Dimerization in Aminergic G-Protein-Coupled Receptors: Application of a Hidden-Site Class Model of Evolution†

Orkun S. Soyer,‡ Matthew W. Dimmic,§ Richard R. Neubig,|| and Richard A. Goldstein*,⊥

Department of Chemistry, Biophysics Research Division, and Department of Pharmacology, University of Michigan, Ann Arbor, Michigan 48109, and Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW71AA, U.K.

Received June 25, 2003; Revised Manuscript Received October 1, 2003

ABSTRACT: G-Protein-coupled receptors (GPCRs) are an important superfamily of transmembrane proteins involved in cellular communication. Recently, it has been shown that dimerization is a widely occurring phenomenon in the GPCR superfamily, with likely important physiological roles. Here we use a novel hidden-site class model of evolution as a sequence analysis tool to predict possible dimerization interfaces in GPCRs. This model aims to simulate the evolution of proteins at the amino acid level, allowing the analysis of their sequences in an explicitly evolutionary context. Applying this model to aminergic GPCR sequences, we first validate the general reasoning behind the model. We then use the model to perform a family specific analysis of GPCRs. Accounting for the family structure of these proteins, this approach detects different evolutionarily conserved and accessible patches on transmembrane (TM) helices 4−6 in different families. On the basis of these findings, we propose an experimentally testable dimerization mechanism, involving interactions among different combinations of these helices in different families of aminergic GPCRs.

G-Protein-coupled receptors (GPCRs)[1] are a superfamily of transmembrane proteins, which facilitate communication between the cell and its environment. They allow the cell to detect external signals and conditions by binding an extracellular ligand and activating a coupled intracellular G-protein (*1*). Found in a range of species from yeast to mammals, GPCRs function in a wide selection of cells from muscle cells to neurons. Their abundance and essential function in many pathways make them a key figure in many diseases, including Alzheimer's, anxiety, depression, and Parkinson's (*2*). All members of the GPCR superfamily share the same topology consisting of seven transmembrane (TM) helices connected with extra- and intracellular loops. On the basis of the similarity of function and sequence, GPCRs are usually classified into six classes [classes A−F (*3*)], the first three of which are found in mammals. Despite their essential role in the biochemistry of the cell, our knowledge of the structure of these proteins is relatively limited. Their nature and location in the cell make them resistant to biophysical methods of structure determination such as NMR spectroscopy and X-ray crystallography. Currently, there is only one determined crystal structure among the more than 4500 available GPCR sequences, that of bovine rhodopsin (PDB entry 1F88) (*4*).

On the basis of many experimental studies on different receptors, dimerization of GPCRs appears to occur quite generally (see ref *5* for a detailed review). Although the existence of homo- and heterodimers from different receptor families has been shown (*6*−*9*), there is no general agreement about the molecular mechanism of dimerization. There are currently two main hypotheses: domain swapping [as proposed by Gouldson *et al*. (*10*)] and contact dimerization. Although some computational studies seemed to support the domain swapping mechanism (*11*, *12*), recent experimental results seem to be inconsistent with this proposal (*13*, *14*). On the other hand, efforts to identify a possible dimer interface in different receptors are still inconclusive. Bouvier *et al*. showed that a peptide derived from the sixth TM region of β2-adrenergic receptor inhibits dimerization of these receptors and proposed a helix−helix interaction involving a conserved GxxxG motif on this helix (*15*). Recently, this view was supported by studies done on α-factor receptors, which have the same motif on the first TM helix (*16*). Other experiments suggested a role for disulfide links in dimerization, involving the extracellular end of the fourth TM helix in dopamine D2 receptors (*17*) and extracellular loops in muscarinic M3 receptors (*18*). Computational studies on different GPCR subgroups also proposed other possible interfaces (*19*), but these remain to be experimentally

† R.A.G. received support from NIH Grant LM0577. M.W.D. received support from NSF Grant 9726427. R.R.N. received support from NIH Grant HL46417. O.S.S. received support from the University of Michigan Bioinformatics Program.

* To whom correspondence should be addressed. E-mail: richard.goldstein@nimr.mrc.ac.uk. Phone: +44 (0)20 8816 2293. Fax: +44 (0)20 8816 2460.

‡ Department of Chemistry, University of Michigan.

§ Biophysics Research Division, University of Michigan. Current address: Department of Biological Statistics and Computational Biology, Cornell, Ithaca, NY 14853.

|| Department of Pharmacology, University of Michigan.

⊥ National Institute for Medical Research.

[1] Abbreviations: 5HT, serotonin receptor; αA, α-adrenergic receptor; βA, β-adrenergic receptor; ACM, muscarinic receptor; AIC, Akaike information criterion; ASA, accessible surface area; DA, dopamine receptor; EM, expectation maximization; GPCR, G-protein-coupled receptor; HH, histamine receptor; TM, transmembrane.

10.1021/bi035097r CCC: $25.00
Published on Web 11/21/2003

verified. Finally, recent work on rhodopsin dimers suggests that this receptor dimerizes through interactions among TM helices 4 and 5 (*20*). Taken together, these data raise the possibility of different dimerization interfaces in different GPCR families (*16*).

While most of our insights into protein structure and function still come from experimental methods such as NMR, X-ray crystallography, and spectroscopy, there has been a rapid development of computational techniques for analyzing proteins on the basis of only their sequences. Several of these techniques aim to detect important locations in proteins that might be involved in functions such as substrate binding or dimerization (e.g., refs *21*−*24*). Most are based on comparative sequence analysis, taking advantage of the availability of a number of different representatives of various types of proteins. The underpinning of such analyses is the fact that the proteins under inspection are evolutionarily related and thus might share common structural and functional features, which might leave detectable patterns in their sequences. A proper representation of the evolutionary relationships is critical in detecting and evaluating such patterns. Proper consideration and modeling of the evolutionary process have led to significant performance improvements in secondary structure prediction (*25*), detection of homologous proteins (*26*, *27*), and creation of multiple-sequence alignments (*28*).

The amount of evolutionary information that can be incorporated into a sequence analysis method is limited by the available evolutionary models. Current popular models of evolution (e.g., refs *29*−*31*) have primarily been developed for inferring phylogenetic relationships among proteins and are of limited use in deriving functional or structural information. This limitation is mainly due to the use of a single substitution matrix to model the evolution over the entire sequence of the protein. We have been developing a novel evolutionary model (*32*−*35*) to relax this limitation and allow substitutions in different parts of the protein to be represented by different substitution matrices. This model allows us to base the entire analysis on the modeling of the molecular evolution instead of ignoring or including such information as a secondary consideration, and can be used directly to gather functional and structural insights about the proteins under study. In previous work, we demonstrated the ability of the model to achieve better phylogenetic inference (*35*) and gather general structural and functional information about the proteins under study (*36*). In this paper, we present an application of this model to aminergic G-protein-coupled receptors (GPCRs).

We first review the evolutionary model as applied to a data set of aminergic GPCRs, demonstrating the correspondence between the information derived in our comparative analysis and what is known about these receptors from other sources of information. We then consider dimerization and use the evolutionary model to predict locations involved in this process, resulting in inference of a dimerization interface among TM helices 5 and 6 in most aminergic families and helices 4 and 5 in the muscarinic and opsin family. On the basis of these inferences and other experimental findings on dimerization mechanisms, we propose a testable hypothesis for GPCR dimerization.

MATERIALS AND METHODS

Evolutionary Model. Sequence evolution occurs at the molecular level through a number of mechanisms, including the accumulation of errors in DNA replication. These errors can lead to a mutant gene, of which some fraction will be accepted in the population and become the dominant genotype. One can create a mathematical model to simulate this process at the DNA or amino acid level and optimize it in light of the given data, that is, a multiple alignment of a family of proteins and a phylogenetic tree that relates them.

The general model for site substitutions is a matrix that specifies the rate at which every possible amino acid or nucleotide substitution occurs on an evolutionary time scale (*37*, *38*). Most available models of this kind use a single substitution matrix for all locations in all protein sequences. The probability that a given mutation would result in a functional protein and be accepted in the population is, however, certainly not constant over all locations in the sequence of a protein. For example, a change from a hydrophilic to a hydrophobic residue in the interior of a soluble protein might be easily accepted, while it would not be accepted so easily if it occurred on the outer surface. There may be more specific restrictions on the acceptable amino acids at locations of structural or functional significance. Thus, the use of a single substitution matrix is an inappropriate model of the evolutionary process, especially if we wish to use these models to improve our understanding of protein structure and function.

There have been attempts to account for this site variation by incorporating absolute rate heterogeneity among locations by multiplying the substitution rates by a location specific scaling factor (*30*). While these models are better able to represent biological data, they cannot account for *qualitative* variations in the type of selection pressure at various sites. Other models have been developed that allow for different locations to be under different types of selective pressure, either due to differences in local structure (*32*, *39*, *40*) or by allowing every location to be described by a different model (*41*). The former method ignores differences in selective pressure due to factors other than local structure, while the latter is limited by the amount of available data, since it is unlikely that all amino acids would be observed at each location of a given set of sequences, no matter how many currently existent organisms are sequenced.

We have been developing a novel, mixture (or "hidden-states") model in a number of publications (*32*−*35*). The novelty of this method is that it allows for variation at different locations by postulating that each location in a set of aligned sequences can be described by one of a number of different *types* of sites, called "site classes", each associated with a specific substitution model. Neither the assignment of locations to different site classes nor the corresponding substitution models are known *a priori*. Rather, these models are optimized using a maximum likelihood formulation. Once the substitution models have been derived, however, it is possible to assign the locations to different site classes *a posteriori*. This approach allows us to identify the locations in the protein which are under similar selective pressure, characterize the nature of these selective pressures, and characterize changes in the selective pressure, while avoiding many unnecessary and potentially

limiting assumptions about the causes underlying these differences. In addition, we are able to include both absolute and relative substitution rate variations in a consistent framework.

The method has been described previously (*32*, *39*, *40*) and is summarized here. The analysis starts with a set of aligned homologous protein sequences and the corresponding phylogenetic tree. As this evolutionary model does not assign locations to site classes *a priori*, we instead define an unknown prior probability $P(k)$ that any given location in the protein belongs to site class k. The set of prior probabilities, that is, our confidence in assigning a given location to any specific site class prior to considering the residues found in the proteins at that location, are the same for all locations in the protein. As all locations must belong to some site class, $\sum_k P(k) = 1$. At each location l, the local likelihood L_l can be calculated as the probability of the observed amino acids at that location (D_l) given the model's parameters Θ and the evolutionary tree topology and branch lengths T. Since each location can be represented by any of the site classes and each site class has distinct parameters θ_k, we must sum over all possible site classes to calculate this likelihood:

$$L_l = \sum_k P(D_l|\theta_k,T)P(k) \qquad (1)$$

where $P(D_l|\theta_k,T)$, the conditional probability that the observed data would result in given site specific model k with parameters θ_k and evolutionary tree T, is calculated using standard techniques (*29*). Summing the log of this likelihood over all locations in the alignment gives us the log likelihood (LL) for the entire set of proteins. We can then adjust all of the various parameters, the prior probabilities as well as the parameters describing the substitution models, to maximize this log likelihood. This is done using a modified form of the expectation maximization (EM) algorithm (*42*).

The hidden-sites model allows us to use all of the data to determine the parameters for a limited number of site classes. Even with this simplification, the limited amount of available sequence data (either currently or in the foreseeable future) makes it necessary to reduce the number of adjustable parameters used to describe the substitution rates. We do this by considering the relative "fitness" $F_k(A_i)$ of amino acid A_i for any location described by a particular site class k. The fitness value is related to the logarithm of the propensity of finding such an amino acid at any location described by this site class. We further assert that the probability of substitution between two amino acids should be a function of the change in fitness values resulting from such a substitution. Thus, for each site class, we define a matrix for all possible substitutions based on an overall substitution rate, the fitness values of the amino acids in that particular site class, and two additional adjustable parameters that define the functional relationship between substitution rate and fitness change. The decision on how many site classes to use to achieve a reasonable model is not trivial and will be discussed in detail (see Results and Discussion).

While we do not know the site class to which a given location belongs *a priori*, following optimization of the model we can calculate *a posteriori* probabilities using Bayes' rule. The conditional probability that a location l belongs to site class k is given by

$$P(k|D_l) = \frac{P(D_l|\theta_k,T)P(k)}{\sum_{k'} P(D_l|\theta_{k'},T)P(k')} \qquad (2)$$

This equation allows us to group locations in the protein that are under similar selective pressure into the same site class, while the parameters of that site class give us insight into the nature of the selective pressure at these locations. In particular, the overall substitution rate provides information about the magnitude of the selective pressure, while the fitness values for each amino acid type can indicate which types of amino acids are preferred for these locations. This information can then be used to obtain structural and functional insight into the proteins under study.

Applications to GPCRs. The evolutionary models used in this study are obtained using a multiple-sequence alignment and phylogenetic tree of 199 aminergic receptors from the class A family of the GPCR superfamily. The alignment is obtained from the September 2002 version of GPCRdb, a GPCR specific database (*43*) (http://www.cmbi.kun.nl/7tm/). The tree for this data set was created with MrBayes (*44*), a phylogenetic tree creation tool employing Bayesian statistics and Markov chain Monte Carlo search algorithms. The final tree was a consensus created from 2000 trees sampled through a run using 2 million generations and four Markov chains.

The physicochemical properties of amino acids used in correlation calculations were obtained from the AAindex database (*45*), which contains 434 different amino acid indices. We avoided indices related to spectroscopic methods and selected 145 physicochemical indices (see the Supporting Information for AAindex database codes of the indices that were used). Surface accessibility calculations for rhodopsin were carried out using the publicly available software GETAREA 1.1 (*46*). These calculations give the ratio of side chain surface area of a residue in the three-dimensional (3D) structure to the "random coil" value (i.e., the average solvent-accessible surface area of X in the Gly-X-Gly tripeptide). The proposed dimerization interactions are visualized using a rhodopsin-based model of pig α2-adrenergic receptor, which is available from http://mosberglab.phar.umich.edu/ resources/. This model is based on the 3D structure of rhodopsin, modified by residue substitution and side chain rotamer optimization (see the above link for more details). Dimers are visualized by manual docking of two copies of the receptor model (and rhodopsin structure), based on the interactions proposed in the text.

The evolutionary models are optimized using a maximum likelihood scheme on a multiple-sequence alignment and phylogenetic tree of proteins under study. The software for optimizing and employing such models is available from the authors upon request.

RESULTS AND DISCUSSION

To assess the statistical behavior of the model with an increasing number of site classes, we optimized a set of models containing between 4 and 17 site classes on the set of 199 aminergic receptors from class A GPCRs, using the associated phylogenetic tree (see Figure 1) and multiple-

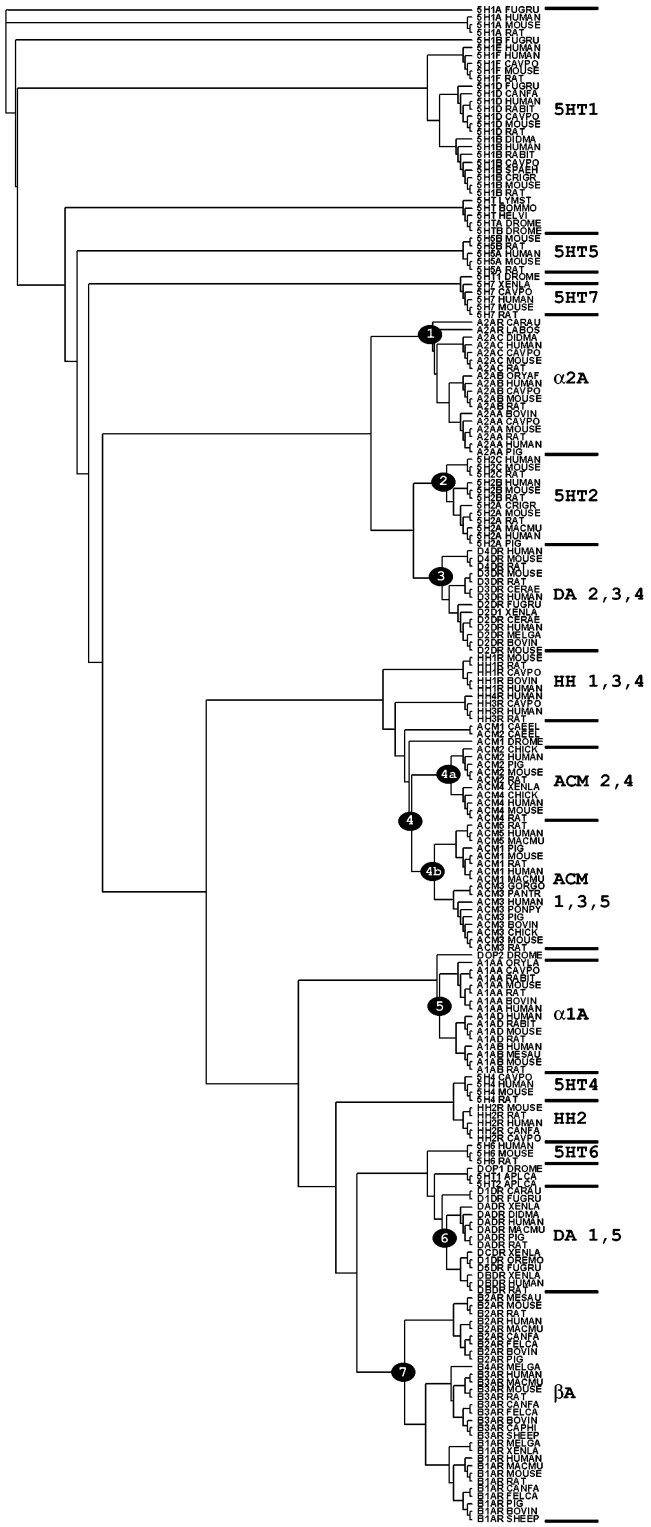

FIGURE 1: Unrooted phylogenetic tree of aminergic GPCRs. Receptors forming a monophyletic family on the basis of their ligand are indicated on the right side of the figure (5HT, serotonin; αA, α-adrenergic; ACM, muscarinic; *β*A, *β*-adrenergic; DA, dopamine; HH, histamine). Sequences falling under the numbered nodes are used in the dimerization analysis. The same node numbers are referenced in Table 3.

sequence alignment. While the choice for the number of site classes is a complex problem, involving questions of both statistical significance and ease of interpreting the results, an incomplete but useful measure of the behavior of the model with an increasing number of site classes can be quantified using the Akaike information criterion (AIC) (*47*) (see ref *48* for a detailed discussion of model comparison). This criterion adjusts the log likelihood value (LL) to take into account the expected dependence on the number of parameters in the model (*K*), giving the distance between the "true" underlying model that created the data and the proposed model fitted from the data.

$$\mathrm{AIC} = -2\mathrm{LL} + 2K \qquad (3)$$

According to this measure, the most informative model among a set of related models is the one with the lowest AIC value. Figure 2 shows the continual increase in log likelihood and decrease in AIC values with an increasing number of site classes. While these evaluations suggest that the most statistically appropriate site class model has even more than 17 classes, the larger number of site classes makes the analysis of the results more difficult. In practice, we found that models with 5−12 site classes are reasonable to analyze and provide good insight into the functional and structural features of proteins under study.

For clarity, we first discuss the results for the model with six site classes. The overall substitution rate and prior probability for the various resulting site classes are listed in Table 1. Also included in this table are the top three correlations between fitness values and amino acid properties for each site class (see Materials and Methods for properties used). One site class (class 1) represents highly conserved locations as indicated by the low overall rate of substitution (ν_k), while two site classes (2 and 3) have a moderate rate of evolution and a positive correlation with properties such as hydropathy, membrane preference, and α-helix propensity. Most locations in the TM helices are assigned to these three site classes (see Figure 3). The remaining three site classes (4−6) exhibit increasing substitution rates, with fitness values showing a positive correlation with properties such as flexibility and hydrophilicity. Locations outside the membrane are almost entirely assigned to these latter site classes. Besides this general correlation between site class distribution and the topology of GPCRs, the evolutionary model provides further insights into locations in the membrane. A careful analysis of the site class distribution over TM helices reveals a correlation with the accessible surface area from rhodopsin structure; almost all residues with high accessible area (facing the lipid or solvent) are assigned to site classes 3 and up in the model, while residues that have a small accessible surface (facing the interior) are generally assigned to site classes 1 and 2.

Increasing the number of site classes in the model can lead to detection of new features of proteins under study as demonstrated in Figure 4. Here we show the site class distribution for a model with 11 site classes, where transmembrane residues are generally assigned to five different site classes (site classes 1−5), while loop residues are distributed among the remaining six site classes. It is interesting to note the distribution of site class 6 (green), which holds mainly residues from intracellular loops 1 and 2, residues at the beginning and end of intracellular loop 3, and residues at the beginning of the (intracellular) C-terminus. This site class has a moderate rate of evolution (0.40), and its fitness values correlate best with properties such as positive charge (correlation coefficient = 0.60) and

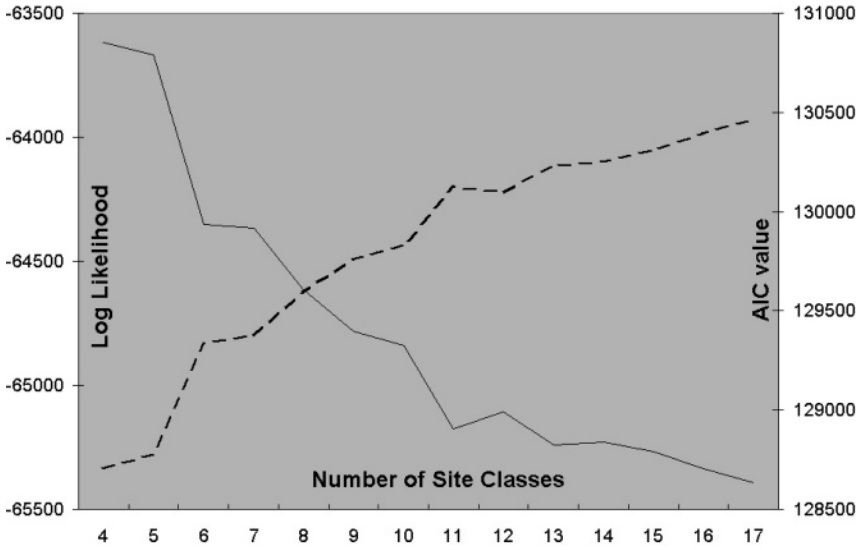

FIGURE 2: Number of site classes in the model vs log likelihood (− − −) and Akaike information criterion (AIC) (—).

Table 1: Parameters for the Six-Site Class Model As Optimized on the Aminergic GPCR Data Set[a]

site	substitution rate (ν)	prior probability	first correlation CC	first correlation property	second correlation CC	second correlation property	third correlation CC	third correlation property
1	0.01	0.10	−0.64	helix frequency (PALJ810109)	−0.55	helix indices (GEIM800104)	−0.52	helix probability (KANM800101)
2	0.08	0.24	0.89	hydropathy (KYTJ820101)	0.81	membrane (ARGP820103)	0.81	helical (ARGP820102)
3	0.27	0.20	0.92	hydropathy (KYTJ820101)	0.82	β-sheet (KANM800104)	0.80	membrane (DESM900102)
4	0.59	0.15	0.84	flexibility (VINM940103)	−0.84	buried (WERD780101)	0.80	water-occluded site (KRMV790102)
5	1.15	0.15	0.72	vap. to Chx E. (RADA880103)	−0.71	hydrophobicity (CIDH920105)	−0.71	polarizable (CHAM820101)
6	2.46	0.16	0.74	charge donor (CHAM830108)	0.74	vap. to Ehx E. (RADA880103)	−0.73	polarizable (CHAM820101)

[a] Also included are three highest correlation coefficients (CC) between fitness values and selected physicochemical properties of amino acids (see Materials and Methods).

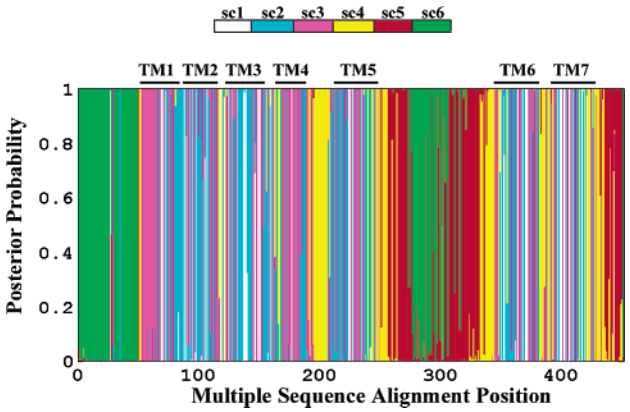

FIGURE 3: Posterior probabilities of site classes for the six-site class model. The posterior probability for each site class at a given location is shown as a colored bar. The bars for each site class are stacked on top of each other starting with site class 1. TM helix locations are given at the top of the figure: site class 1 (white), 2 (turquoise), 3 (magenta), 4 (yellow), 5 (red), and 6 (green).

isoelectric point (correlation coefficient = 0.64). It is clear that this site class picks the residues that are under a selective pressure which seems to satisfy the "positive-inside rule", the high frequency of positively charged residues found in cytosolic loops of membrane proteins (*49*).

Identifying Sites of Functional Significance. We would expect that a large fraction of mutations at locations that are unimportant for the protein's structure and function would be neutral, with the result that mutations accumulate rapidly at these locations because of genetic drift. Conversely, positions that are more important in the protein will be more constrained; fewer mutations will be accepted, and there will be a slower substitution rate. Thus, locations that are conserved in a set of related proteins over the course of evolution are likely involved in structural and functional roles. In the evolutionary model, such locations are assigned to site classes with the slowest rate of substitution (smallest ν_k). Because the evolutionary model considers both the overall substitution rate and the propensities of the various amino acids for different locations, this list of locations is not necessarily the same as the list of locations that are highly conserved in the multiple-sequence alignment. A location could be assigned to a site class with a higher rate of evolution, even though it is highly conserved, if it contains residues that are consistent with that site class. In addition, many of the more generic selective pressures, such as hydrophobicity and flexibility, are included in the distribution of fitness values and thus are removed from the estimation of ν_k. This behavior of the model might make it easier to identify conservation and could allow discrimination among

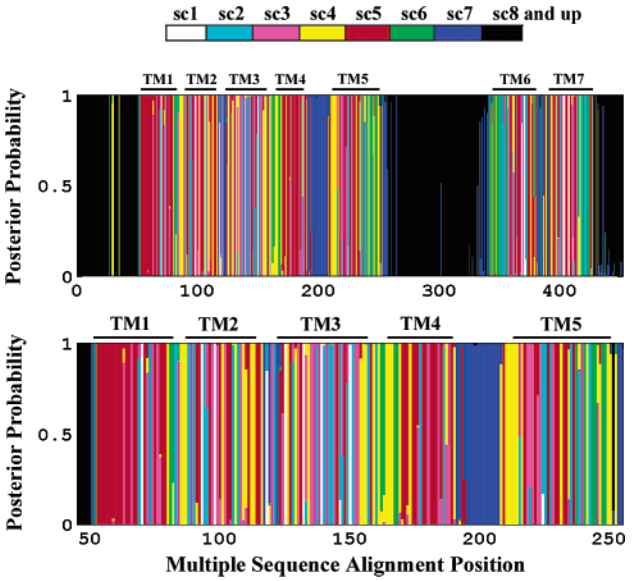

FIGURE 4: Posterior probabilities of site classes for a model with 11 site classes: (top) full view and (bottom) expansion for the first five TM helices. Color coding is as in Figure 3, with additional site classes 7 (blue) and 8 and above (black). TM helices are given at the top of each panel. Locations from the beginning and end of intracellular loops 1−3 (between helices 1 and 2, 3 and 4, and 5 and 6, respectively) and locations from the beginning of the carboxyl tail are primarily assigned to site class 6 (green) as discussed in the text.

conserved locations that are solely functional and those that are under other constraints.

Table 2 lists the locations that have a total posterior probability of 0.7 or higher as belonging to site classes 1 and 2, the slowest evolving site classes, in the 11-site class model. The indexing scheme is that of Ballesteros *et al*. (*50*). Twenty-one of these 56 locations have been experimentally determined to be involved in ligand or G-protein binding and specificity (*51, 52*). These include 16 of 33 locations that have been experimentally defined in ligand binding in aminergic receptors (*52*).

Insights into the Dimerization Interface. As discussed in the introductory section, dimerization of GPCRs could involve different regions in different families, making evolutionary considerations more important for any sequence-based analysis. We used the evolutionary model to detect possible dimerization interfaces in different families of GPCRs by employing the following strategy. First we fixed all parameters of the 11-site class model we derived from the phylogenetic tree of all aminergic receptors. Then we updated the posterior site class distribution of this model at different nodes of the tree, corresponding to monophyletic families based on the ligand specificity of the receptors. Thus, each updated model is a representative of the evolutionary constraints acting on the family defined by such node. Models from different nodes are then analyzed for evolutionarily slow evolving locations (posterior probability of the three most slowly evolving site classes above 0.7) that are in the TM region and accessible according to the rhodopsin structure [accessible surface area (ASA) ratio of >30%]. Although there may be other reasons for conservation of these locations, we argue that identifying such locations (i.e., locations conserved in a family and accessible in the rhodopsin structure) would highlight regions potentially involved in dimerization of the corresponding family. Table 3 summarizes the results of this analysis where each model is designated by the node code from which it was derived (see Figure 1 for node codes) and locations are indexed again by the indexing scheme of Ballesteros *et al*. (*50*).

Table 2: Slowly Evolving Locations that Have a Total Posterior Probability for Site Classes 1 and 2 Equal to or Greater than 0.7 in the 11-Site Class Model[a]

TM index	α2a	TM index	α2a
	Asn14	*4.50*	*Trp158*
1.49	Gly50	*4.59*	*Pro167*
1.50	Asn51	*5.47*	*Phe205*
1.53	Val54	5.48	Phe206
1.63	Leu64	5.50	Pro208
2.40	Asn69	5.58	Tyr216
2.42	Phe71	5.61	Ile219
2.45	Ser74	5.65	Ala223
2.46	Leu75	6.30	Glu331
2.50	*Asp79*	6.32	Arg333
2.57	Val86	*6.33*	*Phe334*
2.59	Phe88	*6.37*	*Leu338*
	Trp99	6.44	Phe345
	Gly102	*6.48*	*Trp349*
3.25	*Cys106*	6.50	Pro351
3.28	*Tyr109*	*6.51*	*Phe352*
3.32	*Asp113*	*6.52*	*Phe353*
3.37	Thr118	*7.40*	*Trp375*
3.39	Ser120	*7.43*	*Tyr378*
3.40	*Ile121*	*7.45*	*Asn380*
3.42	His123	*7.46*	*Ser381*
3.43	Leu124	7.49	Asn384
3.46	Ile127	7.50	Pro385
3.49	*Asp130*	7.53	Tyr388
3.50	*Arg131*	7.57	Asn392
3.51	*Tyr132*	7.60	Phe395
	Ala138	7.63	Ala398
	Tyr141	7.64	Phe399

[a] TM indexing is that of Ballesteros *et al*. (*50*), where the first index identifies the TM helix and the second gives the relative location in the helix, with 50 representing the most conserved residue in each helix. Residue identities are from the human α2a-adrenergic receptor. Locations experimentally determined to be involved in ligand binding or G-protein coupling are given in italics.

As seen in Table 3, we observe a general abundance of conserved locations on TM helices 4−6, with different distributions among different families. Interestingly, these three helices have been predicted to be involved in dimerization by both experimental and computational studies as discussed in the introductory section. More specifically, we observe a number of conserved locations in almost all families. These are 2.59, 3.51, 3.53, 5.48, 5.58, 5.62, 6.42, and 6.50, corresponding to P88, Y132, S134, F206, Y216, Y220, G343, and P351 in the human α2a-adrenergic receptor, respectively. Location 3.51 contains the last residue in the conserved "DRY" motif at the N-termini of the second intracellular loop that has been shown to be important in stabilizing the receptor in a functional conformation (*53*). Locations 2.59 and 6.50 contain proline residues in almost all aminergic receptors and are inferred to induce functional kinks in these helices (*50, 54*). Location 5.48, which contains a phenylalanine in 5HT2 receptors, has been shown to be involved in ligand binding in this subfamily (*55*), but such evidence is not reported for any other subfamily where this location is sometimes replaced with a tyrosine. Thus, the conservation at these four locations in each subfamily is not surprising given these experimental findings. Location 3.53

Table 3: Locations that Are Slowly Evolving According to the 11-Site Class Model (total posterior probability for site classes 1−3 of >0.7) and Are Accessible in the Rhodopsin Structure (ASA value of >30%), for Various Subgroups of Aminergic GPCRs[a]

α2A (1)		5HT2 (2)		DA2,3,4 (3)		ACM2,4 (4a)		ACM1,3,5 (4b)		α1A (5)		DA1,5 (6)		BA (7)	
1.59	T	1.34	W			1.40	G			1.59	C			1.40	A
						1.41	S								
2.55	T					2.48	C	2.48	C						
2.59	*P*	*2.59*	*P*	*2.59*	*P*	*2.59*	*N*	*2.59*	*N*	*2.59*	*P*	*2.59*	*K*	*2.59*	*P*
2.63	A	2.67	Y	2.63	Y										
				3.26	D			3.48	F					3.26	E
3.51	*Y*	*3.51*	*Y*	*3.51*	*Y*	*3.51*	*Y*	*3.51*	*Y*	*3.51*	*Y*	*3.51*	*Y*	*3.51*	*Y*
3.53	*S*	*3.53*	*A*	*3.53*	*A*	*3.53*	*S*	*3.53*	*F*	*3.53*	*G*	*3.53*	*A*	*3.53*	*L*
4.38	T											4.38	T	4.38	T
4.39	P					**4.43**	**G**								
4.54	A	4.55	G					**4.47**	**G**					4.54	A
		4.61	P					4.54	F			4.58	F	4.58	F
5.37	W							5.36	P			5.37	S		
								5.40	F						
5.48	*F*	*5.48*	*F*	*5.48*	*Y*	*5.48*	*Y*	*5.48*	*Y*	*5.48*	*Y*	*5.48*	*Y*	*5.48*	*Y*
5.49	A													5.56	F
5.58	**Y**	**5.58**	**Y**	**5.58**	**Y**	**5.58**	**Y**	**5.58**	**Y**	**5.58**	**Y**	**5.58**	**Y**	**5.58**	**Y**
										5.59	C	5.59	T		
5.62	**Y**			**5.62**	**Y**	**5.62**	**S**	**5.62**	**Y**	**5.62**	**Y**	**5.62**	**Y**	**5.62**	**F**
6.30	E					5.67	S			6.27	F				
6.35	F			6.30	E					6.28	S	6.26	S	6.30	E
												6.31	T		
6.42	**G**			**6.42**	**G**					**6.42**	**G**	**6.42**	**G**	**6.42**	**G**
6.49	F	6.49	C												
6.50	*P*	*6.50*	*P*	*6.50*	*P*			*6.50*	*P*	*6.50*	*P*	*6.50*	*P*	*6.50*	*P*
6.53	F									6.56	P	6.56	C		
6.56	S							6.60	F			6.60	F		
												7.34	T		
7.44	C			7.37	A					7.55	C	7.41	F		
												7.44	A		
												7.58	D		
												7.69	G	7.69	C
		7.70	C			7.70	C	7.70	C			7.70	C		

[a] Subgroup identification and number (in parentheses) match the data in Figure 1. The most common amino acid in each particular subgroup is also included. Locations that are present in almost all models are given in italics, and locations possibly involved in dimerization and discussed in the text are given in bold. TM indexing is as described in the footnote of Table 2. Corresponding locations are approximately aligned.

is highly conserved in a family-dependent manner, where it is occupied by a different residue in different families. Although there is no experimental evidence for its role, this residue is located at the beginning of the second intracellular loop and could be involved in a functional role such as interaction with the G-protein.

The fully conserved tyrosine at location 5.58 and the highly conserved tyrosine and serine residues at location 5.62 in most families are accompanied by the conserved GxxxG motif at locations 6.38 and 6.42 in TM helix 6 in all of these families, except for muscarinic (ACM) and serotonin type 2 (5HT2) receptors (location 6.38, which contains the other glycine of this motif, is also highly conserved among families but does not show up in this analysis because of its rather low ASA ratio of 21.7% in the rhodopsin structure.). In some receptors, one or both glycines from this motif are replaced with a serine or alanine. These data could be explained if dimerization in these receptors is stabilized through a $C_\alpha H\cdot\cdot\cdot O$ hydrogen bond between these conserved entities, a type of interaction known to be an important factor in protein dimerization and TM helix interactions (*56*). As described by Engelman *et al*. (*56*), the key residues involved in such an interaction are alanine, glycine, and serine with their small side chains, and serine and threonine with their extended oxygen atom. Such an interaction could involve a tyrosine as easily. An interaction like this could bring TM helices 5 and 6 closer, creating a dimerization interface explaining the observed conservation on the accessible surface of these helices. The possibility of such an interaction is supported by the observation that the described entities from these two helices are at the same level in the lipid bilayer according to rhodopsin-based models (see Figure 5). Furthermore, some experimental studies on β-adrenergic and α-factor receptors suggest a role for the same GxxxG motif in dimerization (*15*, *16*).

Such a model must explain the absence of a GxxxG-like motif on TM helix 6 in some aminergic families. This motif is even absent in rhodopsin, which has recently been shown to dimerize *in vivo* (*57*), even though the tyrosine at location 5.58 is still conserved in these receptors. In muscarinic receptors, a potential resolution of this question is suggested by the presence of several highly conserved locations on TM helix 4. In fact, locations 4.43 and 4.47, which we detect upon carrying out the above-explained analysis on nodes 4a and 4b, contain an alanine or glycine in all muscarinic receptors. (Because of such interchange between these residues, locations 4.43 in ACM1, -3, and -5 and 4.47 in ACM2 and -4 have a total posterior probability that is slightly less than the set threshold of 0.7.) Thus, a possible resolution is that the interaction in muscarinic receptors is between the conserved tyrosine and these conserved residues on TM helix 4. Interestingly, all rhodopsin sequences show a similar motif composed of a fully conserved alanine at position 4.42 and a highly conserved glycine at position 4.45, supporting a possible interaction between TM helices 4 and 5 as suggested by other studies (*20*). Finally, serotonin type 2 receptors also

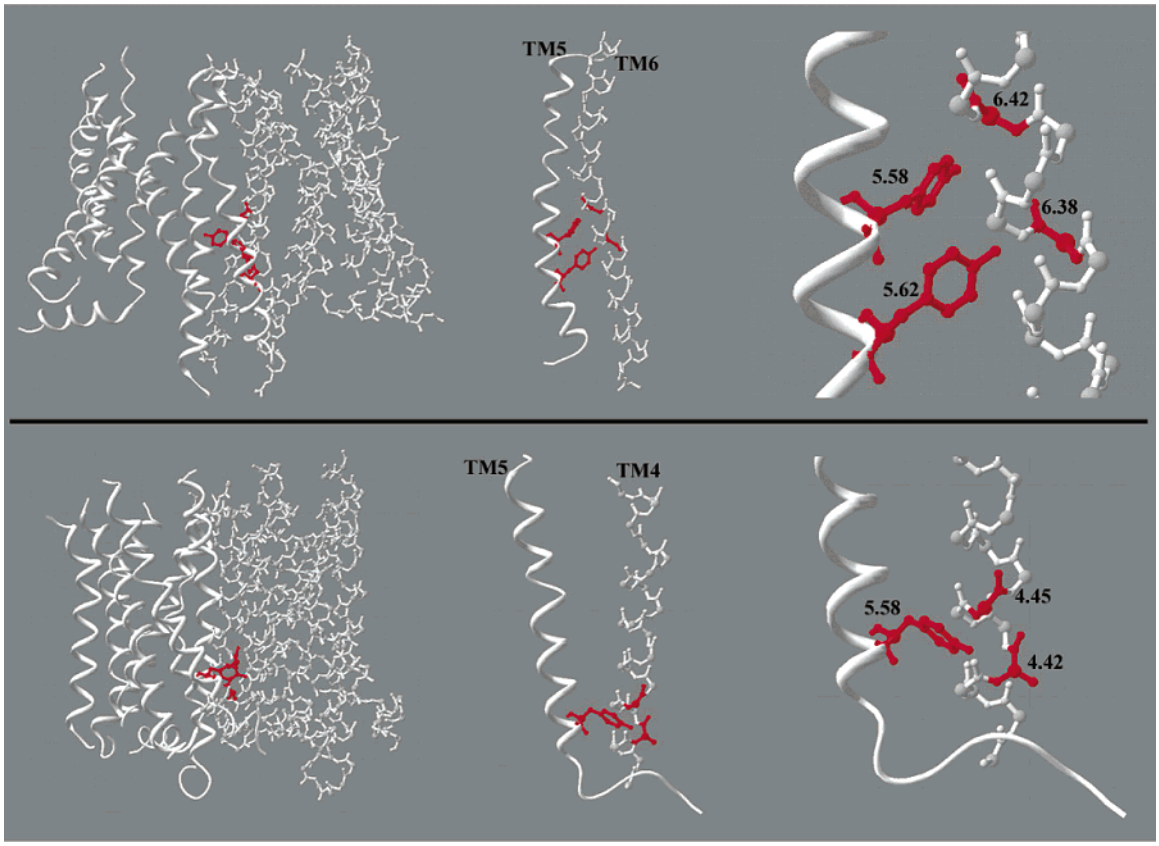

FIGURE 5: Molecular models of two receptor molecules dimerized manually based on the proposed interactions. The two receptors are shown as a ribbon and α-carbon trace to indicate the two separate molecules. The top panel shows the dimer interaction through TM helices 5 and 6 as modeled using a pig α2-adrenergic receptor model (see Materials and Methods for model description). The bottom panel shows the dimer interaction through TM helices 4 and 5 as modeled using the rhodopsin structure. Panels from left to right show increasing detail: complete receptor molecules with all seven TM helices, TM helices 5 and 6 and TM helices 4 and 5 in the top and bottom panels, respectively, and close-up views of the proposed interaction. Residues discussed in the text (4.42, 4.45, 5.58, 5.62, 6.38, and 6.42) are shown in the close-up views.

have a similar motif created by a highly conserved serine and alanine at locations 4.38 and 4.42, respectively (these locations are not listed in Table 3 because of their low accessibility in the rhodopsin structure). Again, an inspection of the rhodopsin structure shows that these entities are at the same level in the lipid bilayer (see Figure 5).

In summary, we propose that the above-mentioned interaction is the main driving force in dimerization of these receptors and involves TM helices 5 and 6 for most of the analyzed families, and TM helices 4 and 5 for the muscarinic, opsin, and possibly serotonin type 2 families. The association of these helices by the interaction of a GxxxG-like motif and a tyrosine or serine residue from location 5.58 or 5.62 could then further be stabilized by other interactions, explaining the relatively high level of conservation on accessible surfaces of these helices. Using rhodopsin-based homology models of aminergic receptors and rhodopsin structure, we found that the proposed interactions result in a possible spatial arrangement for dimers without any structural conflicts, as shown in Figure 5.

Comparison with Other Computational Studies. Previously, there have been several other computational studies on GPCRs for detecting possible dimerization interfaces (*11*, *12*, *19*) and mechanisms (*10*). The studies that looked for possible dimerization interfaces used correlated mutation analysis (*12*, *19*) and the evolutionary trace method (*58*) in conjunction with 3D models to detect functional locations lying on external faces of TM helices. The latter analysis used 52 β-adrenergic receptor sequences and detected two clusters of locations, one on helices 5 and 6 and another on helices 2 and 3. The authors concluded that these findings supported a domain swapping mechanism that they had proposed previously (*10*). The same group also carried out a correlated mutation analysis using sequences from chemokine, neurokinin, opiate, thyrotropin, somatostatin, and aminergic receptors to further support such a mechanism (*12*). While the general results of their studies are similar to ours in terms of the abundance of detected locations on helices 5 and 6, there are many differences. For example, among the 28 locations that make up the two clusters they detected on β-adrenergic receptors, only three are detected in our analysis (locations 5.48, 6.42, and 6.50). The domain swapping mechanism has been challenged recently by some experimental findings with vasopressin (*14*) and dopamine receptors (*13*).

The second correlated substitution analysis was carried out by Weinstein *et al.* (*19*) using δ-, μ-, and κ-opioid receptor sequences. They used the predicted locations of correlated substitutions to construct possible structural models of homodimers of these receptors. Their analysis suggests a possible dimer interface involving TM4 and TM5 or combinations thereof in δ- and κ-opioid receptors. Although this study involves a completely different family of GPCRs, and analyzing correlated substitutions is problematic, it is

interesting that the results suggest helices 4 and 5 as possible candidates for the dimer interface.

The specific differences in the outcome of different analyses can be attributed both to the differences in selected data sets and to methods employed in the analyses. Although comparison of different approaches to detecting functional locations and correlated mutations is beyond the scope of this paper, it is important to note the importance of accounting for evolutionary relations in any type of sequence analysis. Such relations can introduce their own patterns into the sequences, causing possible problems in sequence analysis as demonstrated in the case of detecting correlated mutations (*59*, *60*). It is also very important to consider such relations when evaluating conserved locations and their relation to function. For example, two sequences that seem to be most closely related in the analysis of a family could actually be separated by many sequences that are not included in the study. Not accounting for these evolutionary relations can have vital effects on the results of the analysis, such as defining conservation from a lack of divergence time as functionally important. Such issues related to the method that is used, and to the way it is applied, could explain the differences between the results presented here and those from other studies.

All of the predictions in this study, and some other computational studies, are based on the rhodopsin structure, the only available structure for class A GPCRs. Although rhodopsin and aminergic GPCRs are members of this same class, it is believed that there are structural differences among the two (*50*). Thus, one must keep in mind the possible effects that such differences might have on sequence-based analyses that use the rhodopsin structure to make predictions about aminergic GPCRs. There should be special care exercised with identification of residues with high degrees of surface accessibility, as used in this analysis and others. One should also be aware of the main assumption on which such predictions are based, namely, the direct relation between "functional" locations on external faces of TM helices and dimerization. Such locations do not have to be involved solely in dimerization and could have many other roles such as structural integrity or interactions with the lipid bilayer or other proteins. Still, such predictions could be very helpful in developing experimentally testable hypotheses, as demonstrated in this study.

CONCLUSION

We present a probabilistic model, employing site specific substitution matrices, to simulate the underlying evolutionary process, resulting in a set of related proteins. The approach simultaneously and naturally includes two different types of information, the degree of conservation and the preference for specific amino acid types, in one integrated model. This work along with our earlier results (*35*, *36*) shows that this model is capable of defining the selective pressures acting on different locations in a set of evolutionarily related proteins, and that such selective pressures correlate with structural and functional features of the proteins under study. Such capabilities of the introduced model allow it to be used as a sequence-based tool to analyze protein structure and function.

Applying the evolutionary model to GPCRs, we were able to show that the model's main parameters, the assignment of different locations in the protein to their corresponding site classes on the basis of the posterior probabilities as well as the fitness parameters for each site class, provide valuable information about these proteins. These include information about the environment of locations among the sequence (i.e., locations in the membrane vs the cytosol and TM locations facing the lipid vs the protein interior), about the preferred secondary structure at such locations (i.e., locations with α-helix propensity), and about regions of functional significance (i.e., locations involved in ligand and/or G-protein binding). More significantly, we were able to detect likely dimerization regions in a family-dependent manner, identifying potential differences in the dimerization location between the muscarinic receptors and most of the other aminergic GPCRs.

The explicit model of the evolutionary process allowed us to derive structural and functional information in a straightforward manner without imposing an *a priori* basis for the relationship between these properties and the evolutionary behavior. Thus, the approach presented here can be applied to any set of evolutionarily related proteins, provided one has sequence information for a sufficiently divergent set of members. Although some of the results presented here may be available for GPCRs from other sequence-based or experimental studies, the simplicity and applicability of this method make it quite attractive for studies on other, less-studied protein families.

There may be some potential issues associated with this approach. One concern, as with all sequence-based analyses, is the accuracy of the multiple-sequence alignment. The analysis is vulnerable to this issue as much as any other sequence-based approach. This is a minor problem in the analysis of the TM regions of GPCRs, as these regions are relatively easy to align because of the presence of highly conserved "anchor" residues in the transmembrane helices (*50*). Outside the membrane and especially in the long third intracellular loop and carboxyl tail, the accuracy of the alignment is more questionable. A second concern is the reliability of the phylogenetic tree. To optimize the parameters of the model, we assume that we know the underlying true phylogenetic tree relating the proteins under study. This assumption is probably not entirely correct given the large space of possible trees and the limitations on phylogenetic inference tools. Still, we expect the model to be robust toward small changes in the tree structure, as shown for similar models (*61*).

A final potential issue with this approach is the optimization procedure. We are faced with finding the global minimum in a high-dimensional space. This is certainly not an easy problem to solve, and it is likely that we are not reaching a global optimum for our parameters. Despite this fact, the results presented here show that the optimization procedure is sufficient to provide reliable and biologically interpretable results.

ACKNOWLEDGMENT

Thanks to Sarah E. Ingalls for her contributions in writing the software, Roger K. Sunahara for insightful discussions, and Todd Raeker for computer support.

SUPPORTING INFORMATION AVAILABLE

AAindex database codes of the indices that were used. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Iismaa, P. T., Biden, J. T., and Shine, J. (1995) *G-Protein Coupled Receptors*, Springer-Verlag, New York.
2. Spiegel, A. M. (1998) *G Proteins, Receptors and Disease*, Humana Press, Totowa, NJ.
3. Kolakowski, J. L. F. (1994) *Receptors Channels 2*, 1−7.
4. Palcewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000) *Science 289*, 739−745.
5. Angers, S., Salahpour, A., and Bouvier, M. (2002) *Annu. Rev. Pharmacol. Toxicol. 42*, 409−435.
6. Angers, S., Salahpour, A., Joly, E., Hilairet, S., Chelsky, D., Dennis, M., and Bouvier, M. (2000) *Proc. Natl. Acad. Sci. U.S.A. 97*, 3684−3689.
7. Franco, R., Ferri, S., Agnati, L., Torvinen, M., Gines, S., Hillion, J., Casado, V., Lledo, P. M., Zoli, M., Lluis, C., and Fuxe, K. (2000) *Neuropsychopharmacology 23*, S50−S59.
8. Maggio, R., Vogel, Z., and Wess, J. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 3103−3107.
9. Zawarynski, P., Tallerico, T., Seeman, P., Lee, S. P., O'Dowd, B. F., and George, S. R. (1998) *FEBS Lett. 441*, 383−386.
10. Gouldson, P. R., Higgs, C., Smith, R. E., Dean, M. K., Gkoutos, G. V., and Reynolds, C. A. (2000) *Neuropsychopharmacology 23*, 60−77.
11. Gkoutos, G. V., Higgs, C., Bywater, R. P., Gouldson, P. R., and Reynolds, C. A. (1999) *Int. J. Quantum Chem. 74*, 371−379.
12. Gouldson, P. R., Dean, M. K., Snell, C. R., Bywater, R. P., Gkoutos, G., and Reynolds, C. A. (2001) *Protein Eng. 14*, 759−767.
13. Lee, S. P., O'Dowd, B. F., Ng, G. Y. K., Varghese, G., Akil, H., Mansour, A., Nguyen, T., and George, S. R. (2000) *Mol. Pharmacol. 58*, 120−128.
14. Schulz, A., Grosse, R., Schultz, G., Gudermann, T., and Schoneberg, T. (2000) *J. Biol. Chem. 275*, 2381−2389.
15. Hebert, T. E., Moffet, S., Morello, J. P., Loisel, T. P., Bichet, D. G., Barret, C., and Bouvier, M. (1996) *J. Biol. Chem. 271*, 16384−16392.
16. Overton, M. C., Chinault, S. L., and Blumer, K. J. (2003) *J. Biol. Chem.* (in press).
17. Guo, W., Shi, L., and Javitch, J. A. (2002) *J. Biol. Chem. 278*, 4385−4388.
18. Zeng, F. Y., and Wess, J. (1999) *J. Biol. Chem. 274*, 19487−19497.
19. Filizola, M., and Weinstein, H. (2002) *Biopolymers 66*, 317−325.
20. Liang, Y., Fotiadis, D., Filipek, S., Saperstein, D. A., Palczewski, K., and Engel, A. (2003) *J. Biol. Chem. 278*, 21655−21662.
21. Casari, G., Sander, C., and Valencia, A. (1995) *Nat. Struct. Biol. 2*, 171−178.
22. Landgraf, R., Xenarios, I., and Eisenberg, D. (2001) *J. Mol. Biol. 307*, 1487−1502.
23. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) *J. Mol. Biol. 257*, 342−358.
24. Valdar, W. S. J., and Thornton, J. M. (2001) *Proteins: Struct., Funct., Genet. 42*, 108−124.
25. Goldman, N., Thorne, J. L., and Jones, D. T. (1996) *J. Mol. Biol. 263*, 196−208.
26. Rehmsmeier, M., and Vingron, M. (2001) *Proteins: Struct., Funct., Genet. 45*, 360−371.
27. Qian, B., and Goldstein, R. A. (2003) *Proteins: Struct., Funct., Genet. 52*, 446−454.
28. Holmes, I., and Bruno, W. J. (2001) *Bioinformatics 17*, 803−820.
29. Felsenstein, J. (1981) *J. Mol. Evol. 17*, 368−376.
30. Yang, Z. (1993) *Mol. Biol. Evol. 10*, 1396−1401.
31. Yang, Z., and Rannala, B. (1997) *Mol. Biol. Evol. 14*, 717−724.
32. Koshi, J. M., and Goldstein, R. A. (1995) *Protein Eng. 8*, 641−645.
33. Koshi, J. M., and Goldstein, R. A. (1998) *Proteins: Struct., Funct., Genet. 32*, 289−295.
34. Koshi, J. M., Mindell, D. P., and Goldstein, R. A. (1999) *Mol. Biol. Evol. 16*, 173−179.
35. Dimmic, M. W., Mindell, D. P., and Goldstein, R. A. (2000) *Pac. Symp. Biocomput.*, 18−29.
36. Soyer, O., Dimmic, M. W., Neubig, R. R., and Goldstein, R. A. (2002) *Pac. Symp. Biocomput.*, 625−636.
37. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) *CABIOS, Comput. Appl. Biosci. 8*, 275−282.
38. Eck, R. V., and Dayhoff, M. O. (1966) Atlas of Protein Sequences and Structure. National Biomedical Research Foundation, Silver Springs, Maryland.
39. Goldman, N., Thorne, J. L., and Jones, D. T. (1998) *Genetics 149*, 445−458.
40. Lio, P., and Goldman, N. (1999) *Mol. Biol. Evol. 16*, 1696−1710.
41. Halpern, A. L., and Bruno, W. J. (1998) *Mol. Biol. Evol. 15*, 910−917.
42. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) *J. R. Stat. Soc., Ser. B 39*, 1−38.
43. Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F., and Vriend, G. (1998) *Nucleic Acids Res. 26*, 275−279.
44. Huelsenbeck, J. P., and Ronquist, F. (2001) *Bioinformatics 17*, 754−755.
45. Kawashima, S., Ogata, H., and Kanehisa, M. (1999) *Nucleic Acids Res. 27*, 368−369.
46. Fraczkiewicz, R., and Braun, W. (1998) *J. Comput. Chem. 19*, 319−333.
47. Akaike, H. (1973) in *Second International Symposium on Information Theory*, Budapest, Hungary.
48. Burnham, K. P., and Anderson, D. R. (1998) *Model Selection and Inference*, Springer-Verlag, New York.
49. Heijne, G. V. (1986) *EMBO J. 5*, 3021−3027.
50. Ballesteros, J. A., Shi, L., and Javitch, J. (2001) *Mol. Pharmacol. 60*, 1−19.
51. Wess, J. (1998) *Pharmacol. Ther. 80*, 231−264.
52. Shi, L., and Javitch, J. A. (2002) *Annu. Rev. Pharmacol. Toxicol. 42*, 437−467.
53. Lu, Z. L., Curtis, C. A., Jones, P. G., Pavia, J., and Hulme, E. C. (1997) *Mol. Pharmacol. 51*, 234−241.
54. Javitch, J. A., Ballesteros, J. A., Weinstein, H., and Chen, J. (1998) *Biochemistry 37*, 998−1006.
55. Shapiro, D. A., Kristiansen, K., Kroeze, W. K., and Roth, B. L. (2000) *Mol. Pharmacol. 58*, 877−886.
56. Senes, A., Belandia, U. I., and Engelman, D. M. (2001) *Proc. Natl. Acad. Sci. U.S.A. 98*, 9056−9061.
57. Fotiadis, D., Liang, Y., Filipek, S., Saperstein, D. A., Engel, A., and Palczewski, K. (2003) *Nature 421*, 127−128.
58. Gkoutos, G. V., Higgs, C., Bywater, R. P., Gouldson, P. R., and Reynolds, C. A. (1999) *Int. J. Quantum Chem. 74*, 371−379.
59. Govindarajan, S., Ness, J. E., Kim, S., Mundorff, E. C., Minshull, J., and Gustafsson, C. (2003) *J. Mol. Biol. 328*, 1061−1069.
60. Pollock, D. D., and Taylor, W. R. (1997) *Protein Eng. 10*, 647−657.
61. Yang, Z. (1995) *Genetics 139*, 993−1005.

BI035097R