research articles Journal of proteome research

Prediction of G-Protein-Coupled Receptor Classes

Kuo-Chen Chou*

Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, California 92130

Received April 4, 2005

Being the largest family of cell surface receptors, G-protein-coupled receptors (GPCRs) are among the most frequent targets of therapeutic drugs. The functions of many of GPCRs are unknown, and it is both time-consuming and expensive to determine their ligands and signaling pathways. This forces us to face a critical challenge: how to develop an automated method for classifying the family of GPCRs so as to help us in classifying drugs and expedite the process of drug discovery. Owing to their highly divergent nature, it is difficult to predict the classification of GPCRs by means of conventional sequence alignment approaches. To cope with such a situation, the CD (**C**ovariant **D**iscriminant) predictor was introduced to predict the families of GPCRs. The overall success rate thus obtained by jack-knife test for 1238 GPCRs classified into three main families, i.e., class A-"rhodopsin like", class B-"secretin like", and class C-"metabotrophic/glutamate/pheromone", was over 97%. The high success rate suggests that the CD predictor holds very high potential to become a useful tool for understanding the actions of drugs that target GPCRs and designing new medications with fewer side effects and greater efficacy.

Keywords: GPCR • rhodopsin like • secretin like • metabotrophic • glutamate • pheromone • evolutionary pharmacology • CD predictor • amino acid composition

I. Introduction

One of the largest gene families in the human genome is that encoding the G-protein-coupled receptors (GPCRs), with approximately 450 genes identified to date. GPCRs are plasma membrane receptors, with a trademark of seven-transmembrane helices (Figure 1). They bind to and transduce signals for a huge variety of ligands including neurotransmitters, peptide hormones, growth factors, morphogens, odorants, tastants, photons, and other small molecules. The action mechanism of GPCRs is thru molecules called "second messengers" that relay signals received at receptors on the cell surface—such as the arrival of protein hormones, growth factors, etc.—to target molecules in the cytosol and/or nucleus. In addition to the job as relay molecules, second messengers also serve to amplify the strength of the signal.

Being the largest family of cell surface receptors, GPCRs are a pharmacologically important protein family; pathways involving these receptors are the targets of hundreds of drugs, including antihistamines, neuroleptics, antidepressants, and antihypertensives. GPCRs also mediate the actions of certain medications used to treat disorders as diverse as cardiovascular disease, drug dependency, and mental illness.[1]

The functions of many of GPCRs are unknown, and determining their ligands and signaling pathways is both time-consuming and costly. This difficulty has motivated and challenged the development of a computational method which can predict the classification of the families and subfamilies of GPCRs based on their primary sequences so as to help us classify drugs, a technique which might be called "evolutionary pharmacology".

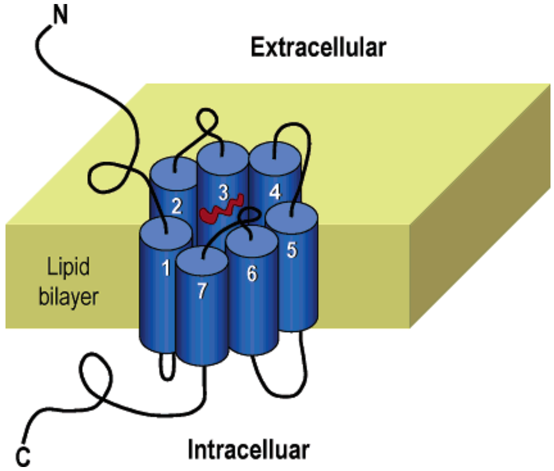

Figure 1. Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The 7-helix bundle thus formed has a central pore on its extracellular surface. The red entity located in the central pore represents a ligand messenger.

Actually, a statistical analysis has been performed for 566 GPCRs within the rhodopsin-like family classified into 7 subfamily classes: (1) adrenoceptor, (2) chemokine, (3) dopamine, (4) neuropeptide, (5) olfactory type, (6) rhodopsin, and (7) serotonin. Each of the 7 subtypes contains at least more than 30 sequences. The results thus obtained were quite encourag-

* To whom correspondence should be addressed. E-mail: kchou@san.rr.com

10.1021/pr050087t CCC: $30.25

Published on Web 06/02/2005

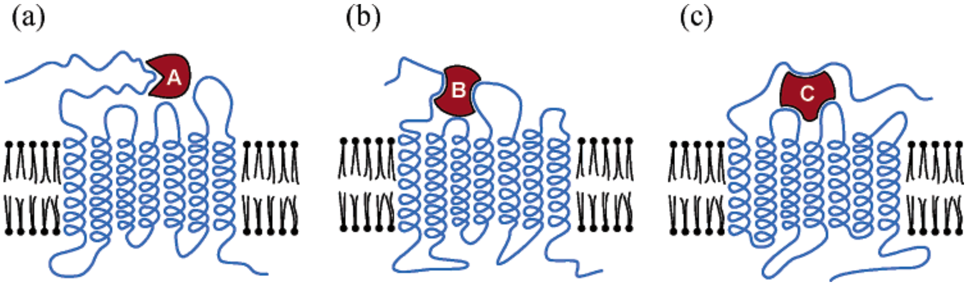

Figure 2. Schematic drawing to show three different main families of GPCRs: (a) class A-rhodopsin like, (b) class B-secretin like, and (c) class C-metabotrophic/glutamate/pheromone.

ing.[2] The present study was initiated in an attempt to extend the statistical analysis from predicting the subfamily classification limited within only a special main family to predicting the classification among several different main families of GPCRs.

II. Materials and Method

The proteins used for this study were collected from the GPCRDB (**G** **P**rotein-**C**oupled **R**eceptor **D**ata **B**ase).[3,4] where GPCRs are classified into the following 6 main families: class A-rhodopsin like; class B-secretin like; class C-metabotrophic/glutamate/pheromone; class D-fungal pheromone; class E-cAMP receptors; and class F−Frizzled/Smoothened family. The sequences of proteins in GPCRDB are derived from the SWISS-PROT and TREMBL Data Banks.[5] All of the incomplete sequences that only contained fragments of the receptors were removed. Meanwhile, the NRDB program[6] was used to check that none of the sequences was identical to any of others in the data set. Next, those families that contain too few sequences to have any statistical significance were dropped for further consideration. The remaining families obtained through such a screening procedure are as follows: (1) class A-rhodopsin like (Figure 2a); (2) class B-secretin like (Figure 2b); and (3) class C-metabotrophic/glutamate/pheromone (Figure 2c). They each contain at least more than 50 sequences. Listed in Table 1 are the accession numbers of the 1238 GPCRs, of which 1103 are of class A, 84 of class B, and 51 of class C. The accession number rather than the SWISS-PROT name is used here because the accession number is more stable for representing a unique protein sequence.

It is instructive to conduct an analysis of the sequence identity for the proteins in a same family subset. The sequence identity percentage between two protein sequences is defined as follows. Suppose one sequence is N_1 residues long and the other N_2 residues long ($N_1 \geq N_2$), and the maximum number of residues matched by sliding one sequence along the other is M. The sequence identity percentage between the two sequences is defined as (M/N_1)%. The treatment for gaps is according to ref 7. The sequence matches performed between all members in each subset of Table 1 have indicated that the average sequence identity percentages for classes A, B, and C are 18.05%, 22.67%, and 26.94% with a standard deviation of 8.43%, 17.44%, and 15.66%, respectively. These numbers indicate that the majority of pairs in each of the subsets concerned have low relative sequence identities.

The amino acid composition[8] is used to represent the sample of GPCR, and the CD (**C**ovariant **D**iscriminant) predictor adopted to perform the prediction of the GPCR families. For readers' convenience, a brief introduction of the CD predictor is given below. For the details about the predictor and its development, refer to a series of previous papers.[9−14] Suppose the GPCRs in classes A, B, and C are categorized into classes 1, 2, and 3, respectively. Thus, class 1 contains only GPCRs rhodopsin like, class 2 only secretin like, class 3 only metabotrophic/glutamate/pheromone. Suppose the kth GPCR in the class m is represented by the following vector

$$\mathbf{R}_k^{\;m} = \begin{bmatrix} a_{k,1}^{\;\;m} \\ a_{k,2}^{\;\;m} \\ \vdots \\ a_{k,20}^{\;\;m} \end{bmatrix} \tag{1}$$

where $a_{k,1}^{\;\;m}$, $a_{k,2}^{\;\;m}$, ..., $a_{k,20}^{\;\;m}$ are the amino acid-composition[8−10] for the kth GPCR of class m, and n_m the total number of GPCRs in class m. The *standard vector* for class m is defined by[10]

$$\bar{R}^m = \begin{bmatrix} \bar{a}_1^m \\ \bar{a}_2^m \\ \vdots \\ \bar{a}_{20}^m \end{bmatrix}, \quad (m = 1, 2, ..., \mu) \tag{2}$$

where

$$\bar{a}_i^m = \frac{1}{n_m}\sum_{k=1}^{n_m} a_{k,i}^m, \quad (i = 1, 2, ..., 20) \tag{3}$$

Suppose **R** is a query GPCR whose family class is to be identified. It can also be represented by a point or vector in the 20-D (dimensional) space with the components of (a_1, a_2, ..., a_{20}), where a_i has the same meaning as $a_{k,i}^m$ of eq 1 but is associated with the receptor **R** instead of $\mathbf{R}_k^m$. The scale in measuring the difference between the query receptor **R** and the norm $\bar{\mathbf{R}}_m$ of class m is by the following covariant discriminant function, as defined by Chou et al.:[12]

$$\Delta(\mathbf{R},\bar{\mathbf{R}}^m) = D_{\mathrm{M}}^{\;2}(\mathbf{R},\bar{\mathbf{R}}^m) + \ln|\mathbf{B}^m|, \; (m = 1, 2, ..., \mu) \tag{4}$$

where

$$D_{\mathrm{M}}^{\;2}(\mathbf{R},\bar{\mathbf{R}}^m) = (\mathbf{R} - \bar{\mathbf{R}}^m)^{\mathbf{T}}\mathbf{B}_m^{\;-1}(\mathbf{R} - \bar{\mathbf{R}}^m) \tag{5}$$

is the squared Mahalanobis distance,[10,15,16] **T** is the transposition operator, while $|\mathbf{B}^m|$ and $\mathbf{B}_m^{\;-1}$ are respectively the determinant and inverse matrix of $\mathbf{B}_m$, which is the covariance matrix for class m and given by

$$\mathbf{B}_m = \begin{bmatrix} b_{1,1}^m & b_{1,2}^m & \cdots & b_{1,20}^m \\ b_{2,1}^m & b_{2,2}^m & \cdots & b_{2,20}^m \\ \vdots & \vdots & \cdots & \vdots \\ b_{20,1}^m & b_{20,2}^m & \cdots & b_{20,20}^m \end{bmatrix} \tag{6}$$

Table 1. List of the Accession Numbers for the 1238 GPCRs Classified into Three Families

(1) 1,103 Class A: Rhodopsin Like

O97666	P35414	Q9WV08	P70058	P79960	P16395	P17200	P30544	P04761	P08482
P11229	P12657	P56489	P06199	P08172	P10980	P30372	P08483	P11483	P20309
P41984	P49578	P08173	P08485	P32211	P08911	P08912	P56490	P11616	P25099
P28190	P30542	P34970	P47745	P49892	O13076	P11617	P29274	P29275	P29276
P30543	P46616	Q60613	Q60614	O02667	P28647	P33765	P35342	Q28309	O08766
Q9W6C4	O02666	O02824	P15823	P18130	P18841	P23944	P25100	P35348	P35368
P43140	P97714	P97717	Q91175	P08913	P18089	P18825	P18871	P19328	P22086
P22909	P30545	P32251	P35369	P35405	Q01337	Q01338	Q28838	Q60474	Q60475
Q60476	Q91081	O54913	O60451	Q13675	Q13729	O42574	P07700	P08588	P18090
P34971	P47899	P79148	Q28927	Q28998	Q9TST6	Q9TT96	P04274	P07550	P10608
P18762	P54833	Q28044	Q28509	Q28997	Q9TST5	O02662	P13945	P25962	P26255
P46626	Q28524	Q9TST4	Q9XT57	Q9XT58	P43141	O96716	P34974	P70115	Q01718
Q64326	Q9Z1S9	O57317	O35210	O77590	P25095	P25104	P29089	P29754	P29755
P30555	P30556	P32303	P33396	P34976	P35373	P43240	P79785	Q13725	Q9WV26
P35351	P35374	P50052	Q9Z0Z6	O54798	O54799	O97967	P21729	P24053	P28336
P30550	P32247	P35371	P47751	P52500	O70526	P25023	P30411	P32299	P46663
P48748	Q28642	O42402	P97583	Q61125	P32246	P51675	P51676	P56482	P46092
O55193	P41597	P51683	O54814	P51677	P51678	P56483	P56492	Q9Z2I3	P51679
P51680	O08556	O62743	O97878	O97879	O97880	O97881	O97882	O97883	P51681
P51682	P56439	P56440	P56441	P56493	P79436	O54689	P51684	P32248	P47774
O97665	P51685	P56484	P51686	Q9WUT7	O00421	O75307	Q9XSD7	O00590	O18793
O75303	O77776	O77833	O97724	O97774	O97962	O97975	Q9XS35	Q9XS99	Q9XT12
Q9XT13	Q9XT14	Q9XT76	O88410	P49682	O08565	O62747	P25930	P30991	P56491
P56498	P70658	P79394	Q28474	P32302	P34997	Q04683	O42445	O60835	O77488
O93247	Q62973	Q9TSQ8	Q9YGC3	P35411	P49238	Q9Z0D9	O09047	O55197	O70129
O88680	P21730	P30992	P30993	P97520	Q16581	O08786	O97772	P30551	P32238
P70031	Q63931	P30552	P30553	P30796	P32239	P46627	P56481	P79266	Q16144
O02777	P20272	P21554	P34972	P47746	P47936	P56971	Q98894	Q98895	Q9PUI7
Q9QZN9	O35786	O75388	O75748	O88416	P97468	Q99788	Q9Z2J6	P41596	Q24563
O77680	P18901	P21728	P21918	P25115	P35406	P42288	P42289	P42290	P42291
P47800	P50130	P53452	P53454	O73810	P13953	P14416	P20288	P24628	P52702
P53453	P19020	P30728	P35462	P52703	P21917	P30729	P51436	O44198	O02146
O42315	O42316	O42317	Q98841	Q98842	Q98843	Q98844	Q9YHA5	Q13167	O42321
O42322	O62709	P21450	P21451	P24530	P25101	P26684	P28088	P32940	P35463
P48302	Q29010	O73739	Q16433	Q91548	O08790	P21462	P25089	P25090	P33766
Q05394	O88535	O88536	O88537	O88538	P20395	P23945	P32212	P35376	P35379
P47799	P49059	P79763	Q95179	Q64183	O08726	O43603	O60755	O88626	O88853
O88854	P47211	P56479	Q62805	O18821	O42329	P30968	P30969	P32236	P32237
P49922	Q01776	Q9YGN8	Q9YGN9	O08725	Q92847	Q95254	O43193	O93412	O93413
P30546	P31389	P31390	P35367	P70174	P17124	P25021	P25102	P47747	P97292
P21109	P25024	P55919	P55920	P70612	O97571	P25025	P35343	P35344	P35407
Q28003	O93237	O93239	O02721	P16235	P16582	P22888	P30730	Q28005	Q14751
Q15996	O97504	P32244	P32245	P33032	P33033	P35345	P41149	P41968	P41983
P56451	P70596	O73667	O73671	O93259	O19037	O77616	P47798	P55167	P56442
P56443	P56444	P56445	P56446	P56447	P56448	Q01726	Q01727	Q29154	O88495
P48039	P48040	P49217	P49219	P49285	P49286	P49288	Q13585	Q28558	Q61184
P87496	P87499	O97512	P16177	P29371	P30098	O02813	O02835	O02836	O62729
O70342	O97969	P21555	P25929	P25931	P34992	P49146	P50391	P79113	P79217
P97295	Q04573	Q15761	Q61041	Q61212	Q63447	Q63634	Q9WVD0	Q9Z2D5	O57463
O73733	O73734	O97505	Q99463	Q99647	Q9YHX1	Q9Z2D4	O88319	O95665	P20789
P30989	P70310	Q63384	O01670	O77408	P22270	Q17232	Q25188	Q25321	Q25322
Q93126	Q93127	O61730	O77254	O97171	P23269	P23271	P23272	P23273	P23274
P30953	P30955	P47887	O95222	P23267	P23270	P30954	O43749	P23266	P47890
Q9Y585	P37067	P37068	P37069	P37070	P37071	P37072	P23265	P23268	P34987
Q95157	O95371	P23275	Q13607	Q15062	Q95156	Q13606	Q95154	Q95155	P34982
P47884	P47881	P47883	P47888	P47893	P70526	O13036	O57597	O95007	Q62944
Q9WU86	Q9Z1V0	O60403	O60404	O70265	O70266	O70267	O70268	Q62007	Q9Y4A9
O60431	Q62942	O14581	O60412	O76100	Q15622	O35434	O76000	O76001	O76002
O95006	O95047	O95499	O95918	Q63394	Q9WV11	Q9WV13	Q9WV14	Q9Y3N9	O35184
O77756	O77757	O77758	Q62943	Q63395	Q9WU91	O70269	O70270	O70271	P32300
P33533	P41143	P33534	P34975	P41144	P41145	P33535	P35372	P42866	P79350
Q95247	P35370	P35377	P41146	P47748	P79292	O57585	O42324	O43613	O43614
P56718	P56719	Q9Y5X5	P30559	P32306	P56449	P56494	P70536	P97926	Q28756
Q90252	Q90334	Q90352	P21556	P25105	P46002	Q62035	O00254	O08675	P55085
P55086	Q63645	O76067	O88634	P32250	P43657	Q15722	Q99677	P35383	P41231
P41232	P34996	P47900	P48042	P49650	P49651	P49652	O93361	P79928	Q15077
Q63371	Q98907	P51582	O00398	O15132	O88855	Q9WTK1	O57466	O35811	P43119
P43252	P43253	P79393	P34995	P35375	P70597	P30557	P34979	P34980	P43115

Table 1 (Continued)

P46069 P50131 P32240 P35408 P43114 Q28691 P43116 P70263 Q13258 Q62053
Q62928 Q9XT82 P37289 P43088 P43117 P43118 O28905 O00325 O15191 O46657
O35932 O01668 P06002 P08099 P22269 P28678 P28679 P35356 P35360 P35361
P35362 Q17053 Q17292 Q17296 Q94741 O61303 P04950 P08255 P17646 P28680
P29404 P90680 P91657 Q26495 Q25157 Q25158 O15973 O15974 O16005 P09241
P24603 P31356 O13018 O14718 O35214 O42266 O42490 P23820 P47803 P47804
P51475 P51476 Q9Z2B3 O13227 O18766 O42604 O62791 O62792 O62793 O62794
O62795 O62796 O62798 O93441 O93459 P02699 P02700 P08100 P15409 P22328
P22671 P28681 P29403 P31355 P32308 P32309 P35359 P35403 P41590 P41591
P49912 P51470 P51488 P51489 P52202 P56514 P56515 P56516 P79756 P79812
P79848 P79863 P79898 P87369 Q28886 Q90214 Q90215 Q90245 Q98980 Q9YGY9
Q9YGZ0 Q9YGZ1 Q9YGZ2 Q9YGZ3 Q9YGZ4 Q9YGZ5 Q9YGZ6 Q9YGZ7 Q9YGZ8 Q9YGZ9
Q9YH00 Q9YH01 Q9YH02 Q9YH03 Q9YH04 Q9YH05 O12948 O18910 O18913 O35476
O35478 O35599 P04000 P04001 P22329 P22330 P22331 P22332 P32313 P35358
P41592 P87367 Q95170 Q9R024 O13092 P03999 P28684 P51473 P51490 P51491
P87368 Q63652 Q90309 P28682 P32310 P51472 P87365 P28683 P32311 P32312
P35357 P51471 P51474 P87366 O02464 O76123 O76124 O76125 O02465 O61473
O61474 O96107 Q9W6K3 Q9W6I4 Q9W6S0 O62860 O97901 Q90226 Q9W684 Q9W6A7
Q9W771 Q9XSF1 Q9XSX3 Q9YI52 O46554 O57605 O70363 Q9W6A9 Q9W6J6 Q9W773
Q9W7K8 Q9XS34 Q9YI51 Q9W609 Q9W6A8 Q9W772 Q9W7C1 Q9YI53 Q9W685 Q9W6A5
Q9W6A6 Q9W6I5 Q9W6S1 Q9YGY7 P20905 Q17239 Q25190 P28285 P28286 Q16950
Q16951 Q25414 O08890 O08892 O42384 O42385 P08908 P11614 P19327 P28221
P28222 P28334 P28564 P28565 P28566 P30939 P30940 P35404 P46636 P49144
P49145 P56496 P79748 Q02284 Q60484 Q61224 Q64264 P08909 P14842 P18599
P28223 P28335 P30994 P34968 P35363 P41595 P50128 P50129 Q02152 O70528
P97288 Q62758 P30966 P31387 P35364 P35365 P47898 P31388 P50406 Q9R1C8
P32304 P32305 P34969 P50407 Q91559 O17470 O76267 Q21034 Q98998 Q63004
P97842 P28646 P30872 P30873 P30680 P30874 P30875 P34993 P34994 P30935
P30936 P32745 P30937 P31391 P49660 O08858 P30938 P35346 P05363 P16610
P21452 P30549 P51144 P79218 Q64077 P14600 P25103 P30547 P30548 Q98982
Q9W6I3 P30974 P30975 Q03566 Q94736 P14763 P16473 P21463 P47750 P56495
Q27987 O46639 O93603 P21761 P34981 Q01717 Q28596 Q27986 O88820 P25116
P26824 P30558 P47749 P56488 Q00991 P21731 P30987 P34978 P56486 Q95125
O75228 P30518 P30560 P32307 P37288 P47901 P48043 P48044 P48974 Q00788
Q62463 Q9WU02 O43192 O77808 O88721 Q9WTV8 Q9WTV9 O12000 P09703 P09704
P16849 P52380 P52381 P52382 P52383 P52542 Q01035 Q98146 O90387 Q9QEV2
Q9QEV3 Q9WRM0 Q9WT52 Q9YTJ2 O08878 Q99527 O43494 O00574 O18983 O19024
Q9XT45 O15218 P31392 P43142 O14842 O14843 O15529 O15552 O14768 O88313
Q9Z0G3 O00155 O00270 O18982 O97663 O97664 P30951 P35412 P35413 P46089
P46090 P46091 P46093 P46094 P46095 P47775 P48145 P48146 P49683 P49685
P50132 P51651 P56412 P97639 Q13304 Q15760 Q61121 Q64121 Q99678 Q99679
Q99680 Q99705 Q9Y2T5 P49681 P70585 Q14330 Q91178 O35797 O75194 Q9UE21
P04201 P12526 P30554

(2) 84 Class B: Secretin Like

P32215 P41586 P70205 Q29627 O73769 O14514 O60241 O60242 O08893 P25117
P30988 P32214 P79222 Q16602 Q60755 Q63118 Q9WUP2 O42602 O42603 O62772
P34998 P35347 P35353 P47866 Q13324 Q60748 Q90812 P43218 P43219 P48546
O35659 P30082 P32301 P43220 P47871 Q61606 O95838 Q9Z0W0 P32082 P34999
Q02643 Q02644 O73768 Q9WU99 P48960 Q14246 Q61549 O00718 Q9Z0M6 O88917
O88923 O88927 O94910 O95490 O97813 O97817 O97822 O97824 O97827 O97830
O97831 Q9Z173 Q9Z174 P25107 P25961 P41593 P49190 P50133 P70555 Q03431
O46502 P23811 P47872 P30083 P32241 P35000 P41587 P41588 Q28992 Q90308
Q9YHC6 P30650 Q09460 O00406

(3) 51 Class C: Metabotrophic/Glutamate/Pheromone

Q9WU48 Q9QY96 Q9PW88 Q93564 Q62916 Q14833 Q14832 Q14831 Q14416 Q13255
Q09630 P91685 P70579 P48442 P47743 P41594 P41180 P35400 P35384 P35349
P31424 P31423 P31422 P31421 P23385 O95975 O93553 O93552 O88871 O75899
O73640 O73639 O73638 O73637 O73636 O73635 O70410 O70409 O35271 O35269
O35268 O35267 O35266 O35265 O35202 O35192 O35190 O35189 O15303 O08620
O00222

where the matrix elements are given by

$$b_{i,j}^{m} = \frac{1}{n_m - 1}\sum_{k=1}^{n_m}[a_{k,i}^{m} - \bar{a}_{i}^{m}][a_{k,j}^{m} - \bar{a}_{j}^{m}],\ (i, j = 1, 2, ..., 20) \quad (7)$$

According to the principle of similarity, the smaller the difference between the query receptor **R** and the norm of class *m*, the higher the likelihood that receptor **R** belongs to class *m*. Accordingly, the identification rule can be formulated as follows

$$\Delta(\mathbf{R},\bar{\mathbf{R}}^{\Lambda}) = \mathbf{Min}\{\Delta(\mathbf{R},\bar{\mathbf{R}}^{1}), \Delta(\mathbf{R},\bar{\mathbf{R}}^{2}), ..., \Delta(\mathbf{R},\bar{\mathbf{R}}^{\mu})\} \quad (8)$$

where Λ can be 1, 2, ..., or μ, and the operator **Min** means

Table 2. Success Rates in Identifying the Main Families of GPCRs

Class A Rhodopsin like	Class B Secretin like	Class C Metabotrophic/ glutamate/pheromone	Overall
	Re-substitution test[a]		
1092/1103 = 99.00%	83/84 = 98.81%	51/51 = 100%	1226/1238 = 99.03%
	Jack-knife test[a]		
1092/1103 = 99.00%	74/84 = 88.10%	40/51 = 78.43%	1206/1238 = 97.42%
	Random re-substitution test[b]		
84/84 = 100%	84/84 = 100%	51/51 = 100%	219/219 = 100%
	Random jack-knife test[b]		
83/84 = 98.81%	79/84 = 94.05%	42/51 = 82.35%	204/219 = 93.15%

[a] Prediction was made on the data set given in Table 1. The CD predictor (see eqs 1−8) was used to perform the prediction. [b] Prediction was made for the data set that consists of 84 class A GPCRs randomly picked from the 1103 class A GPCRs of Table 1, as well as its 84 class B and 51 class C GPCRs. See the above footnote for further explanation.

Table 3. List of the Accession Numbers for the 84 GPCRs Randomly Picked from the 1103 GPCRs of Class A in Table 1

O00254	O02666	O08766	O13092	O15974	O35210	O42317	O43193	O54814	O60431
O62792	O70270	O75228	O76267	O88313	O88820	O93459	O97504	O97880	P04001
P08173	P09704	P14600	P17200	P19328	P21554	P22328	P23270	P25021	P25105
P28222	P28679	P29754	P30549	P30728	P30940	P30992	P32236	P32300	P32745
P34971	P34994	P35358	P35372	P35408	P37288	P41596	P43116	P46089	P47745
P47804	P48039	P49146	P49682	P50406	P51490	P51684	P53453	P56443	P56486
P56516	P70596	P79436	P87367	P97639	Q01727	Q13607	Q16144	Q25158	Q28474
Q28998	Q61212	Q63004	Q64264	Q91548	Q95179	Q98998	Q9QZN9	Q9W6A6	Q9W772
Q9WV08	Q9XT13	Q9YGN8	Q9YGZ9						

taking the minimal one among those in the brackets. The value of the superscript Λ derived from eq 8 indicates which class the query receptor $\mathbf{R}$ belongs to. If there is a tie case, then Λ is not uniquely determined, but that did not happen for the datasets studied here.

Before using the above equations for practical calculations, the following point should be realized. Owing to the normalization condition imposed by the definition of amino acid-composition, of the 20 components in eq 1, only 19 are independent,[10] and hence the covariance matrix $\mathbf{B}_m$ as defined by eq 7 must be a singular one.[9] This would lead the Mahalanobis distance defined by eq 5 and the covariant discriminant function by eq 4 to be divergent and meaningless. To cope with such a situation, the dimension-reducing procedure[10] was adopted in practical calculations; i.e., instead of 20-D space, a receptor is defined in a (20−1)-D space by leaving out one of its 20 amino acid components. The remaining 19 components would be completely independent, thereby the corresponding covariance matrix $\mathbf{B}_m$ being no longer singular. In other words, the Mahalanobis distance (eq 5) and the covariant discriminant function (eq 4) based on such a 19-D space can be uniquely defined without any trouble. However, a question might be raised: which one of the 20 components can be left out? The answer is: any one of them. Will it lead to a different predicted result by leaving out a different component? The answer is: no. According to the *invariance theorem* given in Appendix A of Chou,[10] both the value of the Mahalanobis distance and the value of the determinant of $\mathbf{B}_m$ will remain exactly the same regardless of which one of the 20 components is left out. Accordingly, the final value of the covariant discriminant function (eq 4) can be uniquely defined through such a dimension-reducing procedure.

III. Results and Discussion

Now let us use the predictor formulated in the last section to examine the success rates in identifying the family classes for the 1238 GPCRs listed in Table 1. The examinations were conducted by two different approaches, the re-substitution test and the jack-knife test, as reported below.

Re-Substitution Test. The re-substitution test is used to examine the self-consistency of a prediction method. During the re-substitution process, the class for each of the GPCRs in the data set is in turn identified using the rule parameters derived from the same data set, the so-called training data set. The success rates thus obtained for the 1238 GPCRs in Table 1 are given in Table 2, from which we can see that the overall success rate is 99.03%, indicating that the current prediction method is highly self-consistent. It should be pointed out that during the above process the rule parameters derived from the training data set include the information of the query GPCR later plugged back for testing itself. This will certainly enhance the success rate because the same samples are used to derive the rule parameters and to test themselves. Therefore, the success rate thus obtained merely represents some sort of optimal estimation.[9,10,14,17] Nevertheless, the re-substitution test is useful because it reflects the self-consistency. A predictor with a poor self-consistency certainly cannot be deemed as a good one. However, to really reflect the power of a predictor, a cross-validation test by excluding the tested samples from the training data set is needed.

Jack-knife Test. Three different examinations are often used in statistical prediction for cross-validation. They are independent data set test, sub-sampling test, and jack-knife test. Of these three, however, the jack-knife test is deemed as the most rigorous and objective one [see ref 18 for a comprehensive discussion about this, and ref 19 for the underlying mathematical principle]. For the cross-validation by jack-knifing, each of the proteins in the data set is in turn singled out as a tested sample and all the rule-parameters are calculated based on the remaining proteins without including the one being identified. Therefore, both the training data set and testing data set during the jack-knifing process are actually open, and a sample will in turn move from one to the other. The results of jack-knife test thus obtained for the 1238 GPCRs are also given in Table

2, from which we can see the following. As expected, the success identification rates by jack-knife test are lower than those by the re-substitution test, particularly for the smallest subset of class C. This is because the cluster-tolerant capacity[20] for small subsets is usually low. Therefore, the information loss due to jack-knifing will have a greater impact on the small subsets than the large ones. Nevertheless, the overall success rate by jack-knife test for the data set of 1238 GPCRs is still as high as 97.42%. It is anticipated that the success rate for identifying class C of GPCRs can be enhanced by adding into its subset more newly found proteins that have been found belonging to this class.

Because the number of samples for class A is overwhelming in the current dataset, the following argument might be brought up against the above high success rates. If the identification was made by always choosing class A, the overall success rate thus obtained could also reach as high as $1103/1238 = 89.10\%$, implying that the high overall success rate was resulted from the extreme uneven distribution of the dataset investigated but not the power of the predictor. To address this problem, a size-reduced subset for class A was formed by randomly picking 84 samples from the 1,103 GPCRs of the original subset for class A. The accession numbers for the 84 GPCRs thus generated are given in Table 3. Now let us use the data of class A in Table 3 as well as the data for class B and class C in Table 1 to form a new working dataset, which contains $84 + 84 + 51 = 219$ GPCRs. For such a new dataset, the same jack-knife test was performed and the results are also given in Table 2, from which we can see that the overall success rate could reach over 93%. In contrast to this, if the identification was made by blindly sticking to class A, the overall success rate would be only $84/219 = 38.35\%$, which is more than 50% lower than that by the CD predictor.

IV. Conclusion

GPCRs are the largest family of cell surface receptors, accounting for >1% of the human genome. They play a key role in cellular signaling networks that regulate various physiological processes. The critical physiological roles of GPCRs have made them among the most frequent targets of therapeutic drugs. Many efforts in pharmaceutical research have been aimed at understanding their structure and function. Unfortunately, so far, very few GPCR structures have been determined by either X-ray or NMR technique because it is difficult to crystallize them and most of GPCRs will not dissolve in normal solvents. In contrast, more than thousand GPCR sequences are known, and much more are expected to come soon. To timely use the uncharacterized GPCRs for drug discovery and basic research, it is highly desirable to develop a computational method that can rapidly and accurately predict the classification of their families.

It is difficult to predict the classification of GPCRs by using the conventional sequence alignment approach owing to the nature of their high divergence. To tackle the sequence divergent problem, the CD predictor is introduced that is formulated based on a series of discrete numbers such as those constituting the amino acid composition.

The high success rates obtained in this study imply that the families of GPCRs are closely correlated with their amino acid composition, and that the CD predictor is quite promising and may become a powerful tool in this area.

Acknowledgment. The author wishes to thank the three anonymous referees whose constructive suggestions are very helpful in strengthening the presentation of this paper.

References

(1) Roth, B. L.; Willins, D. L.; Kroeze, W. K. G protein-coupled receptor (GPCR) trafficking in the central nervous system: relevance for drugs of abuse. *Drug Alcohol Depend.* **1998**, *51*, 73–85.

(2) Chou, K. C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **2002**, *1*, 429–433.

(3) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **1998**, *26*, 275–279.

(4) Horn, F.; Vriend, G.; Cohen, F. E. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.* **2001**, *29*, 346–349.

(5) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* **2000**, *25*, 31–36.

(6) Gish, W. http://blast.wustl.edu/pub/nrdb/**1999**.

(7) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice *Nucleic Acids Res.* **1994**, *22*, 4673–4680.

(8) Chou, J. J.; Zhang, C. T. A joint prediction of the folding types of 1490 human proteins from their genetic codons *J. Theor. Biol* **1993**, *161*, 251–262.

(9) Chou, K. C.; Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.* **1994**, *269*, 22014–22020.

(10) Chou, K. C. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 319–344.

(11) Liu, W.; Chou, K. C. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *J. Protein Chem.* **1998**, *17*, 209–217.

(12) Chou, K. C.; Liu, W.; Maggiora, G. M.; Zhang, C. T. Prediction and classification of domain structural classes. *Proteins: Struct. Funct. Genet.* **1998**, *31*, 97–103.

(13) Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.* **1999**, *12*, 107–118.

(14) Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729–738.

(15) Mahalanobis, P. C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49–55.

(16) Pillai, K. C. S. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1985; Vol. 5, pp 176–181.

(17) Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 57–59.

(18) Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.

(19) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979; Chapter 11 Discriminant Analysis; Chapter 12 Multivariate analysis of variance; Chapter 13 cluster analysis; pp. 322–381.

(20) Chou, K. C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.* **1999**, *264*, 216–224.

PR050087T