Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs

Bin Qian[a], Orkun S. Soyer[b], Richard R. Neubig[c], Richard A. Goldstein[d,*]

[a] Biophysics Research Division, University of Michigan, Ann Arbor, MI 48105, USA
[b] Department of Chemistry, University of Michigan, Ann Arbor, MI 48105, USA
[c] Department of Pharmacology, University of Michigan, Ann Arbor, MI 48105, USA
[d] Division of Mathematical Biology, National Institute for Molecular Biology, The Ridgeway, Mill Hill NW7 1AA, UK

Accepted 24 September 2003

First published online 7 October 2003

Edited by Robert B. Russell

Abstract Related proteins with similar biological functions generally share common features, allowing us to extract the common sequence features. These common features enable us to build statistical models that can be used to classify proteins, to predict new members, and to study the sequence–function relationship of this protein function group. Although evolution underlies the basis of multiple sequence analysis methods, most methods ignore phylogenetic relationships and the evolutionary process in building these statistical models. Previously we have shown that a phylogenetic tree-based profile hidden Markov model (T-HMM) is superior in generating a profile for a group of similar proteins. In this study we used the method to generate common features of G protein-coupled receptors (GPCRs). The profile generated by T-HMM gives high accuracy in GPCR function classification, both by ligand and by coupled G protein. © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Signalling pathway; Phylogenetic analysis; G protein; Multiple sequence analysis; Ligand binding; Protein coupling

1. Introduction

G protein-coupled receptors (GPCRs) are a large superfamily of receptors, generally consisting of a bundle of seven transmembrane helices connected by six loops of varying lengths. Different GPCRs respond to a wide variety of different external stimuli and activate a number of different GTP binding proteins (G proteins), initiating a broad spectrum of intracellular responses [1]. GPCRs play important roles in cellular signalling networks involving such processes as neurotransmission, cellular metabolism, secretion, cellular differentiation and growth, inflammatory and immune responses, smell, taste and vision. This superfamily of proteins is of importance in understanding many human diseases, and has been proved to be one of the most attractive targets for pharmaceutical intervention. The breadth of physiological activities involving these receptors is due to the wide range of ligands that interact with receptors of this class, as well as the number of different G proteins to which they can be coupled. Although there are thousands of GPCR sequences available [2], only one high-resolution structure has been solved, that of bovine rhodopsin [3].

GPCRs have been divided into six principal classes generally based on sequence: class A (rhodopsin-like), class B (secretin-like), class C (metabotropic glutamate/pheromone), class D (fungal pheromone), class E (cAMP receptors), and the Frizzled/Smoothened class [2]. Each class is further divided into families based on their ligand specificity, with some families combined into larger groups based on closely related ligands. For example, the class A GPCRs include groups such as amine binders, peptide binders, prostanoid receptors, and olfactory receptors. The amine binding group, for instance, is formed by seven families (acetylcholine receptor, adrenoceptor, dopamine receptor, histamine receptor, serotonin receptor, octopamine receptor and trace amine receptor). The G proteins to which GPCRs couple are themselves divided into four categories (Gi/o, Gq/11, Gs, and G12) based on structure and function, of which the first three are the most abundant [4].

Characterizing the role of any GPCR involves the identification of both the activating ligand and the activated G protein. This is a difficult problem given the complicated nature of the interactions, as GPCRs that bind to the same ligand may couple to different G proteins and GPCRs that couple to the same G protein may bind to different ligands. There are considerable efforts made to predict the placement of new GPCRs into the appropriate ligand-based classification. Some of the work includes the support vector machine (SVM) method by Karchin et al. [5], a classifier based on principal chemical properties of primary amino acid sequences developed by Lapinsh et al. [6], the construction of a 'fingerprint' in the PRINTS database [7] by Attwood et al. [8], and Daeyaert's *n*-tuple method [9]. While it is not always simple to compare the accuracy of different methods due to various test sets used, the highest reported accuracy rate for these various approaches is 97.4% for GPCR group level classification by the method of Lapinsh et al.

Less has been done classifying GPCRs based on the associated G protein. Horn et al. [10] used correlated mutation analysis to study the basis of G protein specificity of GPCRs and found weak sequence signal which determines the specificity. Möller et al. [11] tried to predict the G protein coupling specificity of a GPCR using a pattern recognition algorithm. They achieved high prediction accuracy in their study, albeit with a small test set. Cao et al. [12] used a naive Bayes model

*Corresponding author. Fax: (44)-20-8816 2460.
E-mail address: richard.goldstein@nimr.mrc.ac.uk (R.A. Goldstein).

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)01112-8

to predict G protein coupling specificity of GPCRs and achieved a 72% accuracy rate in their study.

In this paper, we describe the construction of a phylogenetic tree-based hidden Markov model (T-HMM) using a method described by Qian et al. [13] and evaluate its performance for GPCR classification. Our classifier achieves an accuracy of 99.9% for ligand group-based classification (i.e. amine vs. peptide binding), and over 99% for ligand family-based classification. In addition, the G protein coupling specificity-based classification gives 83% accuracy over a large data set. Some applications of our model for functional inference are also discussed.

2. Materials and methods

2.1. Use of the T-HMM for sequence classification

Many methods for classifying a target protein are based on the development of statistical models of the various classes, followed by identifying which statistical model best describes the target protein. Of the various approaches that have been developed, profile HMMs have proved to be one of the more sensitive methods [14]. There are, however, a number of complications in implementing this approach [13]:

- The currently available sequences are biased towards specific taxonomic groups, requiring the use of one of a set of ad-hoc weighting procedures.
- The limited available data require the use of statistical approaches [15], which generally do not reflect the evolutionary relationships and processes behind the observed sequences.
- Most methods assume that each location in the set of aligned sequences evolved independently. The shared phylogenetic relationships between all of the locations in the proteins induce correlations that can confound the information present in the multiple sequences.
- Most approaches generate a single statistical model which represent the entire class. There may be a hierarchical ordering of the members of the class, based on the underlying evolutionary process. More information might be provided by constructing a set of models, each representing some part of this substructure.

We have developed methods for addressing these problems through the explicit modeling of the underlying phylogenetic relationships and the sequence changes that occurred during the evolutionary process [13], based on the T-HMM method of Mitchison and Durbin [16,17]. The strength of this method is that it naturally incorporates both site substitutions and indels (including of multiple amino acids) in a consistent framework.

The method of constructing a T-HMM is discussed in more depth elsewhere [13]. Briefly, given a set of aligned homologous sequences and a corresponding phylogenetic tree, we can calculate the probability of the various amino acids in the ancestral proteins [18]. These posterior probabilities can be viewed as an amino acid profile of the protein family at that node. Profiles at different nodes can be viewed as common features of the protein family at different evolution stages. All combined, they can serve as a description of various subclasses of the protein family. In particular, we can construct a profile HMM at each node of the phylogenetic tree where the emission probabilities are given by the posterior probabilities resulting from the phylogenetic reconstruction of ancestral nodes. Using the T-HMM approach, we can similarly use a probabilistic reconstruction of the path of the ancestral sequence through the HMM to calculate the transition probability between various states of the HMM. We can thus calculate an HMM for every node in the phylogenetic tree. The score can be calculated by forward algorithm or Viterbi algorithm [19]. As we do not know where the putative new member of the family will be on the existing tree, we can compare each target sequence with the HMM representing every node, and use the highest score as the classification score. If we have derived a number of T-HMM models, each of which corresponds to one of the known protein classes, we can calculate the score of a query sequence with each T-HMM, and assign the query sequence to the class that provides the highest score.

2.2. Ligand-based classification

Both extracellular loop regions and transmembrane helices are important in determining a GPCR's ligand binding specificity [20]. This result suggests that best classification power will be achieved when the whole GPCR sequences are used for the ligand-based classification.

The score of a random protein in the HMM will depend upon the length of the protein, the composition of the protein, and various regularities and periodicities in the sequence. In order to account for these effects in the statistical analysis, we use the reverse HMM null model introduced by Karplus et al. [21]. In a reverse HMM null model, an alternative HMM is constructed, the reverse of the model HMM. The score of the protein sequence in the reversed model is then used to represent a random sequence of the same length, composition, and periodic structure. The logarithm of the ratio of the probabilities that the sequence aligned with a model and the corresponding null model is used as the score; Score = log [p(sequence | HMM)/p(sequence | reverse HMM)]. Conceptually, this score measures how much more likely it is that a sequence matches the HMM compared to a random event.

Our analysis is based on the GPCRDB (March 2002 release, http://www.gpcr.org/7tm/) [2] database, a collection of all the currently known GPCR sequences, along with putative sequences and fragments. All putative/orphan sequences and fragments are excluded from our data set, as are all ligand-based families with fewer than six sequences. This results in 1749 sequences and 57 families. Half of each family is chosen at random to form group I, while the remaining sequences form group II.

In group I and II, respectively, a multiple sequence alignment is constructed for each family using T-Coffee [22]. Then PAML (phylogenetic analysis by maximum likelihood) [23] is used to construct the phylogenetic tree for each family. Finally T-HMMs are constructed. Each of the sequences in group II is compared in turn with the 57 models derived from the group I sequence sets, aligned to each model using Viterbi algorithm [19] and assigned to the model that yields the highest score. Similarly, each of the sequences in group I is assigned to one of the 57 families based on the models derived from group II sequence sets.

2.3. G protein coupling-based classification

Much less is known about the basis for G protein coupling specificity. From the currently known GPCR coupling specificity reported by the latest *Trends in Pharmacological Science* Nomenclature Supplement [24], we generated a database of 470 GPCRs with an identified G protein coupling specificity. Each GPCR in the database is assigned to one of the three G protein coupling categories which correspond to binding to the three most abundant classes of G proteins: Gio, Gs and Gq11. Known dual coupling and triple coupling cases are excluded from this study.

Experimental studies have shown that G proteins bind to GPCRs mainly through the intracellular loop regions which connect the seven transmembrane helices. In addition, work by Möller et al. [11] and Cao et al. [12] suggests that the intracellular loop regions are sufficient to define the G protein coupling specificity. For these reasons, we use the intracellular parts of the GPCR sequences to build our models for G protein coupling classification,

Each of the 470 GPCR sequences is submitted to the TMHMM server [25] which predicts the location of transmembrane helices in a given GPCR. Of these, 418 GPCRs are predicted by TMHMM as having seven transmembrane helices. In these 418 sequences, there are different copies of the same protein from different species. We eliminate those repetitive sequences from our data set based on the Swiss-Prot protein ID. This procedure left us with 95 protein sequences. Among them, 54 belong to the Gio coupling category, 27 belong to the Gq11 category, and 17 belong to the Gs category. From each of these 95 sequences, intracellular loop region 1 (i1, between helix 1 and helix 2), region 2 (i2, between helix 3 and helix 4), region 3 (i3, between helix 5 and helix 6), and the carboxy tail region (tail) are extracted based on the TMHMM prediction result.

As with the ligand specificity classification, the proteins in each category are divided randomly and evenly into two subsets, with one subset from each category combined to form group I and the remaining sequences forming group II. A multiple sequence alignment is constructed for each loop region (i1, i2, i3 and tail) in group I of every G protein coupling category. Phylogenetic trees and T-HMMs are built based on these alignments using the same methods as in the ligand-based classification case. From group I, we construct 12 T-HMMs corresponding to all combinations of the four protein regions with the three G protein specificities. Every loop or tail region identified in the group II receptors is compared in turn with the three corresponding T-HMMs using the Viterbi algorithm [19].

The results from using the tail region alone were not promising, so we first discarded the tail region. Adding the scores of the remaining elements in various combinations was attempted, with the highest accuracy achieved by summing the results for loops i2 and i3. A category score divided by the sum of all category scores gives the probability that a sequence belongs to a G protein category. Then the group II sequences are used as a model training set to compute a similar set of T-HMMs, and the group I sequences are used as test sequences to perform the same classification procedure.

3. Results

3.1. Ligand-based classification performance

For the ligand-based two-fold classification problem, of the 1749 total sequences, the T-HMM approach yielded a correct family prediction in all but 17, an error rate of less than 1%. Only two of the 1749 sequences (0.1%) were misclassified into the wrong ligand-based group. Both of these (Q9YHY2 and Q9YHY3) are classified in the GPCRDB as olfactory proteins yet predicted using the T-HMM approach to be serotonin receptors. Both are listed in the SwissProt database [26] as ‘putative’ odorant receptors, even though GPCRDB does not assign them to the ‘putative family’.

To ensure the consistency of the estimate of prediction accuracy, another random split of the 1749 sequences was performed and the two-fold classification was repeated. This time there were 18 misclassifications at the family level, and the same two misclassifications at the group level.

Note that in these comparisons, the models are built from half of the available sequences. We expect to observe better results if we use all the currently available data to build profiles and classify the newly discovered proteins.

A GPCR classification example is shown in Fig. 1. Each of the 83 adrenoceptor family GPCRs is aligned to each of the 57 T-HMM models. Each protein sequence renders 57 scores, one score from each model, shown as 57 data points in each column of the plot. The solid circles correspond to the scores for the adrenoceptor model. As shown in Fig. 1, all scores from the adrenoceptor model are higher than scores from the other models, indicating an accurate assignment of all the adrenoceptor sequences. As adrenoceptors belong to the group of amine binders, scores generated with models from other families belonging to the amine receptor group (open circles) are generally higher than those scores generated by families not belonging to the amine receptor group (squares).

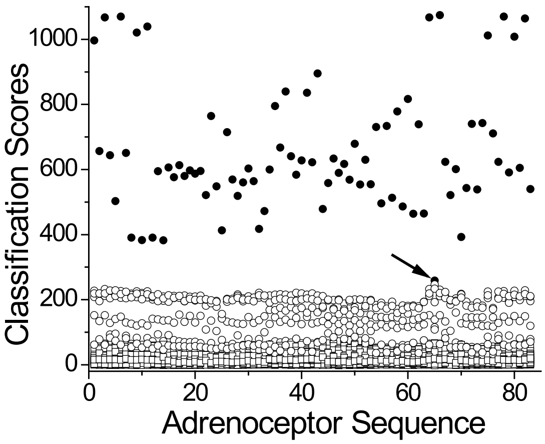

Fig. 1. Sequences belonging to the adrenoceptor family (members of the amine group) are fitted to all of the ligand-based family T-HMMs. Each column of points corresponds to a sequence scored on the 57 various models. Solid circles: scores generated by adrenoceptor family model; open circles: scores generated by amine receptor models other than the adrenoceptor model. Squares: scores generated by family models not belonging to the amine receptor group.

In Fig. 1, there is one protein (O96716, indicated by an arrow) which has an adrenoceptor score only slightly higher than the score generated from the dopamine receptor family model. The SwissProt database [26], however, indicates that this protein is a dual-specificity ‘dopamine D1/β-receptor’, binding both dopamine and adrenaline. For comparison purposes, we also submitted this sequence O96716 to the GPCR classification server by Kachin et al. [5]. Their server gave a score of 0.85 to adrenoceptor, and −0.71 to dopamine receptor, missing the dual specificity.

3.2. G protein coupling-based classification performance

We randomly split the 95 GPCRs used in our G protein coupling specificity classification experiment and perform a two-fold validation as described in Section 2. The best performance was achieved when we only combined scores for i2 and i3 loop regions (see Section 2): of the 95 total GPCRs tested, 13 misclassifications were observed. To ensure the consistency of the estimate of prediction accuracy, a second random split of the 95 sequences was performed and the two-fold classification was repeated, again using only the i2 and i3 scores. This time there were 19 misclassifications, yielding an average success rate of 83%.

The method of G protein coupling specificity prediction is based on a phylogenetic tree of the loop regions, which are difficult to align. Loop regions 1 and 2 can be aligned with relative confidence as they are rather short (approximately 15 residues), and in particular, loop 2 has a conserved DRY motif. Loop 3 is more problematic, and the tail region is particularly difficult. The success of the T-HMM method, even given the problems with generating accurate alignments and phylogenetic trees, indicates that the profiles are still able to represent the common features of those sequence groups. It is possible that difficulties in aligning the tail region may explain why including this region in the analysis did not improve the predictions, and this suggests that use of better alignments and phylogenetic trees could yield even higher accuracies in this and similar phylogenetically based predictive schemes.

Fig. 2 shows the separation of the three G protein coupling categories. The probability shown in the plot is the probability of a GPCR coupled by a specific G protein class. There are some GPCRs with much higher probability of interacting with one specific G protein than the other two, as shown by those data points close to the three corners of the triangle. More common, however, are the GPCRs with higher probability of interacting with one specific G protein, and still with substantial probabilities of interacting with the other two classes of G proteins. This observation agrees with the experimental results that GPCRs are generally capable of binding with more than one class of G proteins, and their preference for G proteins can be changed when the experimental conditions are changed [27], as well as suggesting that GPCRs with only one known G protein partner can have higher degrees of promiscuity. It may be the competitive nature of the binding

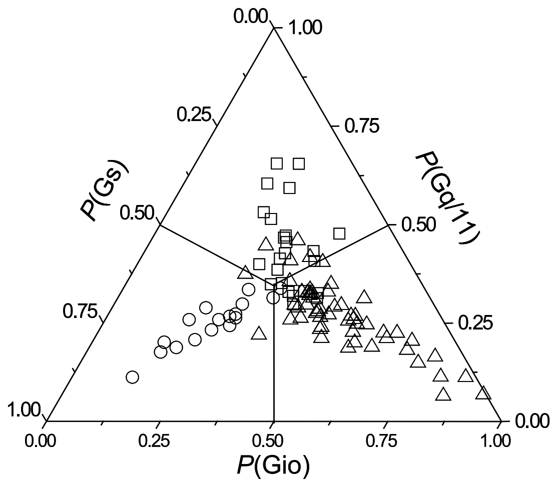

Fig. 2. Classification of the GPCRs based on G protein coupling specificity. The three coordinates show the probability of a GPCR interacting with a specific class of G protein. Data points show GPCRs that specifically couple to a class of G protein: Gs (circles), Gio (triangles), Gq11 (squares). Internal lines separate regions where receptors are assigned based on predicted coupling: Gs (lower left), Gio (lower right), Gq11 (top).

that determines the actual G protein coupling specificities of GPCRs.

3.3. G protein coupling specificity-based profile confirms crucial residues in the interaction

In addition to the prediction of the G protein coupling or ligand binding specificity of a newly found GPCR, our models can also be used to gain insight into the sequence–function relationship of GPCRs. For example, our models can be used to identify or confirm the crucial residues playing important roles in G protein–GPCR interaction. Blin et al. [28] analyzed a set of m2 and m3 type muscarinic receptors and found that four residues can determine the G protein specificity of these muscarinic receptors. They did an array of point mutations on these four residues and found it is possible to convert the Gi/o-coupled m2 type muscarinic receptor to a Gq/11-coupled receptor. We repeated their mutations in the m2 type muscarinic receptor and our classification result changed from Gio to Gq11, as observed in the biochemical experiments.

3.4. Comparison with predictions from phylogenomics

An alternative phylogenetic-based approach to protein classification is the use of a phylogenomics approach, as developed by Eisen [29]. In this method, a phylogenetic tree of the sequences is constructed. The assumption is then made that changes in function are parsimonious, that is, the best model is the one that minimizes the number of changes in function consistent with the phylogenetic tree. A standard parsimony analysis can then be performed on the tree to minimize such changes. With such an approach, there will be situations where a number of models are equally parsimonious, and no clear prediction can be made.

A phylogenetic tree of the training and test protein sequences was constructed using both the distance-based ClustalW method [30] and the parsimony-based MEGA method of Nei and co-workers [31]. Although the unambiguous predictions were of high accuracy (ligand classification accuracy 99.6% and 99.2% with the ClustalW and MEGA trees, respectively; G protein classification accuracy 80.6% and 64%, respectively), the percentage of cases where such unambiguous predictions could be made was not high for the ligand-based classification scheme (83.2% and 61.6%, respectively), and substantially lower for the prediction of coupled G protein (40.9% and 44.3%, respectively).

3.5. Annotation of the human GPCRs

With the accuracy demonstrated in the above analysis, our classifier can be used to annotate the ligand binding and G protein coupling specificity of GPCR sequences in protein databases. The International Protein Index (IPI) [32] combines all protein sequences from different resources, thus is a good starting point for our annotation effort. We first used every sequence in GPCRDB to run a BLAST search of the IPI human protein sequences, and saved all the resultant sequences with BLAST *E*-values smaller than 1.0 as putative GPCRs. Then we used the 57 GPCR ligand family models to classify these putative GPCRs. Since some of these putative GPCRs are not GPCRs, we needed a null statistic to measure the confidence that a sequence is indeed a GPCR.

We used the score distribution of all the non-members of the 57 families as a null statistic. If a sequence achieved a score larger than 95% of the null statistic scores, i.e. $P < 0.05$, then we considered this sequence a significant hit. We found 930 sequences from the IPI human protein database with $P < 0.05$, and 736 of them with $P < 0.01$. Most of these significant hits are known GPCRs. We then annotated the ligand binding and G protein coupling properties of these 930 sequences using our ligand-based and G protein coupling-based classifier. Not all sequences, however, can be classified with G protein coupling specificity, since the seven transmembrane regions of some of these sequences cannot be correctly identified. The annotated results can be obtained from http://mathbio.nimr.mrc.ac.uk/goldstein/GPCR/.

4. Discussion

In the pharmacological community, it is useful to classify GPCRs based on their predicted ligand binding properties [33]. Most sequence classification methods assume, at some level, that proteins that are more similar or more closely related are more likely to share functional attributes. This assumption holds reasonably well for the prediction of the GPCR ligand specificity. (This is not always the case; for example, most α2-adrenergic receptors show higher sequence similarity to dopamine type 2 receptors than to α1-adrenergic receptors.) Such agreement between evolutionary relationships and ligand specificity could explain the high accuracy of computational methods in ligand-based GPCR classifications as achieved in our study and others' [5,6]. There is much less agreement between G protein coupling specificity and sequence similarity, making the prediction of the coupled G protein a much more difficult problem, consistent with the generally lower accuracy achieved by G protein-based classification schemes, and in particular, the small percentage of unambiguous predictions that can be made based on a phylogenomics approach.

The underlying structural and physiological reasons for the G protein coupling specificity are still not clear. The specificity of a GPCR to different G proteins could change under different experimental conditions. Various experimental results show that GPCRs are generally capable of binding with different G proteins. Also, the posttranslational modification of GPCRs may play some role in G protein specificity [27]. All these factors further hinder the efficiency of G protein coupling specificity-based classification of GPCRs, and suggest that new possibilities for GPCR classification will emerge as more becomes known about the phenomenology and mechanisms of the binding specificity.

There are a number of different ways to undertake the task of protein classification. Based on whether or not we know the number of classes in our target protein group, these methods can be categorized into supervised and unsupervised classification [34]. In this GPCR classification case, since we know there are certain numbers of functional groups in our data, we are performing a supervised classification. Many different supervised classification methods have been used in protein classification, including Bayesian networks [12], SVMs [5], density estimation [9,6], etc. Our method can be regarded as along the lines of density estimation: parametric models of different protein families are built, then the probability of each sequence given any family model is calculated and these probabilities from different models are compared. The strength of our method lies in the way we extract the sequence features, by incorporating evolutionary information into the sequence profiles.

As mentioned in Section 1, there are many other interesting works associated with GPCR classification. Lapinsh et al. [6] used principal chemical properties of the primary amino acid sequences to construct a GPCR classifier. This method is promising because the explicit use of chemical/physical properties of an amino acid sequence can probably reveal more properties of a protein family, when the pure sequence-based methods fail to find some detailed information hiding beyond the discriminative power of the 20 characters. Lapinsh et al. were able to achieve a correct ligand group-based assignment of 521 GPCRs out of a test set of 535 for an error rate of 2.6%, although they did not attempt to apply their approach to the more difficult problem of ligand family classification. Attwood's PRINTS database incorporates expert knowledge of the differences between protein families, providing specific sequence fingerprints for individual GPCR families and for receptor subtypes [8]. This information can be used to search the protein database for new members of various GPCR families. Daeyaert's *n*-tuple method [9] considers only the amino acid composition of protein families and uses various statistics to adjust the discrimination power of the method. The prediction accuracies of these last two methods are, unfortunately, not evaluated. Karchin et al. [5] used a SVM to process the scores from SAM-T2K, an HMM-based iterative database searching method. The trained SVM is then used to classify GPCRs. They showed improved performance by SVM treatment. In principle, any HMM can potentially be improved by having its scores refined through a discriminant SVM analysis, suggesting that applying such a method to the T-HMM approach might yield further improvements.

Acknowledgements: We would like to thank Sarah Ingalls, Matt Dimmic and David States for helpful discussions, and Todd Raeker and Michael Kitson for computational assistance. Financial support was provided by NIH Grant LM0577, and NSF Equipment Grant BIR9512955.

References

[1] Iiri, T., Farfel, Z. and Bourne, H.R. (1998) Nature 394, 35–38.
[2] Horn, F., Mokrane, M., Beukers, M.W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F. and Vriend, G. (1998) Nucleic Acids Res. 26, 277–281.
[3] Palczewski, 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 2000, 739–745.
[4] Simon, M.I., Strathmann, M.P. and Gautam, N. (1991) Science 252, 802–808.
[5] Karchin, R., Karplus, K. and Haussler, D. (2002) Bioinformatics 18, 147–159.
[6] Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T. and Wikberg, J.E. (2002) Protein Sci. 11, 795–805.
[7] Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003) Nucleic Acids Res. 31, 400–402.
[8] Attwood, T.K., Croning, M.D. and Gaulton, A. (2002) Protein Eng. 15, 7–12.
[9] Daeyaert, F., Moereels, H. and Lewi, P.J. (1998) Comput. Methods Programs Biomed. 56, 221–233.
[10] Horn, F., van der Wenden, E.M., Oliveira, L., Izerman, A.P. and Vriend, G. (2000) Proteins 41, 448–459.
[11] Moller, S., Vilo, J. and Croning, M.D.R. (2001) Bioinformatics 17, 1–8.
[12] Cao, J., Panetta, R., Yue, S., Steyaert, A., Young-Bellido, M. and Ahmad, S. (2003) Bioinformatics 19, 234–240.
[13] Qian, B. and Goldstein, R.A. (2003) Proteins 52, 446–453.
[14] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) J. Mol. Biol. 284, 1201–1210.
[15] Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S. and Haussler, D. (1996) Comput. Appl. Biosci. 12, 327–345.
[16] Mitchison, G. and Durbin, R. (1995) J. Mol. Evol. 41, 1139–1151.
[17] Mitchison, G. (1999) J. Mol. Evol. 49, 11–22.
[18] Koshi, J.M. and Goldstein, R.A. (1996) J. Mol. Evol. 42, 413–420.
[19] Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.J. (Eds.) (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Amino Acids, Cambridge University Press, Cambridge.
[20] Sun, J., Ember, J.A., Chao, T.H., Fukuoka, Y., Ye, R.D. and Hugli, T.E. (1998) Protein Sci. 8, 2304–2311.
[21] Karplus, K., Barrett, C. and Hughey, R. (1998) Bioinformatics 14, 846–856.
[22] Notredame, C., Higgins, D. and Heringa, J. (2000) J. Mol. Biol. 302, 205–217.
[23] Yang, Z. (2002) Phylogenetic Analysis by Maximum Likelihood (PAML), version 3.12, http://abacus.gene.ucl.ac.uk/software/paml.html, March.
[24] Alexander, S., Peters, J. and Mathie, A. (2001) Trends Pharmacol. Sci., Nomenclature Supplement, 12th edn.
[25] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) J. Mol. Biol. 305, 567–580.
[26] Bairoch, A. and Apweiler, R. (2000) Nucleic Acids Res. 28, 45–48.
[27] Wess, J. (1998) Pharmacol. Ther. 80, 231–264.
[28] Blin, N., Yun, J. and Wess, J. (1995) J. Biol. Chem. 270, 17741–17748.
[29] Eisen, J.A. (1998) Genome Res. 8, 163–167.
[30] Thompson, J.D., Higgins, D.G. and Gibson, T.J.. (1994) Nucleic Acids Res. 22, 4673–4680.
[31] Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) Bioinformatics 17, 1244–1245.
[32] IPI (2003) The International Protein Index, http://www.ebi.ac.uk/IPI/, Feburary.
[33] Watson, S. and Arkinstall, S. (1994) G Protein Linked Receptor Facts Book, Academic Press, New York.
[34] Webb, A.R. (2002) Statistical Pattern Recognition, 2nd edn., John Wiley and Sons, Chichester.