0026-895X/03/6306-1256–1272$7.00
MOLECULAR PHARMACOLOGY
Copyright © 2003 The American Society for Pharmacology and Experimental Therapeutics
Mol Pharmacol 63:1256–1272, 2003

Vol. 63, No. 6
2356/1068327
Printed in U.S.A.

The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints

ROBERT FREDRIKSSON, MALIN C. LAGERSTRÖM, LARS-GUSTAV LUNDIN, and HELGI B. SCHIÖTH

Department of Neuroscience, Uppsala University, Uppsala, Sweden (R.F., M.C.L., L.-G.L., H.B.S.); and Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden (R.F.)

Received December 23, 2002; accepted March 11, 2003

This article is available online at http://molpharm.aspetjournals.org

ABSTRACT

The superfamily of G-protein-coupled receptors (GPCRs) is very diverse in structure and function and its members are among the most pursued targets for drug development. We identified more than 800 human GPCR sequences and simultaneously analyzed 342 unique functional nonolfactory human GPCR sequences with phylogenetic analyses. Our results show, with high bootstrap support, five main families, named glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin, forming the GRAFS classification system. The rhodopsin family is the largest and forms four main groups with 13 subbranches. Positions of the GPCRs in chromosomal paralogons regions indicate the importance of tetraploidizations or local gene duplication events for their creation. We also searched for "fingerprint" motifs using Hidden Markov Models delineating the putative inter-relationship of the GRAFS families. We show several common structural features indicating that the human GPCRs in the GRAFS families share a common ancestor. This study represents the first overall map of the GPCRs in a single mammalian genome. Our novel approach of analyzing such large and diverse sequence sets may be useful for studies on GPCRs in other genomes and divergent protein families.

The superfamily of G-protein-coupled receptors (GPCRs) is one of the largest families of proteins in the mammalian genome (Lander et al., 2001; Venter et al., 2001). It has been estimated that more than half of all modern drugs are targeted at these receptors (Flower, 1999), and several ligands for GPCRs are found among the worldwide top-100-selling pharmaceutical products. It is also evident that drugs have still only been developed to affect a very small number of the GPCRs, and the potential for drug discovery within this field is enormous.

The ligands for the GPCRs have tremendous variation; ions, organic odorants, amines, peptides, proteins, lipids, nucleotides, and even photons are able to mediate their message through these proteins. The GPCR proteins are also highly variable. There are two main requirements for a protein to be classified as a GPCR. The first requirement relates to seven sequence stretches of about 25 to 35 consecutive residues that show a relatively high degree of calculated hydrophobicity. These sequences are believed to represent seven α-helices that span the plasma membrane in an counter-clockwise manner, forming a receptor, or a recognition and connection unit, enabling an extracellular ligand to exert a specific effect into the cell. The second principal requirement is the ability of the receptor to interact with a G-protein. There is a great diversity in the functional coupling of the GPCRs; they have a number of alternative signaling pathways, interacting directly with a number of other proteins. Interaction with G-proteins has not been demonstrated for most GPCRs, in particular for those whose genes have just recently been sequenced. It may therefore be more technically correct to term this superfamily "seven transmembrane (TM) receptors", but the GPCR terminology is more established.

Several classification systems have been used to sort out this superfamily. Some systems group the receptors by how their ligand binds, and others have used both physiological and structural features. One of the most frequently used systems uses clans (or classes) A, B, C, D, E, and F, and subclans are assigned using roman number nomenclature (Attwood and Findlay 1994; Kolakowski, 1994). This A–F system is designed to cover all GPCRs, in both vertebrates and invertebrates. Some families in the A–F system do not exist in humans. Examples of this are clans D and E, which represent fungal pheromone receptors and cAMP receptors, family IV in clan A, which is composed of invertebrate opsin receptors, and clan F, which contains archaebacterial opsins.

This work was supported by the Swedish Research Council (VR, medicine), the Swedish Society for Medical Research, Åke Wibergs Stiftelse, Svenska Läkaresällskapet, Petrus och Augusta Hedlunds Stiftelse, and Melacure Therapeutics AB, Uppsala, Sweden.

R.F. and M.C.L. contributed equally to this work.

ABBREVIATIONS: GPCR, G-protein-coupled receptor; TM, transmembrane; HMM, Hidden Markov Models.

The overall classification of the GPCRs has been hampered by the large sequence differences between mammalian and invertebrate GPCRs. The GPCRs in *Drosophila melanogaster* show in many cases little resemblance to those in mammals (Broeck, 2001). Certain species show also a high difference in the numbers of receptor genes in different classes. *Caenorhabditis elegans*, a worm, has, for example, developed a remarkable number of chemosensory (olfactory) GPCRs related to the creature's specific lifestyle. Those chemosensory receptors, as well as the olfactory receptors in *D melanogaster*, do not show any clear resemblance to the olfactory receptors in humans.

Gene duplication occurs both by individual duplication, which often leaves the new gene near the parent gene, and by block duplications involving chromosomal regions or entire chromosomes. Large-scale duplications, including polyploidizations, are believed to be an important mechanism of vertebrate evolution. Two rounds of large-scale duplications are thought to have occurred in early vertebrate ancestry (Lundin, 1993; Holland et al., 1994), resulting in up to four copies of each gene in mammals, which originate from a common ancestor gene in a cephalochordate. It is now known as the "2R hypothesis" or the "one-to-four model". This has led to the construction of maps that contain paralogous chromosomal regions, or *paralogons* (Lundin, 1993; Holland et al., 1994; Katsanis et al., 1996; Popovici et al., 2001), in vertebrates, which in combination with phylogenetic analysis can provide valuable information on gene relationships and origins.

In this study, we collected a large set of GPCR sequences in the human genome and performed multiple phylogenetic analyses. The first task was to compile a comprehensive data set with just a single copy of each gene. We wanted to avoid polymorphism, pseudogenes, duplicates (resulting from the same gene having multiple names), and other related problems. We identified more than 800 GPCRs in databases and simultaneously analyzed sequences of 342 unique functional nonolfactory human GPCRs and grouped them by phylogenetic analysis. The chromosomal localization and positioning in paralogous groups of the genes were studied to give insight into the mechanism involved in creating the receptor genes. The different families were also analyzed for common sequence motifs, and we discuss the evidence for common descent of the families.

Materials and Methods

Data Retrieval. Approximately 200 GPCRs, both orphans and characterized receptors, known from the literature were downloaded from the GenBank database using the Entrez data-retrieval tool (http://www.ncbi.nlm.nih.gov/Entrez/). This data set was considered the start set, and all the genes were manually searched against the human genome database using BLASTP (Altschul et al., 1997) on the protein database. New receptors that were not already in the data set were saved and included. At least 20 of the most significant BLAST hits (sorted by E-Value), for each receptor, were checked to further extend the data set obtaining the first crude database. Duplicates were removed from this data set using a crude phylogenetic analysis. Thereafter, Entrez was used in keyword searches to identify orphan receptors, which are usually named GPR*nnn*, where *nnn* is a number. In our case searches were made with *nnn* ranging from 1 to 150.

To extend the data set, searches were made with all receptor sequences in the data set against the human genome protein database at NCBI. All genes were screened against the first version of the database to avoid duplicates. To identify possible novel receptors, not yet annotated in the human genome database at NCBI, we searched with a diverse set of GPCR receptors at the nucleotide level using BLASTX against the Genescan data set. A P value of 0.001 was used as a threshold or a maximum of 100 BLAST hits were analyzed for each search.

The genes were named according to the convention used in the human genome database at NCBI, although several orphan GPCRs, which recently had their ligands identified, were subsequently renamed according to recent literature. If no name was assigned to a specific sequence in the database, these were assigned GPR numbers as provided by the HUGO nomenclature committee. Sequences not present in the human genome database were given either an accepted name from the literature or the GenBank accession number. Accurate chromosomal positions were obtained from the University of California Santa Cruz "the golden gate" human genome database (http://genome.ucsc.edu), the Dec 2001 assembly. If not present in the public genome assembly, we used the chromosomal position from the Celera database (http://www.celera.com).

Alignment. Each data set was randomized 20 times with regard to sequence input order using a program called Randfasta (http://www.neuro.uu.se/medfarm/schiothSoft.html), because the input order of sequences is known to affect the resulting alignment. These 20 data sets, containing the full set of sequences but in different order, were all aligned using the Win32 version of ClustalW 1.81 (Thompson et al., 1994). The default alignment parameters were applied.

Sequence Bootstrapping and Randomization. The 20 alignments were all bootstrapped 50 times using SEQBOOT from the Phylip package (Felsenstein, 1993) to obtain a total of 1000 different alignments from each dataset.

Neighbor-Joining Trees. Protein distances were calculated using Protdist from the Win32 version of the Phylip package. For the calculation, the Dayhof PAM matrix was used. The trees were calculated on the 20 different distance matrixes, previously generated with Protdist, using neighbor from the Phylip package, resulting in 20 files with 50 trees each. All trees were unrooted. Because of limitations in the Consense program (version 3.5; Felsenstein, 1993), a consensus tree for the complete rhodopsin family could not be calculated; therefore, 300 bootstrap replicas were used. The trees were plotted using Treeview (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Maximum Parsimony Trees. Maximum parsimony trees were calculated from the same input files that were used for Protdist using Protpars from the Phylip package. The trees were unrooted and calculated using ordinary parsimony, and the topologies was obtained using the built-in tree search procedure. As above, consensus trees were calculated using Consense 3.5 from Phylip and trees were plotted using Treeview.

Calculating the Overall Relationship of the Main GPCR Families Using Random Selection of Genes. These calculations are based on all members from four of the main groups: secretin, frizzled, glutamate, and adhesion, together with 20 randomly selected rhodopsin receptors, selected using Randfasta. Randfasta was used to randomize the input order of sequence 20 times. The 20 datasets were aligned, sampled using SEQBOOT (50 replicas each), and 1000 parsimony trees were calculated using Protpars and consensus trees were calculated using Consense 3.5.

Fingerprint Analysis. For the fingerprint/motif analyses an approach using Hidden Markov Models (HMM) was applied as implemented in the HMMR 2.1 package (Eddy, 1998), recompiled for WIN32 using Visual C++ 6.0. From the secretin, adhesion, glutamate, rhodopsin and frizzled families, alignments of the entire coding regions were constructed using ClustalW 1.81; from these alignments, one HMM per family was calculated using the HMMbuild. The model allowed local alignments within the HMM, global alignments with respect to the query sequence, and multiple domains per

sequence to hit. All HMMs were calibrated using HMMcalibrate. To define the transmembrane regions statistically described by the HMMs, the transmembrane region as described in the literature for one of the members of each family was aligned to the respective HMM using HMMsearch. The sequences used were FZD3, GRM1, GLP1, LEC1, and ADRB2. The identified TM regions from the HMMs were subsequently aligned to each other, region by region, using ClustalW 1.81, and conserved motifs were identified in the HMM alignments by manual inspection.

Results

The schematic presentation of the approach used for retrieving sequences and the overall phylogenetic analysis is

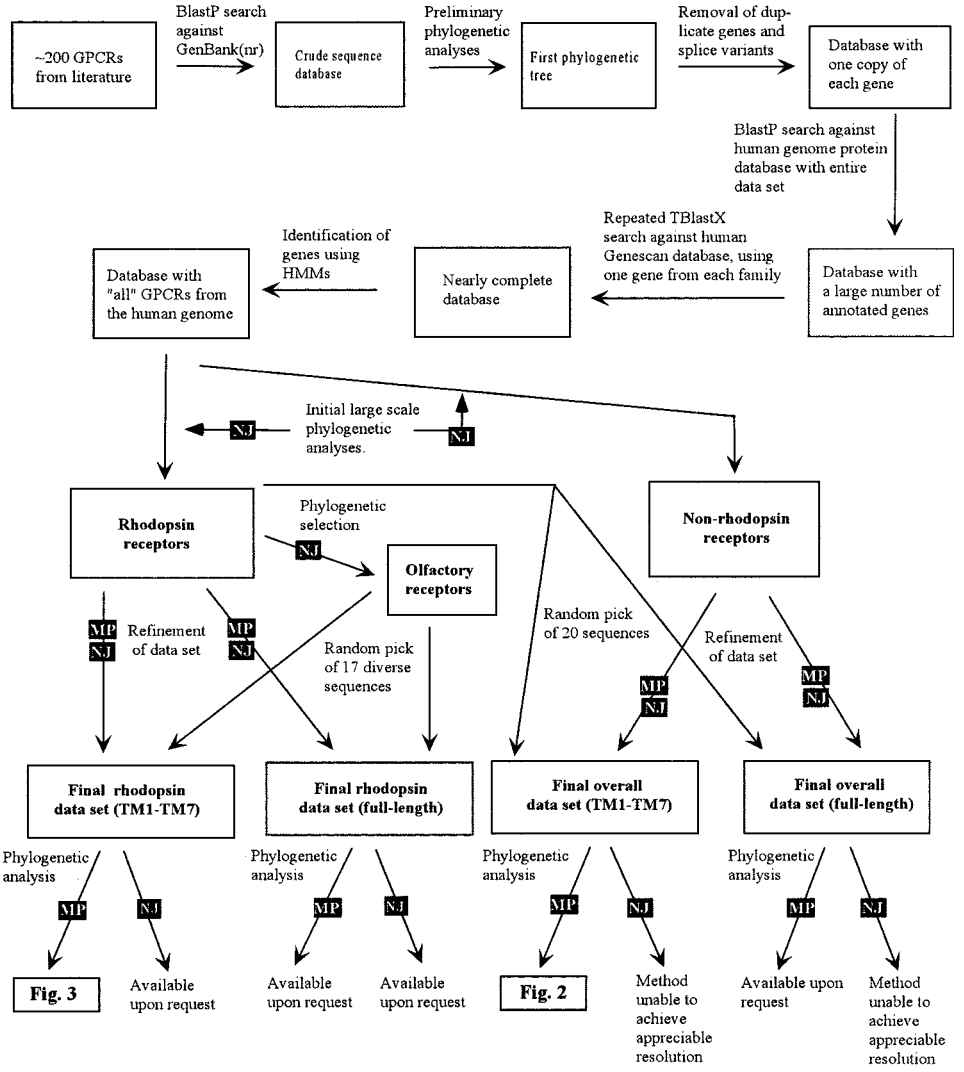

Fig. 1. Flowchart describing the sequence analysis strategy used in this work. The first step was to construct a database of GPCRs in the human genome. Using the Entrez online data retrieval tool and keyword searches, we downloaded approximately 200 GPCRs known from literature. Most GPCRs have several names, and they have also been deposited in database under several entries. Therefore, the database was carefully checked to remove any duplicated genes throughout the process. The approximately 200 human GPCRs were considered our "seeding" set, and the sequences were manually searched against the human genome database and the NR database to extend the GPCR database. Primary phylogenetic analyses, using a small number of bootstrap replicas and no randomization of the input file, were performed on the sequences in the database to identify splice variants, polymorphism, and duplicates. The final step was to search against the human genome GeneScan database, which contains genes predicted from the genome sequence by the GeneScan algorithm, to obtain possible nonannotated genes. The large phylogenetic analyses were carried out as described under *Materials and Methods*. Briefly, the Fasta file containing all sequences in the database was randomized using the Randfasta program to randomize the input order of the sequences, because the input order of the sequences can influence the resulting alignment. The sequences in these files were subsequently aligned and each of the sequence files was bootstrapped using the SEQBOOT software to obtain 1000 replicas of the alignment. Neighbor joining trees were constructed using the Protdist, Neighbor, and Consense programs. From this initial tree, the rhodopsin-like GPCRs and the nonrhodopsins were identified. The rhodopsin family was analyzed as one unit using the same strategy as above; from that analysis, the olfactory receptors and the four rhodopsin groups were identified. This analysis was carried out several times using both maximum parsimony and neighbor joining methods, and the groups that were finally defined were consensus groups from all these trees. A few receptors did not show stable topology in any group and these are discussed separately under Results. The nonrhodopsin receptors were analyzed both as full-length receptors and with the N- and C-termini removed, as shown in Fig. 2. To investigate how the rhodopsin family is related to the nonrhodopsins, 20 rhodopsins were randomly selected and included in the calculations. These analyses were repeatedly performed using the maximum parsimony method, with the dataset randomized as above, using Protpars and Consense. The four rhodopsin groups were also analyzed using maximum parsimony in the same way as described for the nonrhodopsin, but also using neighbor joining and maximum likelihood methods as described under *Materials and Methods*. These trees are not presented in this work but were used to identify instabilities in the topologies and are available upon request.

shown in Fig. 1. Detailed descriptions of the different steps are given under *Materials and Methods*. We assembled a primary data set of 802 unique GPCRs from the human genome. We believe that this data set contains most of the functional GPCRs in the human genome. The results show that the receptors cluster in five main families that we term glutamate (G, with 15 members), rhodopsin (R, 701), adhesion (A, 24), frizzled/taste2 (F, 24) (frequently abbreviated to frizzled hereafter), and secretin (S, 15), to which we apply the acronym GRAFS. Twenty-three protein sequences could not be assigned to any of the five families with appreciable bootstrap values (above 50%); these are discussed separately below under the section "other 7TM receptors". Figure 2 shows trees describing the overall relationship between the five main families of GPCRs. The bootstrap values shown in Fig. 2 separating the respective family from its closest neighbor [secretin (862), adhesion (789), glutamate (839), and frizzled (774)] are high; together with the overall topology, they give good support for each of the GRAFS families. The phylogenetic analysis shown in Fig. 2 is performed on protein sequences in which the N and C termini were deleted (see detailed comments below), whereas analysis on the full-length sequences also provided good support for five main families (data not shown). It should be noted here that the five families represent the smallest number of clusters that the phylogenetic analysis can delineate from the data set with appreciable bootstrap values and that the phylogenetic analysis does not show sufficient bootstrap support to link any of the GRAFS families together. It is possible, however, to further subdivide each family, because several bootstrap values within them show very high values. For example, the GABA receptors could be divided from the other receptors in the glutamate family, but because there are appreciable bootstrap values that link them within the glutamate family, we have decided to stick to this minimum number of families (i.e., five). The rhodopsin family has by far the largest number of receptors and was therefore further subdivided into four main groups and 13 branches (see below).

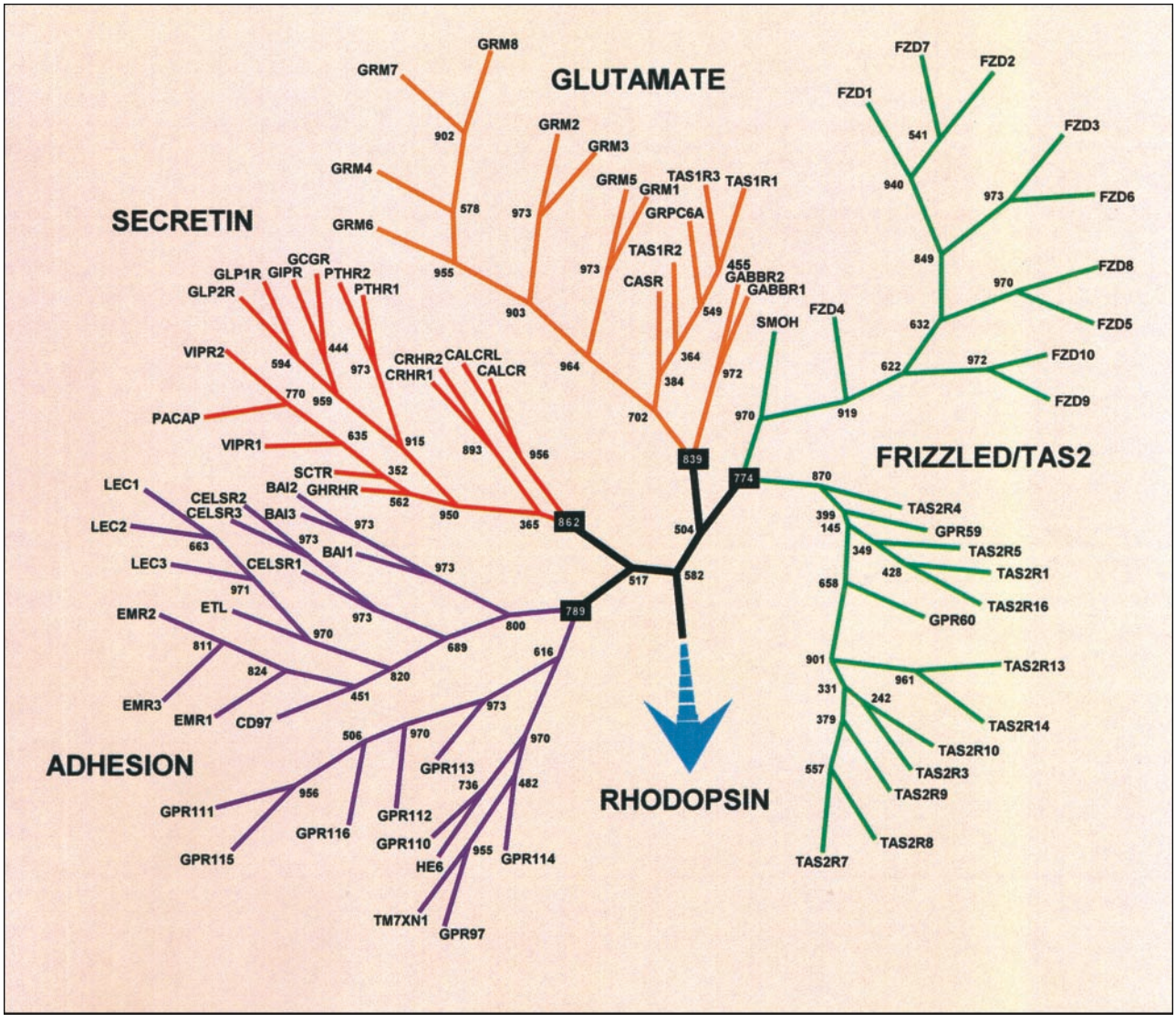

Fig. 2. Phylogenetic relationship between the GPCRs (TMI–TMVII) in the human genome. The tree was calculated using the maximum parsimony method on 1000 replicas of the data set terminally truncated GPCR as described under *Materials and Methods*. The position of the rhodopsin family was established by including twenty random receptors from the rhodopsin family. These branches were removed from the final figure and replaced by an arrow toward the rhodopsin family analysis in Fig. 3.

The receptors in all of the main families, except the rhodopsin family, have long N termini, whereas the rhodopsin family has only a few members with this characteristic. These long N termini are especially evident for the receptors within the adhesion family, but the secretin, glutamate, and frizzled receptors have also rather long N termini that are fairly rich in Cys residues. The only significant common feature of the proteins is the seven TM stretch; from an evolutionary perspective, it could be misguided to include these diverse and long N termini in the analysis. The number of evolutionary events needed for generating long N termini is likely to be more related to, for example, the number of domains than the replacements of single amino acids used for the phylogenetic calculations. Therefore, we decided to use the truncated receptors, where we use the sequence from the start of the TMI to the end of TMVII, for the main tree presented in Fig. 2. Each of the receptors was thus manually cut to provide this data set.

Below we give comments to our results for each of the families. The number of receptors in each family is indicated in parentheses. At the end of each section, we list the receptor names. First, we give the sequence identification name in bold. We provide the HUGO name in parenthesis in those cases in which it is different from the name we found to be most appropriate, for various reasons, except for the chemokine receptors (found in the rhodopsin family). HUGO lists only a few chemokine receptors, and the current naming system is thus not appropriate until it is more complete. We did not add their names in parenthesis, because we would have ended up with the same name for different receptors in our lists. After the name, we list the sequence accession code followed by the chromosomal position. We want the reader to be aware that many of the receptors have multiple additional names; a list with alternative names, which can be found online (http://www.neuro.uu.se/medfarm/schiothArt.html), includes many of the names provided by ENSEMBL (http://www.ensembl.org/).

The Secretin Receptor Family (15)

The receptors in the secretin family bind rather large peptides that share high amino acid identity and most often act in a paracrine manner. The secretin family name is related to the fact that the secretin receptor was the first one to be cloned in this family. The term "secretin-like receptor" has also frequently been used in the literature for receptors in this cluster. This group basically corresponds to clan B of the A-F system. The N terminus, between ~60 and 80 amino acids long, contains conserved Cys bridges and is particularly important for binding of the ligand to these receptors. The N terminus of the vasoactive intestinal peptide receptor (VIPR) and pituitary adenylyl cyclase-activating protein (PACAP) receptors alone constitutes a functional binding site for the ligand. Members of this family are the calcitonin receptor (CALCR), the corticotropin-releasing hormone receptors (CRHRs), the glucagon receptor (GCGR), the gastric inhibitory polypeptide receptor (GIPR), the glucagon-like peptide receptors (GLPRs), the growth hormone-releasing hormone receptor (GHRHR), PACAP, the parathyroid hormone receptors (PTHR), the secretin receptor (SCTR), and VIPR. The tree has four main subgroups: the CRHRs/CALCRLs, the PTHRs, GLPRs/GCGR/GIPR and the subgroup including secretin and four other receptors. Most of these receptors, 11 of 15, belong to the HOX paralogon, 2q/12q/17q/7/(3p) (see Fig. 4):

CALCR, NP_001733.1, 7q21.3; **CALCRL**, NP_005786.1, 2q21.1-q21.3; **CRHR1**, NP_004373.1, 17q21.31; **CRHR2**, NP_001874.1, 7p14.3; **GCGR**, NP_000151.1, 17q25.3; **GHRHR**, NP_000814.1, 7p14; **GIPR**, NP_000155.1, 19q13.3; **GLP1R**, NP_002053.1, 6p21.2; **GLP2R**, NP_004237.1, 17p11.2; **PACAP**, NP_001109.1, 7p14; **PTHR1**, NP_000307.1, 3p21.31; **PTHR2**, NP_005039.1, 2q33; **SCTR**, NP_002971.1, 2q14.1; **VIPR1**, NP_004615.1, 3p22.1; **VIPR2**, NP_003373.1, 7q36.3

The Adhesion Receptor Family (24)

This rather new and peculiar family of GPCRs consists of receptors with GPCR-like transmembrane-spanning regions fused together with one or several functional domains with adhesion-like motifs in the N terminus, such as EGF-like repeats, mucin-like regions, and conserved cysteine-rich motifs (for overview on the N termini in some of these receptors, see Hayflick, 2000; Harmar, 2001). The N termini are variable in length, from about 200 to 2800 amino acids long, and are often rich in glycosylation sites and proline residues, forming what has been described as mucin-like stalks. The family name "adhesion" relates to these long N termini, which contains motifs that are likely to participate in cell adhesion (McKnight and Gordon, 1998; Stacey et al., 2000). Some receptors in this family have been termed secretin-like receptors, and the latrotoxin receptors have previously been placed into clan B (Flower, 1999) or clan B2 (Harmar, 2001), but our analysis clearly shows that they belong to a distinct family of their own. The bootstrap values for the adhesion and the secretin families are also very high at 789 and 862, respectively, indicating clear distinction between the families. The analysis of the full-length proteins also indicates distinction between the secretin and adhesion families (data not shown). Although the phylogenetic analyses by Harmar (2001) does not stretch beyond "clan B" (secretin and adhesion), it basically supports our conclusion of separate clusters of secretin and adhesion receptors. Our analysis shows that several of the receptors appear in clusters of three or four; the CELSRs (EGF LAG seven-pass G-type receptors), the brain-specific angiogenesis-inhibitory receptors (BAIs), the lectomedin receptors (LECs) and the EGF-like module containing (EMRs). CD97 antigen receptor (CD97) and EGF-TMVII-latrophilin-related (ETL) also group with these on a separate main branch. CD97 share highest sequence similarity with EMR2 (56%), which is higher than the level of identity within the EMRs. The EMRs and CD97 are all positioned on 19p31, indicating that they may have arisen through several local gene duplications. The other main branch includes HE6 (TMVIILN2) and GPR56 (TMVIIXN1 or TMVIILN4) and a group of recently discovered receptors, related to GPR56 and HE6, named GPR97 and GPR110 to GPR116 (Fredriksson et al., 2002). The N termini of the receptors in this branch have varying lengths and relatively few identified functional domains compared with the other main branch of the adhesion receptors. Most of the genes of the entire adhesion family are positioned within the paralogon 1/5p-q21/6p21-p25/9/15q11-q26/19p providing support for their common ancestry (Fig. 4):

BAI1, NP_001693.1, 8q24; **BAI2**, NP_001694.1, 1p35; **BAI3**, NP_001695.1, 6q12; **CELSR1**, NP_055061.1, 22q13.3; **CELSR2**, NP_001399.1, 1p21; **CELSR3**, NP_001398.1, 3p21.31; **CD97**, NP_001775.1, 19p13.13; **EMR1**, NP_001965.1, 19p13.3; **EMR2**, NP_038475.1, 19p13.1; **EMR3**, NP_115960.1, 19p13.3; **ETL**,

NP_071442.1, 1p33-p32; **GPR97**, AY140959, 16q13; **GPR110**, AY140952, 6p12.3; **GPR111**, AY140953, 6p12.3; **GPR112**, AY140954, Xq26.3; **GPR113**, AY140955, 2p23.3; **GPR114**, AY140956, 16q13; **GPR115**, AY140957, 6p12.3; **GPR116**, AY140958, 6p12.3; **HE6 (GPR64)**, NP_005747.1, XP22.22; **LEC1**, NP_036434.1, 1p31.1; **LEC2**, NP_055736.1, 19p13.2; **LEC3**, NP_056051.1, 4q13.1; **GPR56 (TMVIIXN1)**, NP_003263.1, 1q42-q43

The Glutamate Receptor Family (15)

This family of receptors consists of eight metabotropic glutamate receptors (GRM), two GABA receptors (e.g., GABAbR1, which has two splice variants, a and b, and GABAbR2), a single calcium-sensing receptor (CASR), and five receptors that are believed to be taste receptors (TAS1). This group basically corresponds to what has been called clan C receptors. Several other GABA receptors are found in the human genome, but these are ion channels. The ligand recognition domain in the metabotropic glutamate is found in the N terminus of ~280 to 580 amino acids, and it has been proposed to share structural homology with bacterial amino acid binding proteins, such as LIVBP. The N terminus is believed to form two distinct lobes separated by a cavity in which glutamate binds, forming a so-called "Venus fly trap" where the glutamate causes the lobes to close around the ligand. The CASR also has a long cysteine-rich N terminus, but it is uncertain if it is involved in the binding of Ca^{2+}, even though it is important for mediating the signal of Ca^{2+}. The N-terminal of the GABA receptors is long and contains the ligand-binding site but lacks the cysteine-rich domain found in the other receptors of this family. The TAS1 receptors also have a long N terminus with a series of conserved Cys residues. They are expressed in the tongue and are likely to mediate taste signals. CASR falls with the TAS1 receptors, whereas the two GABA receptors branch basally in the family. GRM2 and GRM3 share 67% sequence identity and are located in chromosomal regions 3p and 7q, respectively. GRM7 and GRM8 share 74% sequence identity and are also positioned on 3p and 7q. These regions are both part of the postulated 1p/3p/7/22q paralogon, supporting a common ancestry (Fig. 4):

CASR, NP_000379.1, 3q21.1; **GABBR1**, NP_001461.1, 6p21.1; **GABBR2(GPR51)**, NP_005449.1, 9q22.1-q22.3; **GRM1**, NP_000829.1, 6q24.3; **GRM2**, NP_000830.1, 3p21.31; **GRM3**, NP_000831.1, 7q21.12; **GRM4**, NP_000832.1, 6p21.1; **GRM5**, NP_000833.1, 11q21.1; **GRM6**, NP_000834.1, 5q35.3; **GRM7**, NP_000835.1, 3p21.1; **GRM8**, NP_000836.1, 7q31.3-q32.1; **GPRC6A**, NP_683766.1, 6q22.1; **TAS1R1**, NP_619642, 1p36.23; **TAS1R2**, NP_689418.1, 1p36.2; **TAS1R3**, XP_060177.1, 1p36.33

The Frizzled/Taste2 Receptor Family (24)

This group includes two distinct clusters, the frizzled receptors and the TAS2 receptors. We were surprised that the TAS2 receptors clustered together with the frizzled receptors with a high bootstrap value. There are no obvious similarities between the receptors in the frizzed branch and the taste branch of this receptor family. However, when we compared the TAS2 receptors consensus sequence against an HMM model of the frizzled receptor branch, several features may explain why these two groups of receptors cluster together, such as consensus sequence of IFL in TMII, SFLL in TMV, and SxKTL in TMVII. None of these motifs is found in the consensus sequences of the other four families. The TAS2 receptors showed no clear similarities with the TAS1 receptors in the glutamate receptor family. The TAS2 receptors show clearly seven hydrophobic regions in a hydrophobicity plot but they have a very short N terminus that is unlikely to contain a ligand binding domain. Rather little is known about the role and function of the TAS2 receptors except that they are expressed in the tongue and palate epithelium, and it is believed that they function as bitter taste receptors. We found 13 TAS2 receptors in the human databases. Two of the receptors we found were not previously annotated or found in any database. We approached the HUGO Gene Nomenclature Committee at University College London and they confirmed that the sequences were unique and not public. The committee provided these receptors with new GPR numbers (GPR59 and GPR60). These numbers had previously been preliminarily assigned to other receptors but were never used, which explains the low GPR numbers.

The frizzled receptors control cell fate, proliferation, and polarity during metazoan development by mediating signals from secreted glycoproteins termed Wnt. The frizzled name was first used for a receptor cloned from *D melanogaster*, and the frizzled name (referring to the curled and twisted Wnt ligand) has frequently been used for this relatively recently discovered cluster of receptors. It has been shown that Wnt ligand binding to the rat F2DR can induce G-protein coupling (Slusarski et al., 1997), providing evidence that the frizzled proteins are GPCRs. This has also been supported by previous phylogenetic analyses showing some structural relationship to GPCRs (Barnes et al., 1998). The frizzled family of receptors have a 200-amino acid N terminus with conserved cysteines that are likely to participate in Wnt binding. The frizzled family consists of 10 frizzled receptors, FZD1–10, together with SMOH, which is the most divergent receptor of the family, sharing only 24% identity with FZD2 and less with the others. The topology of the tree shows four main clusters of the frizzled branch of receptors; the cluster containing FZD1, -2, and -7 share approximately 75% identity with each other, FZD8 and -5 share 70% identity, FZD 10, 9, and 4 share ~65% identity, and finally, FZD6 and -3 share 50% amino acid identity. The identities shared by receptors from different clusters are between 20 and 40%, indicating that four parental genes from the frizzled family were formed initially and the four clusters of receptors were subsequently formed out of these. All the frizzled genes, except FZD6, -3, and -8, are located in the chromosomal regions belonging to the HOX paralogy group. In addition, the phylogeny does indicate that the frizzled family was expanded in the two genome duplications proposed to have occurred basally in the vertebrate lineage (see Introduction). This is supported by the fact that the FZD7, -1, and -2 genes are located on different paralogous chromosomes, as are FZD9 and -10. However, if this scenario is true, several genes were lost (for example, all other copies of the SMOH gene). Interestingly, all the taste2 receptors from this group are located in the 1p3/3q/7q/12p/17p paralogon, indicating that some of these genes were present early in vertebrate evolution. The fact that the genes are clustered on chromosome 7q31 and 12p13 suggests that this family expanded through several local gene duplications. Noteworthy is that two of the frizzled receptors, FZD9 and SMOH, are also located in the same paralogon:

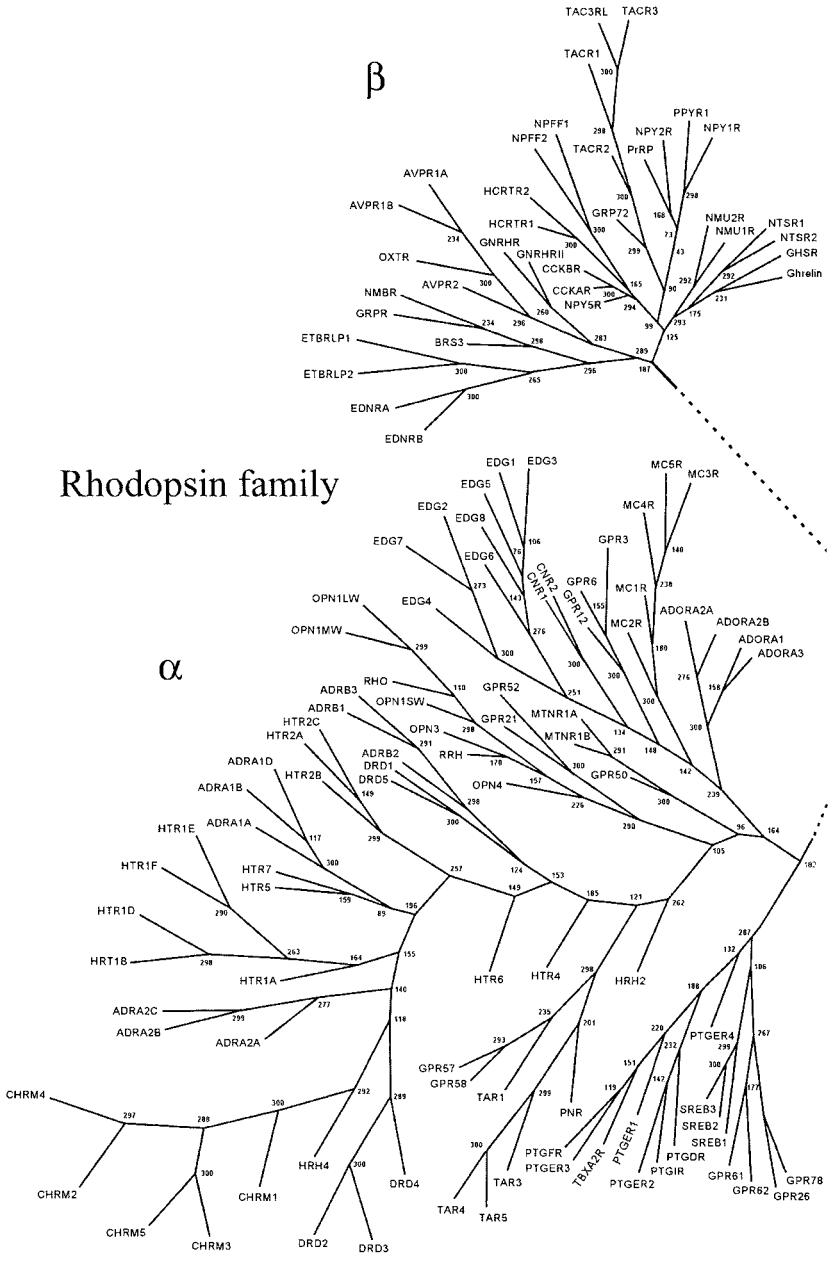

Fig. 3. The phylogenetic relationship between GPCRs (TMI–TMVII) in the human rhodopsin family. The tree was calculated using the maximum parsimony method on 300 replicas. The position of the olfactory cluster was established by including 17 diverse random receptors from the olfactory cluster. These branches were removed from the final figure and replaced by an arrow toward the olfactory receptor cluster.

FZD1, NP_003496.1, 7q21.13; **FZD2**, NP_001454.1, 17q21.31; **FZD3**, NP_059108.1, 8p21.1; **FZD4**, NP_036325.1, 11q14.2; **FZD5**, NP_003459.1, 2q33-q34; **FZD6**, NP_003497.1, 8q22.3-q23.1; **FZD7**, NP_003498.1, 2q33; **FZD8**, NP_114072.1, 10p11.21; **FZD9**, NP_003459.1, 7q11.23; **FZD10**, NP_009128.1, 12q24.33; **SMOH**, NP_005622.1, 7q32.1; **TAS2R13**, NP_076409, 12p13; **TAS2R14**, NP_076411.1, 12p13; **TAS2R7**, NP_076408.1, 12p13; **TAS2R9**, NP_076406.1, 12p13; **TAS2R8**, NP_76407.1, 12p13.2; **TAS2R3**, NP_058639.1, 7q31.3-q32; **TAS2R10**, NP_076410.1, 12p13; **TAS2R5**, NP_061853.1, 7q31.3-q32; **TAS2R4**, NP_058640.1, 7q31.3-q32; **TAS2R1**, NP_062545.1, 5p15; **TAS2R16**, NP_58641.1, 7q31.1-q31.3; **GPR59**, XP_069626, 7q33; **GPR60**, XP_090424, 7q33

The Rhodopsin Family (241 Nonolfactory, Total of 701)

The rhodopsin family has the largest number of receptors and overall analysis is shown in Fig. 3 (except the olfactory cluster; see comments below). The rhodopsin family corresponds to what has previously been called either the rhodopsin-like receptors or clan A in the A-F classification system. The rhodopsin family has several characteristics such as NSxxNPxxY motif in TMVII, the DRY motif or D(E)-R-Y(F) at the border between TMIII and IL2. Only a few receptors do not comply with these motifs, but these have other "fingerprint" elements that clearly link them to the rhodopsin family, apart from the phylogenetic analysis. The crystal structure of bovine rhodopsin has been revealed (Palczewski et al., 2000). Bovine rhodopsin has highest homology to rhodopsin (RHO) in the opsin receptor group. It should be noted that bacteriorhodopsin has no sequence similarity with the GPCR receptors in the human genome (Josefsson, 1999). The ligands for most of the rhodopsin receptors bind within a cavity between the TM regions (Baldwin, 1994). There are, however, important exceptions to this, in particular for the glycoprotein binding receptors (LH, FSH, TSH, and LG), where the ligand-binding domain is in the N terminus. Our analysis showed four main groups. We have opted to call

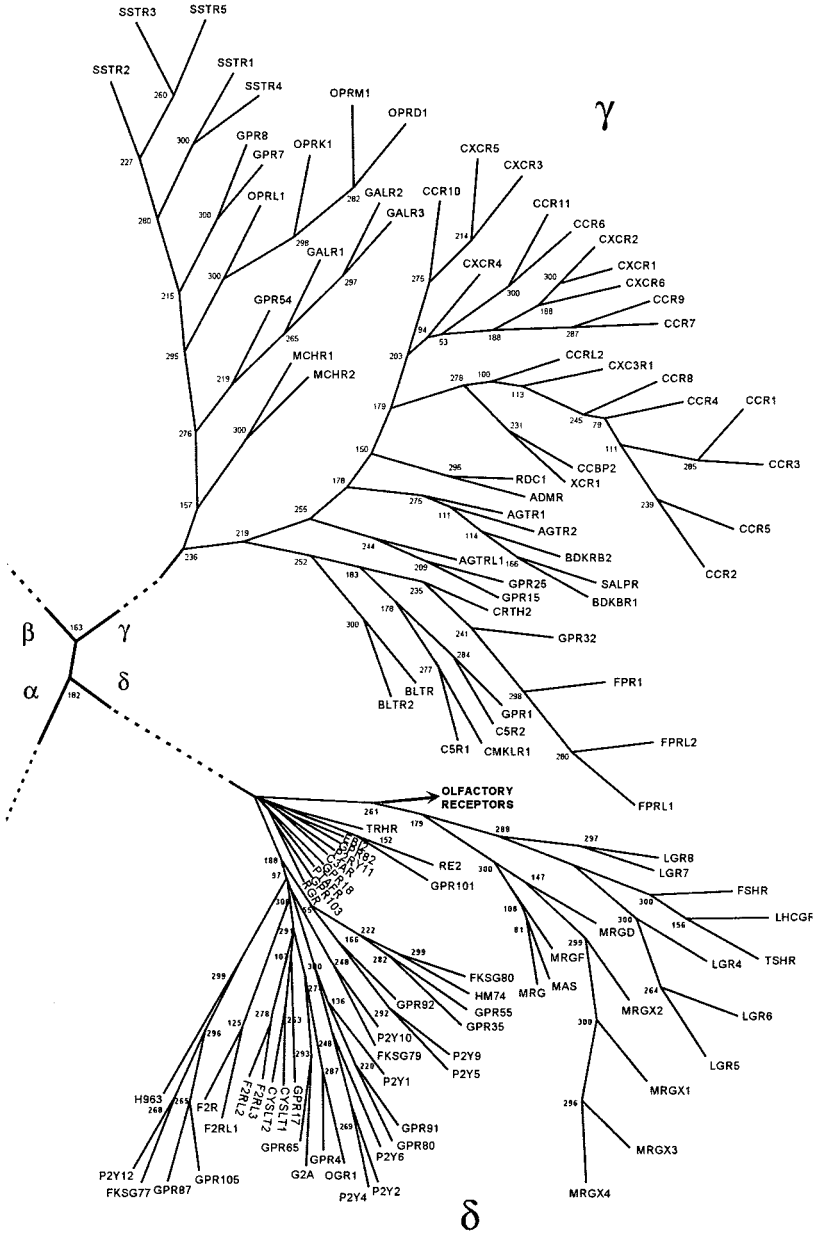

Fig. 3. continued.

these main groups α, β, γ, and δ. Results for each of the groups are described below.

The α-Group of Rhodopsin Receptors (89). This group has five main branches: the prostaglandin receptor cluster, amine receptor cluster, opsin receptors cluster, melatonin receptor cluster, and MECA receptor cluster. The bootstrap values that define these branches are very high (267, 262, 290, 299, and 239 of 300, respectively); these are highlighted in bold in Fig. 3.

The prostaglandin receptor cluster (15). This branch has eight prostaglandin receptors and seven orphan receptors. The prostaglandin receptors (PTGERs) are between 19 and 41% identical and share motifs in TMVII (IXDPW), and in the TMI (LXXTDXXG). The PTGERs, except PTGDR and PTGER4, belong to the paralogous regions on chromosomes 1/5p-q21/6p21-p25/9/15q11-q26/19p, further supporting the likelihood that the receptors in this group share a common evolutionary origin (Fig. 4). PTGDR and PTGER4 belong to the 1q23-q44/2p22-p25/11q13.1-q23.4/14q/15q11-q26/19q/20p paralogon:

TBXA2R, NP_001051.1, 19p13.3; **PTGER3**, NP_000948.1, 1p31; **PTGER2**, NP_000947.1, 1q22.1; **PTGDR**, XP_051711.1, 14q22.1; **PTGER4**, NP_000949.1, 5p12; **PTGIR**, NP_000951.1, 19q13.31; **PTGER1**, NP_000946.1, 19p13.12; **PTGFR**, NP_000950.1, 1p31.1; **SREB3**, NP_061842.1, Xp11; **GPR26**, XP_061555.1, 10q26.2; **SREB1(GPR27)**, NP_061844.1, 3p21-p14; **SREB2(GPR85)**, NP_061843.1, 7q31; **GPR61**, NP_114142, 1p13.3; **GPR62**, NT_005975.6, 3p21.31; **GPR78**, NT_006307.5, 4p16.1

The amine receptor cluster (40). The biogenic amine receptor group contains serotonin receptors (HTR), dopamine receptors (DRD), muscarinic receptors (CHRM), histamine receptors (HRH), adrenergic receptors (ADR), trace amine receptors (TAR), and several orphan receptors. All the known ligands of the receptors in this group are structurally related small amine molecules with a single aromatic ring. The degree of sequence conservation varies among the different

classes. The HTRs display a heterogeneous phylogenetic pattern. Two distinct subgroups can be seen, the HTR2s and HTR1B-1F. The rest of the HTRs branch separately or together with other biogenic amine receptors. These receptors are positioned near each other on chromosome 5q, suggesting early local gene duplication. The ADRs form three clusters in the phylogenetic tree, resulting in branches containing ADRA1, ADRA2, and ADRB, respectively. The three clusters could be a result of the postulated vertebrate genome duplications because the receptor genes, with a few exceptions, are positioned within the MetaHOX paralogon (Lundin, 1993; Coulier et al., 2000). This could explain why the sequence identities within the clusters are more than 45%, whereas the identities between the groups are about 25%. The TAR subgroup shares 37 to 82% sequence identity and the receptors are all positioned on chromosome 6q23, suggesting several early and late local gene duplications. This is evident also in rat, having 14 different TARs with high sequence identity, indicating an ongoing expansion of this gene family in mammals. Two orphan GPCRs, GPR57 and GPR58, share sequence similarities with the TARs. Several motifs, including RKAAKTLG in TMVI and FKQLHXPTN in TMI, together with the chromosomal data, strengthens their relationship to the TARs. CHRMs form the most homogenous cluster within the amine group, sharing between 40 and 50% identity. This can be seen in the tree with the receptors grouping together with strong bootstrap support. The DRDs appear in two clusters in the tree: with DRD2, DRD3, and DRD4 on one branch, placing DRD4 most basal, and DRD1 and DRD5 together with the β-adrenergic receptors. Identities within the dopamine clusters are 38 to 52% and 54%, respectively. The sequence identities between the clusters are ~27%, whereas ADRAB1 and DRD1 are 31% identical. The serotonin receptors are the largest group, with 13 members distributed more or less over the entire amine group tree, in general sharing low sequence identity, often as low as 20%:

HTR1A, NP_000515.1, 5q11.2-q13; **HTR5(HTR5A)**, NP_076917.1, 7q36.3; **HTR7**, NP_000863.1, 10q21-q24; **HRH2**, NP_071640.1, 5q35.2; **HTR4**, NP_000861.1, 5q31-q33; **HTR6**, NP_000862.1, 1p36-q35; **ADRA1A**, NP_000671.1, 8p21.2; **ADRA1D**, NP_000669.1, 20p13; **ADRA1B**, NP_000670.1, 5q33.1; **ADRB1**, NP_000675.1, 10q25.3; **ADRB3**, NP_000016.1, 8p12-p11.2; **ADRB2**, NP_000015.1, 5q32; **DRD5**, NP_000789.1, 4p16.1; **DRD1**, NP_000785.1, 5q35.2; **HTR2B**, NP_000858.1, 2q36.3-q37.1; **HTR2A**, NP_000612.1, 13q14-q21; **HTR2C**, NP_000859.1, Xq24; **TAR1**, AAK71236; 8q23.2; **PNR**, NP_003958.1, 6q23; **TAR3**, AAK71240; 6q23.2; **TAR4**, AAK71243; 6q23.2; **TAR5(GPR102)**, NP_444508.1, 6q23.2; **GPR58**, NP_055441.1, 6q24; **GPR57**, NP_055442.1, 6q23.2; **HTR1B**, NP_000854.1, 6q13; **HTR1D**, NP_008555.1, 1p36.3-p34.3; **HTR1E**, NP_000856.1, 6q14-q15; **HTR1F**, NP_000857.1, 3p12; **ADRA2B**, NP_000673.1, 3p13-q13; **ADRA2A**, NP_000672.1, 10q25.2; **ADRA2C**, NP_000674.1, 4p16; **DRD4**, NP_000788.1, 11p15.5; **DRD3**, NP_000787.1, 3q13.3; **DRD2**, NP_000786.1, 11q23; **HRH4**, NP_067830.1, 18q11.2; **CHRM4**, NP_000732.1, 11p12-p11.2; **CHRM2**, NP_000730.1, 7q31-q35; **CHRM1**, NP_000729.1, 11q13; **CHRM3**, NP_000731.1 1q43; **CHRM5**, NP_036257.1, 15q26

The opsins receptor cluster (9). This cluster of receptors comprises the rod visual pigment (RHO), the three cone visual pigments (OPN1SW, OPN1LW, OPN1MW), the peropsin (RRH), the encephalopsin (OPN3), the melanopsin (OPN4), and the retinal G-protein-coupled receptor (RGR). The opsins are the only GPCRs that are known to respond to light, and none of the receptors are known to bind any physical ligand. OPN1LW and OPN1MW are found in the same chromosomal position, Xq28. These two proteins are more than 96% identical, indicating, together with the fact that they are positioned near one another on Xq, that they share a recent common ancestor. Phylogenetic comparison of opsins in different species also indicates that the duplication is specific for mammals. The phylogenetic analysis divides the group into three branches; RHO/OPN1SW/OPN1LW/OPN1MW, RRH/RGR, and OPN3/OPN4. The chromosomal localization of these receptors is not consistent with any paralogy group, but it is worth noting that RGR and OPN4 are found in the same chromosomal position, 10q23:

GPR21, NP_005285.1, 9q33; **GPR52**, NP_005675.1, 1q24; **RHO**, NP_000530.1, 3q21-q24; **OPN1LW**, NP_064445.1, Xq28; CBP; **OPN1MW**, NP_000504.1, Xq28; **OPN1SW**, NP_001699.1, 7q31.3-q32; **RRH**, NP_006574.1, 4q; **OPN3**, NP_055137.1, 1q43; **OPN4**, NP_150598.1, 10q22

The melatonin receptor cluster (3). The analysis discerns two subgroups in this tree: the melatonin receptors (MTNR1A, MTNR1B) together with the orphan receptor GPR50. GPR50 has an extended C-terminal end compared with the MTNRs, whereas the other regions of the receptors most closely resemble MTNRs, especially in the third TM helix, which is almost identical. GPR50 and MTNR1A both belong to the ParaHOX paralogon (Fig. 4):

GPR50, NP_004215.1, Xq28; **MTNR1A**, NP_005949.1, 4q35.1; **MTNR1B**, NP_005950.1, 11q21-q22

The MECA receptor cluster (22). This group consists of the melanocortin receptors (MCRs), endothelial differentiation G-protein coupled receptors (EDGRs), cannabinoid receptors (CNRs), and adenosin binding receptors (ADORAs). Three orphan receptors also belong to this group (GPR-3, -6, and -12). It is interesting to note that the receptors in this group bind structurally different ligands; melanocyte stimulating hormone (13-residue peptide, MCRs); lysophosphatidic acid (lipid, EDGRs), and anandamide (arachidonylethanolamide, CNRs) and adenosine. The orphan receptors are 55% identical to each other and roughly 25% identical to the MCRs. The orphans share several motifs with the MCRs, such as PM(Y/F)X(F/L)X(C/G)SLAXADXL in TMIII, ALXY(H/Y) in TMIV, and PXIYAFR in TMVII. The CNRs share 39% identity to each other and their chromosomal positions indicate a common ancestor, because both genes are located in the paralogous group involving the positions 1p3 and 6q (Spring, 1997) (Fig. 4). GPR3 and GPR6 share the same chromosomal positions as the CNRs, which may indicate that these orphans share a common ancestor with the CNRs. The MCRs shares between 39 and 56% identity and belong to the 8q/16q/18/20q paralogon, supporting the idea that they share a common ancestor (Fig. 4). The EDG receptors form clusters at chromosome 1p, 9q, and 19p, suggesting two common ancestors together with one extra gene duplication at position 19p, resulting in two EDGRs at 1p and 9q, together with four EDGRs at chromosome 19p. These genes are all positioned in the paralogy group that was first proposed by Katsanis et al. (1996) and subsequently expanded by Popovici et al. (2001) 1/5p-q21/6p21-p25/9/15q11-q26/19p (Fig. 4). All the adenosine receptors except ADORA1 are located in the paralogy group 7/16p/17/22q (Fig. 4):

ADORA3, NP_000668.1, 1p13.3; **ADORA1**, NP_000671.1, 8p21.2; **ADORA2A**, NP_000666.1, 22q11.23; **ADORA2B**, NP_000667.1,

17q12; **GPR3**, NP_005272.1, 1p35.3; **GPR12**, NP_005279.1, 13q12.13; **GPR6**, NP_005275.1, 6q21; **MC2R**, NP_000520.1, 18p11.2; **MC1R**, NP_002377.1, 16q24.3; **MC3R**, NP_063941.1, 20q13.31; **MC4R**, NP_005903.1, 18q22; **MC5R**, NP_005904.1, 18p11.2; **EDG7**, NP_036284.1, 1p22.3; **EDG2**, NP_001392.1, 9q31.3; **EDG4**, NP_004711.1, 19p12; **EDG8**, NP_110387.1, 19p13.2; **EDG5**, NP_004221.1, 19p13.2; **EDG6**, NP_003766.1, 19p13.3; **EDG3**, NP_005217.1, 9q22.1; **EDG1**, NP_001391.1, 1p21; **CNR1**, NP_001831.1, 6q15; **CNR2**, NP_001832.1, 1p36.11

The β-Group of Rhodopsin Receptors (35). This group has no main branches and includes 36 receptors (Fig. 3). All the known ligands to these receptors are peptides. The group includes the hypocretin receptors (HCRTRs), the neuropeptide FF receptors (NPFFs), the tachykinin receptors (TACRs), the cholecystokinin receptors (CCKs), the neuropeptide Y receptors (NPYRs), the endothelin-related receptors (EDNR and ETBRLP1/2), gastrin-releasing peptide receptor (GRPR), the neuromedin B receptor (NMBR), the uterinbombesin receptor (BRS3), the neurotensin receptors (NTSRs), the growth hormone secretagogues receptor (GHSR), the neuromedin receptors (NMURs), the thyrotropin releasing hormone receptor (TRHR), the ghrelin receptor, arginine vasopressin receptors (AVPRs), the gonadotropin-releasing hormone receptors (GNRHRs), and the oxytocin receptor (OXTR) and orphan receptor.

The NPY5R groups with the CCK receptors rather than with the other NPY receptors. This might seem confusing, but it is consistent regardless of the method used (maximum parsimony, neighbor joining). One reason for this topology is that the NPY5R has a large third extracellular loop that is not present in the other NPYRs but is found in the CCK receptors. This feature might be the reason for this seemingly large difference between the NPY5R and the other NPY receptors. If the third extracellular loop of the NPY5R is removed, the NPY5R places on the same branch as NPY2R (data not shown). Surprisingly, the NPY2R has a higher identity to PrRP and GPR72 than to the other NPY receptors. The receptor GPR118 is 27% identical to GPR72 whereas the identity to the other receptors on that branch is below 20%. Several of these receptor clusters (i.e., NPY, NPFF, CCK, TACR) are positioned within the MetaHOX paralogon, consisting of chromosomes 4, 5q, 10q21-26, 8p12-22, and 2p11-23 (see Fig. 4). EDNRA and EDNRB are both positioned in the paraHOX paralogon; 4q/5q/13q/X (Fig. 4). This paralogon also includes BRS3:

AVPR2, NP_000045.1, Xq28; **AVPR1A**, NP_000697.1, 12q14.1; **AVPR1B**, NP_000698.1, 1q32; **EDNRB**, NP_000106.1, 13q22.3; **EDNRA**, NP_001948.1, 4q31.21; **ETBRLP1 (GPR37)**, NP_005293.1, 7q31; **ETBRLP2**, NP_004758, 1q31.3; **BRS3**, NP_001718.1, Xq21-q28; **CCKAR**, NP_000721.1, 4p15.1-p15.2; **CCKBR**, NP_000722.1, 11p15.4; **Ghrelin(GPR38)**, NP_001498.1, 13q14-q21; **GHSR**, NP_004113.2, 3q26.2; **GNRHR**, NP_000397.1, 4q21.2; **GNRHRII**, NP_476504.1, 1q12; **GRPR**, NP_005302.1, Xp22.1-p22.13; **HCRTR2**, NP_001517.1, 6p12.1; **HCRTR1**, NP_001516.1, 1p33; **NTSR1**, NP_002522.1, 20q13; **NTSR2**, NP_036476.1; **NMU2R**, NP_064552.1, 5q33.2; **NMU1R(GPR66)**, NP_006047.1, 2q37.1; **NMBR**, NP_002502.1, 6q24.1; **OXTR**, NP_000907.1, 3p25; **NPFF1**, NP_071429.1, 1q21-q22; **NPFF2(GPR74)**, NP_004876.1, 4q21; **TACR2**, NP_001048.1, 10q22.1; **TACR3**, NP_001050.1, 4q25; **TACR1**, NP_001049.1, 2p13.1; **TAC3RL**, NP_006670.1; **NPY5R**, NP_006165.1, 4q31-q32; **PPYR1**, NP_005963.1, 10q11.21; **NPY1R**, NP_000900.1, 4q31.3; **PrRP (GPR10)**, NP_004239.1, 10q25.3-q26; **GPR72**, NP_057624.1, 11q21; **NPY2R**, NP_000901.1, 4q31

The γ-Group of Rhodopsin Receptors (59). This group has three main branches: the SOG receptor cluster, MCH receptor cluster, and the chemochine receptors cluster. The bootstrap values that define these branches are high (276, 299, and 219, respectively) (Fig. 3).

The SOG receptor cluster (15). This cluster of receptors contains the GALRs that bind to the neuropeptide galanin and the RF-amide binding receptor GPR54, the somatostatin receptors (SSTRs), and the opioid receptors (OPRs). GPR7 and GPR8 have recently been shown to bind neuropeptide W. The known ligands to the receptor in this branch are thus all peptides but they themselves share no structural similarities.

Regarding the somatostatin receptors, we knew that SSTR1 and SSTR4 are more closely related to each other than to other SSTRs, whereas the relationship between the other SSTRs was uncertain. The relationship between SSTR1 and SSTR4 is strengthened by the fact that they share the same paralogous group, involving the chromosomal positions 20p and 14q (Fig. 4). The other three SSTRs belong to the paralogous regions consisting of chromosomes 7, 16p, 17, and 22q. GPR7 has the highest identity to GPR8 (60.4%). Their sequence identity to both SSTRs and OPRs is around 33%. It is intriguing to see that these orphans place at the same positions as the OPRK1 and OPRL1 at chromosomal position 8q11.23 and 20q13.33, respectively. This indicates that these orphans may indeed share an evolutionary origin with the OPRs. The OPRs share 49 to 59% identity, and are all part of the paralogous group consisting of 1p3, 2p, 8q, 6, 16q, 18, and 20q. The MCH1R and MCH2R have 32% identity to each other and 26% to the SSTRs. The structural motifs in TMI and TMVII are conserved in MCH1R, whereas only the motif in TMII is conserved in MCH2R, although several other common features of the group are represented within this receptor as well. The two GALR are positioned within the same paralogous group; 7/16p/17/22q. Motifs such as CCVPFXA in TMII and YLLP in TMV, together with a relatively high sequence identity to the GALR, strongly connect GPR54 to this cluster of GPCRs:

GPR54, NP_115940.1, 19p13.3; **GALR1**, NP_001471.1, 18q23; **GALR2**, NP_003848.1, 17q25.3; **GALR3**, NP_003605.1, 22q13.1; **GPR8**, NP_000836.1, 7q31.3-q32.1; **GPR7**, NP_000835.1, 3p26.1; **OPRL1**, NP_000904.1, 20p13.3; **OPRD1**, NP_000902.1, 1p36.1-p34.3; **OPRM1**, NP_000905.1, 6q25.2; **OPRK1**, NP_000903.1, 8q11.23; **SSTR3**, NP_001042.1, 22q13.1; **SSTR5**, NP_001044.1, 16p13.3; **SSTR2**, NP_001041.1, 17q25.1; **SSTR1**, NP_061842.1, Xp11; **SSTR4**, NP_001043.1, 20p11.2

The MCH receptor cluster (2). Two receptors branch off the SOG cluster with very high bootstrap value. The ligand is the melanin-concentrating hormone (MCH), which is a cyclic neuropeptide of 19 amino acids that is involved in regulation of feeding behavior:

MCHR2, NP_115892.1, 6q16.2; **MCHR1 (GPR24)**, NP_005288.1, 22q13.2

The chemokine receptor cluster (42). This branch consists of the classic chemokines (CCRs, CXCRs), the angiotensin (AGTRs)/bradykinin (BDKRBs)-related receptors, and a large number of orphan GPCRs. Most of the ligands are peptides (chemokine, cystenyl-leukotriene, angiotensin, bradykinin). The topology of the tree and the fact that large numbers of these receptors appear in clusters on several chromosomes both point toward a common ancestral origin.

This could be a result of several local gene duplications or, in the case of receptors appearing in paralogous regions, genome duplications. A combination of these events might be the reason for the relatively diffuse phylogenetic topology of this group.

The AGTR1 and AGTR2 receptors position within the 3q/13q/11q14-q25/17p/19q/Xq paralogon (Fig. 4). The two BDKRBs are both positioned at 14q32.1, indicating possible local gene duplications. The genes for the receptors CCR1-5, CCR8, CCR9(GPR28), CCR11, CCRL2, CX3CR1, CCBP2, and XCR1 are all positioned on chromosome 3p2, indicating several local gene duplications. All the chemokine receptors, except CCR6, CXCR5, and CXCR3, belong to the HOX paralogon 2q/12q/17q/7/(3p) (Holland et al., 1994):

RDC1, NP_051522.1, 2q37.3; **AGTRL1**, NP_005152.1, 11q12.1; **GPR1**, NP_005270.1, 2q33.3; **CRTH2(GPR44)**, NP_004769.1, 11q12.2; **AGTR2**, NP_000677.1, Xq23; **ADMR**, NP_009195.1, 12q32.3; **AGTR1**, NP_000646.1, 3q24; **CCR7**, NP_001829.1, 17q21.2; **CCR6**, NP_004358.1, 6q27; **CXCR6**, NP_006555.1, 3p21; **CCR9**, NP_006632.2, 3p21.31; **CCR11**, NP_057641.1, 3p21.31; **CXCR4**, NP_003458.1, 2q21.3; **CCR8**, NP_005192.1, 3p22.2; **CCRL2**, NP_003956.1, 3p21.31; **CXC3R1**, NP_001328.1, 3p22.2; **CCR4**, NP_005499.1, 3p24; **CCR1**, NP_001286.1, 3p21.31; **CCR3**, NP_001828.1, 3p21.31; **CCR2**, NP_000639.1, 3p21.31; **CCR5**, NP_000570.1, 3p21.31; **XCR1(CCXCR1)**, NP_005274.1, 3p21.3; **CCBP2**, NP_001287.1, 3p21.31; **CXCR5**, NP_001707.1, 11q23.3; **CCR10(GPR2)**, NP_057687.1, 17q21.31; **CXCR3(GPR9)**, NP_001495.1, Xq13; **CXCR1(IL8RA)**, NP_000625.1, 2q35; **CXCR2(IL8RB)**, NP_001548.1, 2q35; **BDKRB1**, NP_000701.1, 14q32.2; **BDKRB2**, NP_000614.1, 14q32.2; **CMKLR1**, NP_004063.1, 12q23.3; **C5L2(GPR77)**, NP_060955.1, 19q13.3; **C5R1**, NP_001727.1, 19q13.32; **GPR32**, NP_001497.1, 19q13.3; **FPR1**, NP_002020.1, 19q14.4; **FPRL2**, NP_002021.1, 19q13.3; **FPRL1**, NP_001453.1, 19q13.3; **GPR25**, NP_005289.1, 1q32.1; **GPR15**, NP_005281.1, 3q12.1; **BLTR2**, NP_062813.1, 14q11.2; **BLTR(LTB4R)**, NP_000743.1, 14q11.2; **SALPR**, NP_057652.1, 5p15.1-p14

The δ-Group of Rhodopsin Receptors (58, Plus an Estimated 460 Olfactory). This group has four main branches: MAS-related receptor cluster, glycoprotein receptor cluster, purin receptor cluster, and the olfactory receptor cluster (not shown in Fig. 3).

The MAS-related receptor cluster (8). This group contains the MAS1 oncogene receptor (MAS) and the MAS-related receptors (MRGs and MRGXs). The MRGX family has high (over 65%) sequence identity. MRGD and MRGF share 30% identity with the MRGXs, whereas MAS has 25% to MRGXs. All the MRGX genes together with MRGF and MRGD are located on chromosome 11 and are likely to have arisen in several very recent gene duplications. MAS, MRG, and the hypothetical protein are all located on chromosome 6. In a recent publication, six novel genes, SNSR1–6, were presented (Lembo et al., 2002). We find that SNSR1–2 are 98% identical to MRGX3, SNSR3–4 share 98~99% identity to MRGX1, and SNSR5–6 are 98~99% identical to MRGX4. All the SNSRs are localized on the same chromosomal position as the respective MRGX. We have been unable to find the reported SNSRs, despite numerous searches in the public genome databases as well as in the Celera database. At present, we are not certain whether these receptors are identical or very similar to the MRGX receptors or if they are simply not present in the assemblies of the human genome, either because of errors or because of missing data. This could also be a result of polymorphisms in the different libraries used during the screening process:

MAS, NP_002368.1, 6q25.3; **MRGF**, AAH16964, 11q12.1; **MRGX2**, NP_473371.1, 11p15.1; **MRGX1**, NP_089843.1, 11p15.1; **MRGX4**, NP_473373.1, 11p15.1; **MRGX3**, NP_473372.1, 11p15.1; **MRGD**, XP_089955.1, 11q12.2; **MRG**, NP_443199.1, 6p21.1

The glycoprotein receptor cluster (8). This cluster of receptors contains the classic glycoprotein hormone receptors (FSHR, TSHR, and LHCGR) and the leucine-rich-repeat–containing G-protein-coupled receptors (LGRs). The phylogenetic tree clearly indicates the presence of three distinct subgroups within this tree: the relaxin binding LGR7–8, the orphans LGR4–6, and the glycoprotein hormone receptors. The sequence identity within these groups is high (54%, 37–52%, and 47–50%, respectively), but the sequence identity among the groups is low (only 15–22%). The LGR7–8 subgroup belongs to the paraHOX paralogon (Coulier et al., 2000) and the LGR4–6 group belongs to the 1/11/12 paralogon (Fig. 4). LHCGR and FSHR positions are in close proximity on chromosome 2, 2p16.3, indicating a possible translocation involving the TSHR gene to chromosome 14:

LGR8, NP_570718.1, 13q13.2; **LGR7**, NP_067647.1, 4q32; **LGR4(GPR48)**, NP_060960.1, 11p14.1; **LGR6**, XP_046692.1, 1q32.1; **LGR5(GPR49)**, NP_003658.1, 12q22-q23; **LHCGR**, NP_000224.1, 2p16.3; **FSHR**, NP_000136.1, 2p16.3; **TSHR**, NP_000360.1, 14q31.1

The purin receptor cluster (42). This branch consists of the formyl peptide receptors (FPRs), the nucleotide receptors (P2Ys), and a large number of orphan GPCRs. The known ligands include extracellular nucleotides for the purin receptors, leukotrienes, and trombins. The nucleotide-binding and related receptors have the most diffuse topology within this group. These receptors contain the nucleotide binding receptors (P2Ys), the formyl peptide binding receptors (FPRs), the thrombin receptors (F2Rs), the cysteinyl leukotriene receptors (CYSLTs), and orphan GPCRs. A proportion of this dispersed receptor group, (i.e., 19 of 38 of these receptors) belongs to the same paralogon: 3q/13q/11q14-q25/17p/19q/Xq (Fig. 4). The phylogenetic pattern suggests that many local gene duplications occurred before the proposed chromosomal duplications. This might explain why the phylogenetic relationship of these receptors is hard to resolve. This is because the receptors would then have appeared during a short period and evolved and diversified over a relatively long period, resulting in a diverse group of receptors without a clear sub-branching resolution. Of the remaining receptors, six are located on 1q, five on 14q, three on 5q, and two on 19p, where 1q, 5q, and 19p belong to the same paralogous group. The sequence identity is in general low (~20%), although several pairs of genes have higher mutual identity:

GPR18, NP_005283.1, 13q32; **PTAFR**, NP_000943.1, 1p36.11; **G2A**, NP_037477.1, 14q32.3; **EBI2**, NP_004942.1, 13q32.3; **P2Y11(P2RY11)**, NP_002557.1, 19p13.2; **GPR92**, NP_065133.1, 12p13.31; **C3AR(C3AR1)**, NP_004045.1, 12p13.31; **P2Y9(GPR23)**, NP_005287.1, Xq21.31; **P2Y5**, NP_005758.1, 13q14.2; **FKSG79**, NP_115942.1, Xq21.1; **P2Y10**, NP_055314.1, Xq21.1; **GPR17**, NP_005282.1, 2q14.3; **F2RL3**, NP_003941.1, 19p13.11; **F2RL2**, NP_004092.1, 5q13.1; **F2R**, NP_001983.1, 5q13.1; **F2RL1**, NP_005233.1, 5q13.1; **GPR87**, NP_076404.1, 3q25.1; **GPR105**, NP_055694.1, 3q25.1; **P2Y12**, NP_073625.1, 3q25.1; **FKSG77(GPR86, GPR94)**, NP_076403.1, 3q25.1; **CYSLT1**, NP_006630.1, Xq21.1; **CYSLT2**, NP_065110.1, 13q14.2; **GPR80(GPR99)**, XP_062888.1, 13q32.1; **GPR91**, NP_149039.1, 3q25.1; **P2Y6(P2RY6)**, NP_004145.1, 11q14.1; **P2Y1(P2RY1)**,

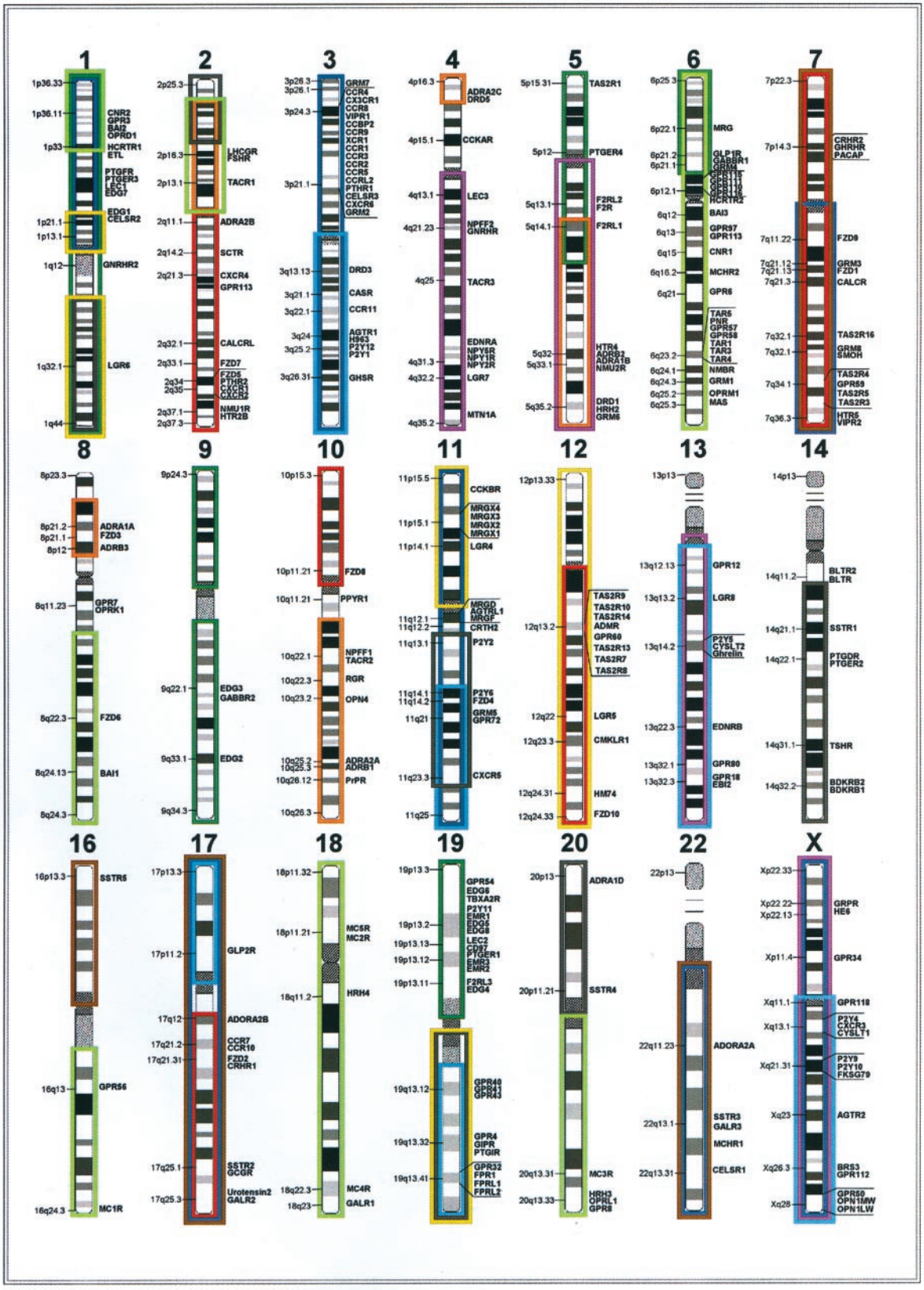

Fig. 4. The positioning of the GPCRs in paralogon groups in the human genome. Frames indicate the paralogons (PGs) according to Lundin (1993), Holland et al. (1994), Katsanis et al. (1996), Sidow (1996), Pebusque et al. (1998), Kasahara (1999), and Holland (1999), further extended in Popovici et al. (2001). Red, 2q/12q/17q/7/(3p) [PG 10 (HOX paralogon)]; dark blue, 1p/3p/7/22q (PG 11), light blue: 3q/13q/11q14-q25/17p/19q/Xq (PG 6/7); dark green, 1/5p-q21/6p21-p25/9/15q11-q26/19p (PG 3), light green: 1p3, 2p, 8q, 6, 16q, 18 and 20q (PG 13/14); orange, 4p16.3, 5q, 10q21-26, 8p12-22/2p11-23 (PG 9 [Meta HOX)]; yellow, 1p21.1-p13.1,1q1-q44/11p/12/19q (PG 1); purple, 4q/5q/13q/X [PG 8 (ParaHOX)]; brown, 7/16p/17/22q (PG 12); black, 1q23-q44/2p22-p25/11q13.1-q23.4/14q/15q11-q26/19q/20p (PG 4).

NP_002554.1, 3q25.2; **P2Y2(P2RY2)**, NP_002555.1, 11q13.1; **P2Y4(P2RY4)**, NP_002556.1, Xq13.1; **FKSG80(GPR81)**, NP_115943.1, 12q24.31; **HM74**, NP_006009.1, 12q24.31; **GPR35**, NP_005292.1, 2q37.3; **GPR55**, NP_005674.1, 2q37; **GPR65**, NP_003599.1, 14q31.3; **OGR1(GPR68)**, NP_003476.1, 14q31; **GPR4**, NP_005273.1, 19q13.3; **H963**, NP_037440.1, 3q25.1; **GPR82**, NP_543007.1, 1; **TRHR**, NP_003292.1, 8p23; **RE2**, NP_031395.1, 1p36.13-q31.3; **GPR103**, NT_006337.5, 4q26; **RGR**, NP_002912.1, 10q22.3; **GPR101**, NP_473362.1, Xq26.3

The olfactory receptor cluster (estimated at 460). Our searches and manual inspection of the resulting data files, looking at each of the genes individually, indicated that there are 460 olfactory receptors in the human genome that we consider likely to represent unique functional receptors (data not shown). Our phylogenetic analysis indicates that these proteins form a stable phylogenetic cluster, without spreading to other groups of the rhodopsin family or other families (data not shown). We do not show phylogenetic analyses of all these genes here because further work is needed to carefully match each of the sequences with expressed sequence tags, do comparative analysis of the NCBI and Celera databases, and annotate all these genes. We randomly picked 17 of these olfactory receptor sequences, one from each of the 17 main branches that formed in our preliminary phylogenetic analysis. This provided us with a diverse olfactory receptor data set that we used in the overall rhodopsin analysis to determine the olfactory node that appears in Fig. 3 in the δ-group in the rhodopsin family.

Three hundred forty-seven putative human full-length odorant receptor genes have previously been identified and physically cloned (Zozulya et al., 2001). It has also been suggested that there are more than 900 olfactory receptor-like sequences in the human genome (Venter et al., 2001). About 60% of these genes are estimated to be pseudogenes. Glusman et al. (2001) reported 322 odorant genes and a number of pseudogenes in the human genome. They also estimate that there were more than 900 olfactory receptor-like genes in the genome. The same number of 322 odorant genes was also reported by Takeda et al. (2002). The large clusters of olfactory receptors are found in paralogous regions distributed on 13 human chromosomes, further supporting the general observation that the human olfactory receptors share a common origin. Moreover, it is worth mentioning that the human olfactory receptors show low or little resemblance to chemosensory receptors in nematodes (Robertson 1998) or the fruit fly (Mombaerts, 1999).

Other 7TM Receptors (23)

Some of the 7TM genes could not be included in any family/group/cluster with appreciable bootstrap values. We have therefore chosen to present these receptors in this section as other 7TM receptors, although they clearly do not belong to the same group. The ligand for most of these receptors is not yet known. The instability in the topology is related to certain atypical parts of their sequences that could be a result of a chimeric origin of the receptors or of evolutionary pressure not shared by their closest phylogenetic neighbors. Most of these receptors give stable topology if they are analyzed with a limited number of sequences (for example, the 5–20 closest BLAST hits), but when analyzed in such a large and diverse data set, the atypical parts are more likely to cause an unstable topology. It is not uncommon in phylogenetic analysis to delete atypical parts from the proteins to avoid such "problems". We did not, however, perform any such manipulation to avoid unbiased handling of the data set. The atypical parts of the proteins are often found in the loops rather than the TM regions. An example of this is the histamine HRH1 and HRH3 receptors, which have a large third intracellular loop of about 170 amino acids, which is significantly longer than in most other rhodopsin family receptors of the α-group (where they obviously belong). When we analyze the amine receptor cluster alone, HRH1 and HRH3 show stable topology; in our large data set, however, they do not, which explains why they have ended up in this section. We also want to mention that at least 53 V1 vomeronasal receptor genes have been reported to be in the human genome (Lane et al., 2002). We approached Dr. Barbara Trask (Columbia University, NY), and she kindly provided us with a file with these 53 genes, which all look like pseudogenes except one (V1RL1). V1RL1 is found here because it does not show clear phylogenetic relationship to any of the main families. Lane et al. (2002) reported that there were three clusters of these genes found on HSA1, HSA7 and HSA19:

GPRC5B, NP_071319, 17q25; **GPRC5C**, NP_016235.1, 16p12; **GPRC5D**, NP_061124.1; **GPR**, NP_009154.1, 15q13.3; **GPR14**, NP_061822.1, 17q25.3; **GPR19**, NP_006134.1, 12p12.3; **GPR20**, NP_005284.1, 8q24.2-q24.3; **GPR22**, NP_005286.1, 7q22-q31.1; **CMKRL2(GPR30)**, NP_001496.1, 7p22; **GPR31:** NP_005290.1, 6q27; **GPR34**, NP_005291.1, Xp11.4-p11.3; **GPR40**, NP_005294.1, 19q13.12; **GPR41(GPR42)**, NP_005295.1, 19q13.12; **GPR43**, NP_005297.1, 19q13.12; **GPR39**, NP_001499.1, 2q21-q22; **GPR63**, NP_110411.1, 6q16.1-q16.3; **GPR75**, NP_006785.1, 2p16; **GPR84**, NP_065103.1, 12q13.13; **HRH1**, NP_000852.1, 3p25; **HRH3**, NP_009163.1, 20q13.33; **SREB2(GPR85)**, NP_061843.1, 7q31; **VLGR1**, XP_057299, 5q13; **V1RL1**, NP_065684, 19q13.43

Discussion

This is the first phylogenetic study of the entire superfamily of GPCRs in a single mammalian genome. The analyses show with high bootstrap support that there are five main families of human GPCRs (Fig. 2). Each of the receptors that we placed in the five families shows appreciable bootstrap value in support of a phylogenetic relationship to the respective family. The results indicate that the members within each family share a common evolutionary origin. We have given the families the following names: glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin, and we refer to them as the GRAFS families or the GRAFS classification, based on the initials of the family names. The rhodopsin receptors make up the largest family, and we show four main groups (Fig. 3) with 13 distinct branches. We chose not to subdivide the other families.

Three of the families, the rhodopsin (A), secretin (B), and glutamate (C) families, correspond to the A-F clan system (Attwood and Findlay, 1994; Kolakowski, 1994), whereas the two other families, adhesion and frizzled, are not included in the clan system. We did not find receptors in the human genome that belong to families that correspond to clans D, E, F, or O. All the receptors, except 23, were designated as members of one of the GRAFS families. We found 342 functional nonolfactory GPCRs in our searches of the human database. Combining this number with the preliminary number of olfactory receptors we identified (460), the total number of functional GPCRs in the human genome is more than

800. Our analysis covers thus about 2% of the genes in the human genome. We are not aware that simultaneous phylogenetic analysis has previously been performed on such a large and complex data set from a single genome. It may seem to be a daunting task to analyze the remaining 98% of the human genome covering the other protein families. We believe, however, that our "manual" approach, inspecting sequence for sequence, group to group, is important to provide clarity into numbers and phylogenetic topology of the proteins in the genome. We believe that our results will be valuable for analyzing the mouse, rat, chicken, fugu, and zebrafish genomes to determine the orthologous relationship of the GPCRs in these other genomes, which are already available or are soon to be completed.

The phylogenetic relationship of the secretin and secretin-like receptors (a term widely used in connection to a variety of receptors) in the human genome has been unclear. Our analysis shows one distinct family of receptors whose ligands are rather large peptides that mainly act in a paracrine manner; we term these the secretin family. However, we also show that there exists another distinct family of receptors that we name, for the first time, the adhesion family. Many of these receptors have very long N termini and most of them have adhesion molecule repeats that are likely to participate in cell-to-cell interactions. Previously, it had been suggested that the metabotropic glutamate receptors belong to the same family as the calcium and GABA receptors (Bockaert and Pin, 1999). Our analysis confirms this and shows that the two GABA receptors branch basally in the glutamate family. A few recently found taste receptors (TAS1) also group into the glutamate family. The fifth family is made up of the frizzled receptors and a number of taste receptors (TAS2). It is important to note that the taste receptors in groups TAS1 and TAS2 do not show any phylogenetic relationship; to add to the confusion, some olfactory receptors have TAS names (probably given by mistake, to our best knowledge).

It has often been stated that the different GPCR families show no structural similarities. Bockaert and Pin (1999) wrote that "There are at least six families of GPCRs showing no sequence similarity". In fact, several 7TM receptors (for example, bacterial rhodopsin, several chemosensory receptors in *C elegans*, and olfactory receptors in *D melanogaster* show very low or no similarities to any GPCR in the human genome (Robertson, 1998; Mombaerts, 1999). Repeated BLAST searches on GPCRs from various species have implied that three overall classes of GPCRs may exist (Josefsson, 1999). A recent study analyzing GPCRs from a number of highly divergent species showed 34 distinct clusters with significant alignment between distantly related clusters (Graul and Sadee, 2001). It is important to note that our *phylogenetic* analysis does not reveal clear evidence of a common descent of the GRAFS families. However, visual inspections of the alignments disclose features that are shared within the families beyond the feature of seven hydrophobic regions. All the families have a conserved Cys between TMI and TMII and another conserved Cys between TMIII and TMIV. These residues are believed to create a disulfide bridge between these loops and to be important for the structural integrity of the protein. The conservation of these two single amino acids does obviously not have an impact in the phylogenetic analysis. This is because of the distance between them, the variability in the length of the receptors, and because these bridges do not seem to need defined structural surroundings, probably because they are found in the flexible extracellular loops. It should be noted that the actual physical presence of these bridges has not been shown for all the different families, although it is very well established that these are functionally crucial for several receptors within the rhodopsin family.

To further analyze the putative similarities between the families, we extended our analysis by generating HMMs for each family. The families may share several regions that are well conserved between the groups that are not evident by looking at the alignment alone. We subsequently tried to align the TM regions of the HMMs. Several motifs shared by some families emerged, as exemplified by the alignments shown in Fig. 5. All the proteins in each family (except the olfactory cluster in rhodopsin) contribute to these HMM consensus sequences in Fig. 5. We found it remarkable that all the consensus sequences derived from the GRAFS families aligned, without generating long or repeated gaps, with their respective TM regions, with only a few exceptions (the glutamate and frizzled families did not align in TMIII and TMVII). The TM consensus sequences could not be aligned to a "wrong" TM region, meaning, for example, that any consensus sequence from TMI could not be aligned with the consensus sequences from TMII, TMIII, TMIV, TMV, TMVI, or TMVII (data not shown). The consensus alignment created "consensus residues", marked by dark shading in Fig. 5. None of these consensus residues is conserved through all five families, but six of them were found in four families. Moreover, the nonidentical residue in the same position as these six consensus residues is also a hydrophobic residue in all cases except one. Furthermore, in three cases, the fifth residue is a valine that is closely related structurally to the consensus residue leucine. The boundaries of the TM regions are defined by hydrophobicity plots (see the Introduction), and it is thus no surprise that the alignable residues are hydrophobic. This could indicate, however, that the sequence similarity may be caused by functional constrains related to the α-helical structure that passes the lipophilic membrane rather than common descent. It should be noted, however, that the hydrophobicity varies notably from one α-helix to another, and none of the sequence similarities is repeated in more than one helix. Visual inspection shows that the numbers of identical residues clearly differs from one helix to another, indicating a nonrandom pattern. Different hydrophobicity patterns from one helix to another could be attributed to different positioning in the seven helical clusters that makes up the receptor, enabling signal transduction through the membrane to the G-proteins. Considering the crystal structure of bovine rhodopsin, the TMIII for example is oriented in the middle of the TM cluster, whereas TMIV and TMV are more exposed to the membrane (Baldwin, 1994; Palczewski et al., 2000). Whether the clustering of these hydrophobic residues is related to common TM orientation or to other important structural features, we are inclined to believe that they add support for a possible common descent of the GRAFS families. We also find it intriguing that although none of the repeated residue motifs are clearly shared by all the five families, they all can be connected through motifs in two or more families. In TMII, the glutamate and the frizzled families align in a seven-residue consensus se-

quence in which five residues are identical and the difference lies in Val and Leu in one position and two polar residues, Thr and Arg, in the other nonidentical position. The adhesion and secretin families share several short motifs in TMI, TMII, TMVI, TMVII, and also an 11-amino acid motif in TMV, where eight residues are identical. The adhesion and secretin families link to the frizzled family in TMIV with a G/AWG/AXPAL/V, where X is always hydrophobic; it should be noted that P and W are rather unusual residues in α-helixes. The rhodopsin family has no long sequence motif that links it clearly to any of the other families. However, two three-amino acid motifs are found in TMIV and TMVI that link the rhodopsin family to the glutamate family and adhesion families, respectively. Moreover, all six positions that have four identical residues include the rhodopsin family; for example, the Trp that is a part of the strong motif in TMIV links the adhesion, frizzled, and secretin families. Thus, the results indicate that primary sequences are shared within the families. The HMM approach applied here and the subsequent alignment is also more sensitive than using simple sequence alignments; further application of such methods could be the key to identifying more conserved motifs between the groups. Considering the direct sequence similarities mentioned above, together with the putative conserved Cys bridge in all families and the TM region-dependent alignment pattern displayed in Fig. 5, we suggest thus that there is confounding evidence that the human GPCRs that we assigned to the GRAFS families share a common ancestor.

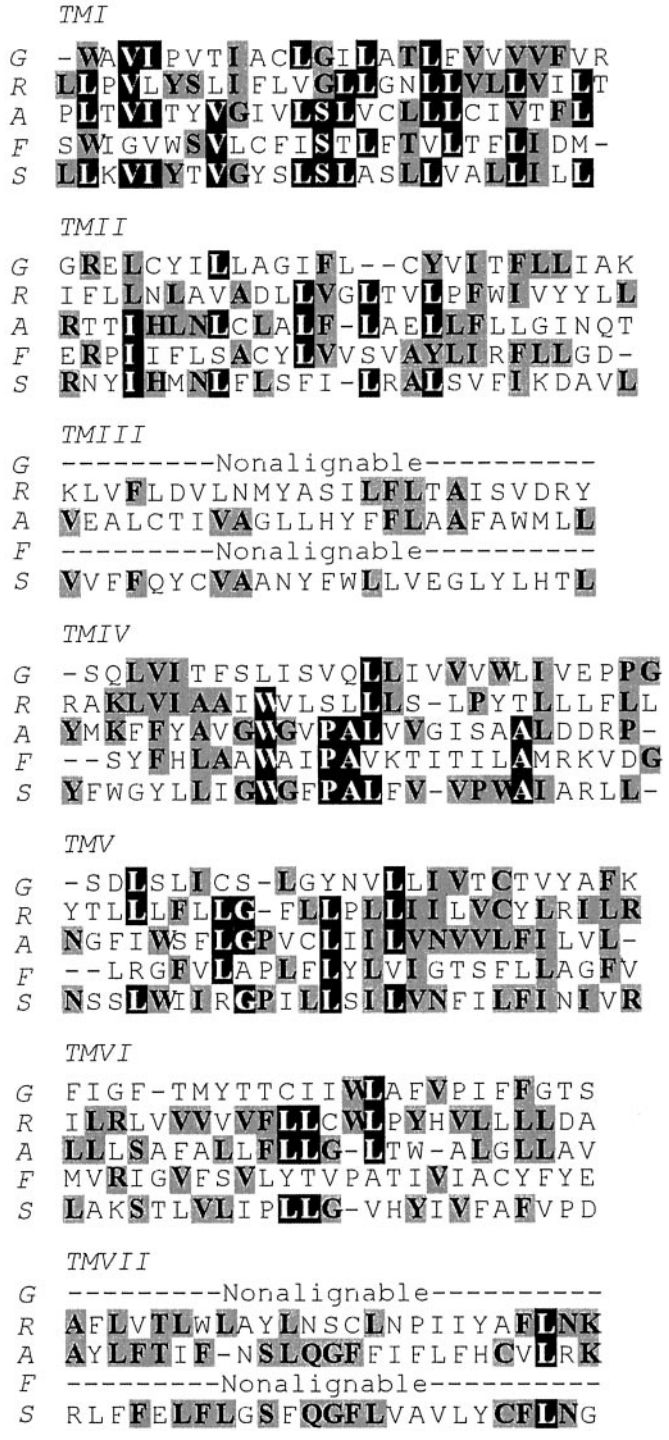

Fig. 5. Alignment of the consensus sequences of the region around the TM-regions for the five families of human GPCRs. The consensus sequences are statistically derived using HMMalign as described under *Materials and Methods*. Black boxes indicates that the residue is conserved between four of the five families, dark gray indicates conservation between three families, and light gray denotes conservation between two families.

We created a chart showing how the GPCRs are found in different paralogy groups (See Fig. 4). This figure shows how several of the GPCRs are located in paralogous regions on the chromosomes. When these groups are studied together with the phylogenetic trees, it demonstrates how a large number of these receptor genes are likely to have been formed through tetraploidizations, whereas others are more likely to have arisen through local gene duplications. Another piece of information that is obtained from the paralogons is the putative mechanism for how the different gene subfamilies in the adhesion family have been composed from different domains. All of the genes in the adhesion family, of course, contain the code for the seven TM regions; apart from this, many of them also have distinct elements in the N termini that can be recognized in various other gene families. We predicted that it might be possible to trace some of the major evolutionary events of putative domain shuffling. We compared the chromosomal locations of these adhesion family genes with the chromosomal locations of the genes that might be supposed to carry the parental domains in question. The three BAI genes are located in the group of paralogous chromosomal regions, 1p3/2p/8q/20q, originally described by Spring et al. (1994) and later extended to contain parts of 6p, 6q, 16q, and 18. Two of the LEC genes, the EMR genes, and CD97, as well as ETL, GPR56 (TMVIIXN1), and one of the CELSR genes, belong to the paralogon 1p-q2/6p/9/19p (Katsanis et al., 1996), later extended to include parts of 5p-q2 and 15q. These two paralogy groups have two human chromosomal regions in common, 1p3 and 6p2, which may give an indication that the ancestral regions of these groups might be have syntenic or arisen from a common region at an earlier stage of vertebrate evolution (Lundin et al., 2002). Furthermore, they share the 1p3 region with a third paralogon, 1p3/3q/7q/12p/17p. It was suggested that parental genes, of the ones found in the 10 main regions included in these paralogy groups plus the likely translocated regions, once could have been syntenic in an early prevertebrate. According to this scenario, this ancestral region duplicated twice as a result of the postulated genome doublings, and these four

newly formed regions must then successively have split up into a larger number of regions, except the one in chromosome 1p3. It is thus interesting to see that most of the genes that are likely to have contributed to the several different domains seen in genes from the adhesion family are also present in these three paralogy groups. Four of the subfamilies (BAI, CD97, EMR, and LEC) contain a mucin domain. Mucin genes are located at 1q22, 3q21.2, 3q29, 6q21, 7q22, and 19p13.2. The LEC subfamily carries an olfactomedin domain, and the two olfactomedin genes mapped in the human genome are located on 9q34.3 and 19p13.2. Genes of the BAI subfamily have several thrombospondin domains, and the three human thromobospondin genes are mapped at 1q21, 6q27, and 15q15. The CELSR genes carry cadherin domains, and no less than 16 cadherin genes are located at 5p14-13, 8q22, 16q21-24, 18q, and 20q13. Furthermore, the CELSR genes contain two laminin A domains, and laminin A genes have been mapped to 6q21-22 (2), 18p11, 18q11, and 20q13. Genes from three of the subfamilies, CD97, EMR, and CELSR, also carry EGF-like domains, and two of the human EGFL genes are found at 1p36.3 and 9q32-33. It does seem likely that all the genes mentioned in this connection were linked in the same chromosomal region in an early metazoan and that unequal crossing-over between parental genes in this region caused exon shuffling, leading to the structures found in extant genes of the adhesion family.

In summary, we have generated the first map for one of the most studied superfamily of proteins found in the human genome. We demonstrated the existence of five distinct families of GPCRs, and we determined the relationship of the genes within subgroups of the large rhodopsin family. This map will be very useful for comparison of GPCRs in other species and will subsequently enhance our understanding of how structural and functional properties evolved. The paralogon analysis presents further evidence for common descent of the phylogenetic clusters and exemplifies how exon shuffling may have played a role in composition of some of the receptor genes. Because of the diversity of structural elements found in this family, it is likely that the examples of evolutionary mechanisms that are predicted here may have a general importance for several other protein families, typically those that share α-helical domains and TM regions that are combined with other functional elements.

Acknowledgments

We thank Dr. David Ardell, Uppsala University, for frequent advice on tree building strategies, alignments, fingerprinting, and programming solutions and for critical assessment of the main conclusions. We also thank David Gloriam and Pär Höglund, both students at Uppsala University, for extensive data mining; professor Anthony J. Harmar, University of Edinburgh, member of the nomenclature committee of the NC-IUPHAR, for comprehensive advice on nomenclature and encouragement; Dr. Hester Wain, HUGO Gene Nomenclature Committee at University College, London, for advice on orphan receptor nomenclature; professor Dan Larhammar, Uppsala University, for valuable criticism; Dr. Magnus Berglund, Uppsala University, for providing the initial data set of GPCRs; and Nils-Einar Eriksson, head of the computer department at BMC, for providing access to computers within the core BMC facilities during nonoffice hours.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:**3389–3402.
Attwood TK and Findlay JB (1994) Fingerprinting G-protein-coupled receptors. *Protein Eng* **7:**195–203.
Baldwin JM (1994) Structure and function of receptors coupled to G proteins. *Curr Opin Cell Biol* **6:**180–190.
Barnes MR, Duckworth DM, and Beeley LJ (1998) Frizzled proteins constitute a novel family of G protein-coupled receptors, most closely related to the secretin family. *Trends Pharmacol Sci* **19:**399–400.
Bockaert J and Pin JP (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO (Eur Mol Biol Organ) J* **18:**1723–1729.
Broeck JV (2001) Insect G protein-coupled receptors and signal transduction. *Arch Insect Biochem Physiol* **48:**1–12.
Coulier F, Burtey S, Chaffanet M, Birg F, and Birnbaum D (2000) Ancestrally-duplicated paraHOX gene clusters in humans. *Int J Oncol* **17:**439–444.
Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14:**755–763.
Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5c. Department of Genetics, University of Washington, Seattle.
Flower DR (1999) Modelling G-protein-coupled receptors for drug design. *Biochim Biophys Acta* **1422:**207–234.
Fredriksson F, Lagerström MC, Höglund PJ, and Schiöth HB (2002) Novel human G-protein coupled receptors with long N-terminals containing GPS domains and Ser/Thr rich regions. *FEBS Lett*, **531:**407–414.
Glusman G, Yanai I, Rubin I, and Lancet D (2001) The complete human olfactory subgenome. *Genome Res* **11:**685–702.
Graul RC and Sadee W (2001) Evolutionary relationships among G protein-coupled receptors using a clustered database approach. *AAPS PharmSci* **3:**E12.
Harmar AJ (2001) Family-B G-protein-coupled receptors. *Genome Biol* **2:**REVIEWS3013.
Hayflick JS (2000) A family of heptahelical receptors with adhesion like domains: a marriage between two superfamilies. *J Recept Signal Transduct Res* **20:**119–131.
Holland PW (1999) Gene duplication: past, present and future. *Semin Cell Dev Biol* **10:**541–547.
Holland PW, Garcia-Fernandez J, Williams NA and Sidow A (1994) Gene duplications and the origins of vertebrate development. *Dev Suppl* 125–133.
Josefsson LG (1999) Evidence for kinship between diverse G-protein coupled receptors. *Gene* **239:**333–340.
Kasahara M (1999) The chromosomal duplication model of the major histocompatibility complex. *Immunol Rev* **167:**17–32.
Katsanis N, Fitzgibbon J, and Fisher EM (1996) Paralogy mapping: Identification of a region I the human MHC triplicated onto human chromosomes 1 and 9 allow the prediction and isolation of novel PBX and NOTCH loci. *Genomics* **35:**101–108.
Kolakowski LF Jr (1994) GCRDb: a G-protein-coupled receptor database. *Recept Channels* **2:**1–7.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome [published erratum appears in *Nature* **412:**565, 2001]. *Nature (Lond)* **409:**860–921.
Lane RP, Cutforth T, Axel R, Hood L, and Trask BJ (2002) Sequence analysis of mouse vomeronasal receptor gene clusters reveals common promoter motifs and a history of recent expansion. *Proc Natl Acad Sci USA* **99:**291–296.
Lembo PM, Grazzini E, Groblewski T, O'Donnell D, Roy MO, Zhang J, Hoffert C, Cao J, Schmidt R, Pelletier M, et al. (2002) Proenkephalin A gene products activate a new family of sensory neuron–specific GPCRs. *Nat Neurosci* **5:**201–209.
Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16:**1–19.
Lundin L-G, Larhammer D, and Hallböök F (2003) Numerous groups of chromosomal regional parologies strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics* **3:**53–63.
McKnight AJ and Gordon S (1998) The EGF-TM7 family: unusual structures at the leukocyte surface. *J Leukoc Biol* **63:**271–280.
Mombaerts P (1999) Seven-transmembrane proteins as odorant and chemosensory receptors. *Science (Wash DC)* **286:**707–711.
Robertson HM (1998) Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement and intron loss. *Genome Res* **8:**449–463.
Slusarski DC, Corces VG, and Moon RT (1997) Interaction of Wnt and a Frizzled homologue triggers G-protein-linked phosphatidylinositol signalling. *Nature (Lond)* **390:**410–413.
Sidow A (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev* **6:**715–722.
Stacey M, Lin HH, Gordon S, and McKnight AJ (2000) LNB-TM7, a group of seven-transmembrane proteins related to family-B G-protein-coupled receptors. *Trends Biochem Sci* **25:**284–289.
Spring J (1997) Vertebrate evolution by interpecific hybridisation—Are we polyploid? *FEBS Lett* **400:**2–8.
Spring J, Goldberger OA, Jenkins NA, Gilbert DJ, Copeland NG, and Bernfield M (1994) Mapping of the syndecan genes in the mouse: linkage with members of the myc gene family. *Genomics* **21:**597–601.
Takeda S, Kadowaki S, Haga T, Takaesu H, and Mitaku S (2002) Identification of G protein-coupled receptor genes from the human genome sequence. *FEBS Lett* **520:**97–101.
Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:**4673–4680.
Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I,

Teller DC, Okada T, Stenkamp RE, et al. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science (Wash DC)* **289:**739–745.
Pebusque MJ, Coulier F, Birnbaum D, and Pontarotti P (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol* **15:**1145–1159.
Popovici C, Leveugle M, Birnbaum D, and Coulier F (2001) Coparalogy: physical and functional clustering in the human genome. *Biochem Biophys Res Commun* **288:** 362–370.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. (2001) The sequence of the human genome [published erratum appears in *Science (Wash DC)* **292:**1838, 2001]. *Science (Wash DC)* **291:** 1304–1351.
Zozulya S, Echeverri F, and Nguyen T (2001) The human olfactory receptor repertoire. *Genome Biol* **2:**RESEARCH0018.

Address correspondence to: Helgi B. Schiöth, Department of Neuroscience, Biomedical Center, Box 593, 75 124 Uppsala, Sweden. E-mail: helgis@bmc.uu.se