Identifying G-protein Coupled Receptors Using Weighted Levenshtein Distance and Nearest Neighbor Method

Jian-Hua Xu*

Department of Computer Science, Nanjing Normal University, Nanjing 210097, China.

G-protein coupled receptors (GPCRs) are a class of seven-helix transmembrane proteins that have been used in bioinformatics as the targets to facilitate drug discovery for human diseases. Although thousands of GPCR sequences have been collected, the ligand specificity of many GPCRs is still unknown and only one crystal structure of the rhodopsin-like family has been solved. Therefore, identifying GPCR types only from sequence data has become an important research issue. In this study, a novel technique for identifying GPCR types based on the weighted Levenshtein distance between two receptor sequences and the nearest neighbor method (NNM) is introduced, which can deal with receptor sequences with different lengths directly. In our experiments for classifying four classes (acetylcholine, adrenoceptor, dopamine, and serotonin) of the rhodopsin-like family of GPCRs, the error rates from the leave-one-out procedure and the leave-half-out procedure were 0.62% and 1.24%, respectively. These results are prior to those of the covariant discriminant algorithm, the support vector machine method, and the NNM with Euclidean distance.

Key words: GPCR, weighted Levenshtein distance, nearest neighbor method

Introduction

G-protein coupled receptors (GPCRs) are a class of seven-helix transmembrane proteins. They play an important role in a cellular signaling network through their extracellular and transmembrane domains. It is known that such a network can regulate many physiological processes, such as neurotransmission, cellular metabolism, secretion, cellular differentiation and growth, inflammatory and immune responses, smell, taste, vision, and so on. Therefore, GPCRs have become the major targets for the development of new drug candidates with potential application in all clinical fields (*1–3*). In pharmaceutics, it is very important to understand their structures and functions. However, there is only one crystal structure of the rhodopsin-like family that has been solved so far (*4*). Moreover, although thousands of GPCRs' amino acid sequences have been acquired, the ligand specificity of many human GPCRs is still unknown and their corresponding types remain undetermined (*5*). Therefore, identifying GRCP types by only using sequence data has become a valuable research issue (*2*, *6–8*).

*** Corresponding author.**
E-mail: xujianhua@njnu.edu.cn

GPCRs are a large and functionally diverse superfamily. According to their bindings with different ligand types, GPCRs are classified into six different families at least, where the rhodopsin-like family is the largest, which constitutes about 90% of all receptors. In the famous open database GPCRDB (*9*), the rhodopsin-like amine GPCRs can be categorized into six classes: acetylcholine, adrenoceptor, dopamine, histamine, serotonin, and octopamine (*6*, *9*). In the December 2000 release of GPCRDB, histamine and octopamine only included ten and six sequences, respectively. Since they were too few to have any statistical significance, such two types were left out for further consideration, thus a total of 167 sequences from other four classes were collected (*6*). For some classification algorithms, the necessary preprocessing step is to convert each sequence into a 20-dimensional feature vector, in which each feature is described by using its amino acid composition (*6*). In the covariant discriminant algorithm (*6*), the overall error rate was 16.77% according to the leave-one-out procedure or the jackknife test. By using the support vector machine (SVM), an overall error rate of 5.99% was achieved with ten-fold cross-validation (*3*).

Xu and Zhang (*10*) reported that, while each

DNA sequence was considered as a string consisting of four bases (A, C, G, T) directly, an average error rate of 5.88% was obtained by combining SVM with kernels based on weighted Levenshtein distance (WLD). Through transforming each DNA sequence into a numerical vector, the lowest error rate was 9.5% among other five classification methods (*11*). It is noted that, when DNA and protein sequences are converted into numerical vectors, it is possible to lose some useful information in sequence data. In this study, a novel approach is proposed for identifying GPCR types only from sequence data, which combines WLD with the nearest neighbor method (NNM). Such an approach can deal with receptor sequences with different lengths directly. According to the accession numbers in Elrod and Chou (*6*), 162 available sequences from four classes of the rhodopsin-like family in the March 2005 release were collected. The overall error rate for these sequences was 0.62% for the leave-one-out procedure and 1.24% for the leave-half-out procedure, respectively. It demonstrates that our experimental results are prior to those of the covariant discriminant algorithm (*6*), the SVM method (*3*), and the NNM with Euclidean distance.

Results

In this section, we report the identification performance of our novel method combining WLD with NNM by using the leave-one-out and leave-half-out procedures, where 162 GPCR sequences belonging to four classes (acetylcholine, adrenoceptor, dopamine and serotonin) were examined. It is noted that in Elrod and Chou (*6*) the accession numbers of 167 GPCR sequences were listed, where five accession numbers that did not occur in the March 2005 release of GPCRDB were not considered in this study.

Identification performance from the leave-one-out procedure

Table 1 lists the discriminated results of GPCR identification from the covariant disciminant algorithm, the NNM with Euclidean distance between two vectors, and the NNM with WLD between two sequences. The accession numbers of misclassified sequences are given and the corresponding numbers in square brackets denote the class labels to be discriminated. The overall error rate achieved by the covariant discriminant algorithm was 16.77% for the 167 sequences of GPCRs (*6*), whereas the overall error rates based on the NNM with Euclidean distance and the NNM with WLD were 8.02% and 0.62% for the 162 sequences in this study, respectively. In the latter case, only one sequence (O96716) from dopamine was misclassified into adrenoceptor. Therefore, the results of our method are prior to that of the covariant discriminant algorithm (*6*).

Identification performance from the leave-half-out procedure

For the 162 available sequences, we divided them into two subsets, where set 1 was constructed by the sequences located in the odd positions of accession numbers and set 2 was constructed by the remainder sequences. The experimental results obtained by the NNM with Euclidean distance and the NNM with WLD are listed in Table 2, and the overall and class

Table 1 Overall and Class Error Rates and Misclassified Accession Numbers of GPCRs in the Leave-one-out Procedure

Method	Acetylcholine [1]	Adrenoceptor [2]	Dopamine [3]	Serotonin [4]	Overall error rate
Covariant discriminant algorithm	10/31 (32.26%)	5/44 (11.36%)	7/38 (18.42%)	6/54 (11.11%)	28/167 (16.77%)
NNM with Euclidean distance	0/28 (0.00%)	5/43 (11.63%) P35405[3] P32251[3] Q91081[3] P07700[4] P43141[4]	4/37 (10.81%) P24628[1] P21917[2] Q24563[2] O44198[4]	4/54 (7.41%) Q16950[3] P20905[2] Q17239[2] Q25414[3]	13/162 (8.02%)
NNM with WLD	0/28 (0.00%)	0/43 (0.00%)	1/37 (2.70%) O96716[2]	0/54 (0.00%)	1/162 (0.62%)

Table 2 Overall and Class Error Rates and Misclassified Accession Numbers of GPCRs in the Leave-half-out Procedure

Method	Test set	Acetylcholine [1]	Adrenoceptor [2]	Dopamine [3]	Serotonin [4]	Overall error rate
SVM with 10-fold cross-validation		0.00%	9.09%	5.26%	7.49%	5.99%
NNM with Euclidean distance	Set 1	0/14 (0.00%)	1/22 (4.55%) P32251[3]	4/18 (22.22%) P21917[2] P53453[1] O44198[4] Q9PSA7[1]	2/27 (7.41%) Q16950[3] P20905[2]	7/81 (8.64%)
	Set 2	0/14 (0.00%)	2/21 (9.52%) Q91081[3] P07700[4]	3/19 (15.79%) P24628[1] Q24563[4] Q42317[4]	2/27 (7.41%) Q17239[2] Q25414[3]	7/81 (8.64%)
NNM with WLD	Set 1	0/14 (0.00%)	0/22 (0.00%)	0/18 (0.00%)	0/27 (0.00%)	0/81 (0.00%)
	Set 2	0/14 (0.00%)	0/21 (0.00%)	2/19 (10.53%) Q24563[4] O96716[2]	0/27 (0.00%)	2/81 (2.47%)

error rates from the SVM with ten-fold cross-validation (*3*) are also provided. According to Table 2, the average error rates over two test sets from the NNM with Euclidean distance and the NNM with WLD were 8.64% and 1.24%, respectively. The overall error rate achieved by the SVM with ten-fold cross-validation was 5.99%, which was 4.75% higher than that of the NNM with WLD. Generally, the average or overall error rate decreases when the number of k in k-fold cross-validation increases. Therefore, the results of the NNM with WLD are better than that of the SVM method.

According to Tables 1 and 2, it can be seen that the performance of the NNM with WLD is better than those of the covariant discriminant algorithm, the SVM method, and the NNM with Euclidean distance. Additionally, 14 misclassified sequences of GPCRs are listed in Table 1 and 16 sequences in Table 2. Among these sequences, 12 ones are identical, which have to be further examined by us.

Discussion

Protein sequence data are described as the symbolic strings consisting of amino acids. To utilize some classification algorithms for identifying structures and functions of proteins, one has to convert sequence data into numerical vectors (for example, amino acid composition vectors) through a proper transform way. However, it is found out that such a transform procedure would lose some useful information. In this study, a novel discriminant technique for classifying GPCR types is introduced, which combines WLD with NNM. Since the widely used Euclidean distance between two vectors is replaced by the WLD between two sequences, the sequence data can be handled directly. In our experiments, 162 available sequences of four classes collected from the rhodopsin-like family were used to evaluate this method. The experimental results show that the error rate of our method was lower than those of the covariant discriminant algorithm, the SVM method, and the NNM with Euclidean distance. It demonstrates that our method is very effective to identify GPCR types only from sequence data.

Our further work will deal with more protein and DNA sequence data to examine the performance of our method, and fuse more biological information into the weight definition of WLD.

Materials and Methods

Sequence data of the rhodopsin-like family

In Elrod and Chou (*6*), the accession numbers of 167 GPCR sequences from the December 2000 release of GPCRDB were listed, including 31 acetylcholines, 44 adrenoceptors, 38 dopamines, and 54 serotonins (http://www.gpcr.org/7tm/). According to the March 2005 release, five sequences (Q9QYN6, Q9QYN7, Q9W180, P35369, and P13953GP) were

Table 3 Summary of 162 GPCRs from Four Classes of the Rhodopsin-like Family

Class	Number	Minimal length (aa)	Maximal length (aa)	Average length (aa)
Acetylcholine	28	460	805	531.46
Adrenoceptor	43	400	519	448.09
Dopamine	37	363	539	441.35
Serotonin	54	357	834	443.69

deleted. Therefore, we collected 162 GPCR sequences as shown in Table 3.

In this study, we mainly considered these receptor sequences as strings of amino acids directly. In order to examine the NNM with Euclidean distance, we also converted these sequences into 21-dimensional numerical vectors, where the first 20-dimensional features represent amino acid composition and the last feature represents ambiguous symbol composition. To eliminate the influence of different sequence lengths, all features were divided by sequence length.

NNM

NNM is a piecewise linear classification technique (*12*), which can only handle numerical vectors converted from sequences originally. Let l training samples from c classes be:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \tag{1}$$

where $x_i \in R^d$ and $y_i \in \{1, 2, ..., c\}$ represent the i^{th} vector and its class label, respectively. For a new sample x to be classified, we calculate l distances between x and x_i $(i = 1, .., l)$, and find out the training sample x_g with the minimal distance:

$$\|x - x_g\| = \min_{i=1,\ldots,l} \|x - x_i\| \tag{2}$$

In this case, we decided that x and x_g belong to the same class. Here $\|\cdot\|$ usually denotes a certain distance between two vectors, such as Euclidean distance.

WLD between two sequences

WLD can directly measure the similarity between two sequences. Here, we assume two symbolic strings (sequences) $\boldsymbol{a}$ and $\boldsymbol{b}$ with different lengths n and m, respectively, denoted as:

$$\boldsymbol{a} = a_1 a_2 \ldots a_n, \boldsymbol{b} = b_1 b_2 \ldots b_m \tag{3}$$

For these symbols existing in the two strings above, three correction operations can be defined as: (1) Deletion operation: some symbol a_i in the string $\boldsymbol{a}$ is deleted; (2) Insertion operation: some symbol b_j in the string $\boldsymbol{b}$ is inserted into the string $\boldsymbol{a}$; and (3) Substitution operation: some symbol a_i in the string $\boldsymbol{a}$ is replaced by some symbol b_j in the string $\boldsymbol{b}$. By using these correction operations, the string $\boldsymbol{a}$ can be transformed into the string $\boldsymbol{b}$ step-by-step.

The Levenshtein (edit) distance is defined as the smallest number of correction operations converting the string $\boldsymbol{a}$ into the string $\boldsymbol{b}$. Since in many real applications the three operations imply different meanings, it is necessary to determine different weights for the different operations. According to this idea, the WLD is defined as the minimum total weights of single symbol deletion, insertion, and substitution operations required to convert one string into the other (*12–14*). A dynamic programming algorithm was proposed by Wagner and Fischer (*15*) for calculating the WLD. Let d_{ij} be the WLD between two sub-strings consisting of the first i symbols of the string $\boldsymbol{a}$ and the first j symbols of the string $\boldsymbol{b}$, and c_D, c_I, and c_S denote the weights of single symbol deletion, insertion, and substitution operation respectively. We have:

$$d_{ij} = \min(d_{(i-1)j} + c_D,\ d_{i(j-1)} + c_I,\ d_{(i-1)(j-1)} + c_S) \tag{4}$$

where $d_{00} = 0$; $i = 1, \ldots, n$; $j = 1, \ldots, m$. Figure 1 illustrates the computational procedure of the WLD. Finally, d_{nm} implies the WLD. We used such a distance to measure the similarity between two symbolic strings, that is, two sequences of GPCRs. It is noted that, when the weights of insertion and deletion operations are identical, the WLD satisfies three conditions in the distance definition.

In order to eliminate the influence of string lengths, we divided the original WLD by summation of two string lengths. However, it is still referred to the WLD. In this study, such a distance between two strings was used in NNM. That is, we classified GPCR types by combining WLD with NNM. In our experiments, the weights of single insertion and deletion operations were equal to 1. If the two symbols were

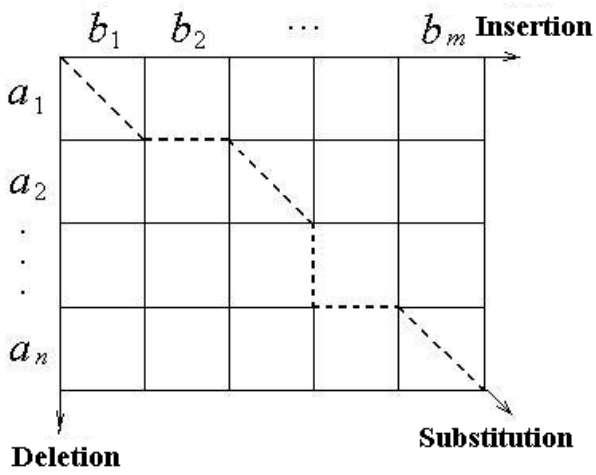

Fig. 1 The computational procedure of WLD.

identical, the weight of substitution operation was 0, otherwise was 3. It implies that there exists no substitution operation between different amino acids.

Identification performance measure

For many classification methods, k-fold cross-validation is a widely used technique for estimating identification performance or generalization error. Generally, the training set is randomly divided into k disjoint subsets of almost equal size. The classifier is trained by using $k-1$ of the subsets and is then tested on the subset left out. This procedure is repeated k times (or trials) and in turn each subset is used for testing once. Averaging the test error over the k trials can give an estimate of the expected generalization error. In real applications, the mean of the k estimates of predication error rate is usually referred to the average error rate. There exist two extreme cases: k=2 and $k = l$. The former is referred as the leave-half-out procedure, and the latter is the leave-one-out procedure or the jackknife test.

In this study, the leave-one-out and leave-half-out procedures were used to measure the identification performance. Since there are four classes of GPCR types in our experiments, we utilized two indexes, overall and class error rates, in order to give more identification details. For the leave-one-out procedure, the overall error rate is defined as the ratio of the number of misclassified receptors to the total number of all receptors, and the class error rate denotes the ratio of the number of misclassified receptors in some class to the receptor number of this class. In the leave-half-out procedure, two corresponding indexes are defined for each test subset and the average error rate is estimated over the two subsets.

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu Province (No. BK2004142) and partly by the National Natural Science Foundation of China (No. 60275007).

References

1. Lameh, J., *et al.* 1990. Structure and function of G protein coupled receptors. *Pharm. Res.* 7: 1213-1221.
2. Hebert, T.E. and Bouvier, M. 1998. Structural and functional aspects of G protein-coupled receptor oligomerization. *Biochem. Cell Biol.* 76: 1-11.
3. Huang, Y. and Li, Y. 2004. Classifying G-protein coupled receptors with support vector machine. *Lect. Notes Comput. Sci.* 3174: 448-452.
4. Palczewski, K., *et al.* 2000. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* 289: 739-745.
5. Schoneberg, T., *et al.* 2002. The structural basis of G-protein-coupled receptor function and dysfunction in human diseases. *Rev. Physiol. Biochem. Pharmacol.* 144: 143-227.
6. Elrod, D.W. and Chou, K.C. 2002. A study on the correlation of G-Protein-coupled receptor type with amino acid composition. *Protein Eng.* 15: 713-715.

7. Karchin, R., *et al.* 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18: 147-159.
8. Bhasin, M. and Raghava, G.P. 2004. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32: W383-389.
9. Horn, F., *et al.* 1998. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 26: 275-279.
10. Xu, J. and Zhang, X. 2004. Kernels based on weighted Levenshtein distance. In *Proceedings of 2004 IEEE International Joint Conference on Neural Networks*, Vol.4, pp.3015-3018. IEEE Press, New York, USA.
11. Mika, S., *et al.* 1999. Fisher discriminant analysis with kernels. In *Proceedings of IEEE Neural Networks for Signal Processing Workshop*, pp.41-48. IEEE Press, New York, USA.
12. Duda, R.O., *et al.* 2002. *Pattern Classif ication* (second edition). John Wiley and Sons, New York, USA.
13. Fu, K.S. 1982. *Syntactic Pattern Recognition and Application.* Prentice-Hall, Englewood Cliffs, USA.
14. Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* 10: 707-710.
15. Wagner, R.A. and Fischer, M.J. 1974. The string-to-string correction problem. *J. ACM* 21: 168-173.