research articles Journal of proteome research

Bioinformatical Analysis of G-Protein-Coupled Receptors

Kuo-Chen Chou* and David W. Elrod

Computer-Aided Drug Discovery/Bioinformatics, Pharmacia, Kalamazoo, Michigan 49007-4940

Received April 9, 2002

G-protein-coupled receptors play a key role in cellular signaling networks that regulate various physiological processes, such as vision, smell, taste, neurotransmission, secretion, inflammatory, immune responses, cellular metabolism, and cellular growth. These proteins are very important for understanding human physiology and disease. Many efforts in pharmaceutical research have been aimed at understanding their structure and function. Unfortunately, because they are difficult to crystallize and most of them will not dissolve in normal solvents, so far very few G-protein-coupled receptor structures have been determined. In contrast, more than 1000 G-protein-coupled receptor sequences are known, and many more are expected to become known soon. In view of the extremely unbalanced state, it would be very useful to develop a fast sequence-based method to identify their different types. This would no doubt have practical value for both basic research and drug discovery because the function or binding specificity of a G-protein coupled receptor is determined by the particular type it belongs to. To realize this, a statistical analysis has been performed for 566 G-protein-coupled receptors classified into seven different types. The results indicate that the types of G-protein-coupled receptors are predictable to a considerable accurate extent if a good training data set can be established for such a goal.

Keywords: rhodopsin-like • adrenoceptor • chemokine • dopamine • neuropepide • olfactory • rhodopsin • amino acid composition • covariant-discriminant algorithm

I. Introduction

The term "receptor" is used in pharmacology to denote a class of cellular macromolecules that are specifically or directly involved in chemical signaling between and within cells. As a receptor, it not only can recognize the particular molecules that activate it but also can alter, when recognition occurs, cell function by causing, e.g., a change in membrane permeability or an alternation in gene transcription. Therefore, the interaction of a hormone, neurotransmitter, or intracellular messenger with its receptor(s) may lead to a change in cellular activity.

Of the cell surface receptors, the G-protein-coupled receptors have occupied a special place of importance. This is because they have the following three remarkable features. (1) Signal transduction function. G-protein-coupled receptors constitute a unifying signal transduction mechanism. A wide range of neurotransmitters, neuropeptides, polypeptide hormones, inflammatory mediators, and other bioactive molecules transmit their signals to the intracellular environment by specific interaction with a class of receptor that relies upon interaction with GTP-binding protein (G-protein) for activation of intracellular effector systems.[1] (2) Seven-helix-bundled structure embedded in the cell membrane. G-protein-coupled receptors consist of a single polypeptide chain of variable length that traverses the lipid bilayer seven times, forming characteristic transmembrane helices and alternating extracellular and intracellular sequences (Figure 1). The significance of the seven transmembrane motif is not fully clear, but it must be uniquely suited for transmitting a signal from the external surface to the internal surface of the plasma membrane through a ligand-induced conformational change, as discussed for bR (see, e.g.,

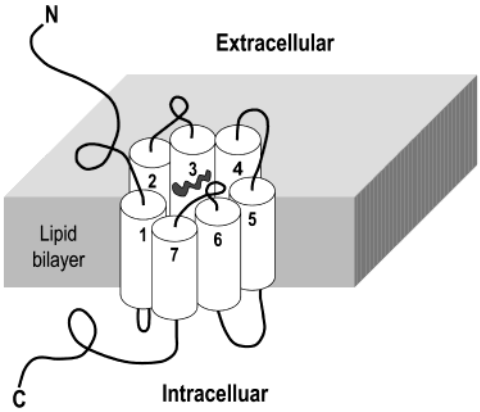

Figure 1. Schematic representation of a G-protein-coupled receptor with putative seven transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The seven-helix bundle thus formed has a central pore on its extracellular surface. The dark-shaded entity located in the central pore represents a ligand messenger.

* To whom correspondence should be addressed.

10.1021/pr025527k CCC: $22.00

Published on Web 06/29/2002

ref 2). (3) One of the largest protein families. The number of G-protein-coupled receptors is increasing rapidly. It is estimated that more than a thousand different G protein-coupled receptors exist in mammals, thus constituting one of the largest protein families in nature.[3]

Although G-protein-coupled receptors exhibit a characteristic hydropathy profile, with a relatively constant length of primary sequence comprising the transmembrane domains, considerable variation is seen in the overall length of members of the receptor superfamily. Variability in the overall length of receptors occurs mainly in the N-terminal extracellular domain, the third intracellular loop and the C-terminal cytoplasmic segment. These domain segments also exhibit the greatest diversity in amino acid composition.[4]

Also, although all known G-protein-coupled receptors are seven-helix transmembrane proteins,[5] they are a large and functionally diverse superfamily. According to their binding with different ligand types, G-protein-coupled receptors are classified into at least six different families; i.e., rhodopsin-like family, secretin-like family, metabotrophic glutamate family, fungal pheromone family, cAMP receptor family, and frizzled-smothened family. Some of these families can be further classified into a number of subfamilies. In the current study, we would like to focus on the rhodopsin-like family because the sequence data available for some of its subfamilies allow us to conduct a statistically significant analysis.

II. Materials and Method

According to the GPCRDB (December 2000 release),[6,7] the rhodopsin-like G-protein-coupled receptors consist of many different subtypes. The sequences in GPCRDB are derived from the SWISS-PROT (Release 39.0, 2000) and TREMBL Data Banks.[8] To collect the sequences used for this study, all of the sequences from GPCRDB that were described as being one of the subtypes of the rhodopsin-like G-protein-coupled receptors were selected. Then, all of the incomplete sequences that only contained fragments of the receptors were removed. Meanwhile, the NRDB program[9] was used to check that none of the sequences was identical to any of others in the data set. Next, those subtypes that contain too few sequences to have any statistical significance were dropped for further consideration. The subtypes obtained through such a screening procedure are as follows: (1) adrenoceptor, (2) chemokine, (3) dopamine, (4) neuropeptide, (5) olfactory type, (6) rhodopsin, and (7) serotonin. They each contain at least more than 30 sequences. Listed in Table 1 are the accession numbers of the 566 G-coupled-protein receptors, of which 66 are of adrenoceptor, 92 of chemokine, 43 of dopamine, 31 of neuropeptide, 84 of olfactory type, 183 rhodopsin, and 67 of serotonin. The accession number rather than the SWISS-PROT name is used because the accession number is more stable for representing a unique protein sequence. To provide an intuitive picture, a schematic drawing to illustrate the seven different types of G-protein coupled receptors is given in Figure 2.

It is instructive to conduct an analysis of the sequence identity for the proteins in a same 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 Thompson et al.[10] The sequence matches performed between all members in each subset of Table 1 have indicated that the average sequence identity percentages for the adrenoceptor, chemokine, dopamine, neuropeptide, olfactory, rhodopsin, and serotonin subsets are 35.93%, 39.92%, 37.49%, 38.29%, 42.23%, 41.71%, and 30.48% with a standard deviation of 18.84%, 21.60%, 20.03%, 20.63%, 8.52%, 21.51%, and 16.37%, respectively. These numbers indicate that the majority of pairs in each of the subsets concerned have the low relative sequence identity.

The covariant discriminant algorithm was used to perform the identification of the types of G-protein-coupled receptors based on their sequences. For the reader's convenience, let us give a brief introduction of the mathematical frame. The details about the algorithm and its development can be found in a series of previous papers.[11−16] Suppose the G-protein-coupled receptors in the types of adrenoceptor, chemokine, dopamine, neuropeptide, olfactory, rhodopsin, and serotonin are categorized into classes 1, 2, 3, 4, 5, 6, and 7, respectively. Thus, class 1 contains only G-protein-coupled receptors of adrenoceptor type, class 2 only chemokine type, class 3 only dopamine type, and so forth. Suppose the kth G-protein-coupled receptor in the class m is represented by the following vector

$$\mathbf{G}_k^m = \begin{bmatrix} g_{k,1}^m \\ g_{k,2}^m \\ \vdots \\ g_{k,20}^m \end{bmatrix}, \quad (k = 1, 2, ..., n_m;\ m = 1, 2, ..., 7) \qquad (1)$$

where $g_{k,1}^m$, $g_{k,2}^m$, ..., $g_{k,20}^m$ are the amino acid composition[11,14] for the kth G-protein-coupled receptor of class m and n_m the total number of G-protein-coupled receptors in class m. The *standard vector* for class m is defined by[11]

$$\bar{\mathbf{G}}^m = \begin{bmatrix} \bar{g}_1^m \\ \bar{g}_2^m \\ \vdots \\ \bar{g}_{20}^m \end{bmatrix}, \quad (m = 1, 2, ..., 7) \qquad (2)$$

where

$$\bar{g}_i^m = \frac{1}{n_m} \sum_{k=1}^{n_m} g_{k,i}^m, \quad (i = 1, 2, ..., 20) \qquad (3)$$

Suppose **G** is a query G-protein-coupled receptor whose type is to be identified. It can also be represented by a point or vector in the 20-D space with the components of (g_1, g_2, ..., g_{20}), where g_i has the same meaning as $g_{k,i}^m$ of eq 1 but is associated with receptor **G** instead of $\mathbf{G}_k^m$. The difference between the query receptor **G** and the norm of class m is measured by the following covariant discriminant function, as defined by Chou et al.:[13]

$$\Delta(\mathbf{G},\bar{\mathbf{G}}^m) = D_M{}^2(\mathbf{G}, \bar{\mathbf{G}}^m) + \ln|\mathbf{S}^m|, \quad (m = 1, 2, ..., 7) \qquad (4)$$

where

$$D_M{}^2(\mathbf{G}, \bar{\mathbf{G}}^m) = (\mathbf{G} - \bar{\mathbf{G}}^m)^{\mathbf{T}} \mathbf{S}_m{}^{-1} (\mathbf{G} - \bar{\mathbf{G}}^m) \qquad (5)$$

is the squared Mahalanobis distance,[11,17,18] **T** is the transposition operator, while $|\mathbf{S}^m|$ and $\mathbf{S}_m{}^{-1}$ are respectively the determinant and inverse matrix of $\mathbf{S}_m$. The latter is the covariance matrix

Table 1. List of the Accession Numbers for the 566 G-Protein-Coupled Receptors Classified into Seven Types

(1) 66 adrenoceptor type receptors

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 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 O54913 O60451 Q13675 Q13729 O96716

(2) 92 chemokine type receptors

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
O88410 P49682 O08565 O62747 P25930 P30991 P56491 P56498 P70658 P79394
Q28474 P32302 P34997 Q04683 P35411 P49238 Q9Z0D9 O08878 Q99527 O35786
O75388 O75748 O88416 P97468 Q99788 Q9Z2J6 O00421 O75307 Q9XSD7 O00590
O18793 O75303 O77776 O77833 O97724 O97774 O97962 O97975 Q9XS35 Q9XS99
Q9XT12 Q9XT13 Q9XT14 Q9XT76 O42445 O60835 O77488 O93247 Q62973 Q9TSQ8
Q9YGC3 O43494

(3) 43 dopamine type receptors

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

(4) 31 neuropeptide type receptors

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

(5) 84 olfactory type receptors

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

(6) 183 rhodopsin type receptors

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

for class m and defined by

$$\mathbf{S}_m = \begin{bmatrix} s^m_{1,1} & s^m_{1,2} & \cdots & s^m_{1,20} \\ s^m_{2,1} & s^m_{2,2} & \cdots & s^m_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ s^m_{20,1} & s^m_{20,2} & \cdots & s^m_{20,20} \end{bmatrix} \quad (6)$$

where the matrix elements are given by

$$s^m_{i,j} = \frac{1}{n_m - 1} \sum_{k=1}^{n_m} [g^m_{k,i} - \bar{g}^m_i][g^m_{k,j} - \bar{g}^m_j], \quad (i, j = 1, 2, ..., 20) \quad (7)$$

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

$$\Delta(\mathbf{G}, \bar{\mathbf{G}}^\mu) = \mathbf{Min}\{\Delta(\mathbf{G}, \bar{\mathbf{G}}^1), \Delta(\mathbf{G}, \bar{\mathbf{G}}^2), \Delta(\mathbf{G}, \bar{\mathbf{G}}^3), ..., \Delta(\mathbf{G}, \bar{\mathbf{G}}^7)\} \quad (8)$$

where μ can be 1, 2, 3, ..., or 7, and the operator **Min** means taking the minimal one among those in the brackets. The value of the superscript μ derived from eq 7 indicates which class the query receptor **G** belongs to. If there is a tie case, μ is not uniquely determined, but that did not happen for the datasets studied here.

Before using the above equations for practical calculations, we would like to draw the reader's attention to the following point. Owing to the normalization condition imposed on amino acid composition, of the 20 components in eq 1, only 19 are independent,[11] and hence, the covariance matrix $\mathbf{S}_m$ as defined

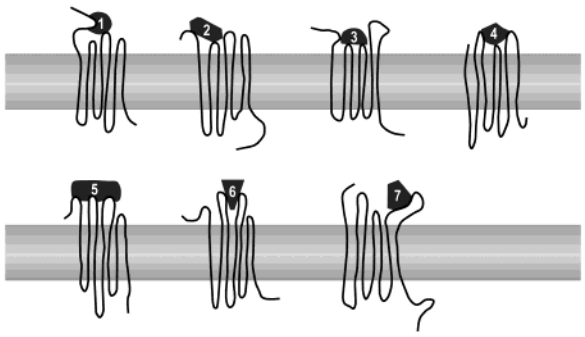

Figure 2. Schematic drawing to show different types of G-protein coupled receptors, where the receptors binding with ligands 1, 2, 3, 4, 5, 6, and 7 represent the adrenoceptor type, chemokine type, dopamine type, neuropeptide type, olfactory type, rhodopsin type, and serotonin type, respectively.

by eq 7 must be a singular one.[14] This implies that the Mahalanobis distance defined by eq 5 and the covariant discriminant function by eq 4 would be divergent and meaningless. To overcome such a difficulty, the dimension-reducing procedure[11] 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 and hence the corresponding covariance matrix $\mathbf{S}_m$ no longer singular. In such a 19-D space, the Mahalanobis distance (eq 5) and the covariant discriminant function (eq 4) can be defined without the divergence difficulty. However, which one of the 20 components can be left out? The answer is: any one. 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 ref 11, the value of the Mahalanobis distance as well as the value of the determinant of $\mathbf{S}_m$ will remain exactly the same regardless of which one of the 20 components is left out. Therefore, the value of the covariant discriminant function (eq 4) can be uniquely defined through such a dimension-reducing procedure.

III. Results and Discussion

By means of the covariant-discriminant algorithm described in the last section, a statistical analysis was performed for the 566 G-protein-coupled receptors listed in Table 1. The analysis was conducted by two different approaches, the re-substitution test and the jackknife test, as reported below.

Re-substitution Test. The so-called re-substitution test is an examination for the self-consistency of an identification method. When the re-substitution test is performed for the current study, the type of each G-protein-coupled receptor in a data set is in turn identified using the rule parameters derived from the same data set, the so-called training data set. The success rate thus obtained for the 566 receptors in Table 1 is summarized in Table 2, from which we can see that the overall success rate is 99.65%, indicating an excellent self-consistency. However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query receptor later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same receptors are used to derive the rule parameters and to test themselves. Accordingly, the success rate thus obtained represents some sort of optimistic estimation.[11,14,19,20] Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. An identification algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating an identification method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of an identification method in practical application. This is important especially for checking the validity of a training database: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

Jackknife Test. As is well-known, the independent data set test, sub-sampling test and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one (see ref 21 for a comprehensive discussion about this and ref 22 for the mathematical principle). During jackknifing, each receptor in the data set is in turn singled out as a tested receptor and all the rule-parameters are calculated based on the remaining receptors. In other words, the type of each receptor is identified by the rule parameters derived using all the other receptors except the one which is being identified. During the process of jackknifing, both the training data set and testing data set are actually open, and a receptor will in turn move from one to the other. The results of jackknife test thus obtained for the 566 G-protein-coupled receptors are also given in Table 2, from which the following phenomena can be observed. First, as expected, the success identification rates by jackknife test are decreased compared with those by the re-substitution test. Such a decrement is more remarkable for small subsets, such as the acetylcholine subset and the dopamine subset. This is because the cluster-tolerant capacity[23] for small subsets is usually low. And, hence, the information loss resulting from jackknifing will have a greater impact on the small subsets than the large ones. Nevertheless, the overall jackknife rate for the data set of 566 G-protein-couple receptors is still as high as 92.05%. It is expected that the success rate for identifying the types of G-protein-coupled

Table 2. Success Rates in Identifying the Types of G-Protein-Coupled Receptors on the Basis of the Amino Acid Composition[a]

success identification rate							
adrenoceptor	chemokine	dopamine	neuropeptide	olfactory	rhodopsin	serotonin	overall
Re-substitution Test							
$\frac{66}{66}$ = 100%	$\frac{91}{92}$ = 98.91%	$\frac{43}{43}$ = 100%	$\frac{31}{31}$ = 100%	$\frac{84}{84}$ = 100%	$\frac{182}{183}$ = 99.45%	$\frac{67}{67}$ = 100%	$\frac{\mathbf{564}}{\mathbf{566}}$ = **99.65%**
Jackknife Test							
$\frac{60}{66}$ = 90.91%	$\frac{85}{92}$ = 92.39%	$\frac{30}{43}$ = 69.77%	$\frac{19}{31}$ = 61.29%	$\frac{82}{84}$ = 97.62%	$\frac{182}{183}$ = 99.45%	$\frac{63}{67}$ = 90.03%	$\frac{\mathbf{521}}{\mathbf{566}}$ = **92.05%**

[a] The covariant discriminant algorithm[12] was used for identification.

receptors can be further enhanced by improving the training data of small subsets by adding into them more new proteins that have been found belonging to the types defined by these subsets.

IV. Conclusion

Let us imagine: if the samples of receptors are completely randomly distributed among the seven possible subsets, the rate of correct identification by randomly assignment would generally be $1/7 \cong 14\%$; if the distribution is weighted according to the sizes of subsets, then the rate of correct identification by the weighted random assignment would be $(66/566)^2 + (92/566)^2 + (43/566)^2 + (31/566)^2 + (84/566)^2 + (182/566)^2 + (67/566)^2 \cong 19\%$. Therefore, the rates of correct identification obtained based on the amino acid-composition in both the re-substitution and jackknife tests are much higher than the corresponding completely randomized rate and weighted randomized rate, implying that the type of G-protein-coupled receptors is considerably correlated with the amino acid-composition. This suggests that the types of G-protein-coupled receptors are predictable to a considerably accurate extent if a complete or quasi-complete training data set can be established for that purpose. The establishment of such a fast and accurate prediction method will speed up the pace of identifying proper G-protein-coupled receptors to facilitate drug discovery for the psychiatric and schizophrenic diseases, as currently conducted in many pharmaceutical laboratories.

Acknowledgment. We thank Raymond B. Moeller, Cynthia A. Ludlow, Wendy Vanderheide, and Katie Crawford of Pharmacia's Graphic Service Group for their help with drawing the figures in this paper.

References

(1) Schwartz, T. W. Molecular Structure of G-Protein-Coupled Receptors. In *Textbook of Receptor Pharmacology*; Forman, J. C., Johansen, T., Eds.; CRC Press: Boca Raton, FL, 1996; pp 65−84.
(2) Chou, K. C. Conformational change during photocycle of bacteriorhodopsin and its proton-pumping mechanism. *J. Protein Chem.* **1993**, *12*, 337−350.
(3) Watson, S.; Arkinstall, S. In *The G-Protein Linked Receptor Facts Book*; Academic Press: London, 1994; pp 1−291.
(4) Iismaa, T. P.; Biden, T. J.; Shine, J. G Protein-Coupled Receptors. In *G Protein-Coupled Receptors*; Springer-Verlag: Heidelberg, Germany, 1995; Chapter 1, pp 1−63.
(5) Voet, D.; Voet, J. G. Biochemistry. In *Biochemistry*, 2nd ed.; John Wiley & Sons: New York, 1995; pp 1276−1278.
(6) 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.
(7) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F. & G., V. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **1998**, *26*, 275−279.
(8) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* **2000**, *25*, 31−36.
(9) Gish, W. http://blast.wustl.edu/pub/nrdb/, 1999.
(10) 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.
(11) 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.
(12) Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.* **1999**, *12*, 107−118.
(13) 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.
(14) 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.
(15) Liu, W.; Chou, K. C. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *J. Protein Chem.* **1998**, *17*, 209−217.
(16) Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729−738.
(17) Mahalanobis, P. C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49−55.
(18) Pillai, K. C. S. Mahalanobis D2. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1985; Vol. 5, pp 176−181.
(19) Cai, Y. D. Is it a paradox or misinterpretation. *PROTEINS: Struct. Funct. Genet.* **2001**, *43*, 336−338.
(20) Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *PROTEINS: Struct. Funct. Genet.* **2001**, *44*, 57−59.
(21) Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.
(22) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979; pp 322 and 381.
(23) Chou, K. C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.* **1999**, *264*, 216−224.

PR025527K

research articles Journal of proteome research

Bioinformatical Analysis of G-Protein-Coupled Receptors *J. Proteome Res.* **2002**, *1*, 429–433,
Kuo-Chen Chou* and David W. Elrod

Page 431. The following type and accession numbers were missing in Table 1:

(7) 67 serotonin type receptors

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

PR030281K