DRUG DISCOVERY TODAY BIOSILICO

Predicting ligands for orphan GPCRs

Enoch S. Huang

G-protein-coupled receptors (GPCRs) represent not only one of the most successful target classes for the pharmaceutical industry, but also one of the largest and most structurally and functionally diverse. Many are 'orphan' GPCRs that have not yet been paired with their cognate ligands. Computational approaches are well suited for this type of classification problem, but most are confounded when the orphan is dissimilar to characterized GPCRs.

The seven-transmembrane (7TM) receptors are among the most abundant families of cell surface receptors, comprising over 800 genes in the human genome [1]. 7TM receptors typically transduce signals through interactions with heterotrimeric G-proteins in response to a large variety of stimuli (e.g. light, organic odorants, nucleotides, peptides, lipids and proteins) and thus are also called G-protein-coupled receptors (GPCRs). Because of their wide expression, historical tractability by small-molecule approaches, and central role in pathophysiology, GPCRs represent the most successful target class in terms of therapeutic benefit and potential sales [2].

Given the track record of GPCRs as validated drug targets, the vast number of potentially untapped targets within this superfamily presents an intriguing challenge for target selection. GPCRs fall into many subfamilies, traditionally defined by the endogenous ligand classes (e.g. aminergic). Of the ~400 non-odorant/sensory GPCRs, which are ignored from a drug discovery standpoint, over 100 are so-called 'orphan GPCRs' (oGPCRs) for which the identity of the natural ligand remains unknown. Pairing of these natural ligands with their cognate receptors is a key step in understanding the biological significance of the orphan GPCRs and their possible role in disease processes. Fortunately, a battery of computational methods has been successfully applied to facilitate the experimental discovery of ligand–receptor matches. In many cases, researchers in bioinformatics have designed algorithms and tools specifically to tackle this important problem of GPCR sub-classification. In this review, we survey the landscape of approaches that are available to group 7TM receptors into ligand-based subfamilies, focusing on the Class A or rhodopsin-like GPCRs, which is the largest and most diverse class of GPCRs.

Pairwise sequence searches

One can often correctly infer the function of an uncharacterized biological sequence (the 'query sequence') by detecting similarity with sequences with known functions. In practice, one takes the entire query sequence and aligns (compares) it against a database of reference sequences using tools such as the Basic Local Alignment Search Tool [3] (BLAST; http://www.ncbi.nlm.nih.gov/BLAST/). The strength of the match is judged by a score based on the similarity of two biological sequences after alignment, typically by using substitution matrices such as PAM [4]. This alignment score is commonly expressed as an 'E-value' which is the expected number of hits that would be expected to occur by chance under the same search criteria. Grundy [5] describes the results of a BLAST-based approach to classify genes into their respective families. In

Enoch S. Huang
Pfizer Research Technology Center,
620 Memorial Drive,
Cambridge,
MA 02139, USA
e-mail: enoch_huang@cambridge.pfizer.com

1359-6446/04/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(04)03349-5

essence, the family (or subfamily) membership of the uncharacterized sequence was set to the family membership of the sequence for which the BLAST E-value was lowest.

Sequence identity, which is the fraction of the pairwise alignment identical between the query and the reference sequence, is an alternative metric for judging the likelihood of two sequences being homologous. It is thought that when sequence identity between two aligned sequences falls below 20–25%, homology, and therefore shared function, can no longer be reliably inferred [6]. Moreover, the accuracy of machine-generated alignments decays as the intrinsic pairwise sequence similarity decreases [7], further complicating matters. Because receptors that share a natural ligand can have pairwise sequence identities below 25% (e.g. histamine receptors [8]), there are no hard cutoffs to positively associate an orphan to a known GPCR, much less a receptor that shares the same ligand. For example, using BLAST [3] to compare the protein sequence of human histamine H4 receptor to that of histamine H1 receptor yields 26% identity over the length of the match, whereas the sequence identity between the human somatostatin receptor type 1 with a receptor with different ligand (nociceptin receptor) is actually higher at 43%. Nevertheless, pairwise sequence comparison is a convenient approach and one that has certainly worked in the identification of putative GPCRs of unknown function and in some cases can provide strong clues as to the natural ligand as well (see [9], and references therein).

Profile methods

Profiles are statistical descriptions of the primary sequence consensus of a gene family derived from a multiple sequence alignment. Whereas pairwise alignment methods such as BLAST use position-independent scoring (i.e. in the case of proteins, the incremental value of aligning two given amino acids is the same irrespective of its location in the overall sequence), profile methods use position-specific scores for the placement of various amino acids. Profiles are commonly represented in a statistical model called a hidden Markov model (HMM) [10]. The HMM is able to estimate the probability that a query sequence was generated by the model itself. Like BLAST, one metric for evaluating a match between a query sequence and the HMM is also the E-value, corresponding to the number of hits that would be expected to have a score equal or better by chance alone.

In practice, one would gather members of each protein family (or subfamily) and train an HMM to represent the group. After building HMMs for all families of interest, one would then match the sequence with unknown function (the query) against them all, and assign membership to the family corresponding to the best E-value (or alternatively, the best score). The predictive power of the model is a function of several variables, such as the exact algorithm used to train the models, as well as the accuracy of any multiple sequence alignment used to guide the model training.

Examples of HMM-based classifiers for GPCRs abound. Shigeta *et al.* [11] describe a library of GPCR HMMs called GPCR-GRAPA-LIB (http://www.affymetrix.com/community/publications/affymetrix/index.affx). They use a custom graph-based scoring function to identify a distinct E-value cutoff suitable for each model in the library to determine positive hits. One can find another web-based HMM classifier for GPCRs at http://bioinformatics.biol.uoa.gr/PRED-GPCR/. This system, dubbed PRED-GPCR, is essentially a front-end to HMMer searches (http://hmmer.wustl.edu/) against a library of profile HMMs for various GPCR subfamilies, combined with a statistically valid method of combining the scores from multiple HMMs into a single subfamily E-value [12,13]. A rather different approach was taken by Qian *et al.* [14] who employed phylogenetic tree-based HMMs [15] to assign GPCRs with 99% accuracy in ligand family-based classification. The impressive performance of their method presumably stems from the way evolutionary information is incorporated into sequence profiles. Finally, Graul and Sadee [16] designed a database approach to sub-classify ~1700 GPCRs into clusters and express the evolutionary relationships within and between the various clusters using a battery of profiling tools, including, but not limited to, HMMs.

Support vector machines

Support Vector Machines (SVMs) are another type of statistical machine learning algorithm that has been successfully used to classify biological sequences [17]. Like HMMs, SVMs are trained from labeled (classified) data. Unlike HMMs, the SVM approach enables training on both positive and negative examples of family membership and hence is trained to discriminate examples from different classes.

Karchin *et al.* [18] applied SVMs to classify GPCRs by training classifiers that recognize ligand-based subfamilies against the rest of the family. Their SVMs, which are online at http://www.soe.ucsc.edu/research/compbio/gpcr-subclass/, demonstrate superior specificity for subfamily recognition compared with BLAST-based and simple HMM-based approaches. Nevertheless, the authors acknowledge that SVMs are more computationally expensive than the other two approaches and are less effective than profile HMMs at class (superfamily) discrimination [18].

Agglomerative clustering methods

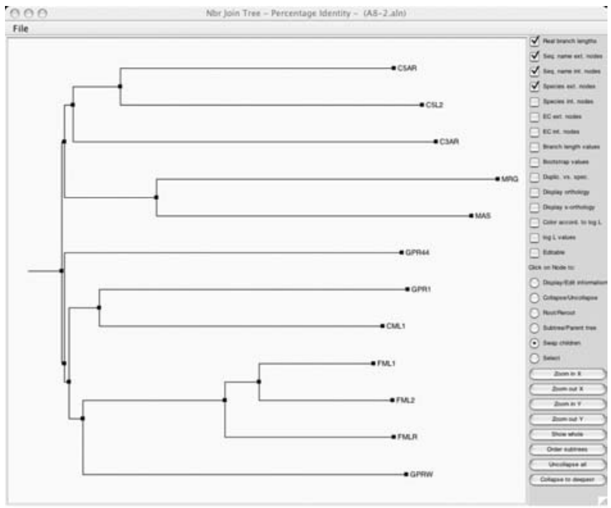
Starting with a set of distances between each pair of sequences, one can construct a phylogenetic tree [10], effectively forming clusters of related sequences. A simple method of clustering sequences begins by joining minimally distant pairs of sequences into clusters, then iteratively joining the minimally distant clusters by creating a new node in the tree.

Practically speaking, the most straightforward approach requires a pairwise distance measure and a formula for computing distances between established clusters. One convenient distance metric can be derived from the overall sequence similarity or identity computed after pairwise alignment. Ideally, all members of a given gene family (or subfamily) cluster in a subtree (beneath a single node) that does not contain members of a different subfamily. In a predictive context, one might construct a phylogenetic tree using a collection of GPCR sequences of known function and one orphan GPCR. Assuming that the method of tree building successfully clusters genes that share a given function, the cluster in which the orphan GPCR resides is predicted to share the same function as its neighbors. The performance of these agglomerative methods is sensitive to the distances inferred, which are dependent on the nature of both the scoring scheme and the quality of the underlying alignment.

Regarding GPCR classification, Joost and Methner [19] have completed a phylogenetic analysis of 277 receptors using ClustalX [20] for multiple alignment and the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) to generate neighbor-joining trees. Out of 19 subgroups, the authors found several that corresponded with those that are well-characterized, including chemokine receptors, bioamine receptors, and P2Y nucleotide receptors. However, the primary shortcoming of hierarchical clustering methods is the interpretation of clusters with mixed ligand labels, such as the 'A8' subtree (Figure 1). This intriguing group contains GPCRs activated by complement factor (proteins), N-formyl methionyl peptides, and GPR44, which is now known to be a second receptor for the lipid prostaglandin D2 [21]. Because of the startling ligand heterogeneity present in this cluster, it is difficult to assign with any confidence the class of ligand, much less the exact ligand, to the four orphans in the A8 group described by Joost and Methner [19]. Similar subgroupings were revealed by a more recent phylogenetic analysis of nearly 800 GPCRs, of which 241 were non-olfactory Class A (rhodopsin-like) receptors [1].

Motif-based classifiers

Subfamilies of GPCRs often have conserved residues that are characteristic for their ligand-based classifications. The patterns of conserved residues, also called motifs or signatures, can be employed as a specific classifier. Typically, motifs are derived by parsing multiple alignments into consensus sequences, the conserved columns of which reflect important structural and/or functional residues. Among the best-known resources include PROSITE [22], PRINTS [23] and BLOCKS [24], the first two of which also contain functional annotation. The strength of a motif-based approach is that by making the patterns very specific and dense, one can essentially eliminate false positives, although at the expense of sensitivity. Moreover, the nature of pattern matching against conserved (and presumably functional) residues is such that one can have intuitive confidence in the validity of database hits, which obviates the need to deal with arbitrary E-value cutoffs.



Drug Discovery Today

FIGURE 1

Cluster 'A8' in the study by Joost and Methner [19] contains 12 GPCRs, including 3 receptors for which no natural agonists have been found (GPR1, GPRW, MRG). Taking for example GPR44, whose natural ligand is prostaglandin D2 [21], it would have been difficult to assign a lipid ligand to this former oGPCR by examining the subtree in which it resides. This subtree only contains known peptidergic GPCRs (e.g. CML1 [40,41] and FMLR [42]) alongside other oGPCRs (e.g. GPR1, GPRW). The protein sequences were obtained from GenBank [43] (http://www.ncbi.nlm.nih.gov) and multiply aligned with ClustalW [20]. The neighbor-joining tree was built using distances derived from the percent sequence identity as implemented in Pfaat [44] (http://pfaat.sourceforge.net).

The aforementioned motif collections often contain fingerprints targeting GPCR subfamilies. For example, Attwood and colleagues [25,26] have compiled a set of ~250 conserved, ungapped fingerprints comprising more than 1000 motifs that effectively serve as a diagnostic resource for GPCRs at various levels of granularity. This searchable database (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/printscontents.html#Receptors) contains several advantages over strict regular expression matching. Because multiple motifs are associated with a GPCR subgroup fingerprint, it is possible to score the strength of the match based on partial matches. However, to provide increased flexibility and sensitivity, the use of E-values is required rather than simply returning the result of an all-or-none match. Finally, the hierarchical nature of the GPCR PRINTS database can highlight the differences between GPCR subtypes rather than be confounded by patterns common to GPCRs in general [9].

Positions involved in ligand binding can be inferred in an automated fashion by performing correlation analysis of positions in a multiple sequence alignment with endogenous ligand affinity [27]. Essentially, the technique recognizes functional residues that are conserved in one ligand-based subfamily but differ across subfamilies. Often, these residues have already been implicated in ligand binding by mutagenesis experiments. The combination of a multiple sequence alignment and a structural template also enables 'evolutionary tracing' methods to predict ligand-binding and other functional residues in GPCRs [28].

Functional, class-specific residues can also be expressed as regular expressions or sequence patterns, thereby conferring on them powerful diagnostic potential. For example, by specifying the identity of only two conserved residues that have been implicated by mutagenesis studies and molecular modeling to be interacting specifically to biogenic amines, one is able to separate all known aminergic GPCRs from the rest of the Class A family [29].

Alignment-independent approaches to GPCR classification

The approaches depicted above are ultimately dependent on the quality of a pairwise or multiple sequence alignment for their classification accuracy. This critical dependence reflects the underlying principle that ligand-binding function can be inferred through sequence similarity, whether it is represented as a profile, phylogenetic tree, motif or pairwise comparison. However, high-quality alignments are very challenging to build when the family is large, divergent and lacking in 3-dimensional structural information, as in the case of GPCRs with ~800 members, low average pairwise similarity of 15% [30], and a crystal structure for only one receptor, that of bovine rhodopsin [31]. In addition to the problem that alignment errors lead to erroneous predictions, homology-based approaches have at least three other serious shortcomings. First, unless the query oGPCR is similar to a reference GPCR with a trusted label or annotation, no useful prediction can be made [32]. Second, it is difficult to conceive a computational approach to pair oGPCRs with natural ligands (putative neuropeptides, for example) that have not yet been identified as possible GPCR agonists. Third, although a rare occurrence, known natural agonists can bind and activate GPCRs in different sequence-based subfamilies. Two interesting examples are prostaglandin D2, discussed earlier as the natural ligand for GPR44/CRTH2 [21], which is found outside the cluster of prostanoid receptors [19] where the original prostaglandin D2 receptor (DP) resides [33], and angiotensin-derived peptides, which activate a pair of very similar AT-2 receptors [34,35] as well as the distantly-related Mas receptor [19,36].

There are, however, classification methods that are alignment independent, which can address the issues related to alignment quality. For example, Inoue and colleagues [37] developed simple 'binary topology patterns' that represent GPCR loop lengths as bit strings, that is, 0 for short loops and 1 for long loops. The length thresholds for each loop were set to give maximum discrimination between functional groups. The authors claim classification accuracies of 94.5%, defined as the square root of the product of sensitivity and specificity, for 15 functional categories including some outside of the Class A clan. An altogether different approach was taken by Lapinsh *et al.* [38], who first translated the primary sequences into vectors of physico-chemical properties, and then applied partial least squares projection (PLS) analysis to classify unlabeled GPCRs. While the classification scheme was certainly effective on aligned sequences, the surprising result was that for unaligned sequences, the cross-validated correlation coefficient Q^2 was 0.895 for ligand-binding class labels in the training set, which was comparable to that of the model built from aligned sequences. The key step was to pre-process the property vectors using an innovative modification of autocross-covariance transformations with centering of the property scales and normalization, before the PLS analyses. Finally, there is the possibility of employing purely structure-based methods, which are attractive in large part because they can be engineered from first principles as opposed to relying on heuristics or training machine-learning algorithms on labeled data. The idea behind one structure-based approach is tantalizingly simple: given a three-dimensional structure of a GPCR, predict the cognate ligand by rank-ordering the docked conformations of a library of natural ligands by their respective energetics. Abagyan and co-workers suggest progress towards this ambitious goal [39] using the crystal structure of bovine rhodopsin [31] and its retinal ligand in a validation study. This encouraging result notwithstanding, one should not unadvisedly apply this strategy to other GPCR subtypes such as oGPCRs for which only homology models are available. Nevertheless, this approach theoretically circumvents all the limitations inherent to sequence-based approaches (listed above), assuming that a complete list of possible GPCR agonists could be generated.

Summary

Now available to the computational biologist is an array of tools effective in pairing orphan GPCRs with their cognate ligand. Most of these classifiers begin with sequence alignments to infer function through homology and/or functional residue identification and hence are dependent on alignment quality. None of the established methods can pair orphan GPCRs with ligands that have not yet been shown experimentally to be agonists, although they can classify by ligand type (peptide, bioamine, nucleotide, etc.). Nevertheless, continued progress in structural modeling, coupled with algorithms to predict or identify secreted proteins and peptides from genomic data, may lead to a breakthrough in the future.

References

1 Fredriksson, R. *et al.* (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* 63, 1256–1272

2 Wise, A. *et al.* (2002) Target validation of G-protein coupled receptors. *Drug Discov. Today* 7, 235–246

3 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

4 Dayhoff, M. *et al.* (1978) A model of evolutionary change in proteins. Matrices for

detecting distant relationships. In. *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–358, National Biomedical Research Foundation, Washington DC.

5 Grundy, W.N. (1998) Homology detection via family pairwise search. *J. Comput. Biol.* 5, 479–491

6 Doolittle, R. (1986) *Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences*. University Science Books.

7 Lesk, A.M. *et al.* (1986) Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng.* 1, 77–78

8 Oda, T. *et al.* (2000) Molecular cloning and characterization of a novel type of histamine receptor preferentially expressed in leukocytes. *J. Biol. Chem.* 275, 36781–36786

9 Gaulton, A. and Attwood, T.K. (2003) Bioinformatics approaches for the classification of G-protein-coupled receptors. *Curr. Opin. Pharmacol.* 3, 114–120

10 Durbin, R. *et al.* (eds.) (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ Press.

11 Shigeta, R. *et al.* (2003) GPCR-GRAPA-LIB–a refined library of hidden Markov Models for annotating GPCRs. *Bioinformatics* 19, 667–668

12 Papasaikas, P.K. *et al.* (2003) A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. *SAR QSAR Environ. Res.* 14, 413–420

13 Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14, 48–54

14 Qian, B. *et al.* (2003) Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett.* 554, 95–99

15 Qian, B. and Goldstein, R.A. (2003) Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins* 52, 446–453

16 Graul, R.C. and Sadee, W. (2001) Evolutionary relationships among G protein-coupled receptors using a clustered database approach. *AAPS PharmSci* 3, 1–18

17 Jaakkola, T. *et al.* (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.* 7, 95–114

18 Karchin, R. *et al.* (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18, 147–159

19 Joost, P. and Methner, A. (2002) Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biol.* 3, RESEARCH0063

20 Thompson, J.D. *et al.* (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882

21 Hirai, H. *et al.* (2001) Prostaglandin D2 selectively induces chemotaxis in T helper type 2 cells, eosinophils, and basophils via seven-transmembrane receptor CRTH2. *J. Exp. Med.* 193, 255–261

22 Falquet, L. *et al.* (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30, 235–238

23 Attwood, T.K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31, 400–402

24 Henikoff, J.G. *et al.* (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* 28, 228–230

25 Attwood, T.K. *et al.* (2002) Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. *Protein Eng.* 15, 7–12

26 Attwood, T.K. (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol. Sci.* 22, 162–165

27 Kuipers, W. *et al.* (1997) Identification of class-determining residues in G protein-coupled receptors by sequence analysis. *Receptors Channels* 5, 159–174

28 Madabushi, S. *et al.* (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.* 279, 8126–8132

29 Huang, E.S. (2003) Construction of a sequence motif characteristic of aminergic G protein-coupled receptors. *Protein Sci.* 12, 1360–1367

30 Horn, F. *et al.* (2000) G protein-coupled receptors, or the power of data. In *Genomics and Proteomics: Functional and Computational Aspects* (Suhai, S. ed.), pp. 191–214, Kluwer Academic/Plenum, New York, NY.

31 Palczewski, K. *et al.* (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289, 739–745

32 Fredriksson, R. *et al.* (2003) Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives. *FEBS Lett.* 554, 381–388

33 Boie, Y. *et al.* (1995) Molecular cloning and characterization of the human prostanoid DP receptor. *J. Biol. Chem.* 270, 18910–18916

34 Mauzy, C.A. *et al.* (1992) Cloning, expression, and characterization of a gene encoding the human angiotensin II type 1A receptor. *Biochem. Biophys. Res. Commun.* 186, 277–284

35 Martin, M.M. and Elton, T.S. (1995) The sequence and genomic organization of the human type 2 angiotensin II receptor. *Biochem. Biophys. Res. Commun.* 209, 554–562

36 Santos, R.A. *et al.* (2003) Angiotensin-(1-7) is an endogenous ligand for the G protein-coupled receptor Mas. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8258–8263

37 Inoue, Y. *et al.* (2004) Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Comput. Biol. Chem.* 28, 39–49

38 Lapinsh, M. *et al.* (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.* 11, 795–805

39 Cavasotto, C.N. *et al.* (2003) Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* 51, 423–433

40 Wittamer, V. *et al.* (2003) Specific recruitment of antigen-presenting cells by chemerin, a novel processed ligand from human inflammatory fluids. *J. Exp. Med.* 198, 977–985

41 Meder, W. *et al.* (2003) Characterization of human circulating TIG2 as a ligand for the orphan receptor ChemR23. *FEBS Lett.* 555, 495–499

42 Boulay, F. *et al.* (1990) Synthesis and use of a novel N-formyl peptide derivative to isolate a human N-formyl peptide receptor cDNA. *Biochem. Biophys. Res. Commun.* 168, 1103–1109

43 Benson, D.A. *et al.* (2004) GenBank: update. *Nucleic Acids Res.* 32, D23–D26

44 Johnson, J.M. *et al.* (2003) Protein family annotation in a multiple alignment viewer. *Bioinformatics* 19, 544–545

Related articles in other Elsevier journals

Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors
Flower, D. and Attwood, T.K. (2004) *Semin. Cell Dev. Biol.* 15, 693–701

Reverse pharmacology and the de-orphanization of 7TM receptors
Kotarsky, K. and Nilsson, N.E. (2004) *Drug Discov. Today: Technol,* 1, 99–104

Novel human G-protein-coupled receptors
Vanti, W.B. *et al.* (2003) *Biochem. Biophys. Res.* 305, 67–71