Bio-Medical Materials and Engineering 26 (2015) S1829–S1836
DOI 10.3233/BME-151485
IOS Press

A novel fractal approach for predicting G-protein–coupled receptors and their subfamilies with support vector machines

Guoping Nie[a], Yong Li[a], Feichi Wang[a], Siwen Wang[a] and Xuehai Hu[b,*]
[a]*College of Science, Huazhong Agricultural University, Wuhan 430070, P.R. China*
[b]*College of Informatics, Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, P.R. China*

Abstract. G-protein-coupled receptors (GPCRs) are seven membrane-spanning proteins and regulate many important physiological processes, such as vision, neurotransmission, immune response and so on. GPCRs-related pathways are the targets of a large number of marketed drugs. Therefore, the design of a reliable computational model for predicting GPCRs from amino acid sequence has long been a significant biomedical problem. Chaos game representation (CGR) reveals the fractal patterns hidden in protein sequences, and then fractal dimension (FD) is an important feature of these highly irregular geometries with concise mathematical expression. Here, in order to extract important features from GPCR protein sequences, CGR algorithm, fractal dimension and amino acid composition (AAC) are employed to formulate the numerical features of protein samples. Four groups of features are considered, and each group is evaluated by support vector machine (SVM) and 10-fold cross-validation test. To test the performance of the present method, a new non-redundant dataset was built based on latest GPCRDB database. Comparing the results of numerical experiments, the group of combined features with AAC and FD gets the best result, the accuracy is 99.22% and Matthew's correlation coefficient (MCC) is 0.9845 for identifying GPCRs from non-GPCRs. Moreover, if it is classified as a GPCR, it will be further put into the second level, which will classify a GPCR into one of the five main subfamilies. At this level, the group of combined features with AAC and FD also gets best accuracy 85.73%. Finally, the proposed predictor is also compared with existing methods and shows better performances.

Keywords: GPCR, chaos game representation, support vector machine, fractal dimension

1. Introduction

G-protein-coupled receptors (GPCRs) are seven membrane-spanning proteins that have the ability to transduce extracellular signals into intracellular reactions. GPCRs are the largest family of cell surface receptors, comprising over 800 genes in the human genome [1]. It is reported that GPCRs could bind to a broad range of ligands, including small organic compounds, eicosanoids, peptides and proteins [1]. By binding these ligands, GPCRs can activate guanine-binding proteins (G-proteins), which lead to a series of celluar reactions. Thereupon, GPCRs regulate many basic physiological processes, such as smell, taste, vision, neurotransmission, cellular differentiation and growth, immune

[*] Address for correspondence: Xuehai Hu, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P.R. of China. Tel.: +8618171282783; Fax: +8687288509; E-mail: huxuehai@mail.hzau.edu.cn.

0959-2989/15/$35.00 © 2015 – IOS Press and the authors.

response and so forth [2]. These GPCR-related pathways are the targets of a large number of drugs, including neuroleptics, antihistamines, antidepressants, antihypertensives and so on [3]. According to statistics, over 50% of marketed drugs target GPCRs, which are among the most frequent targets of therapeutic drugs [4]. Because of these important facts, GPCRs have got close attention and intensive researches by both academic institutions and pharmaceutical industries [1, 2, 5].

Conventional methods for identifying non-annotated protein are experimental means, such as X-ray crystallography or NMR spectroscopy and so on [5, 6]. Unfortunately, GPCRs are difficult to crystallize because of their hydrophobicity. Up to present, very few crystal GPCR structures have been determined, which implies that traditional experimental methods are not suitable to identify large-scale non-annotated proteins to be GPCRs. With the absence of experiment conditions, researchers may choose to run a standard basic local alignment search tool (BLAST) [7] to identify a protein to be GPCR in cases that it has high sequence similarity to annotated GPCR sequences in the database. However, this method will not function effectively because there are low sequence similarities (about 15%) between GPCRs protein sequences [1].

With the rapid development of large-scale genome and proteome sequencing project, huge amounts of biological data begin to accumulate. In the area of GPCR, the GPCRDB is a molecular class-specific information system that collects, combines, validates and disseminates large amounts of heterogeneous data on GPCRs [8]. According to the latest release of GPCRDB, the data are grouped into six families based on the pharmacological classification of GPCRs [9]. These GPCRs families and their structural features are closely correlated with their function [1], it would be significant to develop a powerful computational method to classify GPCRs into particular families for the purpose of understanding their biological function and their potential as future drug targets.

With the rapid development of bioinformatics, lots of new methods, e.g. machine learning, are widely used in data mining and knowledge discovery. It encouraged series of researches on predicting GPCRs based on protein primary sequence information. Among them, in 2004, a web server called "GPCRpred" was developed for prediction of families and subfamilies of G-protein coupled receptors using support vector machine (SVM) [10]. Gao and Wang [11] introduced a nearest neighbor method to discriminate GPCRs from non-GPCRs and subsequently classify GPCRs at four levels on the basis of amino acid composition (AAC) and dipeptide composition of proteins in 2006. In 2008, Xiao, Wang and Chou [12] employed "cellular automaton" to reveal the pattern features hidden in piles of long and complicated protein sequences and predicted GPCRs and subfamilies with CD classifier. In 2010, Peng, Yang and Chen [13] proposed a new method called "PCA-GPCR" to predict GPCRs. They extracted a comprehensive set of 1497 sequence-derived features, and then employed the principal component analysis (PCA) to reduce the dimension of the feature space to 32. The resulting 32-dimensional feature vectors were fed into SVM and random forest (RF) to predict GPCRs at five levels. Very recently, Gao, Ye and He [14] extracted 170 sequence-derived features encapsulating both amino acid composition and physicochemical features of proteins, and used SVM as the engine to classify GPCRs to the finest subtype level.

The chaos game representation (CGR) of DNA sequences was proposed by Jeffrey in 1990 [15]. It effectively excavated fractal concealed patterns in DNA sequences by using a square with ACGT as its' four vertexes. Subsequently, CGR method of DNA sequences had been extended to exhibit protein sequences. In 1997, a similar CGR algorithm was presented by Basu, et al. [16] to represent a protein sequence by using a 12-sided regular polygon, each vertex of which represents a class of amino acid residues on the basis of conservative substitutions. Up to present, CGR method has achieved some applications in the studies of bioinformatics [17-21]. Moreover, fractal dimensions (FD) are important features of highly irregular geometries. One of the important fractal dimensions is the box-counting

Table 1

The detailed GPCRs subfamilies of dataset

GPCR family	Number of proteins from GPCRDB	Number of proteins after CD-HIT	Final positive dataset
Class A Rhodopsin-like	33167	1014	1014
Class B Secretin-like	1722	171	171
Class C Metabotropic glutamate/pheromone	1529	74	74
Class D cAMP receptors	8	1	0
Class E Vomeronasal receptors	1388	14	14
Class F Taste receptors	711	16	16
Overall	38525	1290	1289

dimension, which is widely used in many research fields for its concise mathematical expression [22]. A number of literatures addressed the successful applications of box-counting dimension [23-26]. For the case of CGR situation, the CGR pictures of protein sequences are plane images-two dimensional geometries. Therefore, it is natural to calculate box-counting dimensions of CGR pictures as an important feature. In fact, box-counting dimension has already been successfully employed as a significant fractal feature of DNA and protein sequences [19-21].

2. Materials and methods

2.1. Dataset

In this paper, a new dataset was built from the latest version (updated at September 26, 2013) of GPCRDB [9]. The newly update GPCRDB classify all the protein sequences into six main families, (1) Class A Rhodopsin like, (2) Class B Secretin like, (3) Class C Metabotropic glutamate/pheromone family, (4) Class D cAMP receptors family, (5) Class E Vomeronasal receptors (V1R and V3R) family and (6) Class F Taste receptors T2R family. All the protein sequences of six subfamilies (38525 sequences in total) were downloaded. To reduce the homology bias of prediction, a redundancy reduction procedure was performed on this dataset by CD-HIT program [27]. Here, we choose the segmentation threshold as 30% to abandon those proteins from the main datasets. That means each protein sequence has equal to or greater than 30% sequence similarity to any other in the same subset. In this way, the resulting training dataset contains 1289 GPCR sequences from above approach and 1289 non-GPCR sequences from GDSL [14] (see Table 1).

2.2. Sample representation

For our computational approach, each protein is represented as a numerical vector, so as to be put into SVM for classification. Here we proposed a novel hybrid fractal method that can capture fractal pattern information of GPCR sequences.

2.2.1. Chaos game representation

The chaos game representation algorithm of protein sequence generates a series of points linked together with an iterative procedure in a 12 polygon. It was firstly proposed by Basu, et al. [16] to study the potential as the discriminative and diagnostic signature of a family of proteins. The detailed algorithm of drawing procedure is listed as follows:

Step 1. Draw a 12-sided regular polygon, each vertex of which represents a class of amino acid residues leading to conservative substitutions;

Step 2. Set the central point of polygon to be the initial point;

Step 3. Given a protein sequence with length N, we draw N points in turn on the basis of the following method: We read residues from the protein sequence successively. Each chosen residue belongs to one class of amino acids, which determine a certain vertex of polygon. And then draw the middle point between initial point and the chosen vertex. Once the point is drawn, and then set it to be the new initial point and start the next round. A chosen residue from given protein sequence determines a drawing point in the polygon. From such criterion, one can successively draw N iterated points, which constitute the CGR picture of given protein sequence.

The CGR algorithm generates an image, which contains fractal structure and visually reveals previously unknown structure information for each concatenated amino acid sequences. Furthermore, we extracted the frequency information of each segments by dividing the 12-sided polygon into 24 grids. From CGR and grid-counting algorithm, each protein sequence can be transformed into a 24-dimensional vector.

2.2.2. Fractal dimension

Geometrical complexities about complex geometries can be studied by the theory of fractal geometry, which was firstly introduced by Mandelbrot who used it to study the length of the coast of Britain [28]. Self-similarity is a widespread natural phenomenon whose components are similar to those of the whole. It is also well-known as the important feature of fractal. Another important feature is non-integer fractal dimensions, which usually are used to measure sizes of fractal geometries.

In subsection 2.2.1, we calculate the frequency of points falling into each grid in a CGR picture. These frequency features extract useful information of CGR picture, but leave more fractal structure information lost. In this paper, inspired by the theory of fractal geometry, we adopt box-counting dimension (BCD) to capture the fractal feature of CGR picture. The following is its' mathematical expression [22]:

$$\dim_B F = \lim_{\delta \to 0} \frac{\ln N_\delta(F)}{-\ln \delta} \tag{1}$$

where $N_\delta(F)$ is the minimal number of squares, whose union can cover the set F with each diameter less than δ . And $\dim_B F$ denotes the BCD of the set F . BCD has a broad range of applications, such as the relationship between cancer and fractals [23] and similarity of complex networks [24] and so on. The detailed computational algorithm of fractal dimension can be found in [21].

2.3. Support vector machine

A support vector machine is one of the most important machine learning algorithms based on statistical learning theory. It is widely used for classification and prediction and become an increasingly popular tool in bioinformatics. The basic principle of SVM is to seek an optimal hyperplane to separate positive samples and negative samples with maximizing the margin between support vectors. More precisely, for its' mathematical expression, SVM maps all positive and negative samples from an original euclidean space (R^d) into a more abstract Hilbert (H) space by using a nonlinear mapping, which is called kernel function. And, in the space H, it tries to find the Optimal Separating Hyperplane (OSH), which distinguishes positive samples from negative ones to the utmost [29]. Here we use an open package LibSVM 3.17, which can be freely downloaded at the following link: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

2.4. Evaluation of the prediction performance

There are usually three criteria employed to test a predictor for its effectiveness in statistical prediction. They are resubstitution test, K-fold cross-validation test and jackknife test [30]. In this research, a 10-fold cross-validation approach is chosen to test our hybrid method. Moreover, the performance of predictor is frequently measured by accuracy (ACC) and Matthew's correlation coefficient (MCC) value [12].

3. Results

3.1. Predicting GPCRs and their subfamilies

This work mainly focus on the question that how to use fractal methods to predict G-protein-coupled receptors and their subfamilies. At first level, an un-annotated protein is predicted to be either a GPCR or a non-GPCR. If it is classified as a GPCR, it will be further put into the second level, which will classify a GPCR into one of the five subfamilies.

In order to seek the optimal combined features for predicting, a series of comparative experiments are carried on via 10-fold cross-validation test. The numerical experiments are designed on four groups: amino acid composition (write "AAC" for abbreviation) features (20-dimensional), AAC together with FD features (21-dimensional), CGR together with FD features (25-dimensional), AAC together with CGR and FD features (45-dimensional). The detailed results are listed in Table 2, which include accuracies, MCC values.

Table 2

Results in identifying GPCR proteins and subfamilies (10-fold cross-validation test)

Feature	Dimension	First level		Second level
		ACC	MCC	Success Rate
AAC	20	0.971683	0.943415	0.850272
AAC+FD	21	0.992242	0.984514	0.857254
CGR+FD	25	0.986811	0.973718	0.833980
AAC+CGR+FD	45	0.993018	0.986037	0.856478

Table 3

Comparisons with GPCR-CA and PCA-GPCR at the first level (jackknife test)

Results	GPCR-CA	PCA-GPCR	This paper
ACC	0.9164	0.9521	0.9685

Table 4

Comparisons with GPCR-CA at the second level (jackknife test)

GPCR family	Number of proteins	Number of correct predictions (GPCR-CA)	Success Rate (%) (GPCR-CA)	Number of correct predictions (Our method)	Success Rate (%) (Our method)
Rhodopsin-like	232	224	96.55	225	96.98
Secretin-like	39	29	74.36	26	66.67
Metabotrophic	44	36	81.82	31	70.45
Fungal pheromone	23	2	8.70	10	43.48
cAMP receptor	10	6	60.00	9	90.00
Frizzled/smoothened	17	8	47.06	4	23.53
Overall	365	305	83.56	305	83.56

3.2. Comparison with other methods at the first level

In order to explain the superiority of our fractal methods, we implement our fractal algorithms on the same dataset (D365, 365 GPCRs, six subfamilies) in GPCR-CA [12] and PCA-GPCR [13] via jackknife test. We list detailed comparisons of our method in Table 3. We find that the overall accuracy of our method is higher than GPCR-CA and PCA-GPCR.

3.3. Comparison with GPCR-CA at the second level

We also implement our method on the same dataset (D365, 365 GPCRs, six subfamilies) in GPCR-CA [12] at the second level via jackknife test. All the detailed results and comparisons are list in Table 4. Although we find that the overall accuracy of our method is equal to GPCR-CA, the success predicting rate about each subfamilies of our method are more balanced. All these results mentioned above show that our method has more generalizability than GPCR-CA.

3.4. Further discussion

With the purpose of supporting our method, a further discussion is proposed. The results mentioned in Tables 3 and 4 show that our fractal method is superior to GPCR-CA and PCA-GPCR. Investigates its reason, the FD feature plays a crucial role in predicting GPCRs. According to reports, amino acid composition (AAC) are simplest but effective features in predicting GPCRs [11], however, only AAC features are insufficient with a lake of sequence order information. To compensate for this deficiency, CGR-based method is proposed in this research. From the results of Table 2, the best accuracy achieves in the group with combined features of AAC and FD. Moreover, the detailed comparisons between different features show an interesting phenomenon. On the one hand, we find that AAC are fundamental features and each group with absence of AAC achieves unsatisfied accuracy from the detailed results of Table 4. On the other hand, only AAC features cannot achieve best accuracy, the best result is achieved when FD features carrying sequence order information are combined with AAC features.

4. Conclusions

In this research, a fractal method based on the combined features of CGR and fractal dimension is employed to predict GPCRs, and then to predict GPCRs to their subfamilies. Taking the results into consideration, on one hand, we can find the highest predicting accuracy and MCC value achieved in the combination of AAC and FD, the best average accuracy achieves 99.22% in 10-fold cross-validation test, and MCC value is 0.9845 at the first level. At the second level, the combined features of AAC and FD also got best accuracy 85.73%. These considerable results imply that our fractal approach is a reliable method to predict GPCRs. On the other hand, several papers addressed so far the problem of predicting GPCRs. Among them, Xiao, et al. [12] have developed a predicting model, called GPCR-CA, they employed "cellular automaton" algorithm to reveal the pattern features hidden in piles of long and complicated protein sequences and predicted GPCRs and their subfamilies with covariant-discriminant classifier, and they got 91.64%, 83.56% accuracy at two levels respectively. To compare with GPCR-CA, we predict the same datasets with our hybrid fractal model. We find that accuracies and MCC values increase with the introduction of our fractal method.

Finally, we note that there are some correlations between different features, which limit our further improvements. Therefore, dimensionality reduction based on redundancy elimination algorithm is the further direction. Furthermore, we also expect that our hybrid approach helps to obtain high recognition rates on other bioinformatics problems.

Acknowledgment

Xuehai Hu was partially supported by the Fundamental Research Funds for the Central Universities HZAU: 2013PY116 and the National Natural Science Foundation of China (NSFC) (Grant No. 11371016). Guoping Nie, Yong Li and Feichi Wang were supported by the National Creative Innovation Plan of College Students of China (Grant No. 201310504040).

References

[1] M.C. Lagerström and H.B. Schiöth, Structural diversity of G protein-coupled receptors and significance for drug discovery, Nature Reviews Drug Discovery **7** (2008), 339-357.

[2] R.J. Lefkowitz, The superfamily of heptahelical receptors, Nature Cell Biology **2** (2000), E133-E136.

[3] T. Gudermann, B. Nürnberg and G. Schultz, Receptors and G proteins as primary components of transmembrane signal transduction, Journal of Molecular Medicine **73** (1995), 51-63.

[4] R. Fredriksson and H.B. Schiöth, The repertoire of G-protein–coupled receptors in fully sequenced genomes, Molecular Pharmacology **67** (2005), 1414-1425.

[5] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox and M. Miyano, Crystal structure of rhodopsin: A G protein-coupled receptor, Science **289** (2000), 739-745.

[6] V. Cherezov, D.M. Rosenbaum, M.A. Hanson, et al., High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor, Science **318** (2007), 1258-1265.

[7] S.F. Altschul, T.L. Madden, A.A. Schaffer, et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research **25** (1997), 3389-3402.

[8] F. Horn, E. Bettler, L. Oliveira, et al., GPCRDB information system for G protein-coupled receptors, Nucleic Acids Research **31** (2003), 294-297.

[9] V. Isberg, B. Vroling, R. van der Kant, et al., GPCRDB: An information system for G protein-coupled receptors, Nucleic Acids Research **42** (2014), D422-425.

[10] M. Bhasin and G.P. Raghava, GPCRpred: An SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, Nucleic Acids Research **32** (2004), W383-389.

[11] Q.B. Gao and Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Engineering Design and Selection **19** (2006), 511-516.

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

[13] Z.L. Peng, J.Y. Yang and X. Chen, An improved classification of G-protein-coupled receptors using sequence-derived features, BMC Bioinformatics **11** (2010), 420.

[14] Q.B. Gao, X.F. Ye and J. He, Classifying G-protein-coupled receptors to the finest subtype level, Biochemical and Biophysical Research Communications **439** (2014), 303-308.

[15] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Research **18** (1990), 2163-2170.

[16] S. Basu A. Pan, C. Dutta and J. Das, Chaos game representation of proteins, Journal of Molecular Graphics & Modelling **15** (1997), 279-289.

[17] Z.G. Yu, V. Anha and K.S. Lau, Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses, Journal of Theoretical Biology **226** (2004), 341–348.

[18] J.Y. Yang, Z.L. Peng, Z.G. Yu, R.J. Zhang, V. Anh and D. Wang, Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation, Journal of Theoretical Biology **257** (2009), 618-626.

[19] X.L. Liu, J.L. Lu and X.H. Hu, Predicting thermophilic proteins with pseudo amino acid composition: Approached from chaos game representation and principal component analysis, Protein & Peptide Letters **18** (2011), 1244-1250.

[20] J.L. Lu, X.H. Hu and D.G. Hu, A new hybrid fractal algorithm for predicting thermophilic nucleotide sequences, Journal of Theoretical Biology **293** (2012), 74-81.

[21] X.H. Niu, X.H. Hu, F. Shi and J.B. Xia, Predicting DNA binding proteins using support vector machine with hybrid fractal features, Journal of Theoretical Biology **343** (2014), 186-192.

[22] K.J. Falconer, Techniques in Fractal Geometry, Wiley, Chichester, 1997.

[23] J.W. Baish and R.K. Jain, Cancer, angiogenesis and fractals, Nature Medicine **4** (1998), 984.

[24] C.M. Song, S. Havlin and H.A. Makse, Self-similarity of complex networks, Nature **433** (2005), 392-395.

[25] F. Grizzi, C. Russo, P. Colombo, et al., Quantitative evaluation and modeling of two-dimensional neovascular network complexity: The surface fractal dimension, BMC Cancer **5** (2005), 14.

[26] K. Metze, Fractal dimension of chromatin: potential molecular diagnostic applications for cancer prognosis, Expert Review of Molecular Diagnostics **13** (2013), 719-735.

[27] Y. Huang, B. Niu, Y. Gao, L. Fu and W. Li, CD-HIT Suite: A web server for clustering and comparing biological sequences, Bioinformatics **26** (2010), 680-682.

[28] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, San Francisco, 1982.

[29] V. Vapnik, Statistical Learning Theory, Wiley Interscience, New York, 1998.

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