Interdiscip Sci Comput Life Sci (2010) 2: 180–184
DOI: 10.1007/s12539-010-0080-3

Using Adaptive K-nearest Neighbor Algorithm and Cellular Automata Images to Predicting G-Protein-Coupled Receptor Classes

Xuan XIAO*, Wang-Ren QIU
(Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333000, China)

Received 24 June 2009 / Revised 3 November 2009 / Accepted 7 November 2009

Abstract: G-Protein-Coupled Receptors (GPCRs) are the largest of cell surface receptor, accounting for >1% of the human genome. They play a key role in cellular signaling networks that regulate various physiological processes. The functions of many of GPCRs are unknown, because they are difficult to crystallize and most of them will not dissolve in normal solvents. This difficulty has motivated and challenged the development of a computational method which can predict the classification of the families and subfamilies of GPCRs based on their primary sequence so as to help us classify drugs. In this paper the adaptive K-nearest neighbor algorithm and protein cellular automata image (CAI) is introduced. Based on the CAI, the complexity measure factors derived from each of the protein sequences concerned are adopted for its Pseudo amino acid composition. GPCRs were categorized into nine subtypes. The overall success rate in identifying GPCRs among their nine family classes was about 83.5%. The high success rate suggests that the adaptive K-nearest neighbor algorithm and protein CAI holds very high potential to become a useful tool for understanding the actions of drugs that target GPCRs and designing new medications with fewer side effects and greater efficacy.

Key words: GPCRs, adaptive K-nearest neighbor algorithm, CAI.

1 Introduction

G protein-coupled receptors (GPCRs) are one of the largest superfamilies of membrane proteins in human. GPCRs play a key role in cellular signaling pathways that regulate many basic physiological processes, such as neurotransmission, secretion, growth, cellular differentiation, inflammatory and immune responses (Attwood *et al.*, 2001).

GPCRs are of enormous importance for the pharmaceutical industry because over half of all existing medicines act on a GPCR. Much effort has been invested in GPCR study by both academic institutions and pharmaceutical industries. The functions of many GPCRs are unknown, and determining their ligands and signaling pathways is both time-consuming and costly. With the availability of sequence data of human and other mammalian genomes, as well as their expressed sequence tag data, a computational method can be applied to predicting the classification of the families and subfamilies of GPCRs based on their primary sequences so as to help us classify drugs in the post genomic era (Xiao *et al.*, 2009).

The typical strategies for identifying GPCRs and their types include similarity search based tools, such as BLAST, FASTA and motif finding tools. Although these tools are very successful in searching similar proteins, they fail when members of a subfamily are divergent in nature. To overcome this limitation, a number of tools based on composition and pattens of protein sequences have been developed. In a pioneer study, Chou and Elrod attempted to identify the subfamily classes of the rhodopsin-like GPCR family by using the covariant-discriminant algorithm (Chou and Elrod, 2002). With more data available later, the study was extended to identify the main family classes of GPCRs with a similar approach (Chou, 2005). Stimulated by the encouraged results, some follow-up studies were conducted by using various different approaches as reported (Bhasin and Raghava, 2005; Gao and Wang, 2006; Wen *et al.*, 2007).

In this study, novel approaches, protein cellular automata image (CAI) and adaptive K-nearest neighbor algorithm, are introduced to cope with this problem. The bottom line is that the cellular automaton images can reveal many important features of protein, which are originally hidden in a long and complicated amino acid sequence (Xiao *et al.*, 2005a). The CAI has been applied to predict the effect on the replication ratio by HBV virus gene missense mutation (Xiao *et al.*, 2005b) and predict the protein subcellular location (Xiao *et al.*, 2006). There are several parameters to evaluate the im-

*Corresponding author.
E-mail: xiaoxuan0326@yahoo.com.cn

age feature, such as Markov Random Fields, maximum local entropy, and complex wavelet coefficients. Of the known complexity measure approaches so far, the Ziv-Lempel complexity measure is the most adequate one in reflecting the repeat patterns occurring in the character sequence, and hence was adopted in this study.

The k-nearest neighbor algorithm is one of the simplest and most attractive pattern classification algorithms. However, it faces serious challenges when patterns of different classes overlap in same regions in the feature space. In this paper, we demonstrate that an extremely simple adaptive distance measure significantly improves the predict success rates.

2 Materials

Protein sequences were collected from the Swiss-Prot database release 54.8 of 05-Feb-2008 at http://www.ebi.ac.uk/swissprot/ by the “UniProt Power Search”. GPCRs are generally categorized into the six types, we extended the discriminative classes from six to nine, *i.e.* rhodopsin-like receptor, peptide hormones receptor, glutamate and calcium receptor, fungal mating pheromone receptor, cyclic AMP receptor, odorant receptors in drosopila, gustatory receptor of drosophila, frizzled/smoothened family, T2R family in mammals. Sequences annotated with ambiguous or uncertain terms, such as “potential”, “probable”, “probably”’, “maybe”, or “by similarity”, were excluded. Sequences annotated with “fragment” were excluded; also, sequences with less than 50 amino acid residues were removed because they might just be fragments. To reduce homology bias, a redundancy cutoff was operated to winnow those sequences which have $\geqslant 50\%$ pairwise sequence identity to any others in a same family class.

After strictly following the above procedures, we finally obtained 780 GPCRs, which are distributed among the nine GPCR family classes. Accordingly, the dataset S, thus obtained is a union of the 9 subsets as formulated below:

$$\mathrm{S} = \mathrm{S}_1 \cup \mathrm{S}_2 \cup \mathrm{S}_3 \cup \mathrm{S}_4 \cup \mathrm{S}_5 \cup \mathrm{S}_6 \cup \mathrm{S}_7 \cup \mathrm{S}_8 \cup \mathrm{S}_9 \qquad (1)$$

On the basis of dataset S, two working datasets, *i.e.*, a learning dataset S^{L} and an independent testing dataset S^{T}, are constructed. In order to fully use the data in S and meanwhile guarantee that S^{L} and ST be completely independent of each other, the following condition is imposed:

$$\mathrm{S}^{\mathrm{L}} \cup \mathrm{S}^{\mathrm{T}} = \mathrm{S} \quad \text{and} \quad \mathrm{S}^{\mathrm{L}} \cup \mathrm{S}^{\mathrm{T}} = \varPhi \qquad (2)$$

Where $\cup$, $\cap$, and $\varPhi$ represent the symbols for “union”, “intersection”, and “empty set” in the set theory, respectively. Protein samples were randomly assigned to the corresponding subsets of S^{L} and S^{T}. The accession numbers and sequences for the corresponding proteins in the learning and testing datasets can been downed http://218.65.61.89:8080/bioinfo/GPCR/gpcrgia_data.html.

3 Method

To avoid losing the sequence-order information, a logic approach is to use the entire sequence to represent the protein sample and apply the sequence search-based tools such as BLAST to perform prediction. However, this kind of approach fails to work when the query protein does not have significant homology to proteins of known characteristics. In order to avoid complete losing the sequence-order information and also enable the prediction more effectively for those proteins that do not have significant homology to characterized proteins, a feasible approach is to use the pseudo amino acid (PseAA) composition to represent the protein sample. The PseAA composition was originally proposed for predicting protein subcellular localization and membrane protein type; while the amphiphilic PseAA composition was proposed for predicting the enzyme functional classification. The essence of PseAA composition is to use a discrete model to represent a protein sample yet without complete losing its sequence-order information. According to its definition, the PseAA composition for a given protein sample is expressed by a set of $20 + \lambda$ discrete numbers, where the first 20 represents the 20 components of the classical amino acid composition while the additional λ numbers incorporate some of its sequence-order information via various different kinds of coupling modes. Ever since the concept of PseAA composition was introduced, various PseAA composition approaches have been stimulated to deal with various different problems in proteins, such as structural class such as protein structural class, protein subcellular localization, protein subnuclear localization, protein submitochondria localization, protein oligomer type, conotoxin superfamily classification, membrane protein type, apoptosis protein subcellular localization, mycobacterial protein subcellular localization, enzyme functional classification, protein fold pattern, signal peptide, and other protein-related systems. Owing to its wide usage, recently a very flexible PseAA composition generator, called “PseAAC”, was established at the website http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate 63 different kinds of PseAA composition.

To successfully use the PseAA composition for predicting various attributes of proteins, the key is how to optimally extract the features for the PseAA components. In this study, the approach by combining the “complexity factor” and the “cellular automaton image” was introduced to derive the PseAA components.

3.1 Complexity measure factors of protein CAI

A protein sequence is generally constituted by 20 native amino acids whose single character codes are: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. It is very difficult to find its characteristic vector particularly when the sequence is very long. To cope with this situation, we resort to the images derived from the amino acid sequence thru the space-time evolution of cellular automata. As a first step, the 20 amino acids are coded in a binary mode as given in Table 1, which can better reflect the chemical and physical properties of an amino acid, as well as its structure and degeneracy. Through the above encoding procedure, a protein sequence is transformed to a serial of digital signals. For example, the sequence "MASAA..." is transformed to "100111100101001 1100111001...".

Table 1 Three different types for coding amino acids

Type	Code									
Character	P	L	Q	H	R	S	F	Y	W	C
Decimal	1	3	4	5	6	9	11	12	14	15
Binary	00001	00011	00100	00101	00110	01001	01011	01100	01110	01111
Character	T	I	M	K	N	A	V	D	E	G
Decimal	16	18	19	20	21	25	26	28	29	30
Binary	10000	10010	10011	10100	10101	11001	11010	11100	11101	11110

We adopt the circulating boundary condition, with the iterative formula given below:

$$\mathrm{D(i,j) = F(D(i-1,j-1), D(i-1,j), D(i-1,j+1))}$$
$$(2 \leqslant i \leqslant n, 2 \leqslant j \leqslant 5N-1) \quad (3)$$
$$D(i,1) = F(D(i-1,5N), D(i-1,1), D(i-1,2))$$
$$(2 \leqslant i \leqslant n) \quad (4)$$
$$D(i,5N) = F(D(i-1,5N-1), D(i-1,5N),$$
$$D(i-1,1)) \quad (2 \leqslant i \leqslant n) \quad (5)$$

Where D(1:n,1:5N) is a two-dimensional (2D) array to present the amino acid sequence image, the first row of array D deposit the protein 01 sequence after digital coding, F is the iterative rule, n the iterative time, and N the length of the amino acid sequence. Data derived by the process with the evolving rule are saved in the rows starting from the second, and data in each row are derived from those in its previous row.

The evolution rule for image formation must be able to very obviously distinguish whether the proteins concerned are similar to each other or not. We find the 84th is the best one in serving such a purpose among all the 256 kinds of evolving rules. The time that the rule evolves determines the width of the images. It was found that the image structure is basically steady when the time is 100.

We transform the 2D array (matrix) into an image with visualization techniques. The basic bitmap format is chosen owing to its easily handled property. In this way, if the matrix element is zero, the color of the counterpart pixel bit is white; otherwise, black. For a systematic description of CAI, refer to the paper by Chou and Zhang (1995). Thus, all the existing tools in the area of image processing can be straightforwardly used for the current study.

Image recognition is concerned with the automatic detection and classification of image. Its techniques can be divided into two main categories: (1) those employing geometrical features, and (2) those using gray-level information. The texture characteristics of these gene images are very complicated and it is very difficult to characterize these images using either deterministic or statistical models. Nevertheless, these protein images are saved in 2D arrays, every row of gene images is a 01 sequence in fact. We can simply regard the Ziv-Lempel complexity of these 01 sequences as pseudo amino acid composition.

The Ziv-Lempel complexity measure reflects most adequately repeats occurring in the text (Ziv and Lempel, 1976). The Ziv-Lempel complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process (Gusev and Chuzhanova, 2001).

We can figure out 100 complexity if the image has 100 rows. These complexity all can be regarded as pseudo amino acid component, but we find the best predict accuracy can be gained under the first 5 complexity used. Thus, by following exactly the same procedure as described by Chou and Elrod (1999), a protein can be expressed by a vector or a point in a 25 dimensional space, *i.e.*

$$X = (x_1, x_2, x_3, \cdots, x_{25})^{\mathrm{T}} \quad (6)$$

Where $x_i (i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 amino acids in the protein, arranged alphabetically according to their single letter codes, $x_j (j = 21, 22, \ldots, 25)$ are the complexity measure factors the protein sequence, T represents the transpose operator.

3.2 Adaptive K-nearest neighbor algorithm (k-NN)

The K-nearest neighbor algorithm is one of the oldest and simplest pattern classification algorithms. Given a set of n labeled examples $D_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ with imputs $X_i \in R^{\mathrm{d}}$ and calss labels Y_i, the K-NN algorithm classifies an unseen pattern X to the calss that appears most often among its K nearest neighbors. To define the locally adaptive distance between a query pattern X and a training example X_i, we first construct the largest sphere that is centered on X_i and excludes all training examples from other classes. This can be easily achieved by setting the radius of the sphere to $r_i = \min_{n:Y_n \neq Y_i} d(X_i, X_n) - \varepsilon$, where $\varepsilon > 0$ is an arbitrarily small number and $d(\bullet, \bullet)$ is the normal Euclidean distance measure. The locally adaptive distance between X and the training example X_i is defined as:

$$d_{new}(X, X_i) = d(X, X_i)/r_i^{\lambda} \quad (\lambda > 0) \tag{7}$$

The adaptive k-NN algorithm works exactly the same as the original k-NN algorithm except that it uses the adaptive distance measure to replace the original Euclidean distance measure for identifying the nearest neighbors.

4 Results and discussion

Now let us demonstrate the prediction quality by using adaptive k-NN predictor. In statistical prediction the independent dataset test, sub-sampling test, and jackknife test are often used in literatures for examining the accuracy of a predictor (Chou and Zhang, 1995). However, as elucidated in (Chou and Shen, 2008) and demonstrated by Eq. 50 of (Chou and Shen, 2007), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors. In the jackknife test, each domain in the dataset is singled out in turn as a 'test domain' and all the rule-parameters are determined from the remaining $N - 1$ domains.

The success rates by jackknife test for the aforementioned learning dataset S^{L} 340 GPCRs classified into nine types are given in Table 2, where for facilitating comparison the corresponding rates obtained by the recently developed algorithms, such as the nearest neighbor algorithm, optimized evidence-theoretic (OET) K-nearest neighbor are also listed. It can been form Table 2 that the overall success rate by the current approach is 83.6% in learning dataset S^{L}, which is higher than those by the other approaches. The best success rate obtained in k=5 and $\lambda = 0.6$.

Meanwhile, predictions were also performed with the current algorithm trained by the dataset S^{L} on the independent testing dataset S^{T}. as shown in Table 2, the overall success rate was 93.6%. However, it should be pointed out that the independent dataset test performed here was just for a demonstration of practical application. Because the selection of independent dataset often bears some sort of arbitrariness, the jackknife test is deemed more objective than the independent dataset test. Therefore, the power of a predictor should be measured by the success rate of jackknife test.

Table 2 Success rates obtained with the current predictor by the jackknife test and independent test in identifying nine main GPCR families

Method	Jackknife[a]	Independent[b]
Nearest neighbor algorithm	79.4%	80.9%
OET k-nearest neighbor algorithm	81.0%	89.8%
Adaptive k-nearest neighbor algorithm	83.6%	93.6%

[a] Performed on the dataset S^{L};

[b] Used the predictor trained by the dataset S^{L} to predict the proteins in the dataset S^{T}.

5 Conclusions

It is demonstrated in this study that using the complexity measure factor of protein CAI as the pseudo amino acid components can more effectively reflect the overall sequence-order feature of a protein, and adaptive K-nearest neighbor algorithm leading to higher success rates in predicting the subclass of GPCR, and also demonstrated that the protein CAI is very useful tool for investigating complicated biological sequences. It is anticipated that introduction of the adaptive K-nearest neighbor algorithm may also have impacts on improving the prediction quality for a series of other protein attributes, such as subcellular localization, membrane types, enzyme family and subfamily classes, among many others.

Acknowledgments The work in this research was supported by the grants from the National Natural Science Foundation of China (No. 60661003 and No. 60961003), the Department of Education of Jiangxi Province (No. GJJ09271), and the plan for training youth scientists (stars of Jing-Gang) of Jiangxi province.

References

[1] Attwood, T.K., Croning, M.D., Gaulton, A. 2001. Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. Protein ENG 15, 7–12.

[2] Bhasin, M., Raghava, G.P. 2005. GPCRsclass: A web tool for the classification of amine type of G-protein-

coupled receptors. Nucleic Acids Research 33, W143–147.

[3] Chou, K.C. 2005. Prediction of G-protein-coupled receptor classes. Journal of Proteome Research 4, 1413–1418.

[4] Chou, K.C., Elrod, D.W. 1999. Protein subcellular location prediction. Protein Eng 12, 107–118.

[5] Chou, K.C., Elrod, D.W. 2002. Bioinformatical analysis of G-protein-coupled receptors. Journal of Proteome Research 1, 429–433.

[6] Chou, K.C., Shen, H.B. 2007. Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370, 1–16.

[7] Chou, K.C., Shen, H.B. 2008. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153–162.

[8] Chou, K.C., Zhang, C.T. 1995. Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30, 275–349.

[9] Gao, Q.B., Wang, Z.Z. 2006. Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19, 511–516.

[10] Gusev, V.D., Chuzhanova, N.A. 2001. A rapid method for detecting interconnections between functionally and/or evolutionary close biological sequences. Mol Biol (Mosk) 35, 1015–1022.

[11] Wen, Z., Li, M., Li, Y., Guo, Y., Wang, K. 2007. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32, 277–283.

[12] Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chen, X., Chou, K.C. 2005a. Using cellular automata to generate Image representation for biological sequences. Amino Acids 28, 29-35.

[13] Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chen, X., Chou, K.C. 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. Journal of Theoretical Biology 235, 555–565.

[14] Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C. 2006. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino acids 30, 49–54.

[15] Xiao, X., Wang, P., Chou, K.C. 2009. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. Journal of Computational Chemistry 30, 1414–1423.

[16] Ziv, J., Lempel, A. 1976. On the complexity of finite sequences. IEEE Trans Inf Theory IT 22, 75–81.