsP_3$-mediated calcium release. *Biophys. J.* 83: 1298-1316.
21. Elefsinioti, A.L., *et al.* 2004. A database for G proteins and their interaction with GPCRs. *BMC Bioinformatics* 5: 208.
22. Bailey, T.L. and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3: 21-29.
23. Dempster, A.P., *et al.* 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39: 1-38.
24. Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-763.
25. Robertson, S.E. and Jones, K.S. 1997. Simple, proven approaches to text retrieval. Technical Report, UCAM-CL-TR-356, Computer Laboratory, University of Cambridge, UK.
26. Maglott, D., *et al.* 2005. Entrez Gene: gene-centerd information at NCBI. *Nucleic Acids Res.* 33: D54-58.
27. Brill, E. 1994. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 722-727. AAAI Press, Menlo Park, USA.