0026-895X/05/6705-1414–1425$20.00
Molecular Pharmacology
Copyright © 2005 The American Society for Pharmacology and Experimental Therapeutics
Mol Pharmacol 67:1414–1425, 2005

Vol. 67, No. 5
9001/1198145
Printed in U.S.A.

The Repertoire of G-Protein–Coupled Receptors in Fully Sequenced Genomes[S]

Robert Fredriksson and Helgi B. Schiöth

Department of Neuroscience, Uppsala University, Biomedical Center, Uppsala, Sweden

Received November 11, 2004; accepted February 1, 2005

ABSTRACT

The superfamily of G-protein–coupled receptors (GPCRs) is one of the largest and most studied families of proteins. We created Hidden Markov Models derived from sorted groups of GPCRs from our previous detailed phylogenetic classification of human GPCRs and added several other models derived from receptors not found in mammals. We used these models to search entire Genscan data sets from 13 species whose genomes are nearly completely sequenced. We found more than 5000 unique GPCRs that were divided into 15 main groups, and the largest one, the *Rhodopsin* family, was subdivided into 13 subclasses. The results show that the main families in the human genome, *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled,* and *Secretin,* arose before the split of nematodes from the chordate lineage. Moreover, several of the subgroups of the *Rhodopsin* family arose before the split of the linage leading to vertebrates. We also searched expressed sequence tag (EST) databases and identified more than 20,000 sequences that match GPCRs. Although the GPCRs represent typically 1 to 2% of the Genscan predictions, the ESTs that match GPCRs are typically only 0.01 to 0.001%, indicating that GPCRs in most of the groups are expressed at low levels. We also provide searchable data sets that may be used for annotation and further detailed analysis of the GPCR family. This study provides an extensive overview of the expansion of the gene repertoire for families and subgroups of GPCRs.

A tremendous amount of primary sequence information has been made available from the recent sequencing projects, providing a near-to-full coverage of the entire genome from a diversity of animal species. The fully sequenced genomes include the mammals mouse and human, two species from the bony fish line (pufferfish *Takifugu rubripes* and zebrafish *Danio rerio*), two protochordates from the tunicate linage (*Ciona intestinalis* and *Ciona savignyi*) together with two nematodes (*Caenorhabditis elegans* and *Caenorhabditis briggsae*), and the insects fruit fly (*Drosophila melanogaster*) and mosquito (*Anopheles gambiae*). Moreover, the plants thale cress (*Arabidopsis thaliana*) and rice (*Oryza sativa*) are sequenced, as well as several unicellular species of yeast from the fungi linage such as the bakers' yeast (*Schizosaccharomyces pombe*) and budding yeast (*Saccharomyces cerevisiae*). The quality of these genomes is constantly being improved, with the most recent assembly of the human genome having only 341 gaps (International Human Genome Sequencing Consortium, 2004), and the results of all analyses from the genomic data, including gene predictions, are dependent on the current gene assembly. Other genomes, such as those of fugu and zebrafish, however, will need a considerable amount of work to reach the same quality. Although only a small fraction of the genes from these genomes are actually annotated, all of the genomes mentioned above have gene prediction data sets constructed using Genscan (Burge and Karlin, 1997).

The superfamily of G-protein–coupled receptors (GPCRs) is one of the largest and most diverse families of proteins in mammals (Bockaert and Pin, 1999). GPCRs are in the group of proteins that draws the most attention in the pharmaceutical industry, and it is estimated that 40 to 50% of all

The studies were supported by the Swedish Research Council (medicine), the Swedish Society for Medical Research, Svenska Läkaresällskapet, Åke Wikberg Foundation, The Novo Nordisk Foundation, Thurings Foundation, and Magnus Bergwall Foundation.

[S] The online version of this article (available at http://molpharm.aspetjournals.org) contains supplemental material.

Article, publication date, and citation information can be found at http://molpharm.aspetjournals.org.
doi:10.1124/mol.104.009001.

ABBREVIATIONS: GPCR, G-protein–coupled receptor; EST, expressed sequence tag; TM, transmembrane domain; HMM, Hidden Markov Model; NCBI, National Center for Biotechnology Information; bp, base pair; TAS2, taste receptors type 2; MCH, melanocyte-concentrating hormone; PUR, purine; GUST, gustatory receptor; DMOD, odorant receptor; CHEM, chemokine receptor; PTGER, prostaglandin receptor; NCHM, nematode chemosensory receptor; ADH, *Adhesion* family; SEC, *Secretin* family; FZD, *Frizzled* family; GLR, *Glutamate* family; RHOD, *Rhodopsin* family; AMIN, serotonin/dopamine/adrenergic/trace amines G-protein–coupled receptor; PEP, neuropeptide/peptide G-protein–coupled receptor; SOG, somatostatin/opioid/galanin G-protein–coupled receptor; MECA, melanocortin/endoglin/adenosin/cannabinoid G-protein–coupled receptor; VR, vomeronasal.

current drug targets are GPCRs. The large number of proteins in this gene family and the complex structure of GPCRs have, until recently, made it difficult to systematically study their overall evolution. The common structural feature of all GPCRs is a seven α-helical transmembrane region (7TM) that anchors the receptor to the plasma membrane of the cell, with the N termini exposed to the extracellular space. In addition to the 7TM region, some families of GPCRs have long N termini containing different kinds of functional or ligand binding domains.

GPCRs can be found in almost any eukaryotic organism, including insects (Hill et al., 2002) and plants (Josefsson, 1999), indicating that these proteins are of ancient origin. There is also a light-sensing 7TM protein found in bacteria, the bacterial *Rhodopsin*, but it is presently unclear whether this protein has a common origin with GPCRs in eukaryotic organisms, because it does not signal through G-proteins and lacks significant sequence homology to GPCRs (Okada and Palczewski, 2001). The human repertoire of GPCRs has recently been described in detail (Joost and Methner 2002; Fredriksson et al., 2003c; Vassilatis et al., 2003), although there are still additional new human GPCRs being annotated (Fredriksson et al., 2003a,b). In addition, the entire repertoire of GPCRs in mouse (Vassilatis et al., 2003) and malaria mosquito (Hill et al., 2002) have been described, although other genomes lack such a whole-genome description of the GPCR superfamily. These studies have provided a good overview of the mammalian genomes, but there exists considerable confusion about the relationship of GPCR subgroups among different eukaryotes. Several classification systems for GPCRs have been proposed, such as the A to F system (Kolakowski, 1994) and the 1 to 5 system (Bockaert and Pin, 1999). These systems attempt to cover the entire GPCR repertoire in several developmental lineages but do not include some of the more recently discovered families. We have described previously a classification system for the GPCR superfamily in the human genome (Fredriksson et al., 2003c) in which we divided the receptors into five families using a phylogenetic approach. Here, we use the terminology *Rhodopsin* (also known as A or 1), *Secretin* (B or 2), *Adhesion* (previously included in B or 2), *Frizzled* (F or 5), and *Glutamate* (C or 3), which form the GRAFS classification system.

In this article, we investigated the origin of the human GPCRs by searching the Genscan data sets for GPCRs from 13 species in which a complete genomic sequence is available. We used Hidden Markov Models (HMMs) taken from our recent GRAFS classification system for GPCRs (Fredriksson et al., 2003c) as well as groups of GPCRs that are not found in mammals, identified previously in other classification systems. The aim was to identify “all” GPCRs in these genomes, group these in families, and determine the relationship of GPCRs in distantly related species and thus reveal the origin and expansion of each group.

Materials and Methods

Description of the Original Data Sets

Human. We used the NCBI build 33 of the Genscan data set. This genome is largely contiguous (i.e., free of gaps) and includes more than 99% of the genetic material (http://genome.ucsc.edu/). The predicted gene set from this assembly contains approximately 55,000 genes, whereas the manually reviewed RefSeq data set contains nearly 20,000 protein sequences (http://www.ncbi.nlm.nih.gov/).

Mouse. The current assembly of the mouse genome (NCBI build 30) consists of 38,000 contigs with a predicted gene set of 110,000 proteins (http://www.ncbi.nlm.nih.gov/). Approximately 90 to 95% of the genetic material is present in the assembly (http://genome.ucsc.edu/). The mouse RefSeq data set currently has approximately 16,200 protein sequences.

Fugu. The genome of fugu (*Takifugu rubripes*) used in this study was Ensembl release 17.2.1 which consists of 8597 contigs, covering almost 320 Mbp (i.e., approximately 95% of the nonrepetitive DNA) (Aparicio et al., 2002) and has 29,600 Genscan predicted genes (http://www.ensembl.org/Fugu_rubripes/). Very few genes have been manually annotated from this species, and the nonredundant protein database at NCBI has fewer than 50 GPCRs from fugu.

Zebrafish. The genome of the zebrafish (*Danio rerio*) used here was the Ensemble 17.2.1 release with 1.56 Gbp in 85,700 contigs. The Genscan predicted protein data set has approximately 60,000 gene predictions (http://www.ensembl.org/Danio_rerio/). Also from this species, very few genes have been manually annotated; the RefSeq data set has 1170 proteins and contains in total 72 GPCRs from zebrafish (http://www.ncbi.nlm.nih.gov/).

***Ciona* Species.** The draft genome sequence of the Urochordate *Ciona intestinalis* was published in late 2002 and is an 8.2× whole-genome shotgun assembly. The sequence contains 2500 contigs and covers 116 Mbp of nonrepetitive sequence, approximately 90% of the total nonrepetitive material (Dehal et al., 2002), and an annotation project is ongoing at (http://genome.jgi-psf.org). The genes are mainly automatically annotated. The current Genscan data set contains 15,800 genes (http://www.ncbi.nlm.nih.gov/).

***D. melanogaster*.** The *D. melanogaster* genome was sequenced in 2000 and is believed to contain approximately 98% of the 120 Mbp of the *D. melanogaster* genome (Adams et al., 2000). For this study, we used the NCBI Genscan data set consisting of 14,300 predicted genes. There is a large annotation project at Flybase (http://www.flybase.org/), which lists 13,500 genes and a total of 193 GPCRs.

***A. gambiae*.** The genomic sequence of the mosquito *A. gambiae* was published as approximately 10× whole-genome shotgun in 2002 (Holt et al., 2002), and current assembly contains 278 Mbp in 18,634 contigs (http://www.ensembl.org/Anopheles_gambiae/). The Genscan data set contains 16,000 predicted genes, and there are almost no manually annotated genes for this specie.

***C. elegans*.** The genomic sequence of the nematode *C. elegans* was published previously (The *C. elegans* Sequencing Consortium, 1998). The most current assembly contains 103 Mbp of genomic sequence in 3266 contigs. The current Genscan data set contains 20,200 predicted sequences (http://www.ncbi.nlm.nih.gov/). An annotation project is underway (http://www.wormbase.org/), and the project lists 4609 confirmed genes, but it provides no classification for these.

***O. sativa*.** The genomic sequence of rice (*O. sativa*) is currently in progress (http://rgp.dna.affrc.go.jp/), and the current Genscan data set contains 2400 predicted proteins (http://www.ncbi.nlm.nih.gov/). Very few genes are annotated in this species. The annotations are currently almost exclusively automated or semiautomated.

***A. thaliana*.** The only completed plant genome is the genome of *A. thaliana* (The Arabidopsis Initiative, 2000) with a Genscan data set of 6600 predicted proteins (http://www.ncbi.nlm.nih.gov/). The annotations for this specie are currently almost exclusively automated or semiautomated.

***Plasmodium falciparum*.** The genome of *P. falciparum* was sequenced in 2002, and the current assembly contains 23.1 Mbp. The current Genscan data set contains 5200 predicted proteins. An annotation project is ongoing (http://plasmodb.org/publications.shtml), and the current database contains domain and gene ontology annotations of the predicted protein set but no verified proteins.

***S. pombe*.** The fission yeast *S. pombe* genome was sequenced in 1998, and the current assembly from NCBI consists of 12.6 Mbp. The Genscan data set contains 5000 predicted proteins. There are several large-scale genomics and proteomics projects ongoing aiming at annotating the entire genome of fungi, such as the YPD (https://www.incyte.com/proteome/) and the *S. pombe* gene DB (http://www.genedb.org/genedb/pombe/index.jsp). These databases contain annotation information of various qualities from biochemically well-characterized proteins to annotations derived from only similarity and computer predictions.

S. cerevisiae. The budding yeast (*S. cerevisiae*) genome project is at a state similar to that of the genome project of *S. pombe,* with a current assembly from NCBI consisting of 10.3 Mbp with a Genscan data set of 6300 predicted proteins. Annotations are available (https://www.incyte.com/proteome/ and http://www.yeastgenome.org/) and vary in quality, as do the *S. pombe* data.

Construction of HMM Models

The overall HMMs were constructed from receptors as shown in Table 1. The accession numbers of each receptor and information regarding which HMM each receptor belongs to can be found in Supplemental Table S1. We removed the long N and C termini from some of the receptor sequences, as identified by RPS-BLAST searches (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The receptor sequences were subsequently aligned using ClustalW 1.81 (Thompson et al., 1994) using default settings. From the alignments, HMMs were constructed using the HMMER 2.2 package (Eddy, 1998). The models were constructed using HMMbuild with default settings and were calibrated using HMMcalibrate.

Identification of GPCR Sequences in the Genscan Data Sets

FASTA files containing the protein versions of the Genscan predicted gene sets were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/ (*H. sapiens*, *C. elegans*, *D. melanogaster*, *S. pombe*, *S. cerevisiae*, and *P. falciparum*) and ftp://ftp.ensembl.org/pub/ (*M. musculus*, *T. rubripes*, *D. rerio*, *C. intestinalis*, and *A. gambiae*). These predicted protein sets were searched against the HMMs using HMMsearch from HMMER 2.2 with a cutoff at $E = 10$. All hits with an *E* value lower than 0.01 were considered correct, and all hits with *E* values between 0.01 and 10 were manually inspected to verify that they are true GPCRs using BLASTP searches against the NCBI GenPept data set. Protein sequences from the list of FASTA tags that were to be manually checked were extracted using the fastacmd program from the NCBI BLAST suite. For each predicted protein, the top five hits were manually inspected, and a minimum of four of these had to be a GPCR for inclusion in the particular data set; all other predicted proteins were excluded from further analysis. The criteria used for the *Rhodopsin* family were the presence of conserved amino acids, identified from pairwise alignments with the closest known human GPCR, such as the NPxxY motif at the end of TM7, the DRY motif at the end of TM3, and the conserved N early in TM1. For the other families, which have less-established recognizable motifs, we used multiple alignments of the protein(s) of interest together with a selection of the human family members that allowed for the identification of conserved residues and motifs.

Collection of Data Sets for Each Species

Lists containing the identification numbers from all significant hits from each species were extracted from the HMMER output files and were imported into Microsoft Excel (Microsoft, Redmond, CA). Here, all predicted proteins that hit more than one model were manually inspected, and the hit with the lowest *E* value was kept in the data set. Furthermore, lists containing the “true” GPCR data set were collected in Microsoft Excel, and statistics were calculated. All calculations and graph-plotting were performed in Microsoft Excel. The sequence names and the classification of the receptors can be found in Supplemental Table S2. The sequences in FASTA format are available as Supplemental File S1.

Subdivision of the Rhodopsin GPCRs

A data set consisting of 614 human *Rhodopsin* GPCRs were divided into 13 groups according to the work of Fredriksson et al. (2003c) and were subsequently used to subdivide the *Rhodopsin* GPCRs found in the HMM searches. The human GPCRs were divided as follows, with the group name shown in boldface type and the number of receptors in parentheses: olfactory receptors, **OLF** (347); serotonin/dopamine/adrenergic/trace amines, **AMIN** (42); neuropeptide/peptide, **PEP** (35); chemokine-like, **CHEM** (42); purine-like, **PUR** (42); somatostatin/opioid/galanin, **SOG** (15); opsin-like, **OPN** (9); glycoprotein binding, **LGR** (8); prostaglandin, **PTGER** (15); melanocortin/endoglin/adeno-

TABLE 1

Description of the data set used to build the HMMs used for mining and classification into the 15 main families of GPCRs in the genomes investigated

The source of the dataset is indicated in parentheses.

Model Name	Number of GPCRs	Description
ADHSEC	46	13 human secretin-like GPCRs (Fredriksson et al., 2003c); 33 adhesion-like GPCRs (Fredriksson et al., 2002, 2003a,b)
CAMP	4	cAMP binding GPCRs from *Dictyostelium* (http://www.gpcr.org/7tm/)
DMODOR	39	*D. melanogaster* odorant receptors (http://www.gpcr.org/7tm/)
FZD	11	Human Frizzled 1–10 and smoothened (Fredriksson et al., 2003c)
GUST	72	Gustatory receptors from *A. gambiae* (Hill et al., 2002)
GLR	18	8 metabotropic glutamate receptors, 3 taste receptors type 1, 2 GABA receptors, calcium-sensing receptor, 4 orphan GPCRs (Fredriksson et al., 2003c)
MLO	16	Plant GPCRs of the MLO type (Devoto et al., 2003)
NCHM	34	Nematode chemokine receptors (http://www.gpcr.org/7tm/)
OA1	2	Ocular albinism genes from mouse and human (Fredriksson et al., 2003c)
RHOD	260	Human *Rhodopsin* receptors (Fredriksson et al., 2003c)
STE2	4	Yeast pheromone receptors of the STE2 type (http://www.gpcr.org/7tm/)
STE3	4	Yeast pheromone receptors of the STE2 type (http://www.gpcr.org/7tm/)
TAS2	13	Human taste receptors type 2 (Fredriksson et al., 2003c)
VR	42	Vomeronasal receptors from mouse and rat (http://www.gpcr.org/7tm/)

sin/cannabinoid, **MECA** (22); MRG receptors, **MRG** (8); melatonin, **MTN** (3); and melanocyte-concentrating hormone receptor, **MCHR** (2). A table listing the accession numbers, names, and exact grouping is provided in Supplemental Table S1. A BLAST database was constructed from the 614 human GPCRs, and the *Rhodopsin* GPCRs from the other species were searched against this database. A cutoff value of E = 1e-9 was used, and the five top hits were inspected manually for each receptor to determine to which human GPCR from the database it was most similar. The requirement for being placed in a given group was to have at least four of the five best hits from that specific group. The receptors that did not match these criteria were grouped as unclassified (**UC**). The sequence names and the classification of the receptors into subgroups can be found in Supplemental Table S3. The sequences in FASTA format are available as Supplemental File S1.

Expression Levels by EST Matches

The entire gbest was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genbank and entered into an SQL database using custom-made software. From this database, FASTA files containing all high-quality ESTs from each species were extracted, one file per species. The NCBI-BLAST package was used to construct a searchable database for each specie, and all Genscan GPCRs found in a given specie were searched against the EST database using TBLASTX with a cutoff at $E = -40$. The BLAST results were automatically extracted and converted into tables. Because it is likely that a given EST will be hit by several GPCRs, all hits except for the one with the highest E value were removed from the tables using custom-made software to obtain a nonredundant list. The results were extracted and converted into tables readable by Microsoft Excel using custom-made software. Data were analyzed, and graphs were plotted using Microsoft Excel. The software used is available from the authors upon request as C++ source code and BASH scripts.

Results

Our strategy was to create HMMs derived from well-characterized groups of GPCRs from our phylogenetic classification of the entire set of GPCRs in the human genome (Fredriksson et al., 2003c). These main families are *Adhesion* (ADH), *Secretin* (SEC), *Frizzled* (FZD), *Glutamate* (GLR), and *Rhodopsin* (RHOD). Moreover, we also created HMMs of groups of GPCRs that we did not find in the human genome, such as the cAMP binding receptors from slime molds (*Dictyostelium*), nematode chemoreceptors (Robertson, 1998), the gustatory receptors from insects (Hill et al., 2002), the odorant receptors from *D. melanogaster* (Hill et al., 2002), MLO receptors in plants (Devoto et al., 2003), and STE2 (Marsh and Herskowitz, 1998) and STE3 (Hagen et al., 1986) from yeast. Some of these families lack significant sequence homology with the mammalian GPCRs. There are also some atypical GPCRs with uncertain relation to other GPCRs, like the ocular albinism gene (OA1) (Shen et al., 2001) and the vomeronasal receptors found in vertebrates (Lane et al., 2002). A summary of the receptors used to construct the HMMs is shown in Table 1. A table with the accession numbers and which HMM model the different GPCRs belong to is available as Supplemental Table S1.

These HMMs were subsequently used to extract and classify GPCRs from the 13 species, including the human Genscan data set, to evaluate the quality of the Genscan data sets regarding GPCRs. During our pilot studies, we found that the HMMs for the *Adhesion* and the *Secretin* families did not easily distinguish between the receptors that hit these models in some species, particularly in fish. We therefore decided to merge them and subsequently separated these on the basis of the presence of a GPS domain in the N terminus close to TM1. The GPS domain is found in almost all *Adhesion* GPCRs, although it is not found in any *Secretin* GPCRs in the human and mouse genomes (Harmar 2001; Bjarnadottir et al., 2004). The overall results of the HMM searches are displayed in Table 2, and a detailed description of the results is available as Supplemental Table S2a-l. All hits

TABLE 2

The number of GPCRs in the different main classes in the genomes investigated

The numbers in parenthesis in the human column shows the number of GPCRs from each group as published in Fredriksson et al. 2003c. No entry indidcates that no GPCRs of this class were found.

	H. sapiens	*M. musculus*	*D. reiro*	*T. rubripes*	*C. intestinalis*	*D. melanogaster*	*A. gambiae*	*C. elegans*	*A. thaliana*	*O. sativa*	*S. cerevisiae*	*S. pombe*
ADH	27 (31)	13	22	6	48	5	13	5			1	
CAMP												1
DMODOR						58	85					
FZD	10 (11)	11	14	10	7	7	7	5				1
GLR	24 (18)	112	52	26	8	9	8	6				2
MLO									5	1		2
NCHM								1006				
OA1	1 (1)	1		1			1					
RHOD	752 (614)	1106	591	224	139	76	77	124				1
SEC	20 (15)	28	55	18	6	13	1	5	1			
STE2											1	1
STE3											1	1
TAS2	13 (25)	3	2									
VR	18 (25)	44	1	1								
GUST						42	76					
Total	865	1318	737	286	208	210	268	1149	6	1	3	9
% of total predicted genes	1.60	1.19	1.23	0.97	1.31	1.47	1.66	5.69	0.09	0.04	0.05	0.18

,

with an *E* value better (lower) than 0.01 were in principle considered to be correct, although the 10 "worst" of these hits were manually controlled using BLASTP searches against the protein database at NCBI, and only a few errors were found. The hits between 0.01 and 10, however, were all controlled manually using BLASTP, and the results of these are available in Supplemental Table S4. From this table, it is evident that the number of positives found for this *E*-value range was highly dependent on the HMM model. The specificity of the HMMs at high *E* values is shown in Supplemental Table S4.

The *Rhodopsin* GPCRs were further subdivided into 13 classes as defined under *Materials and Methods* and as summarized in Table 3. This comparison used BLASTP against a database of all human GPCRs. The results of this subdivision of the *Rhodopsin* family can be seen in Table 4. A detailed description of the results is available in Supplemental Table S3, a–h.

Our results for the human genome are in good agreement with what we and others have published earlier. These earlier published numbers for the human genome are shown in parentheses in Tables 2 and 4. The numbers of GPCRs in the different genomes are also displayed graphically in Fig. 1. All numbers published previously are from Fredriksson et al. (2003c), with the exception of olfactory receptors, which are from Zozulya et al. (2001); the vomeronasal, which are from Kouros-Mehr et al. (2001), and the taste receptors type 2 (TAS2), which are from our own unpublished studies. The main discrepancies are in the olfactory and the vomeronasal groups. This is probably related to the number of pseudogenes in the Genscan predictions. It is fairly well established that there are a number of olfactory (Zozulya et al., 2001) and vomeronasal pseudogenes in the Genscan set. The vomeronasal receptors in humans have been shown to be nonfunctional pseudogenes (Kouros-Mehr et al., 2001). It is also not clear how many functional olfactory receptors there are in the human genome. We also noticed that a number of the human olfactory receptors were longer than 400 bases, and when these were investigated further, they were found to correspond to two or three olfactory receptors. These are likely to be erroneously predicted by Genscan because of the small intergenic distance for these genes. If these proteins are counted separately, the number of olfactory receptors increases from 494 to 545. In addition, in the group of TAS2, we found fewer proteins than have been reported in Genbank. BLAST searches with all known TAS2 receptors against the entire Genscan data set showed that many of these proteins are not present in the Genscan data set. This probably reflects the inability of Genscan to predict these proteins, because the sequence of these genes is present in the human genome assembly (data not shown). Considering these exceptions, the overall results indicate that the num-

TABLE 3

Description of the human data set used for subclassification of *Rhodopsin* GPCRs

The classification is based on Fredriksson et al. (2003c), except for OLF, which are from Zozulya et al. (2001).

Model Name	Number of GPCRs	Description
AMIN (α group)	42	Bioamine GPCRs binding 5-hydroxytryptamine, dopamine, histamine, trace amines, adrenalin, and acetylcholine
MEC (α group)	22	Receptors for phospholipids (EDG), melanocortin, cannabinoids, and somatostatin receptors together with three orphan GPCRs
MTN (α group)	3	Melatonin and orphan receptor GPR50
OPN (α group)	9	Opsin/putative opsin receptors and orphan receptors GPR21 and GPR52
PTGR (α group)	15	Prostaglandin receptors and orphan receptors SREB 1–3, GPR26, GPR61, GPR62, and GPR78
PEP (β group)	35	Receptors for NPY, tachykinins, neurotensin, orexin, neuromedin, NPFF, PrRP, GnRH, CCK, etc.
CHEM (γ group)	42	Bradykinin receptors and receptors/putative receptors for chemokines
MCHR (γ group)	2	Receptors for melanocyte concentrating hormone
SOG (γ group)	15	Somatostatin, opsin, and galanin receptors
LGR (δ group)	8	Orphan LGR receptors and receptors for relaxin, FSH, TSH, and LH
MRG (δ group)	8	MRG and MAS receptors
OLF (δ group)	347	Olfactory receptors according to Zozulya et al. (2001)
PUR (δ group)	42	Purin/putative purin receptors, formyl-peptide receptors, retinoic acid receptors, and orphan GPCRs

TABLE 4

Subdivision of *Rhodopsin* GPCRs

The classification is according to Fredriksson et al. (2003c). The numbers in parentheses in the human column shows the number of GPCRs from each group as published in Fredriksson et al. (2003c).

	H. sapiens	*M. musculus*	*D. reiro*	*T. rubripes*	*C. intestinalis*	*C. elegans*	*D. melanogaster*	*A. gambiae*
AMIN (α group)	44 (42)	57	122	28	23	20	21	18
MEC (α group)	18 (22)	23	35	11	17	1	1	2
MTN (α group)	3 (3)	2	6	3	3		2	2
OPN (α group)	11 (9)	6	31	21	3	1	8	12
PTGER (α group)	13 (15)	11	22	13	2			
PEP (β group)	43 (35)	43	67	32	16	31	21	29
CHEM (γ group)	43 (42)	51	77	23	7			
MCHR (γ group)	1 (2)	1	4					
SOG (γ group)	10 (15)	15	34	9	18	10	5	3
LGR (δ group)	7 (8)	8	8	3	36	1	4	3
MRG (δ group)	7 (8)	22						
OLF (δ group)	494 (347)	789	31	20				
PUR (δ group)	35 (42)	49	80	26				
Unclassified	20 (17)	32	74	35	26	60	17	8
Total	749 (607)	1109	591	224	151	124	79	77

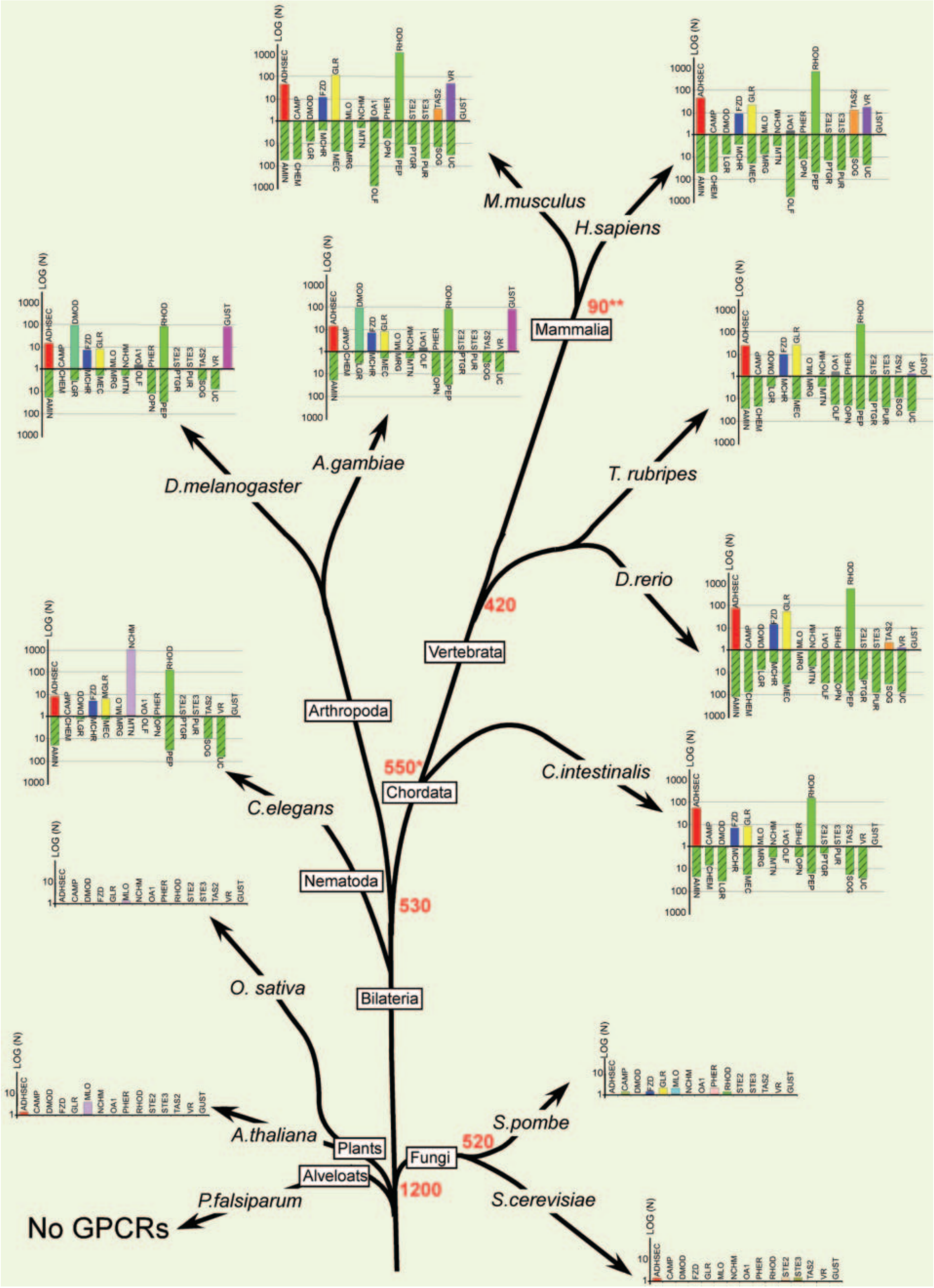

Fig. 1. Evolutionary tree with the number of GPCRs in different species indicated in graphs. The graph displays the logarithm of the number of GPCRs at the *y*-axis, with the main class above the *x*-axis and the subdivision of *Rhodopsin* GPCRs below the *x*-axis (hatched). The numbers at the nodes indicate the time in millions of years since the split at that node occurred, taken from fossil data according to Hedges and Kumar (2003); *, Dehal et al. (2002); **, Springer et al. (2003).

bers of GPCRs that we identified are in good agreement with previous data.

The results for the mouse genome are similar to those for the human genome with only a few exceptions. It is well known that there are much fewer olfactory receptors in the human genome compared with the mouse genome. This has been discussed in a recent article (Young et al., 2002), but we emphasize that there are still large uncertainties regarding the exact numbers in each genome. It is noteworthy that these receptors lie in large blocks on the chromosomes that cause problems with the assembly of the genomes as such, because many of these receptors are very similar, and it is in many cases difficult to distinguish between what is polymorphism and what is truly a unique gene during the assembly of the shotgun data. We also know that we underestimate the number of TAS2 receptors in mouse, most likely for the same reasons as discussed for the human genome. The GLR HMM picks up a considerably higher number of predicted proteins in the mouse genome than in the human. This is not related to a large expansion of classic metabotropic glutamate receptors, because these are found in similar numbers in mouse and human. This is rather related to pheromone receptors that are not found in the human genome. They show similarity to TAS1 receptors (three copies in humans) and have expanded in mouse, resulting in at least 80 receptors that are fairly similar but found on at least 11 chromosomal segments (T. K. Bjarnadottir, unpublished data). In addition, for the mouse, the number of olfactory receptors is underestimated because of multiple proteins joined together by Genscan, and the number of olfactory receptors is increased from 789 to 827 when these are considered. In a recent article, the number of GPCRs for endogenous ligands in mouse was determined to be 392 (Vassilatis et al., 2003), which is in agreement with the data set presented here, which includes 391 *Rhodopsin* GPCRs, excluding the olfactory receptors.

The fish species have not been analyzed previously with regards to the GPCR repertoire, except that it has been estimated from the genome sequencing projects that there are in total approximately 457 *Rhodopsin* GPCRs in fugu (Aparicio et al., 2002). This is significantly higher (twice the number) than what we found in the current Genscan data set. The GPCRs described by Aparicio et al. (2002) was found in a set of 27,779 predicted proteins, whereas the Genscan data set used in our study has 29,625 proteins. The reason for the large difference in the number of GPCRs could be caused by differences in the "gene-building pipeline" used by the fugu genome-sequencing group and the Genscan program. We know, however, that the number of GPCRs in the Genscan data set is similar to the published (verified manually) numbers for mouse and human, and we find it likely that the Genscan sets we used are providing a good estimate of the GPCR numbers also in fugu. It is noteworthy that zebrafish has 2- to 3-fold as many GPCRs in the main families as fugu. Compared with mammals, the zebrafish has up to twice the number of receptors, whereas fugu has approximately half the number. The *Frizzled* family is an exception, because it has approximately the same numbers in all mammals and fish. Among the subgroups of *Rhodopsin* GPCRs, the picture is similar, with zebrafish having approximately 2- to 3-fold more GPCRs than fugu, whereas the mammalian numbers are in between these. An exception to this is the olfactory receptors, which have a very small number in both fish species, and the MRG receptors that seem to be missing in fish. We have performed detailed separate searches in the fish genomes with the human MRG receptors as baits using TBLSTX and PSI-BLAST without finding any MRG-like sequence in fish (data not shown).

The *Ciona* genome has likewise not been investigated with regards to GPCRs. The main groups, which are all present in *Ciona*, contain 3- to 4-fold fewer receptors than the mammalian counterparts, again with the exception of the *Frizzled* family, which has similar numbers in *Ciona* and the vertebrate genomes. Within the *Rhodopsin* family, there seems to be several subgroups missing. It is noteworthy that the *Ciona* family does not seem to have any receptors that match the olfactory receptors in vertebrates, but melanocyte-concentrating hormone (MCH) and purine (PUR) receptors seem to be absent. It is notable that the *Ciona* family has fewer members in the GLR group, and this is partly explained by the fact that there seems to be no TAS1 genes in *Ciona* (R. Fredriksson, unpublished data).

The two insect species have the same five main families of GPCRs as mammals, fish, and *Ciona*. The main difference is that these species also have a large number of gustatory receptors (GUST) and odorant (DMOD) receptors, as reported previously (Hill et al., 2002), and several of the *Rhodopsin* subgroups are missing. As for *Ciona*, the insects seem to be lacking OLF, MCH, and PUR receptors, but also chemokine (CHEM) and prostaglandin (PTGER) binding receptors. A previous analysis described that *A. gambiae* has 276 GPCRs and *D. melanogaster* has 270 (Hill et al., 2002), which fits with the numbers presented here, 260 and 210, respectively. The difference in the number of receptors found in *D. melanogaster* is related to the fact that we found 40 fewer gustatory GPCRs compared with what was reported by Hill and colleagues (2002). In another article, it was found that *D. melanogaster* has 211 GPCRs in total (Adams et al., 2000).

The pattern of GPCRs in *C. elegans* shows that the five main families of GPCRs, ADH, SEC, FZD, GLR, and RHOD, are present, as they are in mammals, *Ciona*, insects, and fish. They do have one additional group, namely the nematode chemosensory (NCHM) receptors. These do not show any significant similarities to the olfactory receptors found in vertebrates, nor to the gustatory or the DMOD found in the arthropods. It is notable that these groups are absent in *C. elegans*. It has been reported previously that *C. elegans* has more than 800 GPCRs of the chemosensory type, of which 550 seem to be functional (Robertson, 1998). In our searches, we found approximately 1000 GPCRs of this type. It should be noted that some of the receptors belonging to chemosensory GPCRs can also be detected by the RHOD HMM as well. These were removed from the RHOD data set manually using BLASTP against a database of known nematode chemosensory receptors and the entire human *Rhodopsin* data set. Most of the subgroups of the *Rhodopsin* family are absent or only represented by one member. Only the AMIN, PEP, and SOG groups have several members in *C. elegans*.

All of the GPCR hits from species that do not belong to bilateria (i.e., plants, yeast, and *P. falciparum*) were investigated individually by additional manual inspection to verify their identity as GPCRs and also to ensure as accurate a classification as possible. In this process, seven putative GPCRs from *S. pombe*, originally identified at low scores in the HMM searches, were removed because BLASTP searches against the nonredundant database at NCBI indicated that these proteins most likely are membrane transporters. That left only two GPCRs, which are the well-known pheromone GPCRs STE2 and STE3. One putative GPCR from *S. cerevisiae*, tentatively placed in the ADH group, were found to correspond to a protein shown previously to be able to bind G-proteins in a yeast two-hybrid system (Yun et al., 1997). This protein does seem to have very low, if any, similarity to the regular 7TM receptors. It consists of a 930 amino acid long open reading frame and seems to have between 12 and 14 hydrophobic segments in a Kyle Doolitle hydrophobicity plot (data not shown). We left this protein in the data set but will not discuss it further in relation to 7TM GPCRs. Apart from this GPCR, we only identify the two STE2 and STE3 GPCRs from *S. cervisa.* No GPCRs were found in the malaria parasite *P. falciparum.* In plants, we found GPCRs of the MLO family (Devoto et al., 2003) but also one GPCR that matched the ADH/SEC model in *A. thaliana*. This GPCR was reported previously (Josefsson and Rask, 1997), but its relationship to vertebrate GPCRs has not been clearly elucidated in terms of the overall classification of GPCRs.

To more clearly visualize the results, we display the distribution of GPCRs in percentages between the different subfamilies of GPCRs in Fig. 1 (top of each diagram). The figure shows that certain families like *Frizzled* receptors have approximately the same percentages of the total number of GPCRs in all species, whereas others, like *Rhodopsin* GPCRs in mammals and *Glutamate* in mouse, have expanded in some lineages. The figure also shows the same kind of analysis of the different groups of the *Rhodopsin* family (bottom of each diagram). The database of GPCR predictions can be obtained from the authors upon request.

The *Rhodopsin* family can be divided into four groups within a total of 13 main branches, or clusters, termed α, β, γ, and δ. The largest of the four main groups (α) contains the large AMIN cluster that includes many receptors that bind monoamines such as adrenalin, serotonin, dopamine, and histamine. The amine cluster (Fig. 1) is highly represented, ranging from 18 to 57 members in bilateral species, with the exception of zebrafish, whose trace amine receptors have undergone large expansion (D. E. I. Gloriam, unpublished data). The other branches in the α group, such as the MECA branch (which includes peptide and lipid binding receptors) and the opsin branch, are also found in all bilateria, albeit not in as high numbers as the receptors in the amine branch. The prostaglandin receptors also belong to the α group, but these were only found in vertebrates, suggesting that these arose later in the evolution than the other branches of the α group. The β group contains many peptide receptors, and several of these are receptors for neuropeptides such as NPFF, NPY, orexin, neurotensin, and TRH. This group contains only one branch, and it is found in fairly similar numbers in all of the bilateria species. This may indicate that the ancestors of the receptors that bind peptides, which regulate many “higher” functions, did not arise later than, for example, the amine receptors, whose ligands are considerably less complex than the peptide ligands of the β group. There are three branches within the γ group; only one of those is found in all bilateria species. This is the SOG branch that contains receptors that bind several peptides such as opioids, RF peptides, neuropeptide W (GPR7 and -8), and somatostatins. This provides further evidence for high representation of receptors that possibly bind complex neuropeptides among prevertebrates. Indeed, there are examples of GPCRs in prevertebrates that bind, albeit with low affinity, peptides of mammalian origin, such as the NPY binding receptor in *D. melanogaster* (Li et al., 1992). This “NPY receptor” is found in our PEP group as expected. It is very difficult to track the origin of most mammalian peptides that bind GPCRs in prevertebrates because of the fact that their conserved motifs are very short and not well-preserved in the primary structure. This analysis, however, shows that it is indeed possible to track the ancestors to the mammalian GPCRs that bind peptides. The other branches in the γ group appear for the “first time” in *C. elegans,* whereas the MCH receptors are only found in vertebrates. This may suggest that the SOG branch includes the ancestors to the entire receptor repertoire in this group. The δ group has four main branches, but only one of them, the LRG, is found in all bilateria species investigated. This group in mammals includes the LH and FSH receptors that have, unlike most other *Rhodopsin* GPCRs, long N termini in addition to the 7TM regions. Indeed, many of the GPCRs that we predict to be the ancestral genes in the invertebrate species do have long N-terminal stretches that show similarities to the mammalian counterparts, even though they do not show recognizable “domains” in, for example, RPS-BLAST (data not shown). The other branches in this group were only found in vertebrates, and it is particularly interesting that the purine and the olfactory receptor branches appear for the “first time” in fish, whereas the MRG group is only found in mammals. To further verify the nature of the sequences classified as purine and olfactory GPCRs from fish, we calculated phylogenetic trees using the neighbor-joining method. All of the sequences classified as purine receptors from fugu were aligned together with the entire δ group of human *Rhodopsin* GPCRs for this calculation. This tree shows that all of the purine sequences from fugu place inside the purine cluster, and it is also clear that most of the fugu sequences have a clear ortholog in the human genome. We took a similar approach for the olfactory receptors and combined all of the fugu and zebrafish sequences classified as olfactory with 20 randomly selected human olfactory sequences and the entire human δ group and calculated a phylogenetic tree. This tree shows that all of the olfactory receptors from fugu place on the same main branch as the human olfactory receptors, although the fugu receptors seem to form clusters of their own, distinct from the human receptors. This phylogenetic analysis strongly supports the conclusion that both olfactory and purine GPCRs are present in teleost fish. These trees are available from the authors upon request. Taken together, it seems evident that the ancestors of many of the peptide and amine receptors that are found in mammals have a long evolutionary history involving multiple members. It does not seem that the gene repertoire for these has taken any drastic changes during the evolution of bilateria; rather, it seems that the numbers have undergone gradual increases in “higher” species.

EST Data. To investigate the level of expression for the different classes of GPCRs, we used the Genscan GPCRs from each species to search against an internal BLAST database containing all ESTs from that particular species. These EST data are collected from hundreds of cDNA projects, and the tissue sampling is highly different between species. We thus do not discuss the tissue origin of the ESTs in any detail but rather focus on the number of ESTs for certain families, as displayed in Figs. 2 and 3. The large numerical range of the results makes it difficult to display these data. In Fig. 2, we show the percentage of GPCRs in the Genscan data set. This percentage is multiplied by a constant (10,000), and the logarithm of that number is plotted as a hatched bar. The solid bar represents the percentage of GPCRs in the EST data set, again multiplied by 10,000 and converted to logarithms. This means that when the hatched bar is higher than the solid bar, the number of GPCRs found in the EST data set is lower than expected compared with the number of predicted GPCR genes found in the Genscan set. Because this is the case for most of the GPCRs in all species investigated, this indicates that GPCRs are generally expressed in low numbers, which is a known fact for many well-studied proteins of this family. Here, we can see that this is a general feature for most groups of GPCRs in most lineages. There are few exceptions to this, because *Rhodopsin* GPCRs in insects seems to be relatively highly expressed. To see the relationship of the expression level of GPCRs between the different families, we plotted the number of ESTs containing GPCRs from each family as a fraction of the total number of GPCR ESTs, as shown in Fig. 3. This analysis shows that the majority of the GPCR ESTs are from the *Rhodopsin* family with two exceptions: the chemosensory receptors in *C. elegans,* which constitutes almost 75% of all GPCR ESTs in this species, and the vomeronasal (VR) receptors in humans. It was a surprise to us to see that there is such a high expression of VR in humans, because these are considered pseudogenes because their mRNA does not contain a full-length open reading frame. It is particularly notable that the expression is much higher in humans compared with mouse, which has VR receptors that are clearly functional. Further investigation showed that these ESTs originate from 186 different libraries and are found in numerous tissues (data not shown).

Discussion

This analysis highlights the tremendous success of GPCRs through the evolution of "higher" species. It is evident that all major GPCR families in the human genome, according to our GRAFS system published previously (the *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled,* and *Secretin*), arose before the split of nematodes from the chordate lineage. Moreover, the majority of the GPCRs in each of the vertebrate species belong to these five families. The overall similarity of GPCR repertoire between the lineage leading to arthropods and chordates are also remarkable, not only for the main families but also for several of the subgroups within the large *Rhodopsin* family. There are only a few of the groups that are clearly lineage-specific in bilateral species. These are the chemosensory receptors in the nematodes that are not found in any other species and represent approximately 87% of the GPCRs in *C. elegans*. The gustatory receptors are only found in the two species of insects in which they represent approximately 20 and 28% of the repertoire in the fruit fly and mosquito, respectively. There are only two families, the VR and the TAS2 genes, that have arisen after the split of *Ciona* from the lineage leading to the vertebrates. Both of these groups are found in fish. No new families of GPCRs seem to have arisen during the last 450 million years in the vertebrate linage.

The *Rhodopsin* family has had the largest evolutionary success, representing approximately 60% of the entire GPCR repertoire in the bilateria species. The *Rhodopsin* family can be divided into four main groups (α-, β- γ-, and δ), with 13 main branches (Fredriksson et al., 2003c). Members within each of the four main groups are clearly found in all eight bilateria species, whereas the representation of each of the main branches is more variable. The four other main families are also found in all of the bilateria species. The *Frizzled* receptors are found in fairly constant numbers ranging from 5 to 14 members. The *Frizzled* receptors control cell fate, proliferation, and polarity that are basic functions within Metazoan development (Gho and Schweisguth, 1998) and that could contribute to the evolutionary pressure that keeps their numbers relatively constant. The *Adhesion* GPCRs were discovered rather recently, and comparatively little is known about their functional role. The long N termini of these receptors are likely to interact with other membrane-bound proteins, perhaps enabling cell-to-cell communications without soluble ligands (Kwakkenbos et al., 2004). The results show that these receptors arose early, having multiple members in nematodes, and they have at least five members in each of the bilateral species. It seems likely that GPCRs evolved the ability of this type of N-terminal–based cell-to-cell interactions long before the presence of vertebrates. The *Secretin* GPCRs have hormone-binding domains in their N termini that interact with rather large peptides. These peptides act, in most cases, in a paracrine manner, whereas the *Glutamate* GPCRs, which also have a ligand binding domain in the long N termini, interact with small molecules such as the neurotransmitter glutamate GABA, Ca^{2+} ions, and taste molecules. Our preliminary analyses of the GPCRs in prevertebrate species, indicates that the characteristic "ligand" domain within the N termini of these four main groups of GPCRs are indeed present in all of these families, at least in some of these proteins (R. Fredriksson, unpublished data). This suggests that not only the 7TM domains but also the specific functional domains within all of these groups of GPCRs appeared in prevertebrates and that their principal functions have been maintained.

There are several groups of GPCRs that have undergone seemingly rapid expansions that are species-specific. These include the olfactory receptor group (in the *Rhodopsin* family) in human and mouse, the chemosensory receptors in *C. elegans*, the gustatory receptors in insects, and the pheromone receptor group in mouse (in the *Glutamate* family). These groups share few inter-relationships, considering their amino acid identity or functional motifs. They do, however, share some general structural features, such as the absence of any functional domains beyond the 7TM regions (according to RPS-BLAST searches), they have short N termini, and they do not show clear motifs within their groups in the TM regions that can easily earmark them (data not shown). Another common feature for these groups of GPCRs that seem to have very "dynamic" gene repertoire is that they bind small ligands such as odor, taste, and chemosensory molecules. Moreover, it is interesting to note that these ligands chemically belong to groups with many structurally similar members within the organism. It is possible that the number of interaction points and the structural constraints of the

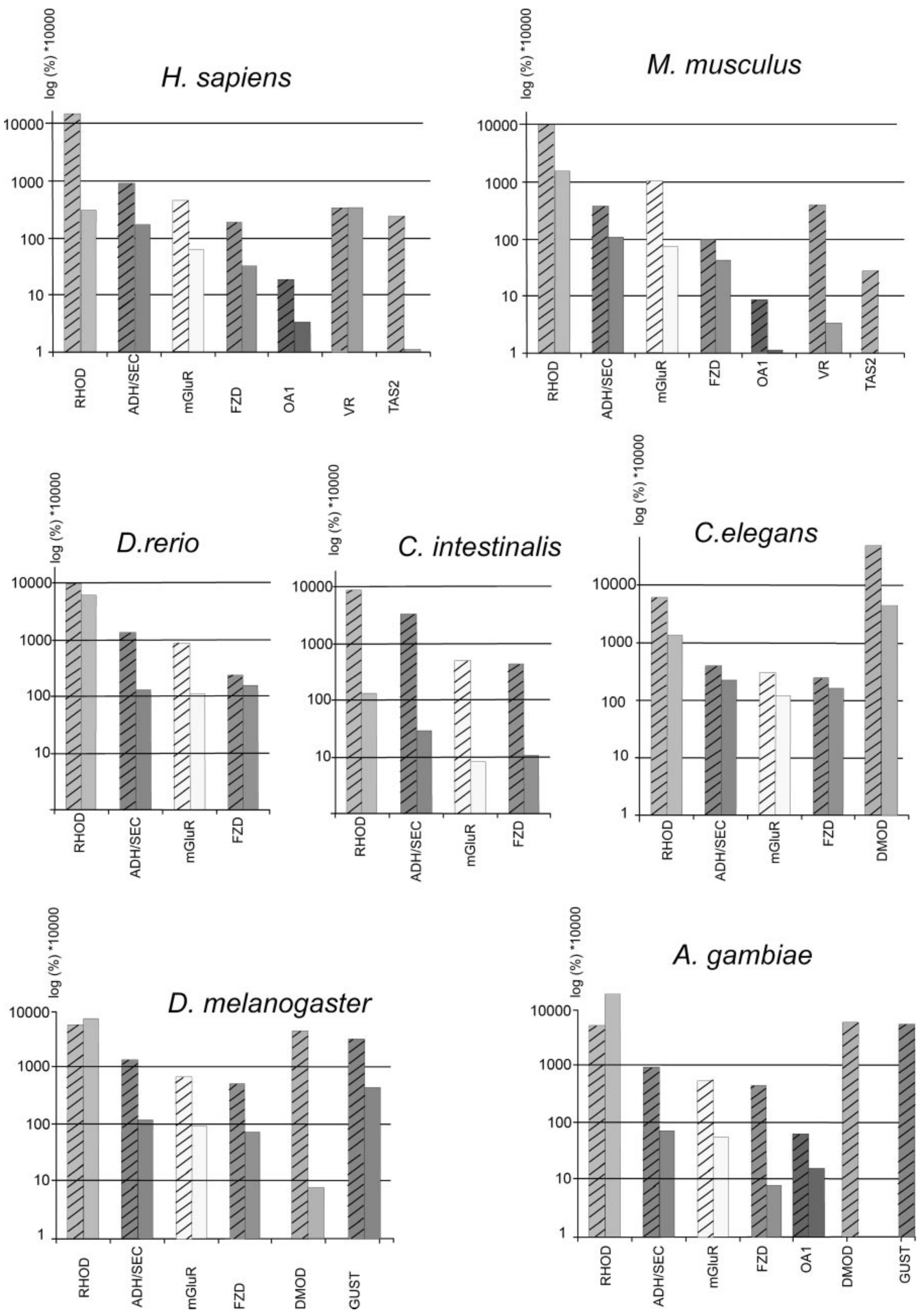

Fig. 2. Bar graphs representing the relative number of Genscan proteins containing GPCRs (hatched) and the relative number of ESTs containing GPCRs (solid) in the seven genomes in which significant amounts of EST data are available. The number of GPCRs was calculated as the percentage of GPCRs in the respective data set, EST, or Genscan data set, multiplied by 10,000 (to obtain a number greater than 1), and then we used the natural logarithm of that number to display the relative number of GPCRs. In all but two cases, the hatched graph is higher than the solid graph (i.e., the relative number of GPCRs in the Genscan data set is higher than the relative number of GPCRs in the EST data set). This would indicate that GPCRs are frequently expressed in low numbers.

ligand binding pocket of the receptors for these small molecules are fewer than for receptors that bind large ligands. Therefore, constraints on the 3D structure of these receptors that bind a variety of small ligands are likely to be relatively low. A duplicated/mutated copy of such relatively "promiscuous" receptors in an environment of multiple ligands that play an important physiological role may affect the ability of these genes to survive after duplication events. It is possible that this could be the reason for the unusual evolution of these groups of GPCRs.

There is only minor sequence homology between the GPCRs in plants and fungi and those in the bilateria. There is obviously a very large evolutionary difference between those species because it is estimated that they diverged more than 1 billion years ago. It is very interesting, however, that there is a sequence in *A. thaliana* that shows a resemblance to the ADH/SEC model. This sequence in *A. thaliana* does not have a long N terminus and is therefore missing all of the domains that are specific for the *Adhesion* receptors in mammals (Bjarnadottir et al., 2004). It is tempting to speculate that this is the only sequence that links the GPCR repertoire in bilateria with more evolutionary distant species, thus providing evidence for a common ancestor of all eukaryotic GPCRs.

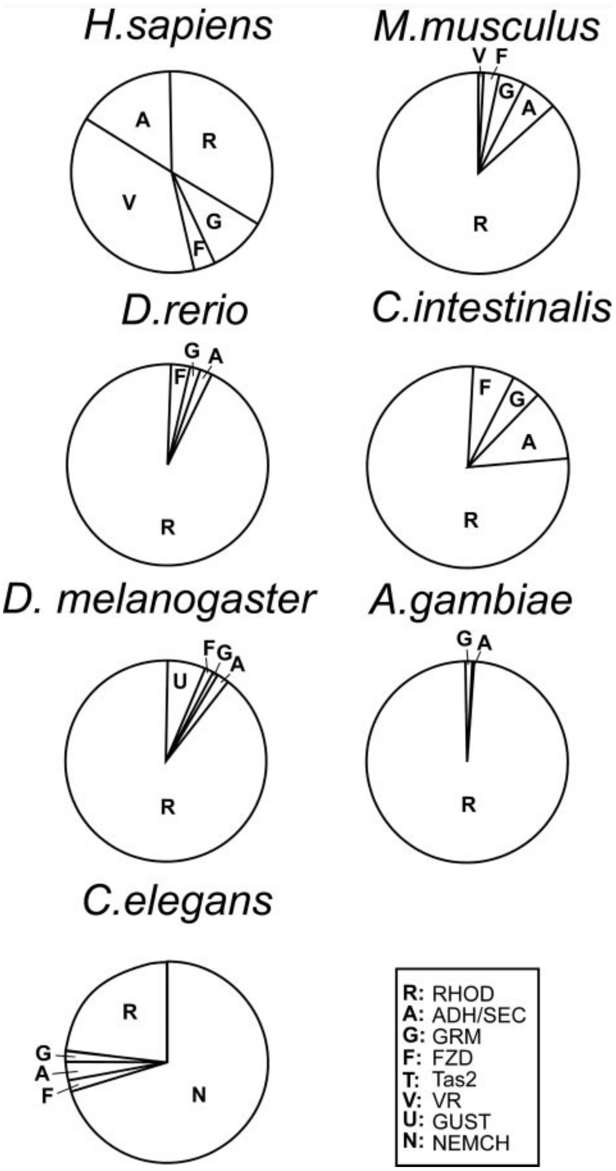

Fig. 3. Circle graphs representing the absolute number of ESTs containing GPCRs in the seven genomes in which significant amounts of EST data are available. One circle represents the total number of GPCRs in that genome (100%), and the different classes are fractions of that total.

Our analysis of the expression pattern using the growing EST databases shows that the relative number or percentage of EST sequences is lower than the percentage of gene predictions, with the exception of the *Rhodopsin* GPCRs in insects. This indicates that GPCRs are generally expressed at low levels, at least when considering the mRNA. It is remarkable that despite the high diversity within the GPCR gene family, this phenomenon of a relatively low expression level is found for all families in all of the species with only a few exceptions. The representation of GPCRs in the EST database correlates fairly well with the relative number of predicted GPCRs (i.e., the larger the number of predicted proteins, the higher the number of ESTs in the database). Another important observation from these data is that the expression level of most of the main families of GPCRs seems to be conserved between the species, even though they are separated by more than 500 million years. One could speculate that the subdivision of the functional roles between the main families of GPCRs has been similar throughout the evolution, even if the number of GPCRs in the different families have changed several-fold. This notion correlates well with the conservation of functional domains within the five main families that we mentioned earlier.

The EST pattern for human and mouse was fairly similar, except for the VR genes that had more than 1800 hits in the human EST database. The high copy number of human VR ESTs is quite remarkable considering the fact that the human VR genes are believed to be pseudogenes (Kouros-Mehr et al., 2001). These VR ESTs are found in 186 different libraries and are found in a large variety of tissues. It seems that these genes have been pseudogenized rather recently and that the promoter has not yet mutated to become nonfunctional. The relatively high expression of the VR pseudogenes in humans compared with the VR expression in mouse can possibly be related to a lack of negative feedback from a functional VR protein that subsequently down-regulates the expression. There are also some examples that the expression of certain genes or entire groups seems to be very low or nonexistent. It was, for example, highly surprising that ESTs for the gustatory and odorant receptors from the mosquito seem to be completely missing in the databases. Because the total number of ESTs is approximately the same in the two species, this could indicate a much lower expression of gustatory and odorant receptors in the mosquito. It was also surprising that the TAS2 receptors in both mammals did only match very few ESTs. This could be caused by the fact that these receptors have a very restricted expression pattern and that these "specific" tissues or cells are not represented in the EST databases.

In summary, this analysis provides an extensive overview of the expansion of the repertoire of GPCRs in many important genomes. The analysis covers a larger number of genes than has been simultaneously analyzed previously in evolutionary perspective for GPCRs and perhaps any protein family. The databases we generated provide a tremendously valuable source for further detailed analysis, assembly, and annotation of individual GPCR genes.

Acknowledgments

We thank Christian A. Murray, Uppsala University, for assistance with handling the EST data. We thank Thóra K. Bjarnadóttir and David E. Gloriam, both from Uppsala University, for providing insight into their preliminary results regarding detailed analysis, assembly, and annotation of several subgroups of GPCRs.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science (Wash DC)* **287:**2185–2195.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science (Wash DC)* **297:**1301–1310.

Bjarnadottir TK, Fredriksson R, Hoglund PJ, Gloriam D, Lagerström MC, and Schiöth HB (2004) Identification of G-protein coupled receptors in mouse belonging to adhesion family; expression pattern, phylogeny and domains. *Genomics* **84:**23–33.

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.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:**78–94.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science (Wash DC)* **298:**2157–2167.

Devoto A, Hartmann HA, Piffanelli P, Elliott C, Simmons C, Taramino G, Goh CS, Cohen FE, Emerson BC, Schulze-Lefert P, et al. (2003) Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family. *J Mol Evol* **56:**77–88.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14:**755–763.

Fredriksson R, Gloriam DE, Höglund PJ, Lagerström MC, and Schiöth HB (2003a) Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions. *Biochem Biophys Res Commun* **301:**725–734.

Fredriksson R, Hoglund PJ, Gloriam DE, Lagerstrom MC, and Schioth HB (2003b) Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives. *FEBS Lett* **554:**381–388.

Fredriksson R, 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.

Fredriksson R, Lagerström MC, Lundin LG, and Schiöth HB (2003c) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups and fingerprints. *Mol Pharmacol* **63:**1256–1272.

Gho M and Schweisguth F (1998) Frizzled signalling controls orientation of asymmetric sense organ precursor cell divisions in *Drosophila*. *Nature (Lond)* **393:**178–181.

Hagen DC, McCaffrey G, and Sprague GF Jr (1986) Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. *Proc Natl Acad Sci USA* **83:**1418–1422.

Harmar AJ (2001) Family-B G-protein-coupled receptors. *Genome Biol* **2:**3013.1–3013.10.

Hedges BS and Kumar S (2003) Genomic clocks and evolutionary timescales. *Trends Genet* **19:4**:200–206.

Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, and Zwiebel LJ (2002) G protein-coupled receptors in *Anopheles gambiae*. *Science (Wash DC)* **298:**176–178.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science (Wash DC)* **298:**129–149.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature (Lond)* **431:**931–945.

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:**63.1–63.16.

Josefsson LG (1999) Evidence for kinship between diverse G-protein coupled receptors. *Gene* **239:**333–340.

Josefsson LG and Rask L (1997) Cloning of a putative G-protein-coupled receptor from *Arabidopsis thaliana*. *Eur J Biochem* **249:**415–420.

Kolakowski LF Jr (1994) GCRDb: a G-protein-coupled receptor database. *Receptors Channels* **2:**1–7.

Kouros-Mehr H, Pintchovski S, Melnyk J, Chen YJ, Friedman C, Trask B, and Shizuya H (2001) Identification of non-functional human VNO receptor genes provides evidence for vestigiality of the human VNO. *Chem Senses* **26:**1167–1174.

Kwakkenbos MJ, Kop EN, Stacey M, Matmati M, Gordon S, Lin HH, and Hamann J (2004) The EGF-TM7 family: a postgenomic view. *Immunogenetics* **55:**655–666.

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.

Li XJ, Wu YN, North RA, and Forte M (1992) Cloning, functional expression and developmental regulation of a neuropeptide Y receptor from *Drosophila melanogaster*. *J Biol Chem* **267:**9–12.

Marsh L and Herskowitz I (1988) STE2 protein of *Saccharomyces kluyveri* is a member of the rhodopsin/beta-adrenergic receptor family and is responsible for recognition of the peptide ligand alpha factor. *Proc Natl Acad Sci USA* **85:**3855–3859.

Okada T and Palczewski K (2001) Crystal structure of rhodopsin: implications for vision and beyond. *Curr Opin Struct Biol* **11:**420–426.

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.

Shen B, Samaraweera P, Rosenberg B, and Orlow SJ (2001) Ocular albinism type 1: more than meets the eye. *Pigment Cell Res* **14:**243–248.

Springer MS, Murphy WJ, Eizirik E, and O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci USA* **100:**1056–1061.

The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature (Lond)* **408:**796–815.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science (Wash DC)* **282:**2012–2018.

Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:**4673–4680.

Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT, Brown A, Rodriguez SS, Weller JR, Wright AC, et al. (2003) The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci USA* **100:**4903–4908.

Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, and Trask BJ (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet* **11:**535–546.

Yun CW, Tamaki H, Nakayama R, Yamamoto K, and Kumagai H (1997) G-protein coupled receptor from yeast *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun* **240:**287–292.

Zozulya S, Echeverri F, and Nguyen T (2001) The human olfactory receptor repertoire. *Genome Biol* **2:**18.1–18.12.

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