# Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions

Robert Fredriksson[a,b,1], Malin C. Lagerström[a,1], Pär J. Höglund[a], Helgi B. Schiöth[a,*]

[a] *Department of Neuroscience, Uppsala University, BMC, Box 593, 751 24 Uppsala, Sweden*
[b] *Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 597, 751 24 Uppsala, Sweden*

Received 22 August 2002; revised 8 October 2002; accepted 9 October 2002

First published online 17 October 2002

Edited by Robert B. Russell

**Abstract** **We report eight novel members of the superfamily of human G protein-coupled receptors (GPCRs) found by searches in the human genome databases, termed GPR97, GPR110, GPR111, GPR112, GPR113, GPR114, GPR115 and GPR116. Phylogenetic analysis shows that these are additional members of a family of GPCRs with long N-termini, previously termed EGF-7TM, LNB-7TM, B2 or LN-7TM. Five of the receptors form their own phylogenetic cluster, while three others form a cluster with the previously reported HE6 and GPR56 (TM7XN1). All the receptors have a GPS domain in their N-terminus and long Ser/Thr-rich regions forming mucin-like stalks. GPR113 has a hormone binding domain and one EGF domain. GPR112 has over 20 Ser/Thr repeats and a pentraxin domain. GPR116 has two immunoglobulin-like repeats and a SEA box. We found several human EST sequences for most of the receptors showing differential expression patterns, which may indicate that some of these receptors participate in reproductive functions while others are more likely to have a role in the immune system.**


*Key words:* Orphan; Human genome; Adhesion; Reproduction

## 1. Introduction

The G protein-coupled receptors (GPCRs) are a large family of integral membrane proteins that generally act as cell surface receptors responsible for the transduction of a remarkable diversity of endogenous signals into a cellular response. The GPCRs share the same basic molecular architecture with seven hydrophobic regions of 25–35 consecutive residues. The defining concept is also that they share a common signalling mechanism, in that they interact with ubiquitous guanine nucleotide binding regulatory proteins (G proteins) to regulate the synthesis of intracellular second messengers. The GPCRs are remarkably diverse at the primary protein sequences, and that is also reflected by their diversity in physiological functions. The variety and importance of the physiological roles undertaken by the GPCR family has resulted in many of their members becoming important targets for drug development. Large number of modern drugs are targeted at GPCRs (for reviews see [1–3]).

The number of known members in the superfamily of GPCRs is continuously increasing. Some of these include receptors that have unusually large N-termini with domains from other well-known proteins. This group of receptors has been assigned various names, including EGF-TM7 [4] to reflect the presence of epidermal growth factor (EGF) domains in the N-termini. The subgroup has also been termed B2 [5], due to some sequence similarity with members of secretin-like receptors or clan B of GPCRs [1]. Others have called the group LNB-TM7 [6] or LN-7TM receptors [7]. LN stands for long N-termini and B denotes the clan. The overall sequence similarity between this LN-7TM and clan B is however fairly low and they differ in many aspects. Clan B receptors bind rather large peptides such as secretin, vasoactive intestinal peptide, pituitary adenylate cyclase-activating polypeptide, glucagon-like peptide, calcitonin, parathyroid hormone and corticotropin-releasing factor with a binding unit that consists of both the N-terminus and other parts of the receptors. The functional modules that are found in the LN-7TM receptors, such as EGF, cadherin, lectin, laminin, olfactomedin, immunoglobulin or thrombospondin, are not found in clan B receptors. Initially, it was believed that expression of LN-7TM receptors was restricted to leukocytes, or other cell types within the immune system. Today, it is well known that a large range of cell types express these complex molecules. The functional roles of these receptors are still obscure but it has been suggested, mostly based on tissue distribution and expression studies, that many of these molecules have a role in the immune system, cell adhesion, cell to cell communications and in the nervous system.

In this study we created hidden Markov models (HMMs) based on previously known clan B and LN-TM7 receptors and searched the human genome database for unique sequences. We identified eight new human GPCRs that have structural similarity to the LN-7TM receptors.

## 2. Materials and methods

### 2.1. Identification of novel receptors

Sequences of known GPCRs from the LN-7TM family were downloaded from the NCBI database using the online BLAST and Entrez tools (http://www.ncbi.nlm.nih.gov/). The sequences were LEC1 (NP_036434.1), LEC2 (NP_055736.1), LEC3 (NP_056051.1), BAI1 (NP_001693.1), BAI2 (NP_001694.1), BAI3 (NP_001695.1), CELSR1 (NP_055061.1), CELSR2 (NP_001399.1), CELSR3 (NP_001398.1), EMR1 (NP_001965.1), EMR2 (NP_038475.1), EMR3 (NP_115960.1), CD97 (NP_078481.1), ETL (NP_071442.1), HE6

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

[1] Should be considered equal first authors.

0014-5793/02/$22.00 
PII: S0014-5793(02)03574-3

(NP_005747.1), GPR56 (TM7XN1) (XP_165649.1) and murine Ig-Hepta receptor (XP_166354.1). We removed the N- and C-terminals, 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 subsequently aligned using ClustalW 1.81 [8]. From the alignments, an HMM was constructed using the HMMER 2.2 package [9]. The model was constructed using HMMbuild with default settings and calibrated using HMMcalibrate. The GeneScan protein dataset, from assembly 28 of the public human genome sequence, was downloaded from the NCBI ftp site, ftp.ncbi.nlm.nih.gov/genbank/, and searched against the HMM using HMMsearch, with a cut-off at $E = 1e\text{-}4$. Novel sequences were confirmed by searching all hits against the public databases using the BLAST package [10].

### 2.2. Identification of human EST clones

The full genomic DNA sequences of the novel 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 expressed sequence tag (EST) sequences were manually inspected to ensure correct identity.

### 2.3. Verification of the predicted coding regions

The machine-predicted coding regions, predicted using GeneScan, 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 by 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.

### 2.4. 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 all 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 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.

## 3. Results

Our strategy was to create HMMs and search the human database for new GPCRs. These search tools are much more sensitive than for example BLAST and other similar tools as the models take into consideration information from a group of proteins and uses a statistical consensus 'fingerprint' as a representation of the group of sequences. The new sequences were not found through previous BLAST searches made by us, or by others [5], with LN-TM7 receptors. The HMMs were constructed through alignment of the transmembrane (TM) regions from a subset of GPCRs, as described in Section 2, and subsequently used to search GeneScan datasets (assembly 28), which were downloaded from the NCBI ftp site at ftp.ncbi.nlm.nih.gov/genbank/. The search resulted in 25 unique human sequences. Sixteen of those sequences were identical to known human sequences, while eight were novel. The sequences were confirmed in Celera genome data base (http://www.celera.com) and all the new sequences were found there as non-annotated protein sequences, except two that were only found through HMM searches in the public genome database. 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, GPR97, had previously been assigned a GPR number under confidentiality to HUGO, by a group unknown to us. The committee provided the other receptors with new GPR numbers (GPR110–116) at our request. We subsequently made all eight new GPCRs, both protein and DNA sequences, public through submission to the NCBI database.

It is known that the GeneScan software used for predicting coding regions for the human genome project has the capacity to predict approximately 80% of the splice sites correctly [12]. Therefore, we decided to verify 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.

GPR97 was identified in build 28 of the human genome GeneScan dataset as open reading frame (ORF) Hs16_10563_28_27_2, and 15 human mRNA sequences were found for this GPCR. When these mRNAs were assembled, at least three-fold coverage of the entire coding region was obtained. After editing, the final coding region of GPR97 was found to differ from the original Hs16_10563_28_27_2 in several ways. In the N-terminal, which does not start with a Met, one additional exon of 57 nucleotides was added. At the seventh splice acceptor, which was found to be wrong, 141 base pairs were added. Also, the last 120 nucleotides in the C-terminus were replaced by 108 nucleotides from mRNA sequences, since these 120 nucleotides from Hs16_10563_28_27_2 were not found to be expressed in any of the four mRNAs found to cover this region. The mouse Q9D6B0 sequence is possibly an orthologue to GPR97.

GPR110 was found in the GeneScan dataset with ORF number Hs6_7559_28_29_4. We identified 29 different human mRNA sequences using Hs6_7559_28_29_4, providing at least five-fold coverage of the entire coding region. We found that the third splice site donor was incorrectly predicted by the GeneScan program, with 42 bp missing, and we added those from mRNA data to give the complete coding region of GPR110.

An ORF with number Hs6_7559_28_36_1 was found in the GeneScan dataset. We could, however, not find any mRNA sequences from any species. The mouse genome project has, on the other hand, a GeneScan-predicted orthologue, which shows 70% identity at the protein level with a high degree of third base substitutions, indicating an evolutionary pressure on this gene. This GPCR was named GPR111.

GPR112 was found as HsX_11876_28_18_4. This is by far the largest protein found in our searches. We found only one

Table 1
Summary of the main features found in the novel GPCRs, their accession numbers and protein IDs from both the public NCBI GeneScan dataset and Celera Discovery system, together with their newly assigned names (GPR numbers)

| Name | Accession number | NCBI, GeneScan ORF | Celera Discovery Protein ID | Chromosome | Size (aa) | Number of exons | Tissue expression (ESTs) |
|---|---|---|---|---|---|---|---|
| GPR113 | AY140955 | Hs2_30849_28_3_1 | hCP1639587 | 2p23.3 | 1077 | 13 | (3) testis |
| GPR111 | AY140953 | Hs6_7559_28_36_1 | hCP1708132 | 6p12.3 | 708 | 6 | [1]lung*, [1]mammary gland, [1]embryo, [1]ductus*, [1]diencephalon, [1](4) gross tissue* |
| GPR115 | AY140957 | Hs6_7559_28_36_2 | hCP1626221 | 6p12.3 | 752 | 10 | (4) pregnant uterus, breast, genitourinary tract |
| GPR110 | AY140952 | Hs6_7559_28_29_4 | – | 6p12.3 | 1054 | 14 | (4) prostate, (6) amnion, (2) uterus, (3) breast, (10) kidney, lung, (2) myeloma* |
| GPR97 | AY140959 | Hs16_10563_28_27_2 | hCP35037 | 16q13 | 538 | 10 | (6) lung*, (5) leukemia cells*, (4) marrow |
| GPR114 | AY140956 | Hs16_10563_28_27_3 | hCP1771754 | 16q13 | 526 | 11 | (5) leukemia cells*, pancreas islets of Langerhans, leukocytes, head |
| GPR112 | AY140954 | HsX_11876_28_18_4 | – | Xq26.3 | 2799 | 15 | retina, [1](3) carcinoma, [1]mammary gland, [1]kidney, [1]blood vessel |
| GPR116 | AY140958 | XP_166354.1 | hCP36563 | 6p12.3 | 1327 | 16 | (5) lung, (4) kidney, placenta, nervous tumor*, (4) brain, (4) gastrointestinal tract, (4) spleen/liver, (2) liver, (2) heart, pelvis, breast, eye, head, testis |

GPR116 was identified in the NR database using BLASTP and therefore has an XP number rather than a GeneScan protein ID.
The tissue expression column summarizes the expression pattern of mRNA/EST sequences from the EST database. The numbers within parentheses denotes the number of clones found; a superscript 1 indicates mouse ESTs and * indicates tumor tissue.

human EST clone in the database representing this GPCR. The putative mouse orthologue is predicted essentially identically in the mouse genome assembly, but the Celera version of the human protein is missing the entire N-terminal. We find it likely that the Celera version is not correctly predicted since the N-terminal has a PTX domain as well as a GPS domain, and a large mucin-like stalk with over 20 glycosylation sites. Also the N-terminal consists largely of one ORF, over 2000 amino acids long.

GPR113 was identified in build 28 of the human genome with ORF number Hs2_30849_28_3_1. Two human EST sequences were found from this receptor, covering approximately 40% of the coding region, showing that the fifth splice site were predicted incorrectly, with 27 bases missing at the donor site, which was subsequently added from the mRNA sequences. The machine-predicted sequences of the mouse orthologues from the mouse genome project at http://genome.ucsc.edu/ were found to be identical, providing some evidence for the correctness of the rest of the splice sites.

GPR114 was found as ORF Hs16_10563_28_29_ 3. It has relatively few mRNA sequences in the database, which may indicate low or specialized expression. We found EST sequences covering the third splice site, which was not predicted correctly in the machine-predicted sequence, and the missing sequence was added from EST data. In total approximately 50% of the protein has EST coverage.

GPR115 was found as ORF Hs6_7559_28_36_2. Approximately 45% of the coding region was covered by EST sequences. We found an error in the machine-predicted coding region in the second last splice site. Here, 154 amino acids were missing and these were added from the EST data. However, we found no EST or mRNA sequences covering the N-terminal region.

GPR116 was identified as the non-annotated sequence XP_166354.1. It is likely to be the human orthologue of the rat Ig-Hepta receptor [13], since they share 72% sequence identity at the amino acid level. We have, however, chosen to give this GPCR the name GPR116, because the orthologous–paralogous relationship between these receptors is at present unclear and it could cause important confusion if we were to assign it wrongly. The machine-predicted sequence of XP_166354.1 was found to be represented by three different and partial mRNAs, which when assembled provided a full coverage of the coding region. The original XP_166354.1 sequence was found to differ from the mRNA sequences at the fourth splice acceptor where 402 nucleotides were missing from XP_166354.1. This sequence was added from the consensus mRNA sequence, and the exon/intron organization of the final sequence of GPR116 was found to correspond exactly to the published mRNA of rat Ig-Hepta [13].

A summary of the results are found in Table 1, where we list the name, accession numbers, chromosomal positioning, GeneScan ORF numbers, number of exons, length in amino acids, and tissue distribution as suggested by EST data. Fig. 1 shows alignment of the TM regions of the eight new GPCRs together with two of the previously known LN-7TM receptors, GPR56 (TM7XN1) [7] and HE6 [14]. Fig. 2 shows phylogenetic analyses, using maximum parsimony, of LN-7TM receptors together with the new receptors we report here. The topology of the tree is largely identical using distance methods (data not shown). Considering this dataset, bootstrap values show five small groups with very high bootstrap

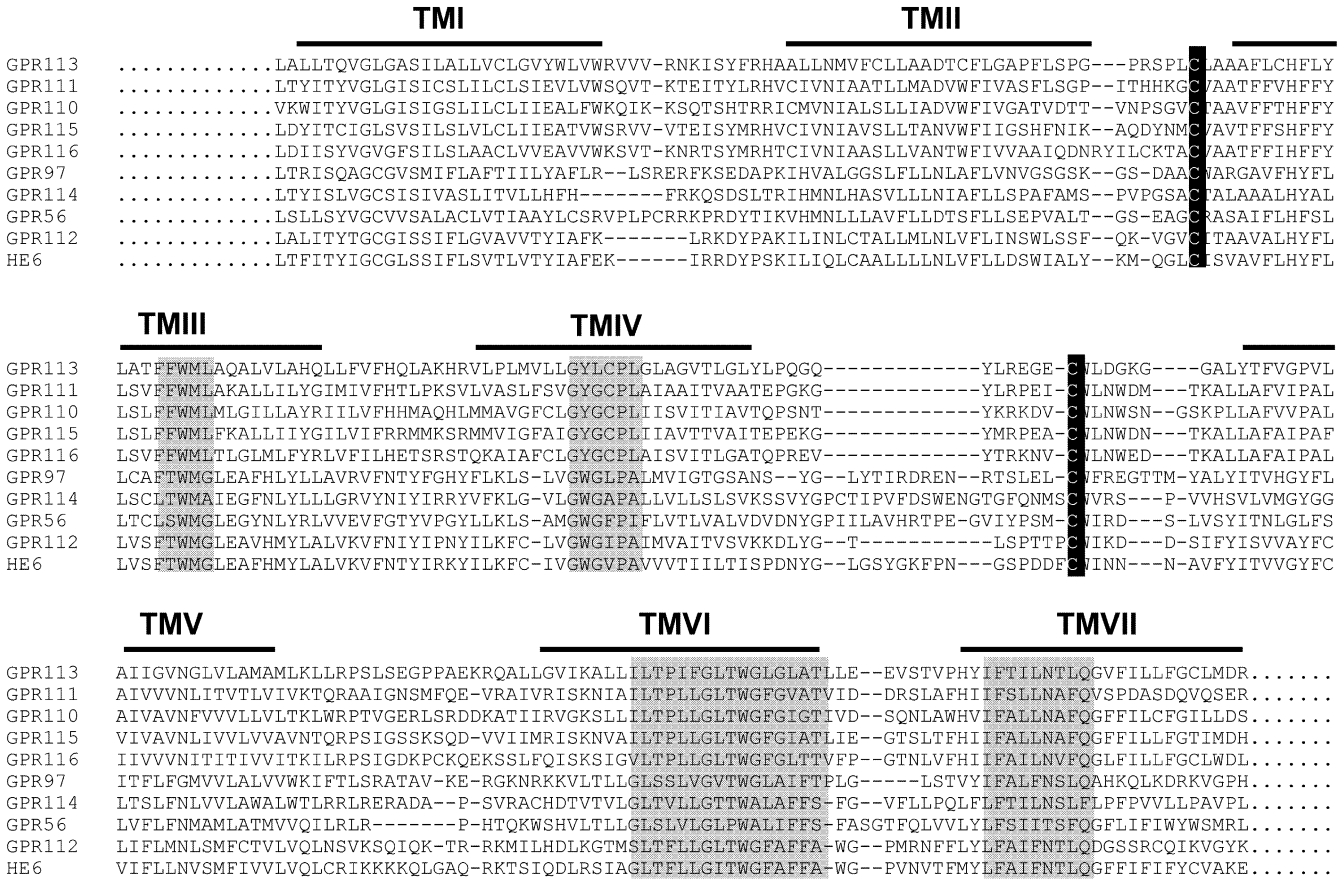


Fig. 1. Amino acid sequence alignment of the region from TM1 to TM7 of the novel human GPCRs, HE6 and GPR56 made using ClustalW (1.82) software and edited by manual inspection. The lines mark putative TM regions, and gray boxes indicate conserved sequence motifs. The accession numbers are listed in Table 1.

support. Phylogenetic analysis of the LN-7TM receptors with other GPCRs (from clans A, B and C) shows that these are evolutionarily distant from the LN-7TM receptors (data not shown). Looking at Fig. 2, and following the clusters in anti-clockwise manner, the first group consists of GPR56 and HE6 together with three of our new GPCRs, GPR97, GPR112 and GPR114 (bootstrap: 952 of 1000). The second group consist of three BAI proteins (bootstrap: 966), while the third group includes five of the new GPCRs, GPR113, GPR116, GPR110, GPR115 and GPR111 (bootstrap: 966). The fourth group consists of three CELSR proteins (bootstrap: 962), while CD97, three EMR proteins, ETL and three LEC proteins form the fifth group (bootstrap: 817). It should be mentioned that we opted to show the phylogenetic analysis of the truncated receptors, i.e. only covering the region from the start of TM1 to the end of TM7. This because it is evident that the proteins in this family are composed of units with different evolutionary origin since the domains in N-termini are derived from other different types of proteins. We did not find any phylogenetic relationship between the N-terminals of these new receptors and any other GPCR beyond the domains that are shown in Fig. 3.

The domains in the N-terminal regions in all of the new receptors were identified using RPS-BLAST at NCBI and HMM domain search at Celera, with the full-length GPCR sequences as query, with the default settings (see Section 2). These search engines predicted the same domains, with excep-

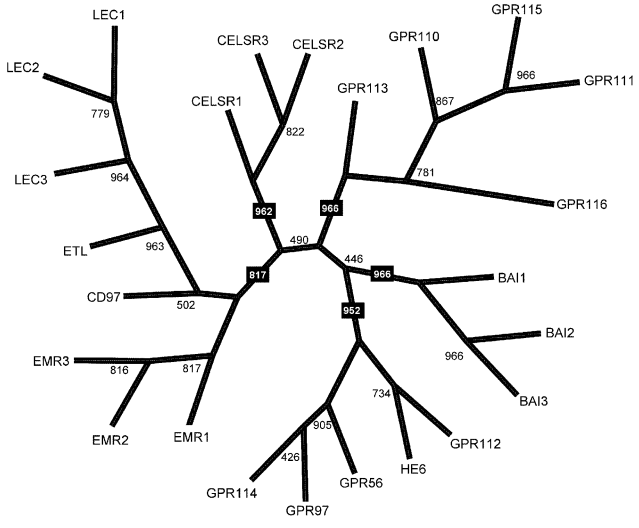


Fig. 2. Phylogenetic analysis of the new human GPCRs with other human LN-7TM receptors using the regions from TM1 to the end of TM7, obtained by maximum parsimony analysis (Phylip 3.6). The numbers above the nodes indicate percentage of bootstrap replicates. The accession numbers are listed in Table 1.

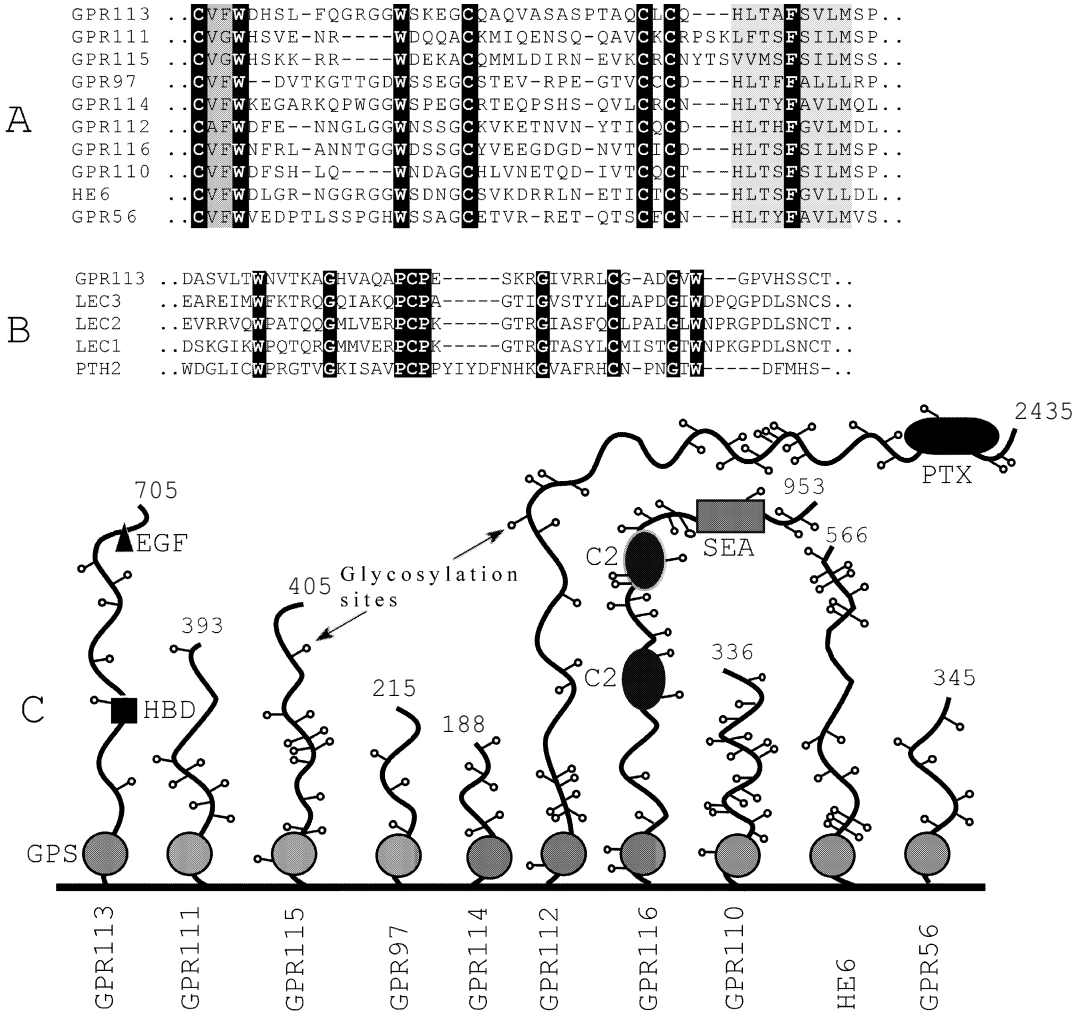


Fig. 3. A: Amino acid sequence alignment of the GPS (Gpcr Proteolytic Site) domain in the novel human GPCRs, HE6 and GPR56 made using ClustalW (1.82) software and edited by manual inspection. The residues in black are conserved through all receptors, while the gray boxes indicate highly conserved sequence motifs. B: Amino acid sequence alignment of the hormone binding domain (HBD) in GPR113 with the HBD in LEC1–3 and PTH2 receptors made using ClustalW (1.82) software and edited by manual inspection. The residues in black are conserved through all receptors. C: Schematic presentation of the domains found in the N-termini. The GPS, EGF, HBD, C2 (C2-set immunoglobulin), SEA (sperm protein, enterokinase, and agrin) and PTX (pentraxin) domains are shown as boxes in black and gray while glycosylation sites (NXS or NXT tripeptide sequences that conform to the consensus sequence for *N*-linked glycosylation) are shown as small circles attached to the N-terminal stretches. The numbers at the end of each N-terminus are the numbers of amino acids found in the N-termini from the predicted start at TM1.

tion of the EGF domain in GPR113, which was not predicted by NCBI's RPS BLAST.

Below we list the human EST hits we found for each new GPCR, listing first the *name of the receptor*; accession number (tissue). *GPR97*: 601809910F1 (lung, carcinoma), AGENCOURT_7785691 (lung, carcinoma), TCAAP1D6491 (pediatric acute myelogenous leukemia cell), TCAAP1D4626 (pediatric acute myelogenous leukemia cell), AGENCOURT_7828938 (lung, carcinoma), TCAAP1E4365 (pediatric pre-B cell acute lymphoblastic leukemia), TCAAP1E0686 (pediatric acute myelogenous leukemia cell), AGENCOURT_7587954 (lung, carcinoma), IL3-MT0267-120101-445-H10 (marrow, adult), hf44e05.x1 (fetal lung NbHL19W, testis NHT, and B-cell NCI_CGAP_GCB1, pooled), wf37e09.x1 (fetal lung NbHL19W, testis NHT, and B-cell NCI_CGAP_GCB1, pooled), TCAAP1E14475 (pediatric acute myelogenous leukemia cell), RC0-MT0059-200600-021-a08 (marrow, adult), PM1-MT0429-200601-009-e07 (marrow, adult), QV1-MT0166-131100-483-g04 (marrow, adult). *GPR110*: 601646506F1 (prostate), xt19d07.x1 (uterus), K-EST0058804 (myeloma), 602492853F1 (kidney), PM1-BN0083-030300-003-h10 (breast_normal), MR3-AN0025-080800-006-g06 (amnion_normal), MR3-AN0025-080800-006-f06 (amnion_normal), MR3-AN0025-250700-001-d12 (amnion_normal), MR3-AN0025-030800-003-b05 (amnion_-normal), ob20b09.s1 (kidney), MR3-AN0025-070800-006-d03 (amnion_normal), IL2-FT0167-070800-123-A09 (prostate_tumor), IL2-FT0167-070800-123-A02 (prostate_tumor), xn18g09.x1 (kidney), MR3-AN0025-030800-003-b08 (amnion_normal), oj15a02.y5 (kidney), hr84a09.x1 (kidney), QV2-BT0616-061200-518-e11 (breast), 602330074F1 (prostate), hw29c08.x1 (kidney), QV4-BN0090-270400-190-h10 (breast_normal), tm56f05.x1 (kidney), K-EST0127034 (myeloma), UI-CF-EN1-adg-c-13-0-UI.s1 (lung), UI-H-BI2-agb-c-

12-0-UI.s1 (kidney), wv02e11.x1 (kidney), 602499188F1 (kidney), tw54b05.x1 (uterus). *GPR111*: none. *GPR112*: ab03a10.r1 (fetal retina). *GPR113*: zu91g01.s1 (testis), 603252694F1 (testis), zu91g02.s1 (testis). *GPR114*: TCAAP1E4325 (pediatric pre-B cell acute lymphoblastic leukemia), TCAAP1Q15716 (pediatric acute myelogenous leukemia cell (FAB M1), 602715392F1 (primary B cells from tonsils (cell line)), TCAAP4E0527 (pediatric acute myelogenous leukemia cell), 603061618F1 (leukocyte), ie67d12.y3 (pancreas islets of Langerhans), CM4-HT0509-110200-095-e09 (head_neck, adult), TCAAP2E5094 (pediatric acute myelogenous leukemia cell). *GPR115*: zl16a07.r1 (uterus, pregnant), zk69a02.r1 (uterus, pregnant), zk27g08.r1 (uterus, pregnant), zk25b11.r1 (uterus, pregnant), PM2-BN0063-210200-003-d06 (breast_normal), hh70c04.y1 (genitourinary tract). *GPR116*: 602591049F1 (lung), 602464465F1 (kidney), AL551728 (placenta), zb53f09.r1 (fetal lung), za74b06.r1 (fetal lung), CM0-NT0186-181100-711-a12 (nervous_tumor), 602501038F1 (kidney), AV725650 (hypothalamus), 602465931F1 (kidney), MR3-CT0465-300800-005-d12 (colon), 603180744F1 (brain), 602564677F1 (lung), MR3-CT0465-300800-005-d12_1 (colon), ya86g04.r1 (fetal spleen), UI-E-EO0-aia-h-11-0-UI.r1 (fetal eye), 602589054F1 (liver), 601861084F1 (lung), IL2-HT0436-250300-044-E09 (head_neck), wn20a12.x1 (stomach), zd24e12.r1 (fetal heart), yj67g03.r1 (breast), UI-H-EZ1-bbe-j-13-0-UI.s1 (left pelvis), xf49h02.x1 (stomach), AV658066 (non cancerous liver tissue), 602017176F1 (brain), yp77c12.r1 (fetal liver spleen), yf43g04.r1 (fetal liver spleen), cp3149.seq.F (human fetal heart), EST66136 (kidney), 603201927F1 (testis), KIAA0757 (brain), AB018301.

## 4. Discussion

Phylogenetic analysis of the new GPCRs shows that they belong to the family of LN-7TM receptors. The overall phylogenetic analysis of the LN-7TM family shows that there exist at least five different groups of LN-7TM receptors (named by us groups 1–5, see Fig. 2). Three of the new receptors show high bootstrap values to group them closely to the previously known HE6 and GPR56 (group 1). Interestingly, the other five receptors form a group of their own (group 3). The high bootstrap values for each of the groups indicate that the receptors that we have assigned to each group are likely to share a common ancestor.

The overall sequence identity among the receptors within groups 1 and 3 is about 28–30% and 45–50% in the TM regions. This is relatively low as compared with for example many of the subgroups of rhodopsin receptors. The well studied monoamine (adrenergic, histamine, serotonin, etc.) and peptidergic (neuropeptide Y, melanocortin, cholecystokinin, etc.) receptors have about 45–60% sequence identity in the TM1 to TM7 region and as high as 60–80% in the TM regions. It is, however, evident that there are some highly conserved features within both groups 1 and 3, indicating that there exists a specific evolutionary pressure on their primary structures, also in the TM regions. There is a highly conserved region in TMIII, including a conserved Trp-Met motif, and another conserved region with a Trp in TMVI (see conserved regions denoted in gray in Fig. 1). There are also conserved regions in TMVII, where all the receptors except GPR114 have Glu positioned in the middle of the TM region. A conserved region in TMIV shows a clear distinction between groups 1 and 3, where group 1 has conserved hydrophobic Trp in a position where group 3 has conserved hydrophilic Tyr. This region also includes a Pro, a residue known to cause kinks in α-helices, which is conserved through both groups [3]. Alignment of the new receptors with some classical clan B receptors (such as the secretin receptor) indicates that these Trp residues in TMIII, TMIV and TMVI are widely found in these receptors (data not shown). Another conserved property in this part of the receptors is Cys residues in the putative extracellular loop 2 and extracellular loop 3. This is a feature that is widely found in the superfamily of GPCRs. There is strong evidence that these residues form a disulfide bridge between these loops and that they are important for the structural integrity of the protein as shown by studies on rhodopsin GPCRs [3].

We searched the protein sequences for functional domains using the NCBI and Celera search tools. The TM domains are proposed to be of type 2 (clan B or secretin-like) in both search machines. It should, however, be noted that the LN-7TM family does not yet have a specific designated domain search tool at either NCBI or Celera. The domains that were found in the N-termini of the five new receptors together with HE6 and GPR56 are shown in Fig. 3. All the new GPCRs have a GPS (Gpcr Proteolytic Site) domain close to the TM regions. The GPS domains are found in all LN-7TM receptors but they are not found in any of the clan B receptors, or any other GPCR as far as we are aware of, but they are found in other types of receptors [15]. One of the LN-7TM receptors, CL1 (LEC1), is shown to be endoproteolytically cleaved at the end of the GPS domain resulting in two subunits that remain non-covalently associated as a heterodimer [15,16]. No clear functional role has yet been ascribed to the GPS. The GPS region in our new receptors has four conserved Cys, two conserved Trp, one conserved Phe, and a number of other residues that are conserved in almost all the receptors (see Fig. 2). It is possible that variation in the sequence in this region has a role as alternative cleavage site for the different receptors that may affect their functional roles. The N-terminals of the receptors do not show any other specific region with high sequence similarities that are shared through all the receptors in either group 1 and 3. Another feature that is shared by all the new receptors is Ser/Thr-rich regions, or multiple repeats of glycosylation sites forming what have also been termed mucin-like stalks. The N-termini of all the new receptors are virtually made up of these stalks. One of the receptors, GPR113, has one EGF domain at the end of the N-terminus (see Fig. 3). The EGF domain is characterized by a set of six conserved Cys residues that typically form disulfide bonds in a 1–3, 2–4, 5–6 arrangement and are often found in connective tissue proteins such as fibrillins and fibulins [17]. GPR113 thus shows a structural similarity to the group of EMR receptors, ETL and CD97 found in group 2 (see Fig. 2) that have multiple EGF domains in their N-termini. The EGF-TM7 receptors are predominantly expressed in myeloid cells (EMR1, EMR2, EMR3, and CD97) and smooth muscle cells (ETL and CD97). It has been suggested that these receptors participate in cellular functions of myeloid leukocytes and cardiac muscle differentiation by interacting with other cell surface proteins or extracellular matrix proteins. Recently, it was demonstrated that CD97 interacts with a defined cellular ligand, CD55 (decay-accelerating factor) solely by the EGF-like domains [18].

GPR113 also has a sequence termed 'hormone binding domain' within its N-terminus. Hormone binding domains are found in many hormone binding receptors, including many of the clan B receptors and some of the LN-7TM receptors, such as LEC and BAI. The four hormone binding domains that are most related to the one in GPR113 are shown in Fig. 3. The hormone binding domain in LEC3 receptor (latrophilin-3) shows the closest similarity but the hormone binding domain in the clan B PTH2 receptor also shows clear similarity. There are nine conserved residues in these five receptors: two conserved Cys, which may form disulfide bridges, two Pro, two Trp and three Gly. There is, however, low similarity between the other residues in this putative domain.

GPR112 also has an interesting domain in the N-terminus or a PTX (pentraxin) domain. The closest hit by BLAST with the PTX domain in GPR112 is the neuronal pentraxin II precursor (NP-II). The pentraxins (earlier also termed pentaxins) are evolutionarily highly conserved proteins that have up to five non-covalently bound identical subunits that are arranged in a flat pentameric disk. Many of them are cytokine-inducible acute phase proteins and some have been implicated in inherited immunity and Alzheimer's disease. C reactive protein and serum amyloid P component are short pentraxins, whose concentrations in the blood increase dramatically upon infection or trauma [19]. We are not aware of other GPCRs that have a PTX domain. The N-terminus of GPR112 is extremely long, about 2400 residues. This is slightly shorter than the longest LN-7TM receptors, CELSR (cadherin, EGF LAG seven-pass G-type receptor), which have up to 2500 residues in their N-termini. It is also much longer than for the most of the other receptors in this phylogenetic cluster. The new receptors we present here are otherwise among the shortest within the group of LN-7TM receptors. It is remarkable that the entire region spanning from the GPS domain to the PTX domain has a number of Ser/Thr glycosylation sites indicating that this long N-termini forms a long stalk that erects from the cell with the PTX domain at its end.

GPR116 has two C2-set immunoglobulin superfamily domains. These immunoglobulin-like repeats are characteristic motifs for the members of the immunoglobulin superfamily of cell surface proteins [20]. GPR116 also has a SEA box more distal in the N-terminus. The SEA box was first described as a motif present in an ectodomain of a number of mucin-like membrane proteins. It was named after the first three proteins in which it was first identified (sperm protein, enterokinase, and agrin). There are several proteins that have a SEA box, whose function, however, is not entirely clear. Its likely function is to serve as a site for proteolytic cleavage [21]. As mentioned in Section 3, GPR116 is possibly the human orthologue of the Ig-Hepta previously cloned in rats and mice [22]. Ig-Hepta is predominantly expressed in the lung, but also in the kidney, and it has been suggested that Ig-Hepta may be involved in pH sensing or pH regulation [13].

The other new GPCRs, GPR97, GPR110, GPR111, GPR114, and GPR115, like the previously cloned HE6 and GPR56, do not have any clear domains beyond the GPS domain that is recognizable with the above-mentioned search tools. It is likely that these receptors have functional domains that we have not yet identified. Their mucin-like stalks indicate that they indeed have functional units in their N-termini, and these could contain amino acid stretches that only recognize specific molecules. It is possible that such domains can only be identified through functional assays. Phylogenetic analysis of these receptors in more distant species, when such sequences become available, may reveal which specific regions are evolutionarily conserved and thus more likely to be of functional importance.

The tissue expression pattern that can be seen in the EST data is variable for the new receptors. GPR110, GPR114, GPR116 and GPR117 are found in multiple tissues with a number of EST hits. Some of the other receptors seem to have a more defined expression pattern perhaps indicating a tissue-specific function. GPR113, the one with a hormone binding and an EGF domain, seems to have its expression restricted to testis. GPR115 also seems to have an expression pattern that could match a protein that has a role in reproduction. GPR110 and GPR111 are also expressed in organs that participate in reproduction, while they are also found in other tissues such as the lung, which is one of the major expression sites for the LN-7TM receptors with EGF domains. The expression of the other receptors, that we found EST sequences for indicates a rather broad expression pattern, including many of the tissues in which the other members of the LN-7TM receptors have previously been found to be expressed.

In summary, we have identified eight new GPCRs that belong to the LN-7TM subgroup of receptors. Several of these receptors have functional domains in their N-termini such as GPS, EGF, SEA box, PTX and immunoglobulin domains that may provide indications of their functional role. The expression patterns of some of the receptors indicate that they may participate in reproductive functions, while others are likely to have a role in the immune system like many other LN-7TM receptors.

*Acknowledgements:* The studies were supported by the Swedish Research Council (VR, Medicine), the Swedish Society for Medical Research (SSMF), Åke Wibergs Stiftelse, Petrus och Augusta Hedlunds Stiftelse and Melacure Therapeutics AB, Uppsala, Sweden.

## References

[1] Flower, D.R. (1999) Biochim. Biophys. Acta 1422, 207–234.
[2] Sadee, W., Hoeg, E., Lucas, J. and Wang, D. (2001) AAPS Pharmacol. Sci. 3, 22.
[3] Lu, Z.L., Saldanha, J.W. and Hulme, E.C. (2002) Trends Pharmacol. Sci. 23, 140–146.
[4] McKnight, A.J. and Gordon, S. (1998) J. Leukocyte Biol. 63, 271–280.
[5] Harmar, A.J. (2001) Genome Biol. 2, REVIEWS3013.
[6] Stacey, M., Lin, H.H., Gordon, S. and McKnight, A.J. (2000) Trends Biochem. Sci. 25, 284–289.
[7] Zendman, A.J., Cornelissen, I.M., Weidle, U.H., Ruiter, D.J. and van Muijen, G.N. (1999) FEBS Lett. 446, 292–298.
[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] 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.
[11] Felsenstein, J. (1993) Phylogenetic Inference Package, distributed by the author. Department of Genetics, University of Washington, Seattle, WA.
[12] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[13] Abe, J., Suzuki, H., Notoya, M., Yamamoto, T. and Hirose, S. (1999) J. Biol. Chem. 274, 19957–19964.
[14] Osterhoff, C., Ivell, R. and Kirchhoff, C. (1997) DNA Cell Biol. 16, 379–389.

[15] Nechiporuk, T., Urness, L.D. and Keating, M.T. (2001) J. Biol. Chem. 76, 4150–4157.
[16] Krasnoperov, V.G., Bittner, M.A., Beavis, R., Kuang, Y., Salnikow, K.V., Chepurny, O.G., Little, A.R., Plotnikov, A.N., Wu, D., Holz, R.W. and Petrenko, A.G. (1997) Neuron 18, 925–937.
[17] Downing, A.K., Knott, V., Werner, J.M., Cardy, C.M., Campbell, I.D. and Handford, P.A. (1996) Cell 85, 597–605.
[18] Lin, H.H., Stacey, M., Saxby, C., Knott, V., Chaudhry, Y., Evans, D., Gordon, S., McKnight, A.J., Handford, P. and Lea, S. (2001) J. Biol. Chem. 276, 24160–24169.
[19] Goodman, A.R., Cardozo, T., Abagyan, R., Altmeyer, A., Wisniewski, H.G. and Vilcek, J. (1996) Cytokine Growth Factor Rev. 7, 191–202.
[20] Smith, D.K. and Xue, H. (1997) J. Mol. Biol. 274, 530–545.
[21] Wreschner, D.H., McGuckin, M.A., Williams, S.J., Baruch, A., Yoeli, M., Ziv, R., Okun, L., Zaretsky, J., Smorodinsky, N., Keydar, I., Neophytou, P., Stacey, M., Lin, H.H. and Gordon, S. (2002) Protein Sci. 11, 698–706.
[22] Abe, J., Fukuzawa, T. and Hirose, S. (2002) J. Biol. Chem. 277, 23391–23398.