

Journal of Computational Chemistry (JCC) Journal

Copy of e-mail Notification

Journal of Computational Chemistry (JCC) Published by John Wiley & Sons, Inc.

Dear Sir or Madam,

Please refer to this URL address

<http://kwglobal.co.in/jw/retrieval.aspx?pwd=e57feaad17f3>

Login: your e-mail address

Password: e57feaad17f3

The site contains 1 file. You will need to have Adobe Acrobat Reader software to read these files. This is free software and is available for user downloading at <http://www.adobe.com/products/acrobat/readstep.html>. If you have the Notes annotation tool (not contained within Acrobat reader), you can make corrections electronically and return them to Wiley as an e-mail attachment (see the Notes tool instruction sheet). Handwritten changes on the proofs returned via fax are also acceptable.

This file contains:

Author Instructions Checklist

Adobe Acrobat Users - NOTES tool sheet

Reprint Order form

A copy of your page proofs for your article

NOTE: Color figure proofs will arrive under separate cover

After printing the PDF file, please read the page proofs carefully and:

- 1) indicate changes or corrections in the margin of the page proofs;
- 2) answer all queries (footnotes A,B,C, etc.) on the last page of the PDF proof;
- 3) proofread any tables and equations carefully;
- 4) check that any Greek, especially "mu", has translated correctly.

Within 48 hours, please return the following to the address given below:

- 1) original PDF set of page proofs,
- 2) print quality hard copy figures for corrections (we CANNOT accept figures on disk at this stage),
- 3) Signed Copyright Transfer Agreement,
- 4) Reprint Order form,
- 5) Return fax form

Return to:

Journal of Computational Chemistry (JCC) Journal

Copy of e-mail Notification

Production Editor, JCC
Cadmus Professional Communications - Ephrata
300 West Chestnut Street
Ephrata, PA 17522
phone: 717-721-2699
fax: 717-738-9478 or 717-738-9479

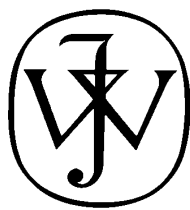
If you experience technical problems, please contact Prasanth/Sankar/Balaji(e-mail: wileysupport@kwglobal.com,
phone: +91 (44) 4205-8888 (ext.310). Be sure to include your article number.

If you have any questions regarding your article, please contact me. PLEASE ALWAYS INCLUDE YOUR
ARTICLE NO. (08-0160.R1) WITH ALL CORRESPONDENCE.

This e-proof is to be used only for the purpose of returning corrections to the publisher.

Sincerely,

Production Editor, JCC
Cadmus Professional Communications - Ephrata
300 West Chestnut Street
Ephrata, PA 17522
phone: 717-721-2699
fax: 717-738-9478 or 717-738-9479



WILEY

Publishers Since 1807

111 RIVER STREET, HOBOKEN, NJ 07030

*****IMMEDIATE RESPONSE REQUIRED*****

Please follow these instructions to avoid delay of publication.

☐ **READ PROOFS CAREFULLY**

- This will be your only chance to review these proofs.
- Please note that the volume and page numbers shown on the proofs are for position only.

☐ **ANSWER ALL QUERIES ON PROOFS** (Queries for you to answer are attached as the last page of your proof.)

- Mark all corrections directly on the proofs. Note that excessive author alterations may ultimately result in delay of publication and extra costs may be charged to you.

☐ **CHECK FIGURES AND TABLES CAREFULLY** (Color figures will be sent under separate cover.)

- Check size, numbering, and orientation of figures.
- All images in the PDF are downsampled (reduced to lower resolution and file size) to facilitate Internet delivery. These images will appear at higher resolution and sharpness in the printed article.
- Review figure legends to ensure that they are complete.
- Check all tables. Review layout, title, and footnotes.

☐ **COMPLETE REPRINT ORDER FORM**

- Fill out the attached reprint order form. It is important to return the form even if you are not ordering reprints. You may, if you wish, pay for the reprints with a credit card. Reprints will be mailed only after your article appears in print. This is the most opportune time to order reprints. If you wait until after your article comes off press, the reprints will be considerably more expensive.

RETURN

- ☐ **PROOFS**
☐ **REPRINT ORDER FORM**
☐ **CTA (If you have not already signed one)**

RETURN WITHIN 48 HOURS OF RECEIPT VIA FAX TO 1-717-738-9478 or 717-738-9479

QUESTIONS?

Production Editor, JCC
Phone: 717-721-2699
E-mail: jrnprodjcc@cadmus.com
Refer to journal acronym and article production number

Softproofing for advanced Adobe Acrobat Users – NOTES tool

NOTE: ADOBE READER FROM THE INTERNET DOES NOT CONTAIN THE NOTES TOOL USED IN THIS PROCEDURE.

Acrobat annotation tools can be very useful for indicating changes to the PDF proof of your article. By using Acrobat annotation tools, a full digital pathway can be maintained for your page proofs.

The NOTES annotation tool can be used with either Adobe Acrobat 6.0 or Adobe Acrobat 7.0. Other annotation tools are also available in Acrobat 6.0, but this instruction sheet will concentrate on how to use the NOTES tool. Acrobat Reader, the free Internet download software from Adobe, DOES NOT contain the NOTES tool. In order to softproof using the NOTES tool you must have the full software suite Adobe Acrobat Exchange 6.0 or Adobe Acrobat 7.0 installed on your computer.

Steps for Softproofing using Adobe Acrobat NOTES tool:

1. Open the PDF page proof of your article using either Adobe Acrobat Exchange 6.0 or Adobe Acrobat 7.0. Proof your article on-screen or print a copy for markup of changes.
2. Go to Edit/Preferences/Commenting (in Acrobat 6.0) or Edit/Preferences/Commenting (in Acrobat 7.0) check “Always use login name for author name” option. Also, set the font size at 9 or 10 point.
3. When you have decided on the corrections to your article, select the NOTES tool from the Acrobat toolbox (Acrobat 6.0) and click to display note text to be changed, or Comments/Add Note (in Acrobat 7.0).
4. Enter your corrections into the NOTES text box window. Be sure to clearly indicate where the correction is to be placed and what text it will effect. If necessary to avoid confusion, you can use your TEXT SELECTION tool to copy the text to be corrected and paste it into the NOTES text box window. At this point, you can type the corrections directly into the NOTES text box window. **DO NOT correct the text by typing directly on the PDF page.**
5. Go through your entire article using the NOTES tool as described in Step 4.
6. When you have completed the corrections to your article, go to Document/Export Comments (in Acrobat 6.0) or Comments/Export Comments (in Acrobat 7.0). Save your NOTES file to a place on your harddrive where you can easily locate it. **Name your NOTES file with the article number assigned to your article in the original softproofing e-mail message.**
7. **When closing your article PDF be sure NOT to save changes to original file.**
8. To make changes to a NOTES file you have exported, simply re-open the original PDF proof file, go to Document/Import Comments and import the NOTES file you saved. Make changes and reexport NOTES file keeping the same file name.
9. When complete, attach your NOTES file to a reply e-mail message. Be sure to include your name, the date, and the title of the journal your article will be printed in.

Publishers Since 1807

PHONE: (201) 748-8789; FAX: (201) 748-6326

PREPUBLICATION REPRINT ORDER FORM

JOURNAL: *JOURNAL OF COMPUTATIONAL CHEMISTRY* VOLUME ISSUE

TITLE OF MANUSCRIPT

MS. NO. _____ NO. OF PAGES _____ AUTHOR(S) _____

No. of Pages	100 Reprints	200 Reprints	300 Reprints	400 Reprints	500 Reprints
	\$	\$	\$	\$	\$
1-4	336	501	694	890	1,052
5-8	469	703	987	1,251	1,477
9-12	594	923	1,234	1,565	1,850
13-16	714	1,156	1,527	1,901	2,273
17-20	794	1,340	1,775	2,212	2,648
21-24	911	1,529	2,031	2,536	3,037
25-28	1,004	1,707	2,267	2,828	3,388
29-32	1,108	1,894	2,515	3,135	3,755
33-36	1,219	2,092	2,773	3,456	4,143
37-40	1,329	2,290	3,033	3,776	4,528

COVERS

100 Covers - \$90	•	200 Covers - \$145	•	300 Covers - \$200
400 Covers - \$255	•	500 Covers - \$325	•	Additional 100s - \$65

☐ Please send me reprints of the above article at..... \$

☐ Please send me _____ Generic covers of the above journal at..... \$ _____

Please add appropriate State and Local Tax {Tax Exempt No. _____} \$ _____

Please add 5% Postage and Handling..... \$_____

TOTAL AMOUNT OF ORDER** \$

***International orders must be paid in U.S. currency and drawn on a U.S. bank*

Please check one: ☐ Check enclosed

 Bill me

☐ Credit Card

If credit card order, charge to: ☐ American Express

☐ Visa

□ Discover

Credit Card No. _____ Signature _____ Exp. Date _____

Bill To:

Ship To:

Name _____

Name _____

Address/Institution

Address/Institution

Purchase Order No. _____

Phone _____ Fax _____

E-mail: _____

Date:

To:

Production/Contribution

ID# _____

Publisher/Editorial office use only

Re: Manuscript entitled _____ (the "Contribution")
for publication in _____ (the "Journal")
published by Wiley Periodicals, Inc. ("Wiley").

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable Wiley to disseminate your work to the fullest extent, we need to have this Copyright Transfer Agreement signed and returned to us as soon as possible. If the Contribution is not accepted for publication this Agreement shall be null and void.

A. COPYRIGHT

1. The Contributor assigns to Wiley, during the full term of copyright and any extensions or renewals of that term, all copyright in and to the Contribution, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution and the material contained therein in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.
2. Reproduction, posting, transmission or other distribution or use of the Contribution or any material contained therein, in any medium as permitted hereunder, requires a citation to the Journal and an appropriate credit to Wiley as Publisher, suitable in form and content as follows: (Title of Article, Author, Journal Title and Volume/Issue Copyright © [year] Wiley Periodicals, Inc. or copyright owner as specified in the Journal.)

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution, and the right to make oral presentations of material from the Contribution.

C. OTHER RIGHTS OF CONTRIBUTOR

Wiley grants back to the Contributor the following:

1. The right to share with colleagues print or electronic "preprints" of the unpublished Contribution, in form and content as accepted by Wiley for publication in the Journal. Such preprints may be posted as electronic files on the Contributor's own website for personal or professional use, or on the Contributor's internal university or corporate networks/intranet, or secure external website at the Contributor's institution, but not for commercial sale or for any systematic external distribution by a third party (e.g., a listserve or database connected to a public access server). Prior to publication, the Contributor must include the following notice on the preprint: "This is a preprint of an article accepted for publication in [Journal title] © copyright (year) (copyright owner as specified in the Journal)". After publication of the Contribution by Wiley, the preprint notice should be amended to read as follows: "This is a preprint of an article published in [include the complete citation information for the final version of the Contribution as published in the print edition of the Journal]", and should provide an electronic link to the Journal's WWW site, located at the following Wiley URL: <http://www.interscience.Wiley.com/>. The Contributor agrees not to update the preprint or replace it with the published version of the Contribution.
2. The right, without charge, to photocopy or to transmit online or to download, print out and distribute to a colleague a copy of the published Contribution in whole or in part, for the colleague's personal or professional use, for the

advancement of scholarly or scientific research or study, or for corporate informational purposes in accordance with Paragraph D.2 below.

3. The right to republish, without charge, in print format, all or part of the material from the published Contribution in a book written or edited by the Contributor.
4. The right to use selected figures and tables, and selected text (up to 250 words, exclusive of the abstract) from the Contribution, for the Contributor's own teaching purposes, or for incorporation within another work by the Contributor that is made part of an edited work published (in print or electronic format) by a third party, or for presentation in electronic format on an internal computer network or external website of the Contributor or the Contributor's employer.
5. The right to include the Contribution in a compilation for classroom use (course packs) to be distributed to students at the Contributor's institution free of charge or to be stored in electronic format in datarooms for access by students at the Contributor's institution as part of their course work (sometimes called "electronic reserve rooms") and for in-house training programs at the Contributor's employer.

D. CONTRIBUTIONS OWNED BY EMPLOYER

1. If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/employer which must sign this Agreement (in addition to the Contributor's signature), in the space provided below. In such case, the company/employer hereby assigns to Wiley, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.
2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, Wiley hereby grants back, without charge, to such company/employer, its subsidiaries and divisions, the right to make copies of and distribute the published Contribution internally in print format or electronically on the Company's internal network. Upon payment of Wiley's reprint fee, the institution may distribute (but not resell) print copies of the published Contribution externally. Although copies so made shall not be available for individual re-sale, they may be included by the company/employer as part of an information package included with software or other products offered for sale or license. Posting of the published Contribution by the institution on a public access website may only be done with Wiley's written permission, and payment of any applicable fee(s).

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government Employees: see note at end.)

F. COPYRIGHT NOTICE

The Contributor and the company/employer agree that any and all copies of the Contribution or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the Journal as published by Wiley.

G. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that the Contribution is the Contributor's original work. If the Contribution was prepared jointly, the Contributor agrees to inform the co-Contributors of the terms of this Agreement and to obtain their signature to this Agreement or their written permission to sign on their behalf. The Contribution is submitted only to this Journal and has not been published before, except for "preprints" as permitted above. (If excerpts from copyrighted works owned by third parties are included, the Contributor will obtain written permission from the copyright owners for all uses as set forth in Wiley's permissions form or in the Journal's Instructions for Contributors, and show credit to the sources in the Contribution.) The Contributor also warrants that the Contribution contains no libelous or unlawful statements, does not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, or contain material or instructions that might cause harm or injury.

CHECK ONE:

☐ Contributor-owned work

Contributor's signature

Date

Type or print name and title

Co-contributor's signature

Date

Type or print name and title

ATTACHED ADDITIONAL SIGNATURE PAGE AS NECESSARY

☐ Company/Institution-owned work
(made-for-hire in the
course of employment)

Company or Institution (Employer-for-Hire)

Date

Authorized signature of Employer

Date

☐ U.S. Government work

Note to U.S. Government Employees

A contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. Government publication is called a "U.S. Government work," and is in the public domain in the United States. In such case, the employee may cross out Paragraph A.1 but must sign and return this Agreement. If the Contribution was not prepared as part of the employee's duties or is not an official U.S. Government publication, it is not a U.S. Government work.

☐ U.K. Government work (Crown Copyright)

Note to U.K. Government Employees

The rights in a Contribution prepared by an employee of a U.K. government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown. In such case, Wiley will forward the relevant form to the Employee for signature.



111 RIVER STREET, HOBOKEN, NJ 07030

Telephone Number:

• Facsimile Number:

To: Production Editor, JCC

Company: Cadmus Professional Communications

Phone: 717-721-2699

Fax: 717-738-9478 or 717-738-9479

From: _____

Date: _____

Pages including this cover
page: _____

Message:

Re:

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

XUAN XIAO,¹ PU WANG,¹ KUO-CHEN CHOU²

¹Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300, China

²Gordon Life Science Institute, San Diego, California 92130

Received 27 March 2008; Revised 12 September 2008; Accepted 7 October 2008

DOI 10.1002/jcc.21163

Published online 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) metabotropic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothed 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.

© 2008 Wiley Periodicals, Inc. J Comput Chem 00: 000–000, 2008

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 receptor (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

© 2008 Wiley Periodicals, Inc.

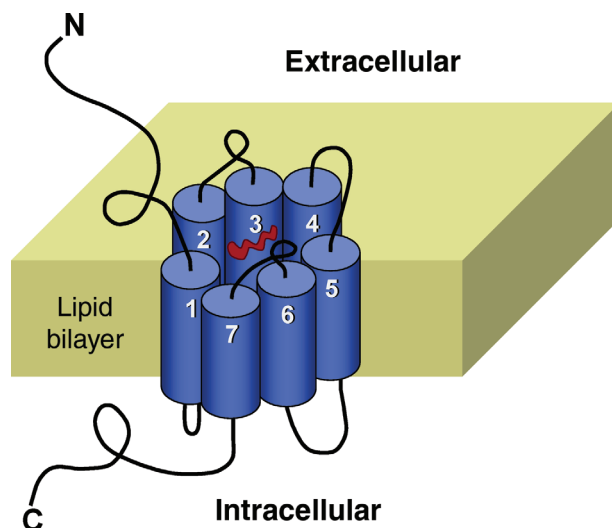


Figure 1. Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The seven-helix bundle, thus formed, has a central pore on its extracellular surface. The red entity located in the central pore represents a ligand messenger. Reproduced from¹⁷ 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

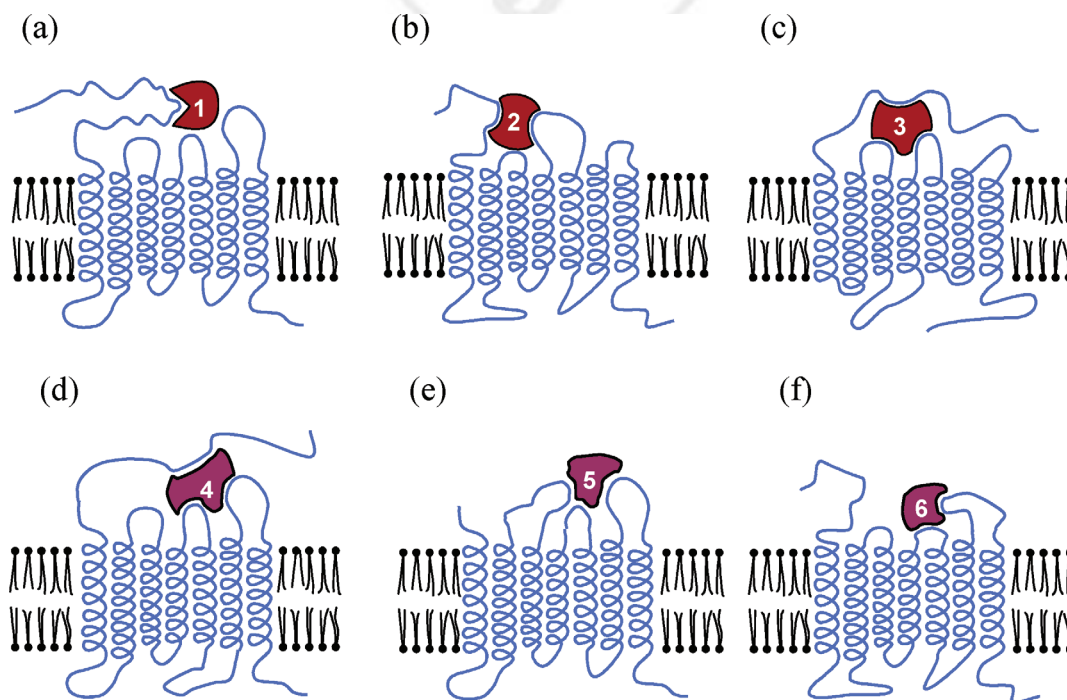


Figure 2. Schematic drawing to show six main function-different family classes of GPCRs: (a) rhodopsin-like, (b) secretin-like, (c) metabotropic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothed family.

design, the approach of structural bioinformatics has been often used to develop the 3D (3-dimensional) structure of a targeted protein (see, e.g.,^{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) algorithm¹⁵ 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) metabotropic/glutamate/pheromone; (d) fungal pheromone, (e) cAMP receptor, and (f) frizzled/smoothed 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-HIT²⁴ 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 metabotropic/glutamate/pheromone, (4) 23 of fungal pheromone, (5) 10 of cAMP, and (6) 17 of frizzled/smoothed 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 composition³¹ was originally proposed for predicting protein subcellular localization and membrane protein type³¹; while the amphiphilic PseAA composition³² 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 localization,

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

^aThe 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 mutation⁸¹ 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.,

$$\mathbf{P} = R_1 R_2 \cdots R_N \quad (1)$$

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.,

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

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

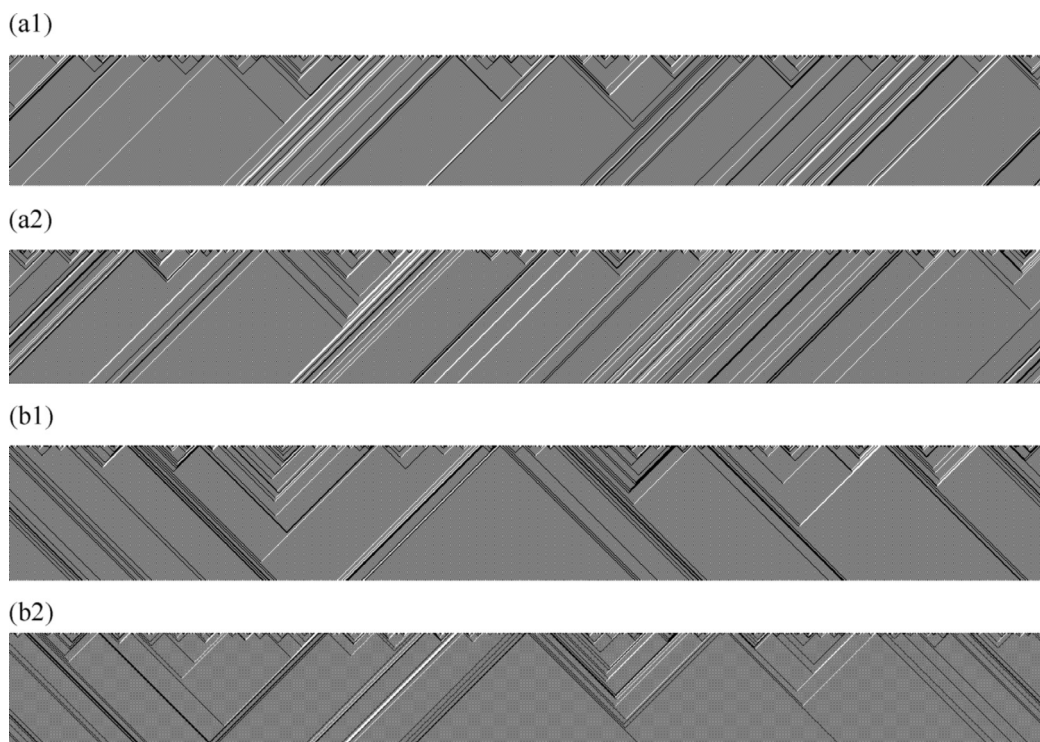


Figure 3. The cellular automaton image generated according to the procedures described in Section III.1 for (a1) the rhodopsin-like family member with accession number P41595; (a2) the rhodopsin-like family member with accession number P18599; (b1) the secretin-like family member with accession number 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 automaton image.

$$P = \begin{bmatrix} P(0) \\ P(1) \\ \vdots \\ P(\Omega) \end{bmatrix} = \begin{bmatrix} g_1(0)g_2(0) & \cdots & g_N(0) & \cdots & g_{5N}(0) \\ g_1(1)g_2(1) & \cdots & g_N(1) & \cdots & g_{5N}(1) \\ \vdots & & \vdots & & \vdots \\ g_1(\Omega)g_2(\Omega) & \cdots & g_N(\Omega) & \cdots & g_{5N}(\Omega) \end{bmatrix} \quad (3)$$

where

$$g_i(t+1)$$

$$= \begin{cases} 0, & \text{if } g_{i-1}(t) = 0, g_i(t) = 0, g_{i+1}(t) = 0 \\ 0, & \text{if } g_{i-1}(t) = 0, g_i(t) = 0, g_{i+1}(t) = 1 \\ 1, & \text{if } g_{i-1}(t) = 0, g_i(t) = 1, g_{i+1}(t) = 0 \\ 0, & \text{if } g_{i-1}(t) = 0, g_i(t) = 1, g_{i+1}(t) = 1 \\ 1, & \text{if } g_{i-1}(t) = 1, g_i(t) = 0, g_{i+1}(t) = 0 \\ 0, & \text{if } g_{i-1}(t) = 1, g_i(t) = 0, g_{i+1}(t) = 1 \\ 1, & \text{if } g_{i-1}(t) = 1, g_i(t) = 1, g_{i+1}(t) = 0 \\ 0, & \text{if } g_{i-1}(t) = 1, g_i(t) = 1, g_{i+1}(t) = 1 \end{cases} \quad (t = 0, 1, \dots, \Omega) \quad (4)$$

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

$$g_0(t) = g_{5N}(t) \text{ and } g_{5N+1}(t) = g_1(t) \quad (5)$$

In this study, the i th grid at t is filled with white color if $g_i(t) = 0$ and black if $g_i(t) = 1$. Accordingly, each $P(t)$ ($t = 0, 1, 2, \dots, \Omega$) 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 in⁸⁹ showed that GLCM was the

F3
F4

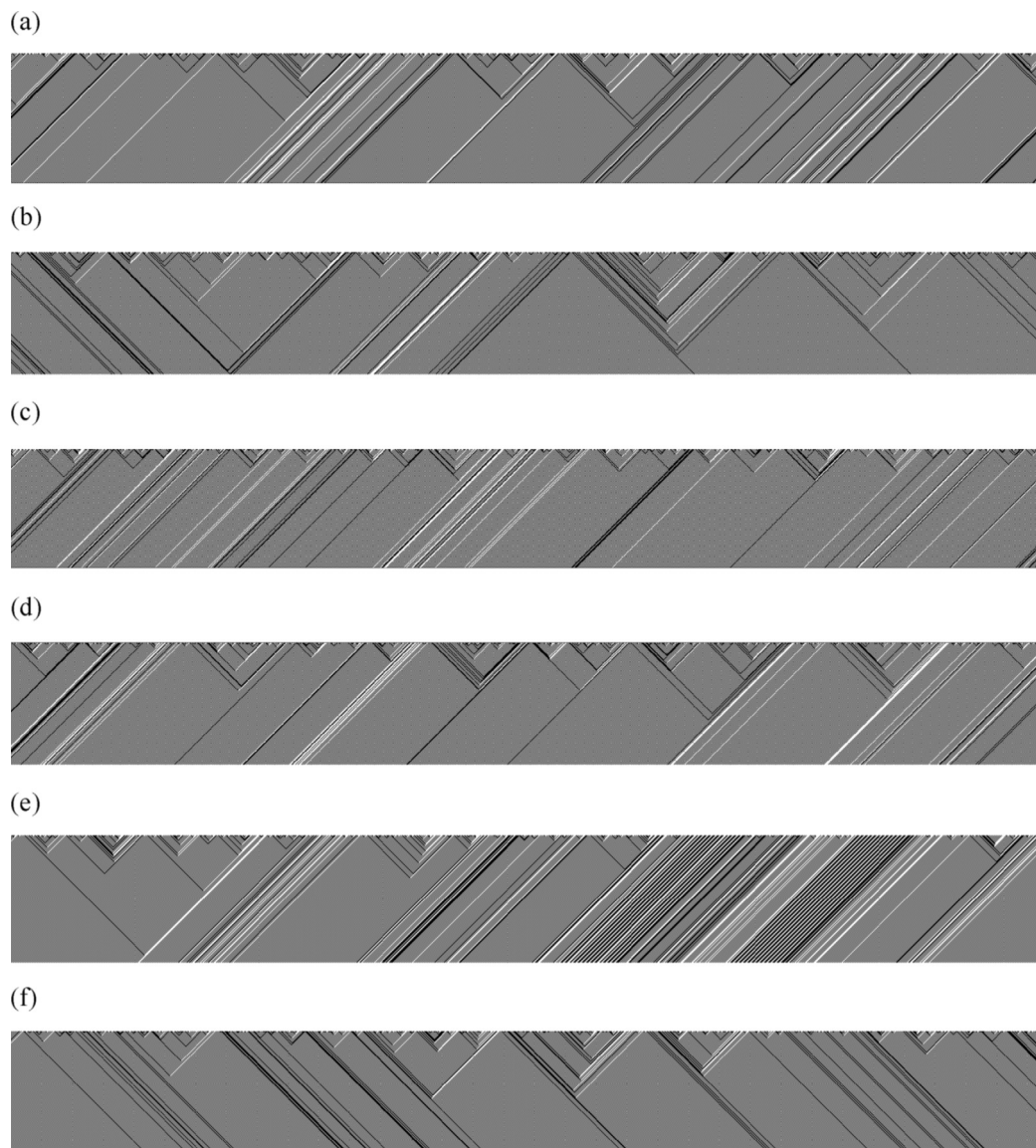


Figure 4. The cellular automaton image generated according to the procedures described in Section III.1 for a protein taken from (a) rhodopsin-like family, (b) secretin-like family, (c) metabotropic/glutamate/pheromone family, (d) fungal pheromone family, (e) cAMP receptor family, and (f) frizzled/smoothed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

$$\mathbf{P}_{\text{GLCM}} = \begin{bmatrix} \Lambda_{00} & \Lambda_{01} \\ \Lambda_{10} & \Lambda_{11} \end{bmatrix} \quad (6)$$

where \mathbf{P}_{GLCM} represents the GLCM for \mathbf{P} in Eq.3; Λ_{00} is the total number of the “00” pairs counted when scanning $g_i(t)$

($i = 1, 2, \dots, 5N$; $t = 1, 2, \dots, \Omega$) 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}_{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:

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

$$\text{CON} = \Lambda_{01} + \Lambda_{10} = \gamma \Phi_2 \quad (8)$$

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

$$\text{ENT} = - \sum_{i=0}^1 \sum_{j=0}^1 \Lambda_{ij} \log \Lambda_{ij} = \gamma \Phi_4 \quad (10)$$

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 to the Chou's PseAA composition,³¹ a protein sequence can be expressed by a vector or a point in a $20 + 4 = 24\text{D}$ space; i.e.,

$$\mathbf{P} = [p_1, p_2, \dots, p_{20}, p_{20+1}, \dots, p_{20+4}]^T \quad (11)$$

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

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

where f_i ($i = 1, 2, \dots, 20$) 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 classifier²¹ 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 ($\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_N$), which have been classified into M subsets (types), i.e.,

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3 \cup \dots \cup \mathbf{S}_M \quad (13)$$

where each subset \mathbf{S}_m ($m = 1, 2, \dots, M$) is composed of proteins with the same type and its size (the number of proteins

therein) is N_m . Obviously, we have $N = N_1 + N_2 + \dots + N_M$. 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 \mathbf{S}_m of Eq. 13 is formulated by (see Eq. 11)

$$\mathbf{P}_m^u = [p_{m,1}^u, p_{m,2}^u, \dots, p_{m,20}^u, \dots, p_{m,24}^u]^T \quad (14)$$

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

$$\bar{\mathbf{P}}_m = [\bar{p}_{m,1}, \bar{p}_{m,2}, \dots, \bar{p}_{m,20}, \dots, \bar{p}_{m,24}]^T \quad (15)$$

where

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

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

$$F(\mathbf{P}, \bar{\mathbf{P}}_m) = D_{\text{Mah}}^2(\mathbf{P}, \bar{\mathbf{P}}_m) + \ln |\mathbf{C}_m|, \quad (m = 1, 2, \dots, M) \quad (17)$$

where

$$D_{\text{Mah}}^2(\mathbf{P}, \bar{\mathbf{P}}_m) = (\mathbf{P} - \bar{\mathbf{P}}_m)^T \mathbf{C}_m^{-1} (\mathbf{P} - \bar{\mathbf{P}}_m) \quad (18)$$

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 $\bar{\mathbf{P}}_m$; P_m

$$\mathbf{C}_m = \begin{bmatrix} c_{1,1}^m & c_{1,2}^m & \dots & c_{1,24}^m \\ c_{2,1}^m & c_{2,2}^m & \dots & c_{2,24}^m \\ \vdots & \vdots & \ddots & \vdots \\ c_{24,1}^m & c_{24,2}^m & \dots & c_{24,24}^m \end{bmatrix} \quad (19)$$

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

$$c_{i,j}^m = \frac{1}{N_m - 1} \sum_{u=1}^{N_m} (p_{m,i}^u - \bar{p}_{m,i}) (p_{m,j}^u - \bar{p}_{m,j}), \quad (i, j = 1, 2, \dots, 24) \quad (20)$$

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

$$\mu = \arg \min_m \{F(\mathbf{P}, \bar{\mathbf{P}}_m)\}, \quad (m = 1, 2, \dots, M) \quad (21)$$

where μ is the argument of m that minimizes $F(\mathbf{P}, \bar{\mathbf{P}}_m)$. If there are two and more arguments leading to a same minimum value for $F(\mathbf{P}, \bar{\mathbf{P}}_m)$, 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 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 procedure⁹⁴ 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 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 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

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

^aThe 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.

^bThe GPCR protein sequences are given in the Online Supporting Information.

^cThe non-GPCR protein sequences are given in the Online Supporting Information.

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
Metabotropic/glutamate/pheromone	44	36	81.82
Fungal pheromone	23	2	8.70
cAMP receptor	10	6	60.00
Frizzled/smoothed family	17	8	47.06
Overall	365	305	83.56

^aSee the footnote "a" of Table 2.

^bThe corresponding sequences are given in the Online Supporting Information.

by investigators to examine the accuracy of various predictors (see, e.g., Ref. 45, 55, 66 68, 70, 96–107).

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 be¹⁵

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

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

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

SGML and CITI Use Only
DO NOT PRINT

AQ4



Author Proof

AQ1: Please confirm that all author names are OK and are set with first name first, surname last.

AQ2: Kindly provide department/division for the second affiliation.

AQ3: Has permission obtained for Figure 1.

AQ4: Kindly provide an appropriate artwork for the graphical abstract (for the table of contents).

AQ5: Please note that the Online Supporting Information A and B given in the table is given as Online Supporting Information only.



Author Proof