GPCR-CA: A Cellular Automaton Image Approach for Predicting G-Protein–Coupled Receptor Functional Classes

XUAN XIAO,[1] PU WANG,[1] KUO-CHEN CHOU[2]
[1]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300, China*
[2]*Gordon Life Science Institute, San Diego, California 92130, USA*

Received 27 March 2008; Revised 12 September 2008; Accepted 7 October 2008
DOI 10.1002/jcc.21163
Published online 25 November 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Given an uncharacterized protein sequence, how can we identify whether it is a G-protein–coupled receptor (GPCR) or not? If it is, which functional family class does it belong to? It is important to address these questions because GPCRs are among the most frequent targets of therapeutic drugs and the information thus obtained is very useful for "comparative and evolutionary pharmacology," a technique often used for drug development. Here, we present a web-server predictor called "GPCR-CA," where "CA" stands for "Cellular Automaton" (Wolfram, S. Nature 1984, 311, 419), meaning that the CA images have been utilized to reveal the pattern features hidden in piles of long and complicated protein sequences. Meanwhile, the gray-level co-occurrence matrix factors extracted from the CA images are used to represent the samples of proteins through their pseudo amino acid composition (Chou, K.C. Proteins 2001, 43, 246). GPCR-CA is a two-layer predictor: the first layer prediction engine is for identifying a query protein as GPCR on non-GPCR; if it is a GPCR protein, the process will be automatically continued with the second-layer prediction engine to further identify its type among the following six functional classes: (a) rhodopsin-like, (b) secretin-like, (c) metabotrophic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothened family. The overall success rates by the predictor for the first and second layers are over 91% and 83%, respectively, that were obtained through rigorous jackknife cross-validation tests on a new-constructed stringent benchmark dataset in which none of proteins has $\geq$40% pairwise sequence identity to any other in a same subset. GPCR-CA is freely accessible at http://218.65.61.89:8080/bioinfo/GPCR-CA, by which one can get the desired two-layer results for a query protein sequence within about 20 seconds.

 J Comput Chem 30: 1414–1423, 2009

Key words: gray level co-occurrence matrix; cellular automaton image; pseudo amino acid composition; covariant-discriminant algorithm; evolutionary pharmacology; G-protein–coupled receptor

Introduction

One of the largest families in the human genome is the encoding the G-protein–coupled receptors (GPCRs), which are cell surface receptors. Because of their characteristic transmembrane topology, GPCRs are also known as heptahelical receptors, seven-transmembrane receptors, 7TM receptors, and serpentine receptors that "snake" across a cell membrane seven times (Fig. 1). The major role of GPCRs is to transmit signals into the cell. GPCR-associated proteins may play at least the following four distinct roles in receptor signaling[1–4]: (1) directly mediate receptor signaling, as in the case of G-proteins; (2) regulate receptor signaling through controlling receptor localization and/or trafficking; (3) act as a scaffold, physically linking the receptor to various effectors; and (4) act as an allosteric modulator of receptor conformation, altering receptor pharmacology and/or other aspects of receptor function.

The importance of GPCRs is also reflected by the fact that GPCR agonists and antagonists occupy approximately one-third of the world small molecule drug market. Much effort has been invested for studying GPCRs by both academic institutions and pharmaceutical industries. The functions of many of GPCRs are

Additional Supporting Information may be found in the online version of this article.

Correspondence to: X. Xiao; e-mail: xiaoxuan0326@yahoo.com.cn

Contract/grant sponsor: National Natural Science Foundation of China; contract/grant numbers: 60661003

Contract/grant sponsor: Province National Natural Science Foundation of Jiangxi; contract/grant numbers: 0611060

Contract/grant sponsor: The plan for training youth scientists (stars of Jing-Gang) of Jiangxi Province

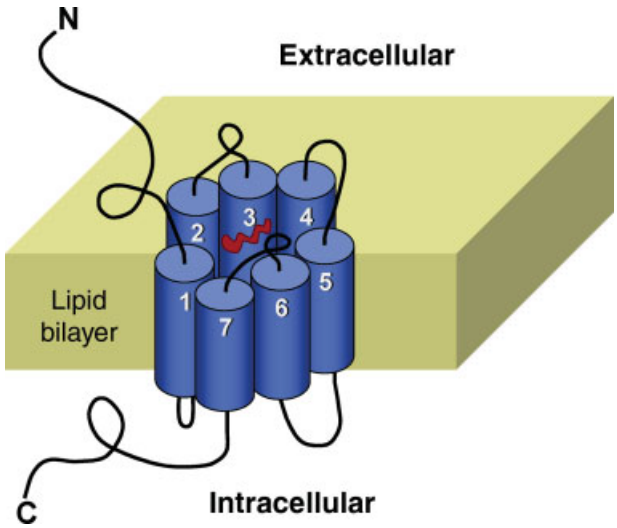

Figure 1. Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The seven-helix bundle, thus formed, has a central pore on its extracellular surface. The red entity located in the central pore represents a ligand messenger. Reproduced from ref. 17 with permission.

unknown and determining their ligands and signaling pathways is both time-consuming and costly. Particularly, as membrane proteins, GPCRs are very difficult to crystallize and most of them will not dissolve in normal solvents. Accordingly, so far very few crystal GPCR structures have been determined. Although the recently developed state-of-the-art NMR technique is a very powerful tool in determining the three-dimensional structures of membrane proteins,[5–8] it is time-consuming and costly. With the avalanche of protein sequence data generated in the post-genomic age, to timely conduct structure-based drug

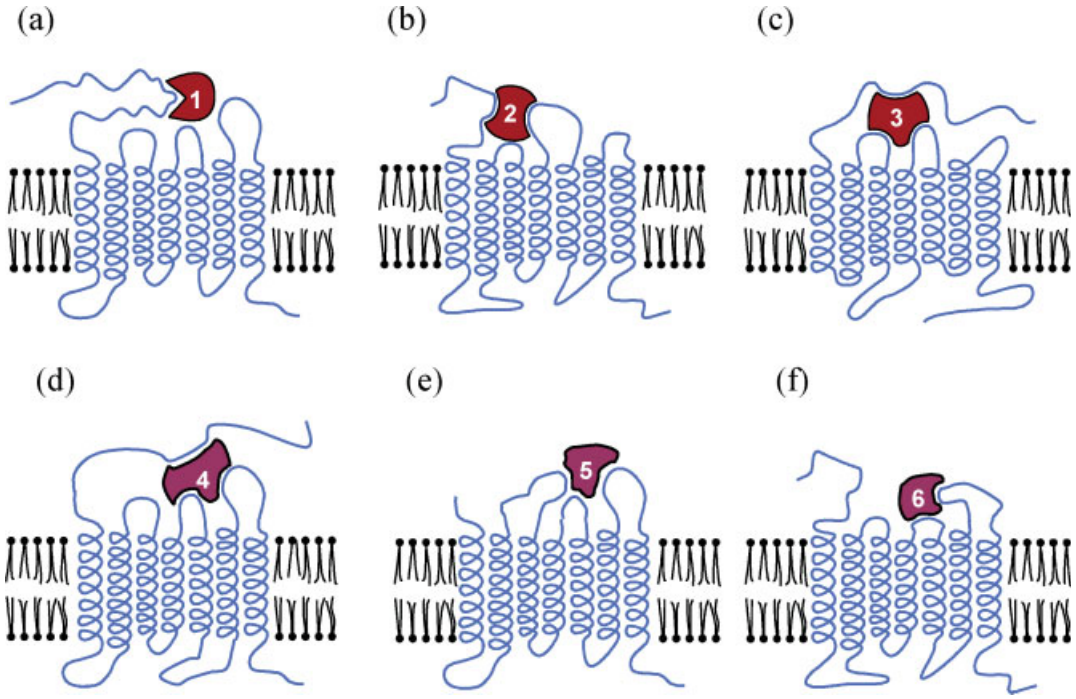

Figure 2. Schematic drawing to show six main function-different family classes of GPCRs: (a) rhodopsin-like, (b) secretin-like, (c) metabotrophic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothened family.

design, the approach of structural bioinformatics has been often used to develop the 3D (3-dimensional) structure of a targeted protein (see, e.g., refs. 9–13). Unfortunately, such an approach fails to work in most cases because very few GPCRs have sufficiently high sequence similarity with existing structure-known proteins, an indispensable condition for developing a reasonable starting structure via structural bioinformatics.[14] Therefore, it is highly desired to develop computational methods that can fast and effectively identify the functional families of GPCRs based on their primary sequences so as to help us classify drugs, a technique called "evolutionary pharmacology" for drug discovery.

Actually, in a pioneer study, the covariant-discriminant (CD) algorithm[15] was introduced to identify the 566 GPCRs within the rhodopsin-like family classified into seven subfamilies classes.[16] Later, a similar approach was used to study the 1,238 GPCRs classified into three main families.[17] Stimulated by the encouraged results, some follow-up studies were conducted as reported in.[18,19]

However, the prediction methods developed by the above investigators need further development because of the following reasons: (1) the reported success rates were derived based on a benchmark dataset without being rigorously screened by a clear data-culling operation to avoid redundancy and homologous bias, and hence the reported success rates therein might be overestimated; (2) none of these methods has provided a web-server for the public usage, and hence their practical application value is quite limited; and (3) they were established basically based on the amino acid (AA) composition. As is well known, using the AA composition to represent a protein sample would lose all of its sequence-order information, and hence affect the prediction quality.

The present study was initiated in an attempt to address these problems.

Materials

A statistical predictor needs two things: a benchmark dataset and an algorithm or operation engine. The benchmark dataset usually consists of a learning (or training) dataset and an independent testing dataset.[20] The former is for training the predictor's engine, whereas the latter for examining its accuracy via cross-validation. However, if the cross-validation is performed by the sub-sampling or jackknife approach, one dataset can serve both the training and testing purposes.[21]

To construct a high-quality benchmark dataset, protein sequences were collected from the G-protein–coupled receptor data base (GPCRDB) at http://www.gpcr.org/7tm/.[22] The GPCRDB is a molecular class-specific information system that collects, combines, validates, and disseminates heterogeneous data on GPCRs, and is updated automatically once every 4–5 months according to the Swiss-Prot and TrEMBL Data Banks.[23] According to the GPCRDB (release 10.0), GPCRs are classified into the following six main families: (a) rhodopsin-like, (b) secretin-like, (c) metabotrophic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothened family (Fig. 2). To guarantee the quality, the data were screened strictly according to the following criteria. First, all of the incomplete sequences (such as fragments) were removed. Second, to avoid any homology bias, a redundancy cutoff was imposed with the program CD-HIT[24] to winnow those sequences which have $\geq$40% pairwise sequence identity to any other in a same subset except for the fifth class (E-cAMP receptor), because it contained only 10 GPCR proteins. If the redundancy-cutoff operation was also executed on this class, the samples left would be too few to have any statistical significance.

After strictly following the above procedures, we finally obtained 365 GPCRs, of which (1) 232 are of rhodopsin-like, (2) 39 of secretin-like, (3) 44 of metabotrophic/glutamate/ pheromone, (4) 23 of fungal pheromone, (5) 10 of cAMP, and (6) 17 of frizzled/smoothened family. The accession number and sequence for each of the proteins in the six GPCR benchmark datasets are given in the *Supplementary Materials A*. Meanwhile, to train a statistical predictor to distinguish GPCR proteins from non-GPCR proteins, a non-GPCR benchmark dataset was also constructed by randomly collecting 365 non-GPCR proteins from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/ according their annotations in the CC (comment or notes) and ID (identification) fields. The corresponding accession numbers and sequences are given in the *Supplementary Materials B*, in which none of the entries has $\geq$40% pairwise sequence identity to any other.

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[25,26] 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.[27–30] In order to avoid complete losing of the sequence-order information and also enable the prediction more effective 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[31] was originally proposed for predicting protein subcellular localization and membrane protein type[31]; while the amphiphilic PseAA composition[32] 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 represent 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 protein structural class,[33–40] protein subcellular localization,[27,41–53] protein subnuclear localization,[54,55] protein submitochondria localization,[56] protein oligomer type,[57] conotoxin superfamily classification,[58,59] membrane protein type,[60–64] apoptosis protein subcellular localiza-

Table 1. Three Different Types for Coding Amino Acids.[a]

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

[a]The binary digital codes listed were derived using a model based on the similarity rule, complementarily rule, molecular recognition theory, and information theory. Such a model can better reflect the amino acid chemical physical properties and their degeneracy.

tion,[65–67] mycobacterial protein subcellular localization,[68] enzyme functional classification,[32,69–71] protein fold pattern,[72] signal peptide,[73,74] and other protein-related systems.[75,76] Owing to its wide usage, recently a very flexible PseAA composition generator, called "PseAAC",[77] 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, a novel approach by combining the "grey accumulative modeling" and the "cellular automaton image"[78,79] was introduced to derive the PseAA components. The CA images can reveal many important features of proteins, which are hidden in a long and complicated amino acid sequence.[80] The CA images have been applied to predict the effect on the replication ratio by HBV virus gene missense mutation[81] and predict the protein subcellular localization.[82] Meanwhile, various parameterization approaches have been used to characterize the image feature, such as Markov random fields, maximum local entropy, complexity measure, and complex wavelet coefficients.[35,43,83–85]

According to Wolfram's theory,[79] each protein sequence is corresponding to a CA image with its own textural feature. Thus, those proteins which belong to a same GPCR class must have some similar textures in their CA images. However, how do we extract the textural features that can give the greatest information pertaining to each texture? In other word, how to optimally extract these features and formulate them as a set of parameters is an important problem yet to be solved. Here, we are to introduce the gray level co-occurrence matrix (GLCM) approach to deal with this problem.

The essence of GLCM approach is to characterize image textures by a set of statistics for the occurrences of each gray level at different pixels and along different directions. Here, the term "feature" is used in texture classification to describe a set of statistics extracted for a co-occurrence matrix, characterizing the texture. For instance, energy, entropy and contrast can all be used as features. We chose four features derived from the GLCM approach as the PseAA components for a protein sample, and high success rates were observed in identifying the GPCR classification.

Cellular Automaton Image

A protein sequence is formed 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 pattern particularly when the sequence is very long. To cope with this situation, we resort to the images derived from the amino acid sequence by means of the space-time evolution of cellular automata.[78,79] As a first step, each of the 20 native amino acids is coded in a binary mode according to Table 1, which can better reflect the chemical and physical properties of an amino acid, as well as its structure and degeneracy.[80] Through the above encoding procedure, a protein sequence is transformed to a series of digital signals. For example, the sequence "MASAA…" is accordingly transformed to "1001111001010011100111001…."

Cellular automata are discrete dynamical systems whose behavior is completely specified in terms of a local relation. A CA can be thought of as a stylized universe consisting of a regular grid of cells, each of which is in one of a finite number of possible states, updated synchronously in discrete time steps according to a local, identical interaction rule.[79] The concept of cellular automata has attracted a great deal of interest in recent years because many extremely complex patterns can be evolved by just repeatedly applying some very simple rules. This is particularly useful for emulating complicated physical, social, and biological systems.

In this study, the practical approach to generate the CA image for a given protein sequence can be described as follows.

Suppose a protein **P** consists of N amino acids; i.e.,

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_N \tag{1}$$

where R_1 represents the first residue of the protein, R_2 the second residue, and so forth. According to Table 1, the residue chain of Eq.1 is initially converted to a sequence with $5N$ digits; i.e.,

$$\mathrm{P}(t) = g_1(t)g_2(t)\cdots g_N(t)\cdots g_{5N}(t), \quad (t=0) \tag{2}$$

where $g_i\ (t) = 0$ or 1 ($i = 1, 2, \ldots, 5N$) as defined by Table 1. Suppose the time for each updated step is consecutively expressed by $t = 0, 1, 2, \ldots, \Omega$, we have

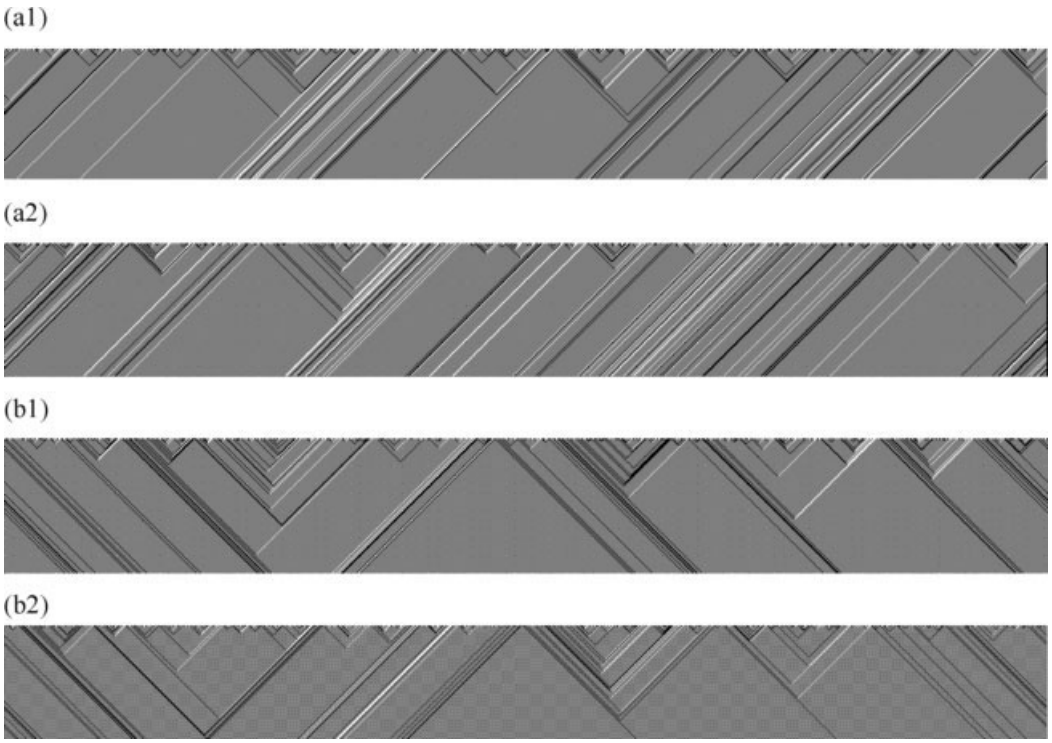

Figure 3. The cellular automaton image generated according to the procedures described in Section III.1 for (a1) the rhodopsin-like family member with accession number P41595; (a2) the rhodopsin-like family member with accession number P18599; (b1) the secretin-like family member with accession number O95838; and (b2) the secretin-like family member with accession number Q02644. As we can see, GPCR members in a same family share a quite similar texture in the cellular automation image.

$$\mathbf{P} = \begin{bmatrix} \mathrm{P}(0) \\ \mathrm{P}(1) \\ \vdots \\ \mathrm{P}(\Omega) \end{bmatrix} = \begin{bmatrix} \mathrm{g}_1(0)\mathrm{g}_2(0) & \cdots & \mathrm{g}_N(0) & \cdots & \mathrm{g}_{5N}(0) \\ \mathrm{g}_1(1)\mathrm{g}_2(1) & \cdots & \mathrm{g}_N(1) & \cdots & \mathrm{g}_{5N}(1) \\ & & & \vdots & \\ \mathrm{g}_1(\Omega)\mathrm{g}_2(\Omega) & \cdots & \mathrm{g}_N(\Omega) & \cdots & \mathrm{g}_{5N}(\Omega) \end{bmatrix} \tag{3}$$

where

$$g_i(t+1) = \begin{cases} 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=1 \end{cases} \quad (t = 0, 1, \cdots, \Omega) \tag{4}$$

with the spatially periodic boundary conditions; i.e.,

$$\mathrm{g}_0(t) = \mathrm{g}_{5N}(t) \text{ and } \mathrm{g}_{5N+1}(t) = \mathrm{g}_1(t) \tag{5}$$

In this study, the ith grid at t is filled with white color if $g_i(t) = 0$ and black if $g_i(t) = 1$. Accordingly, each $\mathbf{P}(t)$ ($t = 0, 1, 2, \ldots, \Omega$) in Eq.3 corresponds to a narrow ribbon mixed with white and black colors. Scanning these ribbons successively on to a screen or sheet will generate a 2D (2-dimensional) black-and-white image. It has been observed that the image texture is basically steady when $t = \Omega = 100$. The image thus evolved is called the CA image, and is represented by P of Eq.3. Its advantage is that it can help us visualize some special features hidden in a long and complex sequence.[80] For instance, the cellular automata images for proteins with a same biochemical attribute generally share a similar texture pattern as illustrated in Figure 3, while those with different biochemical attributes have different texture patterns, as illustrated in Figure 4.

Thus, all the existing tools in the area of image processing can be straightforwardly used for the current study.

GLCM and PseAA Components

To extract texture features from an image, there are numerous approaches available, such as those based on GLCM,[86] multiplicative autoregressive random fields, frequency domain filtering in terms of Fourier transform, and fractal dimension.[87,88] These methods each have different advantages and drawbacks in terms of computational burden and the capability of discriminating texture patterns. The studies in[89] showed that GLCM was the

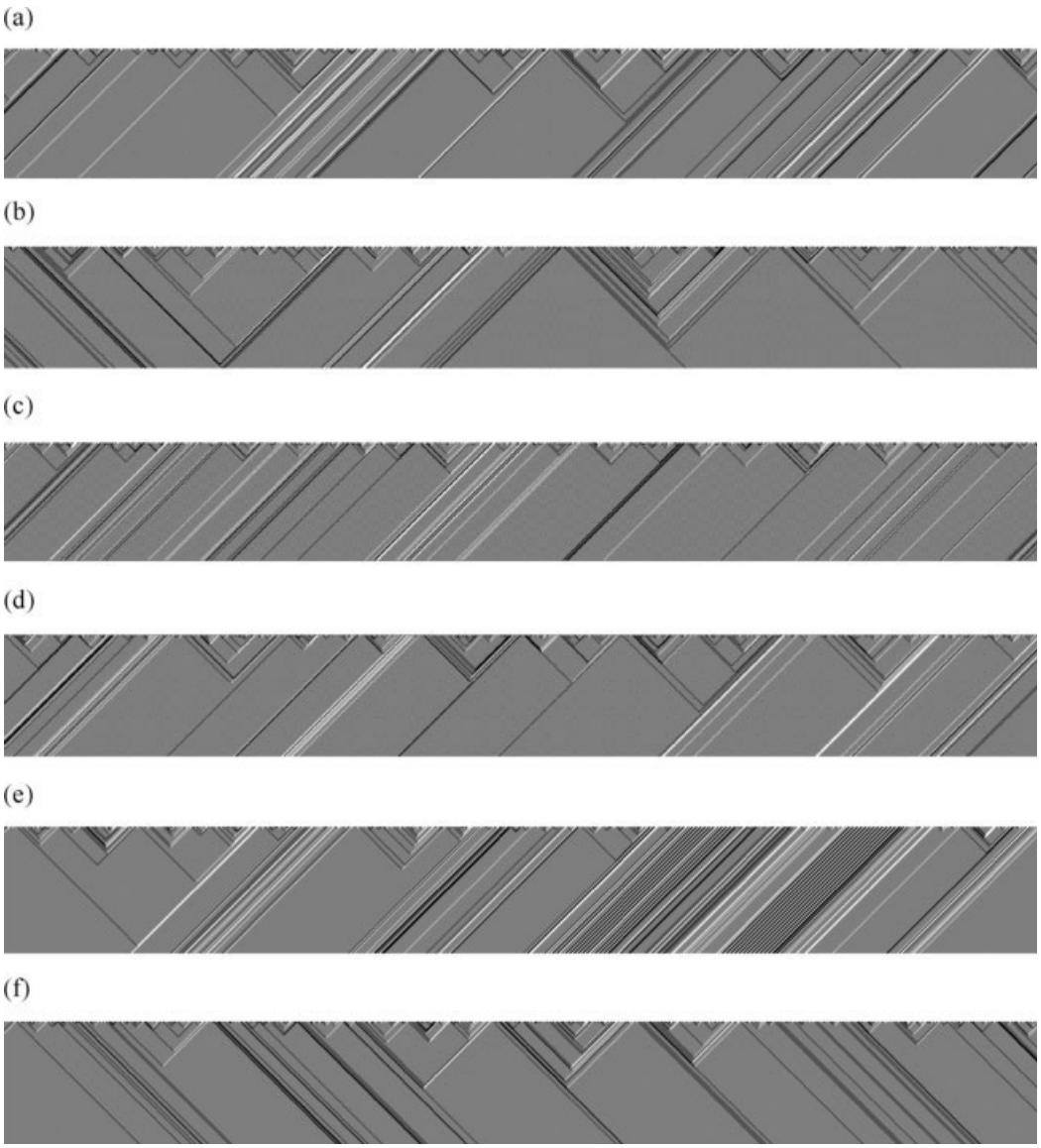

Figure 4. The cellular automaton image generated according to the procedures described in Section III.1 for a protein taken from (a) rhodopsin-like family, (b) secretin-like family, (c) metabotrophic/glutamate/pheromone family, (d) fungal pheromone family, (e) cAMP receptor family, and (f) frizzled/smoothened family, respectively.

best approach in terms of its higher texture discrimination accuracy.

According to the definition by Haralick et al.,[86] the GLCM is a matrix of frequencies at which two pixels, separated by a certain vector, occur in the image. The GLCM is a tabulation of how often different combinations of pixel grey level occur in an image, as can be formulated by

$$\mathbf{P}_{\mathrm{GLCM}} = \begin{bmatrix} \Lambda_{00} & \Lambda_{01} \\ \Lambda_{10} & \Lambda_{11} \end{bmatrix} \tag{6}$$

where $\mathbf{P}_{\mathrm{GLCM}}$ represents the GLCM for $\mathbf{P}$ in Eq.3; Λ_{00} is the total number of the "00" pairs counted when scanning $g_i(t)$ ($i = 1, 2, \ldots, 5N$; $t = 1, 2, \ldots, \Omega$) in Eq.3 row-by-row with each row scanned from left to right and then from right to left (i.e., according to the "round-trip" mode); Λ_{01}, the total number of the "01" pairs thus counted; and so forth. For example, when using the "round-trip" mode to scan the sequence "00011", the "00" pair number counted is 4, the "01" pair number is 1, the "10" pair number is 1, and the "11" pair number is 2.

Although $\mathbf{P}_{\mathrm{GLCM}}$ can provide a quantitative description of a spatial pattern, they are too unwieldy for practical image analysis. Thus, Haralick et al.[86] proposed a set of scalar quantities for summarizing the information contained in a GLCM. He originally proposed a total of 14 quantities, or features; however, typically only a subset of them is often used.[90] Here, we consider

the following four GLCM-derived features: (1) angular second moment (ASM), (2) contrast (CON), (3) inverse different moment (IDM), and (4) entropy (ENT), as formulated below:

$$\mathrm{ASM} = \sum_{i=0}^{1}\sum_{j=0}^{1}(\Lambda_{ij})^2 = \gamma\Phi_1 \tag{7}$$

$$\mathrm{CON} = \Lambda_{01} + \Lambda_{10} = \gamma\Phi_2 \tag{8}$$

$$\mathrm{IDM} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{\Lambda_{ij}}{1+(i-j)^2} = \gamma\Phi_3 \tag{9}$$

$$\mathrm{ENT} = -\sum_{i=0}^{1}\sum_{j=0}^{1}\Lambda_{ij}\log\Lambda_{ij} = \gamma\Phi_4 \tag{10}$$

where γ is the adjustment factor, which is for making Φ_j (j = 1,2,3,4) within the range easier to be handled. In the current study, $\gamma = 10^{15}$. According to Eqs. 7–10, we can compute Φ_j (j = 1,2,3,4) for any protein sequence as formulated by $\mathbf{P}$ of Eq. 3. These four quantities were used as the PseAA components.[31] The main advantage of introducing the PseAA components is that they can reflect some important features of a protein sequence through a discrete model.[27] Thus, according the Chou's PseAA composition,[31] a protein sequence can be expressed by a vector or a point in a 20 + 4 = 24D space; i.e.,

$$\mathbf{P} = \left[p_1, p_2, , \ldots, p_{20}, , p_{20+1}, \ldots, p_{20+4}\right]^{\mathrm{T}} \tag{11}$$

where $\mathbf{T}$ is the transpose operator, and

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{4} w_j\Phi_j}, & (1 \le k \le 20) \\ \dfrac{w_{(k-20)}\Phi_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{4} w_j\Phi_j}, & (21 \le k \le 24) \end{cases} \tag{12}$$

where f_i (i = 1, 2, …, 20) are the occurrence frequencies of the 20 native amino acids in the protein, arranged alphabetically according to their single letter codes, Φ_j (j = 1,2,3,4) are the characteristic quantities of the CA image of protein $\mathbf{P}$ as given by Eqs. 7–10, and w_j are the weight factors.

Covariant-Discriminant Classifier

Now the augmented CD algorithm[15,91] or CD classifier[21] was adopted to perform the prediction. For reader's convenience, a brief introduction about the CD classifier is given below.

Suppose a system containing N proteins ($\mathbf{P}_1$, $\mathbf{P}_2$, …, $\mathbf{P}_N$), which have been classified into M subsets (types), i.e.,

$$\mathbf{S} = S_1 \cup S_2 \cup S_3 \cup \ldots \cup S_M \tag{13}$$

where each subset S_{m} (m = 1, 2, …, M) is composed of proteins with the same type and its size (the number of proteins therein) is N_m. Obviously, we have $N = N_1 + N_2 + \cdots + N_M$. Now, for a query protein $\mathbf{P}$ (Eq.1), how can we identify which subset it belongs to? According to the CD classifier,[21] we can suppose without losing generality that the u-th protein in the subset S_{m} of Eq. 13 is formulated by (see Eq. 11)

$$\mathbf{P}_m^u = \left[p_{m,1}^u \;\; p_{m,2}^u \;\; \cdots \;\; p_{m,20}^u \;\; \cdots \;\; p_{m,24}^u\right]^{\mathbf{T}} \tag{14}$$

where $p_{m,j}^u$ (j = 1, 2, …, 24) is the j-th component of the u-th protein in S_m, and the standard vector for the subset S_m is defined by

$$\overline{\mathbf{P}}_m = \left[\overline{p}_{m,1} \;\; \overline{p}_{m,2} \;\; \cdots \;\; \overline{p}_{m,20} \;\; \cdots \;\; \overline{p}_{m,24}\right]^{\mathbf{T}} \tag{15}$$

where

$$\overline{p}_{m,i} = \frac{1}{N_m}\sum_{u=1}^{N_m} p_{m,i}^u, \quad (i = 1, 2, \ldots, 24) \tag{16}$$

Actually, $\overline{\mathbf{P}}_m$ as defined above can be deemed as a standard protein for the subset S_m. Thus, the similarity between a query protein $\mathbf{P}$ and $\overline{\mathbf{P}}_m$ is defined by the following covariant discriminant function:

$$\mathbf{F}(\mathbf{P}, \overline{\mathbf{P}}_m) = D_{\mathrm{Mah}}^2(\mathbf{P}, \overline{\mathbf{P}}_m) + \ln|\mathbf{C}_m|, \quad (m = 1, 2, \ldots, M) \tag{17}$$

where

$$D_{\mathrm{Mah}}^2(\mathbf{P}, \overline{\mathbf{P}}_m) = (\mathbf{P} - \overline{\mathbf{P}}_m)^{\mathbf{T}}\mathbf{C}_m^{-1}(\mathbf{P} - \overline{\mathbf{P}}_m) \tag{18}$$

is the squared Mahalanobis distance[92–94] (Reference 93 also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics) between $\mathbf{P}$ and $\overline{\mathbf{P}}_m$;

$$\mathbf{C}_m = \begin{bmatrix} c_{1,1}^m & c_{1,2}^m & \cdots & c_{1,24}^m \\ c_{2,1}^m & c_{2,2}^m & \cdots & c_{2,24}^m \\ \vdots & \vdots & \ddots & \vdots \\ c_{24,1}^m & c_{24,2}^m & \cdots & c_{24,24}^m \end{bmatrix} \tag{19}$$

is the covariance matrix for the subset S_m; the 24 × 24 elements in $\mathbf{C}_m$ are given by

$$c_{i,j}^m = \frac{1}{N_m - 1}\sum_{u=1}^{N_m}\left(p_{m,i}^u - \overline{p}_{m,i}\right)\left(p_{m,j}^u - \overline{p}_{m,j}\right), (i, j = 1, 2, \ldots, 24) \tag{20}$$

and $|\mathrm{C}_m|$ is the determinant of the matrix $\mathbf{C}_m$. The smaller the value of $\mathrm{F}(\mathbf{X},\overline{\mathbf{X}}_m)$, the higher the similarity between $\mathbf{P}$ and $\overline{\mathbf{P}}_m$. Therefore, the query protein is predicted belonging to the subset S_μ or the μ-th type if

$$\mu = \arg\min_m\left\{\mathbf{F}(\mathbf{P}, \overline{\mathbf{P}}_m)\right\}, \quad (m = 1, 2, \ldots, M) \tag{21}$$

where μ is the argument of m that minimizes $\mathbf{F}(\mathbf{P}, \overline{\mathbf{P}}_m)$. If there are two and more arguments leading to a same minimum value for $\mathbf{F}(\mathbf{P}, \overline{\mathbf{P}}_m)$, the query protein will be randomly assigned to one

of the subcellular locations associated with these arguments although this kind of tie case rarely happens. Note that owing to the normalization condition imposed by Eq. 12, of the 24 components in Eq. 15, only 23 are independent, and hence the covariance matrix $\mathbf{C}_m$ as defined by Eq. 19 must be a singular one.[95] This would lead the Mahalanobis distance defined by Eq. 18 and the covariant discriminant function by Eq. 17 to be divergent and meaningless. To cope with such a situation, the dimension-reducing procedure[94] was adopted in practical calculations; i.e., instead of 24D space, a protein sample is defined in a (24-1)D space by leaving out one of its 24 components. The remaining (24-1) components would be completely independent, thereby the corresponding covariance matrix $\mathbf{C}_m$ being no longer singular. In other words, the Mahalanobis distance (Eq. 18) and the covariant discriminant function (Eq. 17) based on such a (24-1)D space can be uniquely defined without any trouble. However, a question might be raised: which one of the 24 components can be left out? The answer is: anyone. Will it lead to a different predicted result by leaving out a different component? The answer is: no. According to Chou's invariance theorem (see Appendix A of Ref. 94), both the value of the Mahalanobis distance and the value of the determinant of $\mathbf{C}_m$ will remain exactly the same regardless of which one of the 24 components is left out. Accordingly, the final value of the covariant discriminant function (Eq. 17) can be uniquely defined through such a dimension-reducing procedure. For more details of the CD classifier, the reader is referred to the articles.[15,21,32]

The predictor thus formed is called GPCR-CA, where CA stands for Cellular Automaton, meaning that the characters of its image have been utilized in prediction.

Results and Discussion

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling test, and jackknife test.[20] However, as elucidated in,[27] among the three cross-validation methods, the jackknife test is deemed the most objective and can always yield a unique result for a given benchmark dataset, and has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., Ref. 45, 55, 66 68, 70, 96–107).

Table 2. Success Rates Obtained With the GPCR-CA Predictor[a] by Jackknife Test in Identifying the GPCR Proteins and Non-GPCR Proteins.

Protein type	Number of proteins	Number of correct predictions	Success rate (%)
GPCR	365[b]	337	92.33
Non-GPCR	365[c]	332	90.96
Overall	730	669	91.64

[a]The weight factors (see Eq. 12) for the four PseAA components are $w_1 = 1.4$, $w_2 = 1.2$, $w_3 = 1.8$, and $w_4 = 1.0$, respectively.
[b]The GPCR protein sequences are given in the Online Supporting Information A.
[c]The non-GPCR protein sequences are given in the Online Supporting Information B.

Table 3. Success Rates obtained With the GPCR-CA Predictor[a] by the Jackknife Test in Identifying Six Main GPCR Families.

GPCR main family	Number of proteins[b]	Number of correct predictions	Success rate (%)
Rhodopsin-like	232	224	96.55
Secretin-like	39	29	74.36
Metabotrophic/glutamate/pheromone	44	36	81.82
Fungal pheromone	23	2	8.70
cAMP receptor	10	6	60.00
Frizzled/smoothened family	17	8	47.06
Overall	365	305	83.56

[a]See the footnote "a" of Table 2.
[b]The corresponding sequences are given in the Online Supporting Information A.

The jackknife success rates obtained with the current GPCR-CA predictor in identifying proteins as GPCR or non-GPCR are given in Table 2, whereas those in identifying the GPCR proteins among their six main functional classes are given in Table 3.

It can be seen from the two tables that the overall success rate in identifying GPCR or non-GPCR proteins is about 92% whereas that in identifying GPCR proteins among their six main functional classes is about 84%, indicating that, even for the stringent benchmark dataset in which none of protein samples has $\geq$40% pairwise sequence identity to any other in a same subset, the GPCR-CA predictor can yield quite reliable results.

Meanwhile, we have also noticed that, the prediction rate for rhodopsin-like class is remarkably higher than those of the other classes. This is because in the current benchmark dataset the numbers of GPCRs in the other classes are still not sufficiently large. Even though, the overall success rate achieved by the current approach is still quite high. Let us imagine, if the GPCR samples are completely randomly distributed among the six possible categories, the overall success rate by random assignments would generally be 1/6 = 16.7%; if the random assignments are weighted according to the number of GPCRs in each class (see column 2 of Table 3), then the overall success rate would be[15]

$$\frac{1}{365^2}(232^2 + 39^2 + 44^2 + 23^2 + 10^2 + 17^2) = 43.7\% \qquad (22)$$

which is about 40% lower than the overall success rate by the current GPCR-CA predictor. It is anticipated that with more GPCR data available in future, particularly for those small subsets, the prediction quality by GPCR-CA will be further improved.

Conclusions

Playing a key role in cellular signaling networks and being the largest family of cell surface receptors, GPCRs regulate various physiological processes and are among the most frequent targets

of therapeutic drugs. However, the functions of many GPCRs are unknown, and determining their ligands and signaling pathways is both costly and time-consuming. This situation has motivated us to develop computational methods to predict the functional classification of GPCRs according to their primary sequences. The information thus obtained can help us classify drugs, a technique called "comparative and evolutionary pharmacology".

It has been demonstrated through this study that using the GLCM factors extracted from the CA images of proteins can more effectively reflect their overall sequence patterns so as to enhance the power in identifying GPCR functional classes. It is anticipated that the novel approach can also be used to improve the prediction quality for a series of other protein attributes, such as subcellular localization, membrane types, enzyme family and subfamily classes, and among many others. The web-server for the GPCR-CA predictor is freely accessible to the public at http://218.65.61.89:8080/bioinfo/GPCR-CA.

Acknowledgments

The authors thank the two anonymous reviewers whose constructive comments were very helpful for strengthening the presentation of this article.

References

1. Heuss, C.; Gerber, U. Trends Neurosci 2000, 23, 469.
2. Milligan, G.; White, J. H. Trends Pharmacol Sci 2001, 22, 513.
3. Hall, R. A.; Lefkowitz, R. J. Circ Res 2002, 91, 672.
4. Chou, K. C. J Proteome Res 2005, 4, 1681.
5. Oxenoid, K.; Chou, J. J. Proc Natl Acad Sci USA 2005, 102, 10870.
6. Call, M. E.; Schnell, J. R.; Xu, C.; Lutz, R. A.; Chou, J. J.; Wucherpfennig, K. W. Cell 2006, 127, 355.
7. Douglas, S. M.; Chou, J. J.; Shih, W. M. Proc Natl Acad Sci USA 2007, 104, 6644.
8. Schnell, J. R.; Chou, J. J. Nature 2008, 451, 591.
9. Chou, K. C. Biochem Biophys Res Commun 2004, 316, 636.
10. Chou, K. C. Biochem Biophys Res Commun 2004, 319, 433.
11. Chou, K. C. J Proteome Res 2004, 3, 1284.
12. Wei, D. Q.; Du, Q. S.; Sun, H.; Chou, K. C. Biochem Biophys Res Commun 2006, 344, 1048.
13. Wang, S. Q.; Du, Q. S.; Chou, K. C. Biochem Biophys Res Commun 2007, 354, 634.
14. Chou, K. C. Curr Med Chem 2004, 11, 2105.
15. Chou, K. C.; Elrod, D. W. Protein Eng 1999, 12, 107.
16. Chou, K. C.; Elrod, D. W. J Proteome Res 2002, 1, 429.
17. Chou, K. C. J Proteome Res 2005, 4, 1413.
18. Gao, Q. B.; Wang, Z. Z. Protein Eng Des Sel 2006, 19, 511.
19. Wen, Z.; Li, M.; Li, Y.; Guo, Y.; Wang, K. Amino Acids 2007, 32, 277.
20. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
21. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1.
22. Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. Nucleic Acids Res 1998, 26, 275.
23. Bairoch, A.; Apweiler, R. Nucleic Acids Res 2000, 25, 31.
24. Li, W.; Godzik, A. Bioinformatics 2006, 22, 1658.
25. Altschul, S. F. In Theoretical and Computational Methods in Genome Research; Suhai, S., Ed.; Plenum: New York, 1997, pp. 1–14.
26. Wootton, J. C.; Federhen, S. Comput Chem 1993, 17, 149.
27. Chou, K. C.; Shen, H. B. Nature Protocols 2008, 3, 153–162.
28. Emanuelsson, O.; Brunak, S.; von Heijne, G.; Nielsen, H. Nat Protoc 2007, 2, 953.
29. Garg, A.; Bhasin, M.; Raghava, G. P. J Biol Chem 2005, 280, 14427.
30. Nair, R.; Rost, B. Protein Sci 2002, 11, 2836.
31. Chou, K. C. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol. 44, 60) 2001, 43, 246.
32. Chou, K. C. Bioinformatics 2005, 21, 10.
33. Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P. Anal Biochem 2006, 357, 116.
34. Chen, C.; Tian, Y. X.; Zou, X. Y.; Cai, P. X.; Mo, J. Y. J Theor Biol 2006, 243, 444.
35. Xiao, X.; Shao, S. H.; Huang, Z. D.; Chou, K. C. J Comput Chem 2006, 27, 478.
36. Lin, H.; Li, Q. Z. J Comput Chem 2007, 28, 1463.
37. Ding, Y. S.; Zhang, T. L.; Chou, K. C. Protein Pept Lett 2007, 14, 811.
38. Zhang, T. L.; Ding, Y. S.; Chou, K. C. J Theor Biol 2008, 250, 186.
39. Xiao, X.; Lin, W. Z.; Chou, K. C. J Comput Chem 2008, 29, 2018.
40. Xiao, X.; Wang, P.; Chou, K. C. J Theor Biol 2008, 254, 691. doi:10.1016/j.jtbi. 2008.1006.1016.
41. Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. J Protein Chem 2003, 22, 395.
42. Chou, K. C.; Cai, Y. D. J Cell Biochem 2004, 91, 1197.
43. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Amino Acids 2005, 28, 57.
44. Chou, K. C.; Shen, H. B. J Cell Biochem 2006, 99, 517.
45. Li, F. M.; Li, Q. Z. Protein Pept Lett 2008, 15, 612.
46. Chou, K. C.; Shen, H. B. J Proteome Res 2007, 6, 1728.
47. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 355, 1006.
48. Chou, K. C.; Shen, H. B. J Cell Biochem 2007, 100, 665.
49. Chou, K. C.; Shen, H. B. BBRC 2006, 347, 150.
50. Shen, H. B.; Yang, J.; Chou, K. C. Amino Acids 2007, 33, 57.
51. Chou, K. C.; Shen, H. B. J Proteome Res 2006, 5, 3420.
52. Shen, H. B.; Chou, K. C. Protein Eng Des Sel 2007, 20, 39.
53. Shen, H. B.; Chou, K. C. Biopolymers 2007, 85, 233.
54. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2005, 337, 752.
55. Mundra, P.; Kumar, M.; Kumar, K. K.; Jayaraman, V. K.; Kulkarni, B. D. Pattern Recognit Lett 2007, 28, 1610.
56. Du, P.; Li, Y. BMC Bioinformatics 2006, 7, 518.
57. Chou, K. C.; Cai, Y. D. PROTEINS: Structure, Function, and Genetics 2003, 53, 282.
58. Mondal, S.; Bhavna, R.; Mohan Babu, R.; Ramakumar, S. J Theor Biol 2006, 243, 252.
59. Lin, H.; Li, Q. Z. Biochem Biophys Res Commun 2007, 354, 548.
60. Liu, H.; Wang, M.; Chou, K. C. Biochem Biophys Res Commun 2005, 336, 737.
61. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 288.
62. Wang, S. Q.; Yang, J.; Chou, K. C. J Theor Biol 2006, 242, 941.
63. Shen, H. B.; Yang, J.; Chou, K. C. J Theor Biol 2006, 240, 9.
64. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2007, 360, 339.

65. Jiang, X.; Wei, R.; Zhang, T. L.; Gu, Q. Protein Pept Lett 2008, 15, 392.
66. Chen, Y. L.; Li, Q. Z. J Theor Biol 2007, 248, 377.
67. Chen, Y. L.; Li, Q. Z. J Theor Biol 2007, 245, 775.
68. Lin, H.; Ding, H.; Feng-Biao Guo, F. B.; Zhang, A. Y.; Huang, J. Protein Pept Lett 2008, 15, 739.
69. Chou, K. C.; Cai, Y. D. Protein Sci 2004, 13, 2857.
70. Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. J Theor Biol 2007, 248, 546.
71. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 364, 53.
72. Shen, H. B.; Chou, K. C. Bioinformatics 2006, 22, 1717.
73. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2007, 357, 633.
74. Shen, H. B.; Chou, K. C. Biochem Biophys Res Commun 2007, 363, 297.
75. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Curr Top Med Chem 2007, 10, 1015.
76. Gonzalez-Díaz, H.; Gonzalez-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics 2008, 8, 750.
77. Shen, H. B.; Chou, K. C. Anal Biochem 2008, 373, 386.
78. Wolfram, S. Nature 1984, 311, 419.
79. Wolfram, S. A New Kind of Science; Wolfram Media: Champaign, Illinois, 2002.
80. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. Amino Acids 2005, 28, 29.
81. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. J Theor Biol 2005, 235, 555.
82. Xiao, X.; Shao, S. H.; Ding, Y. S.; Huang, Z. D.; Chou, K. C. Amino Acids 2006, 30, 49.
83. Gusev, V. D.; Nemytikova, L. A.; Chuzhanova, N. A. Bioinformatics 1999, 15, 994.
84. Gusev, V. D.; Nemytikova, L. A.; Chuzhanova N. A. Mol Biol (Mosk) 2001, 35, 1015.
85. Xiao, X.; Shao, S. H.; Chou, K. C. Biochem Biophys Res Commun 2006, 342, 605.
86. Haralick, R. M.; Shanmugam, K.; Dinstein, I. IEEE Trans Syst Man Cybern 1973, 3, 610.
87. Qiu, H. L.; Lam, N. S.; Quartrochi, D. A.; Gamon, J. A. Photogrammetric Eng Remote Sens 1999, 65, 63.
88. Fan, G.; Xia, X. G. IEEE Trans Geosci Remote Sens 2001, 39, 2680.
89. Tso, B.; Mather, P. M. IEEE Trans Geosci Remote Sens 1999, 37, 1255.
90. Newsam, S.; Wang, L.; Bhagavathy, S.; Manjunath, B. S. Appl Opt 2004, 43, 210.
91. Chou, K. C. Biochem Biophys Res Commun 2000, 278, 477.
92. Mahalanobis, P. C. Proc Natl Inst Sci India 1936, 2, 49.
93. Pillai, K. C. S. InEncyclopedia of Statistical Sciences, Vol. 5; Kotz, S., Johnson, N. L., Eds.; John Wiley: New York, 1985; pp. 176–181.
94. Chou, K. C. Proteins: Structure, Function & Genetics 1995, 21, 319.
95. Chou, K. C.; Zhang, C. T. J Biol Chem 1994, 269, 22014.
96. Kedarisetti, K. D.; Kurgan, L. A.; Dick, S. Biochem Biophys Res Commun 2006, 348, 981.
97. Jahandideh, S.; Abdolmaleki, P.; Jahandideh, M.; Asadabadi, E. B. Biophys Chem 2007, 128, 87.
98. Diao, Y.; Li, M.; Feng, Z.; Yin, J.; Pan, Y. J Theor Biol 2007, 247, 608.
99. Chen, K.; Kurgan, L. A.; Ruan, J. J Comput Chem 2008, 29, 1596.
100. Chen, C.; Chen, L. X.; Zou, X. Y.; Cai, P. X. J Theor Biol 2008, 253, 388.
101. Ding, Y. S.; Zhang, T. L. Pattern Recognit Lett 2008, 29, 1187. doi:10.1016/j.patrec. 1006.1007.
102. Du, P.; Li, Y. J Theor Biol 2008, 253, 579.
103. Jahandideh, S.; Sarvestani, A. S.; Abdolmaleki, P.; Jahandideh, M.; Barfeie, M. J Theor Biol 2007, 249, 785.
104. Shi, M. G.; Huang, D. S.; Li, X. L. Protein Pept Lett 2008, 15, 692.
105. Niu, B.; Cai, Y. D.; Lu, W. C.; Zheng, G. Y.; Chou, K. C. Protein Pept Lett 2006, 13, 489.
106. Jin, Y.; Niu, B.; Feng, K. Y.; Lu, W. C.; Cai, Y. D.; Li, G. Z. Protein Pept Lett 2008, 15, 286.
107. Niu, B.; Jin, Y. H.; Feng, K. Y.; Liu, L.; Lu, W. C.; Cai, Y. D.; Li, G. Z. Protein Pept Lett 2008, 15, 590.