# Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives

Robert Fredriksson, Pär J. Höglund, David E.I. Gloriam, Malin C. Lagerström, Helgi B. Schiöth*

*Department of Neuroscience, Uppsala University, BMC, Box 593, 751 24 Uppsala, Sweden*

Received 27 June 2003; revised 9 September 2003; accepted 9 October 2003

First published online 22 October 2003

Edited by Robert B. Russell

**Abstract** **We report seven new members of the superfamily of human G protein-coupled receptors (GPCRs) found by searches in the human genome databases, termed GPR100, GPR119, GPR120, GPR135, GPR136, GPR141, and GPR142. We also report 16 orthologues of these receptors in mouse, rat, fugu (pufferfish) and zebrafish. Phylogenetic analysis shows that these are additional members of the family of rhodopsin-type GPCRs. GPR100 shows similarity with the orphan receptor SALPR. Remarkably, the other receptors do not have any close relative among other known human rhodopsin-like GPCRs. Most of these orphan receptors are highly conserved through several vertebrate species and are present in single copies. Analysis of expressed sequence tag (EST) sequences indicated individual expression patterns, such as for GPR135, which was found in a wide variety of tissues including eye, brain, cervix, stomach and testis. Several ESTs for GPR141 were found in marrow and cancer cells, while the other receptors seem to have more restricted expression patterns.**


*Key words:* Orphan; Human genome; Rhodopsin; G protein-coupled receptor; Evolution

## 1. Introduction

G protein-coupled receptors (GPCR) are integral membrane proteins with seven α-helices. The superfamily of GPCRs is among the largest and most diverse families of proteins in mammals [1,2]. Their functions are highly variable as they play an important role in the physiology of all major peripheral organ systems and also in the brains of 'higher' vertebrates. GPCRs are located at the cell surface and are responsible for translation of an endogenous signal into an intracellular response through heterotrimeric G proteins that target other proteins, often enzymes that influence the level of intracellular messengers. The vital role of GPCRs in various central and peripheral physiological events has made them important targets for drug discovery. It has been estimated that 40–45% of all modern drugs are targeted at these receptors [3].

The rhodopsin family (clan A) is the largest group within the superfamily of GPCRs [4]. Their natural ligands are highly diverse, comprising biogenic amines (such as adrenaline, dopamine, histamine, and serotonin), peptides (such as angiotensins, bradykinins, somatostatins, and melanocortins), large proteins (such as luteinizing hormone, follicle-stimulating hormone, and thyroid-stimulating hormone), nucleosides and nucleotides (adenosine, ATP, UTP, and ADP), lipids and eicosanoids (such as leukotrienes, prostaglandins, and cannabinoids) and photons. Moreover, it has also been suggested that there exist over 900 genes for olfactory receptors in the human genome [5]. A large number of these are pseudogenes and the specific roles of only a few of the receptors are known.

The rhodopsin family of GPCRs has been very much studied because of the intense pharmaceutical interest in amine binding receptors. The number of drugs for other GPCRs is increasing, in particular for those receptors that bind peptides. The therapeutic potential of most rhodopsin GPCRs has, however, not yet been exploited. Many of these receptors are still orphans, without any known ligand. The diversity of the known genes that encode GPCRs is so large that it cannot be excluded that more such genes could be found in the human genome. Today, more than 2 years after the first presentation of the draft human genome sequence [1,2], the assemblies are still being adjusted allowing better predictions of putative proteins, in particular for those with complex genomic structures. The number of expressed sequence tags (ESTs) has also increased rapidly during the last 2 years and currently there are over 5 million human EST sequences in the NCBI database. The EST information provides crucial information on whether predicted genes are functional and for prediction of their physiological role.

Recently, we performed a large-scale charting of the GPCRs in the human genome [6]. These sequences provided a large and highly variable sequence dataset that we used to create hidden Markov models (HMM) to search for additional genes. In this study we searched the human NCBI and Celera genome databases for new members and identified seven new human GPCRs that belong to the rhodopsin (class A) family. We identified several orthologues of these receptors in mouse, rat, fugu (pufferfish) and zebrafish. We also studied the genomic structure of these proteins, predictive protein structures, phylogenetic relationships and EST expression patterns.

*Corresponding author. Fax: (46)-18-51 15 40.
*E-mail address:* helgis@bmc.uu.se (H.B. Schiöth).

0014-5793/03/$22.00 
doi:10.1016/S0014-5793(03)01196-7

## 2. Materials and methods

### 2.1. Identification of receptors in the Celera database

Sequences of known GPCRs from the rhodopsin family PPYR1 (NP_005963), TACR2 (NP_001048), GPR3 (AAH32702), EDG8 (AAH34703), HTR2B (S43687), HTR1D (P28221), CHRM2 (NP_000730), DRD3 (1705199A), TA3 (AAK71240), SSTR3 (AAA60592), GALR1 (NP_001471), and SLT (JC7695) were downloaded from the GenBank database at http://www.ncbi.nlm.nih.gov. The Celera database http://www.celera.com was searched using the amino acid sequence from each of these known GPCRs as bait for BLASTP [7] searches. The novelty of the new sequences was confirmed by searching all hits against our internal and the public database at NCBI.

### 2.2. Identification of human receptors in the NCBI database

A set of 262 human rhodopsin GPCRs [6] was used as a seeding material for this study. We removed the N- and C-termini, as identified by RPS-BLAST searches at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, from these sequences. The truncated receptor sequences were aligned using ClustalW 1.82 [8]. From the alignments, a HMM was constructed using the HMMER 2.2 package [9]. The model was constructed using HMMbuild with default settings and calibrated using HMMcalibrate. The Genscan protein dataset, from assembly 28 of the public human genome sequences, was downloaded from the NCBI ftp site ftp.ncbi.nlm.nih.gov/genbank/ and searched against the HMMs, using HMMsearch, with a cut-off at $E = 1e\text{-}4$. The new sequences were confirmed by searching all hits against the public databases at NCBI and against the Celera database using the BLAST package [7].

### 2.3. Identification of human EST clones

The genomic DNA sequences of the new GPCRs were searched against the human EST database at www.ncbi.nlm.nih.gov/BLAST/ using BLASTN and against http://genome.ucsc.edu/ using BLAT with a cut-off at $E = 1e\text{-}12$. The alignments with the identified EST sequences were manually inspected to ensure correct identity.

### 2.4. Identification of orthologous receptors from other species

We used the same parameters for all the BLAST searches. The sequences were masked for low complexity regions. The cut-off value for the hits was $E = 10$, and the lists of positive hits were manually inspected, mainly based on the sequence alignments, to ensure the identity of the hits. No sequences with $E$ values above $10^{-9}$ were included for further analyses.

*2.4.1. Mouse.* The mouse orthologues were identified at http://genome.ucsc.edu using the amino acid sequences of all the new human GPCRs as baits. Searches were performed using BLAT against the translated version of the DNA sequence of the assembled genome, in order to reduce the effect of mis-predictions in the Genscan set. The splice sites were assumed to be conserved between mouse and human sequences. These were subsequently verified, to best extent, using EST or cDNA sequences from http://www.ncbi.nlm.nih.gov.

*2.4.2. Rat.* The amino acid sequences of the new human GPCRs were used as baits in BLASTP searches at http://www.ncbi.nlm.nih.gov/BLAST/ in the nr database to identify rat orthologues.

*2.4.3. Zebrafish.* For zebrafish, searches were performed at http://www.ensembl.org/Danio_rerio/ against the 3× whole genome shotgun assembly using the amino acid sequences of all the new human receptors as baits using TBLASTX against the Genscan-predicted cDNA database and TBLASTN against the assembled genome. The three best hits from each query were searched against a local database, containing a nearly complete set of human GPCRs and the seven new receptors presented here, to identify the zebrafish receptors. A zebrafish receptor that hit the receptor used as bait was considered a positive hit.

*2.4.4. Fugu.* For identifying fugu orthologues, searches were performed at http://www.ensembl.org/Fugu_rubripes/ against the 319 Mb assembly using the amino acid sequences of the new human GPCRs with TBLASTX against the Genscan-predicted cDNA database (36 000 proteins) and TBLASTN against the assembled genome. Positive hits were verified the same way as for zebrafish.

### 2.5. Verification of the predicted coding regions

The machine-predicted coding regions, predicted using Genscan [10], were verified by assembling the human EST sequences and the full genomic DNA sequence using SeqMan from the DNASTAR package. Here, the DNA sequences from the human genome were considered correct, while the EST and mRNA sequences were used to correct the predicted exon–intron boundaries. When sufficient coverage could not be obtained using human EST or mRNA sequences, a combination of other vertebrate mRNA and EST sequences as well as the machine-predicted mouse orthologues from http://genome.ucsc.edu/ were used to verify exon–intron boundaries. Protein alignments with closely related receptors were also used. This process is described in detail in Section 3 for each of the sequences.

### 2.6. Phylogenetic analysis

To avoid input order bias, the dataset was randomized 20 times with regard to sequence input order using a program called Randfasta (http://www.medfarm.neuro.uu.se/schioth.html). These 20 datasets, containing the full set of sequences but in different order, were all aligned using the UNIX version of ClustalW 1.82 [8]. The default alignment parameters were applied. The 20 alignments were also bootstrapped 50 times using SEQBOOT from the Win32 version of the PHYLIP 3.6 package [11] to obtain a total of 1000 different alignments. Protein distances were calculated using PROTDIST from the Win32 version of the PHYLIP 3.6 package. The Jones–Taylor–Thornton matrix was used for the calculation. The trees were calculated on the 20 different distance matrices, previously generated with PROTDIST, using NEIGHBOR from the Win32 version of the PHYLIP 3.6 package, resulting in 20 files with 50 trees each. The 20 files were merged using the Gnu UNIX cat command and the resulting file was analyzed using CONSENSE from the Win32 version of the PHYLIP 3.5 package to get a bootstrapped consensus tree. The trees were plotted using TREEVIEW (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). Maximum parsimony (MP) trees were calculated from the same input files that were used for PROTDIST using PROTPARS from the Win32 version of the PHYLIP 3.6 package. The trees were unrooted and calculated using ordinary parsimony and the topologies were obtained using the built-in tree search procedure. Consensus trees were calculated and plotted as described above.

### 2.7. Construction of transmembrane (TM) region alignments

The new receptors were aligned with the five most closely related human GPCRs, as obtained by BLASTP searches in the human genome database at http://www.ncbi.nlm.nih.gov/BLAST/ using MEGAlign from the DNASTAR package (see Table 2). In MEGAlign, the ClustalW method was used for multiple alignments with a gap penalty of 10, a gap length penalty of 0.20 and delay divergent sequences of 30%. The slow-accurate method was used for the initial pairwise alignments. The protein weight matrix was Blossom 30. When necessary, alignments were optimized by manual editing. To determine the borders of the TM regions the protein sequence of bovine rhodopsin (SwissProt P02699) was included in each alignment. From these alignments the borders of the TM regions were assigned as defined by the crystal structure of bovine rhodopsin [12]. The N- and C-termini, together with all loops, were then manually removed from the alignments, and the sequence of bovine rhodopsin was subsequently removed.

## 3. Results

Our strategy to find new GPCRs was (1) to use BLASTP searches in the Celera database using individual rhodopsin GPCR genes and (2) to create HMMs and search the human Genscan datasets that were downloaded from the NCBI ftp site. Both methods have been successful in finding new genes encoding adhesion GPCRs [13,14]. The HMMs were constructed through alignment of the TM regions of 262 human rhodopsin-like GPCRs [6]. The new sequences found in the NCBI dataset were confirmed in the Celera genome database and vice versa. The searches resulted in seven new human sequences. We approached the HUGO, Gene Nomenclature Committee at University College London and they confirmed that the sequences were unique and not public. One of the sequences, GPR100, had previously been assigned a GPR

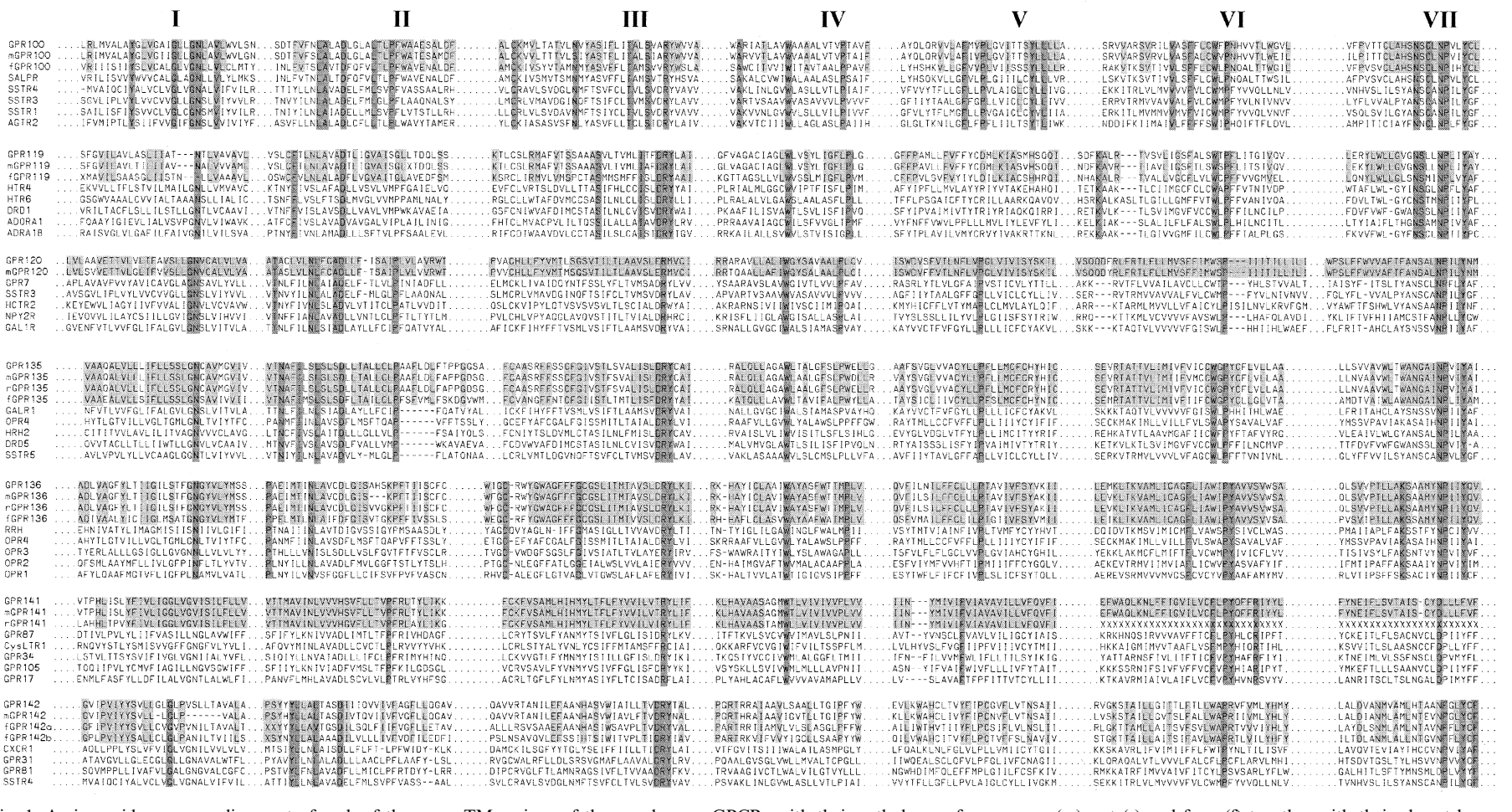


Fig. 1. Amino acid sequence alignment of each of the seven TM regions of the new human GPCRs with their orthologues from mouse (m), rat (r) and fugu (f) together with their closest human (h) BLAST hits made using the ClustalW algorithm using MEGAlign from the DNAStar package. The extra- and intracellular loops are replaced with dots. The alignments were made together with the sequence of the bovine rhodopsin receptor and the boundaries of the TM regions were determined according to those predicted by the crystal structure [12]. We used a well conserved residue in each TM region according to the following list to lock (or align) the sub-alignments to each other: N (in gNs) in TM1; D (sD1) in TM2; Y (DRY) in TM3; W in TM4; P (LLPF) in TM5; P (CWLP) in TM6; Y (NPxxY) in TM7. Light gray marks conserved residues among orthologues proteins while dark gray shows residues conserved between all proteins within the alignment. X denotes an unknown amino acid.

number under confidentiality to HUGO, by a group unknown to us. The committee provided the other receptors with new GPR numbers upon our request. We subsequently made the seven new GPCRs, both protein and DNA sequences, public through submission to the NCBI database.

It is known that the Genscan software used for predicting coding regions for the human genome project has the capacity to predict approximately 80% of the splice sites correctly [10]. Therefore, we verified the coding regions, to the extent it could be done, by using mRNA and EST sequences from a variety of vertebrates, mainly rodents and primates. The full genomic sequences of the predicted GPCRs were used as baits to identify mRNA and EST sequences using BLASTN at www.ncbi.nlm.nih.gov/BLAST/ and BLAT at http://genome.ucsc.edu/. Below we show the origin and how each protein was assembled.

The predicted sequences of hGPR100 consist of one single coding exon. We found only one EST from this receptor, covering amino acids 71–211, 34% of the coding region. In the mouse genome assembly the full-length mGPR100 was identified as a predicted protein having 78% amino acid identity to hGPR100. As can be seen from Fig. 1, the TM regions are almost 100% conserved between the two species, while the main differences are in the N- and C-termini. In the fugu genome assembly fGPR100 was identified as a full-length predicted protein with 35% amino acid identity to hGPR100, fGPR100 has the closest similarity to SALPR (somatostatin- and angiotensin-like peptide receptor), although just marginally higher then GPR100 ($E = 1\text{-e}91$ vs. $E = 1\text{e-}80$, 55% vs. 35%). This is also seen in the phylogenetic analysis (Fig. 2) where fGPR100 is placed closest to SALPR but still close to hGPR100. One speculative interpretation is that the predicted fugu protein represents a common ancestor to both SALPR and GPR100. Sequences from intermediate species, like the chicken, will be needed to resolve this question phylogenetically. The zebrafish genome was mined using fGPR100 as bait in repeated TBLASTN and BLASTP searches without identification of the zebrafish orthologue.

hGPR119 has one EST in the public databases. This EST covers about 60% of the coding region and since hGPR119 seems to have one coding exon only, wrongly predicted splice sites are not an issue with this protein. mGPR119 was identified as a full-length predicted protein in the Genscan dataset from the mouse genome assembly, with 82% amino acid identity to hGPR119. The fugu orthologue of hGPR100, fGPR119, was identified as a full-length Genscan predicted protein, with 37% amino acid identity to hGPR119. The zebrafish genome was mined using fGPR119 as bait in repeated TBLASTN and BLASTP searches without identification of the zebrafish orthologue.

hGPR120 consists of four coding exons, and we found two ESTs corresponding to this receptor, covering the first 180 amino acids of the receptor. The first splice site is located 189 amino acids downstream in the receptor and none of

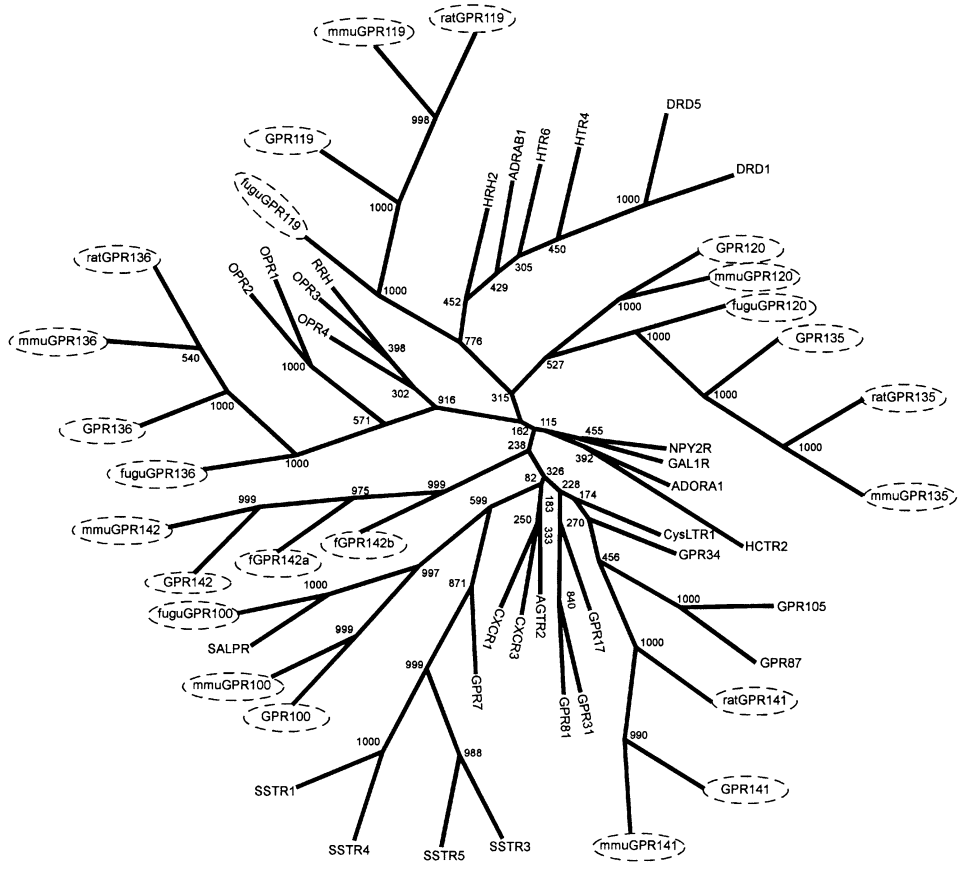


Fig. 2. Phylogenetic analysis of the new human GPCRs and their species orthologues, together with their closest BLAST hits. The alignment was constructed using ClustalW 1.82, and the tree calculated using the neighbor joining method with PROTDIST, NEIGHBOR and CONSENSE from the PHYLIP 3.6 package. The alignments were bootstrapped 1000 times as described in Section 2. The numbers on the branches are bootstrap replicas. The new receptors described in this work are marked with circles.

the splice sites could therefore be verified, although alignment with closely related receptors (Fig. 1) indicates that the splice sites are predicted correctly in the original Genscan protein. Also, the predicted protein mGPR120, with 86% amino acid identity to hGPR120, has the splice sites predicted identically to hGPR120. Despite TBLASTN searches against the assembled genomes of both fugu and zebrafish, together with BLASTP searches against the Genscan sets from the two species, no teleost orthologue of the mammalian GPR120 was identified.

From GPR135 we identified six EST or cDNA clones that together provided at least three-fold coverage of most of the predicted coding region, excluding the first exon. When compared to mGPR135, as identified as a predicted protein, it was clear that the first exon in the predicted hGPR135 was not included in mGPR135. When this exon from hGPR135 was searched against the assembled mouse genome, no significant hits were found, while all other parts of GPR135 were clearly conserved between the two species. The first exon of the predicted hGPR135 was therefore removed. mGPR135 has 81% amino acid identity to hGPR135 and 96% identity to rGPR135, which was identified as a predicted sequence in the nr database at http://www.ncbi.nlm.nih.gov/BLAST/. fGPR135 was found as a fragment of about 140 amino acids in the predicted protein dataset. The remaining part of this protein was identified using TBLASTX searches in the fugu scaffolds from the genome assembly using human, rat and mouse GPR135 as baits. TBLASTN searches in the zebrafish genome assembly using fGPR135 as bait gave no significant hits.

From GPR136 we found four cDNA clones in the databases that when assembled covered everything except the first 150 amino acids of the protein. We found an obvious error in the predicted sequence where one cryptic exon of 60 amino acids at position 146 was included by Genscan. From protein alignments with related proteins, including species orthologues, this error was discovered and the exon was removed from the sequence, leaving only the first 90 amino acids and the first two splice sites unverified. The other splice sites were verified and corrected by the cDNA data. About 150 amino acids of mGPR136 were identified as a predicted protein in the mouse assembly at http://genome.ucsc/edu/ and the rest of the protein was obtained by performing TBLASTN searches against the assembled genome using hGPR136 as a bait. The protein mGPR136 has 96% amino acid identity to hGPR136. rGPR136 was identified as a full-length cDNA clone in the nr database at http://www.ncbi.nlm.nih.gov/, having 92% sequence identity to hGPR136. As for mGPR136, fGPR136 was found as a partial sequence in the predicted protein dataset. The rest of the protein was identified by TBLASTN searches against the fugu scaffolds and the fGPR136 was found to be 72% identical to hGPR136. TBLASTN searches against the zebrafish assembly at ENSEMBL revealed two fragments representing different parts of zGPR136. In total only around 80 amino acids were identified from this protein and this protein is not used for further analysis.

GPR141 has nine EST clones in the databases. These cover the first 100 amino acids of the protein. Since GPR141 seems to have one coding exon only, wrongly predicted splice sites are not an issue with GPR141. mGPR141 was found as a full-length predicted protein in the UCSC database. The protein mGPR141 has 67% amino acid identity to hGPR141. rGPR141 has also 67% identity to hGPR141 and was found as a predicted protein in the nr database at NCBI. TBLASTN and BLASTP searches in the fugu and zebrafish databases revealed only 89 amino acids of fGPR141 but no zebrafish orthologue of GPR141. The fGPR141 was located at the border of a scaffold in the assembled genome and no other scaf-

Table 1
Summary of some of the main features found in the novel GPCRs, their accession numbers and protein IDs from both Celera Discovery system and the public NCBI Genscan dataset among others

| Name | Celera number | NCBI Genscan number | GenBank accession number | Length (aa) | Number of exons | Chromosome position |
|---|---|---|---|---|---|---|
| hGPR100 | not present | Hs1_5015_31_3_1 | AY288415 | 374 | 1 | 1q22 |
| mGPR100 | not present | Mm3_39274_30_194_2 | Ay288422 | 412 | 1 | chr3:89443832–89445180 |
| fGPR100 | – | Scaffold_243 | AY288410 | 408 | 1 | – |
| hGPR119 | not present | HsX_11943_31-37_1 | AY288416 | 335 | 1 | Xp26.1 |
| mGPR119 | not present | MmX_39742_30_78_1 | AY288423 | 335 | 1 | chrX:34322403–34323407 |
| rGPR119 | – | XM_229126 | AY288429 | 468 | – | – |
| fGPR119 | – | scaffold_615 | AY28841 | 393 | 2 | – |
| hGPR120 | not present | Hs10_30314_31_12_1 | AY288417 | 377 | 3 | 10q23.33 |
| mGPR120 | mCP5702 | Mm19_39729_30_51_10 | AY288424 | 361 | 3 | chr19:37744355–37761464 |
| hGPR135 | hCP1629103.1 | Hs14_26604_31_172_1 | AY288418 | 494 | 1 | 14q23.1 |
| mGPR135 | mCP26319 | Mm12_39591_30_109_4 | AY288425 | 457 | 1 | chr12:66714313–66715683 |
| rGPR135 | – | XM_234276 | AY288430 | 458 | – | – |
| fGPR135 | – | scaffold_3906 | AY301619 | 444 | 1 | – |
| hGPR136 | hCP1626176 | Hs6_7559_28_36_4 | AY288419 | 472 | 6 | 6p12.3 |
| mGPR136 | mCP48018 | Mm17_39695_30_13_3 | AY288426 | 351 | 6 | chr17:41575341–41610708 |
| rGPR136 | – | XP_236960 | AY288431 | 497 | – | – |
| fGPR136* | – | scaffold_3420 | AY288412 | 259 | 4 | – |
| hGPR141 | hCP1781674 | Hs7_7976_31_328_8 | AY288420 | 299 | 1 | 7p14.1 |
| mGPR141 | mCP9530 | Mm13_39618_30_47_2 | AY288427 | 305 | 1 | chr13:19157113–19158027 |
| rGPR141* | – | Xm_225424.1 | AY288432 | 247 | – | – |
| hGPR142 | not present | Hs17_10798_31_10_3 | AY288421 | 462 | 4 | 17q25.1 |
| mGPR142 | mCP32737 | Mm11_39561_30_546_4 | AY288428 | 365 | 3 | chr11:115705944–115713762 |
| fGPR142a* | – | scaffold_3880 | AY288413 | 371 | 2 | – |
| fGPR142b* | – | scaffold_140 | AY288414 | 383 | 2 | – |

All the genes are full-length except those denoted by an asterisk.

folds containing this protein could be found. This indicates that the rest of fGPR141 is not present in the current genome sequence of fugu. This protein is not used for further analysis.

GPR142 has no corresponding ESTs in the public databases, although alignment with closely related proteins suggests correct prediction of the splice sites in the original Genscan protein. Also alignments with species orthologues shows the predicted exons are the actual conserved parts and the splice sites have been predicted essentially the same way in the different species. mGPR142 was found as a full-length predicted protein in the UCSC database, being 68% identical to hGPR142. Searches in the fugu predicted proteins database revealed two partial variants of this protein in the predicted protein dataset at ENSEMBL, both equally identical to hGPR142 and mGPR142. These two variants had identical exon–intron organization and were designated fGPR142a and fGPR142b. The fugu proteins were extended using TBLASTN searches against the fugu scaffolds in the ENSEMBL database and everything except the first exon, 55 amino acids in humans, was recovered.

Additionally, the genomes of two tunicates, *Ciona intestinalis* and *Ciona savignyi*, and the fruit fly (*Drosophila melanogaster*), the malaria parasite (*Plasmodium falciparum*), the African malaria mosquito (*Anopheles gambiae*), the nematode *Caenorhabditis elegans*, and the plant *Arabidopsis thaliana* were searched for the receptors but we did not find any clear orthologues of our new genes in these genomes (data not shown).

A summary of the results is given in Table 1, where we list the name, accession numbers, chromosomal positioning, Genscan ORF numbers, number of exons, length in amino acids, and tissue distribution as suggested by EST data. In Table 2 we list the closest BLAST hits with BLAST score, percentage identity, divergence and an overview of their tissue expression pattern.

Fig. 1 shows alignments with the TM regions of the new human proteins together with their orthologues from other species and the five closest BLASTP hits from the nr database. It is obvious from the alignments that all these proteins have ancient origin and lack other family members in any of the public databases, while considering the relatively high degree of conservation among the orthologues and the relatively few conserved residues between the new proteins and the closest BLASTP hits. Phylogenetic trees were constructed using the MP and neighbor joining (NJ) methods. When the receptors were analyzed with clusters of the five main families of GPCRs (glutamate, rhodopsin, adhesion, frizzled/taste2 and secretin) they clearly belonged to the rhodopsin family (data not shown). We therefore selected the five closest human BLAST hits to compile a phylogenetic tree of these new rhodopsin receptors.

In Fig. 2 we show the NJ tree. The trees calculated with the different methods are essentially identical. The only difference between the trees is that in the MP tree the branching order of ADORA1, HCTR2, GAL1R and NPY2R is not resolved and also the larger group containing the chemokine-like receptors,

Table 2
Overview of the most related genes to the new receptors in the human genome and their tissue expression

| Name | Five best BLAST hits | BLAST score | % Identity | Divergence | Overview of tissue expression |
|---|---|---|---|---|---|
| hGPR100 | SALPR (NP_057652.1) | 2e-70 | 38.1 | 101 | Marrow |
| | SSTR4 (NP_001043.1) | 2e-37 | 27.0 | 174 | |
| | SSTR3 (NP_001042.1) | 2e-37 | 26.8 | 174 | |
| | SSTR1 (NP_001040.1) | 2e-34 | 23.8 | 186 | |
| | AGTR2 (NP_000676.1) | 7e-34 | 22.3 | 200 | |
| hGPR119 | HTR4 (NP_000861) | 7e-22 | 25.7 | 214 | Pancreas |
| | HTR6 (NP_000862.1) | 2e-21 | 24.2 | 195 | |
| | DRD1 (NP_00785.1) | 6e-21 | 23.3 | 204 | |
| | ADORA1 (NP_009195.1) | 5e-20 | 20.6 | 202 | |
| | ADRAB1 (NP_000670.1) | 7e-19 | 18.2 | 245 | |
| hGPR120 | HCRTR2 (NP_001517.1) | 2e-21 | 20.7 | 228 | Stomach |
| | GALR1 (NP_001471.1) | 3e-19 | 18.0 | 213 | |
| | SSTR3 (NP_001042.1) | 1e-18 | 19.9 | 218 | |
| | GPR7 (NP_005276.1) | 7e-18 | 18.8 | 206 | |
| | NPY2R (NP_000901.1) | 2e-17 | 23.3 | 194 | |
| hGPR135 | GALR1 (NP_001471.1) | 5e-24 | 26.6 | 153 | Neuronal, stomach, eye, reproductive organs (male and female) |
| | OPR4 (NP_150598.1) | 1e-22 | 20.1 | 222 | |
| | HRH2 (NP_071640.1) | 3e-22 | 21.7 | 202 | |
| | DRD5 (NP_00789.1) | 4e-21 | 20.3 | 194 | |
| | SSTR5 (NP_001044.1) | 6e-20 | 23.9 | 209 | |
| hGPR136 | RRH (NP_006574.1) | 9e-34 | 24.9 | 170 | Reproductive organs (male) |
| | OPR4 (NP_150598.1) | 3e-31 | 22.1 | 195 | |
| | OPR3 (NP_055137.1) | 7e-28 | 19.1 | 201 | |
| | OPR2 (NP_00530.1) | 1e-23 | 20.1 | 226 | |
| | OPR1 (NP_001699.1) | 1e-21 | 19.8 | 258 | |
| hGPR141 | GPR87 (NP_076404.1) | 1e-16 | 21.0 | 273 | Marrow |
| | CysLTR1 (NP_006630.1) | 5e-16 | 19.0 | 221 | |
| | GPR34 (NP_005291.1) | 2e-15 | 18.4 | 253 | |
| | GPR105 (NP_055694.1) | 4e-15 | 18.4 | 246 | |
| | GPR17 (NP_005282.1) | 4e-14 | 17.4 | 229 | |
| hGPR142 | CXCR1 (NP_001286.1) | 7e-10 | 15.5 | 232 | None |
| | GPR31 (NP_005290.1) | 2e-9 | 14.4 | 252 | |
| | GPR81 (NP_115943.1) | 1e-8 | 15.3 | 298 | |
| | SSTR4 (NP_001043.1) | 1e-8 | 17.8 | 253 | |
| | CXCR3 (NP_001495) | 2e-8 | 14.1 | 274 | |

for example CXCR1, CXCR3, GPR34 and CysLTR1, has a few unresolved branches. Apart from that, the topology of the two trees is identical and most importantly the position of the new receptors presented here is identical with NJ and MP. As can be seen GPR120 and GPR135 form a branch of their own, with the closest neighbor among the classic receptors being the bioamine receptors. Continuing around the tree in an anticlockwise manner we see that GPR119 is placed as a new subgroup within the bioamine receptors. Further ahead GPR136 forms a new subgroup within the group of classic opsin receptors. GPR142 forms a novel branch extending from the center of the star-shaped tree. This group has two fugu receptors, fGPR142a and fGPR142b. One likely explanation for this, given that they are around 40% identical to each other at the amino acid level, is that they are a result of the basal large-scale duplication event in the teleost lineage [15]. The GPR100 subgroup is placed close to the somatostatin receptors and especially close to the orphan receptor SALPR. The actual identity of fGPR100 is still unclear. From the phylogeny, and also from the level of sequence identity, it is indicated that fGPR100 is most closely related to SALPR, but on the other hand the difference is marginal, so an equally likely possibility is that human GPR100 is a recent copy of a pre-mammalian receptor that is the ancestor of both mammalian GPR100 and SALPR. This would mean that fugu is likely to have only one GPR100/SALPR receptor. And finally, GPR141 groups with the chemokine-like receptors, but on a sub-branch of its own.

Below we list the human EST hits we found for each new GPCR, listing first the **name of the receptor**, accession number (tissue). **GPR100**, CM1-MT0238-051200-622-g09 (marrow). **GPR119**, CA841236 (pancreas). **GPR120**, BM739118 (stomach), BM757151 (stomach). **GPR135**, UI-H-BW1-aoc-b-08-0-UI.s1 (eye), UI-E-EJ0-ahi-1-19-0-UI.r2 (eye), AW517245 (cervix), AI537485 (stomach), BQ179274 (brain), BG772522 (testis). **GPR136**, AI810121 (pooled), BG721121 (testis). **GPR141**, BG221739 (HT1080 cell line), BF896644 (marrow), BG461295 (HT1080 cell line), BE786005 (large cell carcinoma), AL598654 (?), BF896644 (?). **GPR142** (none).

## 4. Discussion

The results show that there exist seven additional human GPCRs. It is evident from the phylogenetic analysis that they belong to the family of rhodopsin GPCRs. This is further supported by short sequences that show similarities to the DRY motif placed in the intracellular side of TM3, which is one of the most characteristic motifs of the rhodopsin GPCRs. GPR119, GPR135, GPR136, and GPR142 all share a DRY motif while GPR100 has ARY, GPR120 has ERM, and GPR141 has TRY. It has been suggested that the DRY motif in some rhodopsin GPCRs is important to keep the receptors in the inactive state as mutations in the DRY motif have frequently caused receptors to be constitutively active. The DRY motif is not found in the other main families of GPCRs (glutamate, adhesion, frizzled/taste2 or secretin). Another of the main rhodopsin GPCR-specific motifs, the NSxxNPxxY motif in TM7, is also present in GPR100, GPR119, GPR120, and GPR135. GPR136 has NPxxY in TM7 and GPR142 has NPxxY while this motif could interestingly not be found in GPR141. Both the mouse and human GPR141 sequences are missing this motif indicating that these receptors are atypical within the rhodopsin family of GPCRs.

We also searched for similar genes in other species and were able to find orthologues for all the genes in at least one other vertebrate including a mouse orthologue for all the new genes. We found orthologues for all of them in fish with the exception of GPR120 and GPR141, using the draft scaffold assemblies of the genomes of zebrafish [16] and fugu (pufferfish) [17]. The data suggest that most of these receptors arose early in vertebrate evolution at least more than 450 million years ago. The receptors show several examples of a remarkable level of conservation. GPR135 and GPR136, for example, show 69% and 74% amino acid identity between the human and fugu orthologues within the TM regions. There is also a very high conservation between the rat, mouse and human GPR141 with over 95% amino acid identity within the TM regions. It was therefore somewhat surprising that we did not find any orthologous genes in fish for this receptor. This could be due to the lack of completeness in the fugu and zebrafish genomes or to the fact that these genes arose late in vertebrate evolution.

Intriguingly, none of the receptors except GPR100 (see below) have any close primary sequence relative in the human genome. This is clearly shown in the phylogenetic analysis in Fig. 2. Moreover, the phylogeny, BLAST score, percent identity and divergence results display ambiguity in the degree of similarity toward different subgroups of rhodopsin receptors for the new receptors. This variability, resulting in a different suggestion of the 'most similar genes', is likely to be due to the different alignment algorithms used in the analysis tools, a phenomenon often seen when the test proteins are highly divergent, forming no easily agreeable alignment with the closest relative in the dataset. Taken together it seems clear that six of the new genes encode a 'single gene family' receptor without any other subtypes in the family. This is remarkable considering the '2R hypothesis' or the 'one to four model' that suggest two rounds of large-scale duplications are proposed to have occurred in early vertebrate ancestry [18,19], resulting in up to four copies of each gene in mammals. If the 2R hypothesis is valid, the other three copies were lost for all these receptors. Furthermore, it is interesting that we did not find subtype copies of these receptors in any other genome with one exception, GPR142, which is found in two copies in fugu. This is particularly noteworthy for the fugu and zebrafish genomes that are believed to have undergone one additional tetraploidization, meaning that several copies were lost in these species. This suggests that there could exist stringent evolutionary pressure on these genes, to only exist in single copies. This is contrary to many of the most studied rhodopsin GPCRs such as the amine (adrenergic, serotoninergic, dopamine, etc.) and neuropeptide (NPY, melanocortin, somatostatin) or chemokine binding receptors. Several of these amine and peptide receptors are known to have multiple functions that are believed to have expanded in 'higher' vertebrates, mainly for 'higher' or central functions. Very little is known about single member genes within the GPCR family and there are only a very few of these that have known ligands or functions. Examples of such are the peptide binding prolactin-releasing peptide receptor (PrRP), the ghrelin receptor, the ACTH receptor (MC2) and the PP receptor (NPY4R). These are receptors that do not share their main ligands with other GPCRs but have a high sequence identity with related GPCRs

(see the phylogenetic clusters in [6]). These similarities are, however, much higher than for the new 'single copy receptors' highlighting the special character of these new GPCRs. It is also obvious from the alignment in Fig. 1 that the new receptors, even though they are well conserved within the orthologues, have very low sequence identity to their closest BLASTP hit. If, for example, the PrRP receptor is used as query against the nr database, six human GPCRs are found with an *e*-value higher than −34. This is the highest value that any of the single copy GPCRs (i.e. all new receptors except GPR100) presented here has to any other GPCR (see Table 2). The equivalent value for other well-known GPCRs is 25 for the ACTH receptor, 23 for the serotonin receptor 1B, 13 for rhodopsin, 11 for the galanin receptor 1, 70 for chemokine receptor 3 and 10 for the purine receptor 10. This clearly highlights the exceptionally low sequence identity of the new receptors to other GPCRs. To our best knowledge, the only GPCRs with such a low degree of sequence identity to other GPCRs are other orphan receptors such as GPR43, GPR63 and GPR84 [6], but we are not aware of any such example among the GPCRs with known ligands. It is possible that these receptors have a role in important functions with high demands of specificity where redundancy in alternative pathways is not tolerated.

One of the receptors has a clear, previously known relative in the human genome. This is GPR100, which has 38% amino acid identity to SALPR [20]. GPR100 has a clear orthologue in the mouse while we also found a full-length fugu gene that shows similarities to both SALPR and GPR100 probably representing a gene that represents a common ancestor to both SALPR and GPR100. The function of or ligand to SALPR is not known. SALPR mRNA is predominantly expressed in brain regions, particularly the substantia nigra and pituitary, although the mRNA can also be detected in low levels in peripheral tissues. GPR100 is only found in marrow according to the EST searches. We also searched for the EST for SALPR but found only three from a pooled library that included mRNA from melanocyte, fetal heart, and pregnant uterus (data not shown). Taken together the EST expression pattern indicates that the functional roles of SALPR and GPR100 may not be similar.

The tissue distribution that can be read from the EST results for the other receptors shows a highly individual pattern for each of the receptors. GPR135 and GPR141 had the highest number of EST sequences. GPR135 seems to be expressed in a wide variety of tissues including eye, brain, and peripheral tissues such as cervix, stomach and testis while GPR141 was found in marrow and cancer cells such as human fibrosarcoma HT1080 cells. The other receptors (GPR100, GPR119, GPR120, GPR136) seem to have very restricted expression patterns, perhaps reflecting that their putative functional role is more cell-specific and/or that the expression levels are low.

In summary, we have identified seven new GPCRs that belong to the rhodopsin subgroup of GPCRs. Six of the receptors do not seem to have any close evolutionary relative in the human genome while one of the receptors is related to SALPR. Most of the receptors are highly conserved through several vertebrate species remaining in single copies without subtypes. The expression patterns of some of the receptors indicate that some of them may be restricted to a single type of tissue, while others have a broader tissue distribution.

*Acknowledgements:* The studies were supported by the Swedish Research Council (VR, medicine), the Swedish Society for Medical Research (SSMF), Petrus och Augusta Hedlunds Stiftelse, Svenska Läkaresällskapet, the Novo Nordisk Foundation, Magnus Bergwalls Stiftelse and Melacure Therapeutics AB, Uppsala, Sweden.

## References

[1] Lander, E.S. et al. (2001) Nature 409, 860–921.
[2] Venter, J.C. et al. (2001) Science 291, 1304–1351.
[3] Flower, D.R. (1999) Biochim. Biophys. Acta 1422, 207–234.
[4] Attwood, T.K. and Findlay, J.B. (1994) Protein Eng. 7, 195–203.
[5] Lane, R.P., Cutforth, T., Axel, R., Hood, L. and Trask, B.J. (2002) Proc. Natl. Acad. Sci. USA 99, 291–296.
[6] Fredriksson, R., Lagerström, M.C., Lundin, L.G. and Schiöth, H.B. (2003) Mol. Pharmacol. 63, 1256–1272.
[7] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[8] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[9] Eddy, S.R. (1998) Bioinformatics 14, 755–763.
[10] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[11] Felsenstein, J. (2003) PHYLIP, phylogenetic inference package, Distributed by the author, Department of Genetics, University of Washington, Seattle, WA.
[12] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and Miyano, M. (2000) Science 289, 739–745.
[13] Fredriksson, F., Lagerström, M.C., Höglund, P.J. and Schiöth, H.B. (2002) FEBS Lett. 531, 407–414.
[14] Fredriksson, R., Gloriam, D.E., Höglund, P.J., Lagerström, M.C. and Schiöth, H.B. (2003) Biochem. Biophys. Res. Commun. 301, 725–734.
[15] Taylor, J.S., Van der Peer, Y., Braash, I. and Meyer, A. (2001) Phil. Trans. R. Soc. Lond. B Biol. Sci. 356, 1661–1679.
[16] Thomas, J.W. and Touchman, J.W. (2002) Trends Genet. 18, 104–108.
[17] Aparicio, S. et al. (2002) Science 297, 1301–1310.
[18] Holland, P.W. (1999) Semin. Cell Dev. Biol. 10, 541–547.
[19] Lundin, L.G. (1993) Genomics 16, 1–19.
[20] Matsumoto, M., Kamohara, M., Sugimoto, T., Hidaka, K., Takasaki, J., Saito, T., Okada, M., Yamaguchi, T. and Furuichi, K. (2000) Gene 248, 183–189.