The Gene Repertoire of G protein-coupled Receptors

New Genes, Phylogeny, and Evolution

ÞÓRA KRISTÍN BJARNADÓTTIR
Dissertation presented at Uppsala University to be publicly examined in Room B21, BMC, Husargatan 3, Uppsala, Wednesday, April 12, 2006 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract
ISBN 91-554-6489-0.

The superfamily of G protein-coupled receptors (GPCRs) is one of the largest protein families of mammalian genomes and can be divided into five main families: Glutamate, Rhodopsin, Adhesion, Frizzled, and Secretin. GPCRs participate in most major physiological functions, contributing to the fact that they are important targets in drug discovery. In paper I we mined the human and mouse genomes for new Adhesion GPCR genes. We found two new human genes (GPR133 and GPR144) and 17 mouse Adhesion genes, bringing the number up to 33 human and 31 mouse genes. In paper II we describe 53 new splice variants for human Adhesion receptors supported by expressed sequence tags (EST) data. 29 of these variants seem to code for functional proteins, several of which lack one or more functional domains in the N-termini. Lack of certain domains is likely to affect ligand binding or interaction with other proteins. Paper III describes the Glutamate GPCR in human, mouse, Fugu, and zebrafish. We gathered a total of 22 human, 79 mouse, 30 Fugu, and 32 zebrafish sequences and grouped these into eight clans using phylogenetic methods. The report provides an overview of the expansion or deletions among the different branches of the Glutamate receptor family. Paper IV focuses on the trace amine (TA) clan of Rhodopsin GPCRs. We identified 18 new rodent genes, 57 zebrafish genes, and eight Fugu genes belonging to the clan. Chromosomal mapping together with phylogenetic relationships suggests that the family arose through several mechanisms involving tetraploidisation, block duplications, and local duplication events. Paper V provides a comprehensive dataset of the GPCR superfamily of human and mouse containing 495 mouse and 400 human non-olfactory GPCRs. Phylogenetic analyses showed that 329 of the receptors are found in one-to-one orthologous pairs, whereas other receptors may have originated from species-specific expansions.

Keywords: Bioinformatics, Evolution, GPCR, Phylogeny

Dóra Kristin Bjarnadóttir, Department of Neuroscience, Box 593, Uppsala University, SE-75124 Uppsala, Sweden

© Dóra Kristin Bjarnadóttir 2006

ISSN 1651-6206
ISBN 91-554-6489-0
urn:nbn:se:uu:diva-6627 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-6627)
Til Pabha


Contents

Introduction.................................................................................................................. 11
The methods of bioinformatics.................................................................................. 11
  Data banks .............................................................................................................. 11
  Search tools and sequence alignments ............................................................... 12
  Expressed sequence tags .................................................................................... 13
  Phylogenetic methods ......................................................................................... 14
The complexity of the human genome ...................................................................... 16
Evolution ................................................................................................................... 17
The superfamily of G protein-coupled receptors .................................................. 17
Molecular structure and function of GPCRs ......................................................... 18
G-proteins ................................................................................................................ 19
Classification of GPCRs in mammalian species ....................................................... 20
  The Glutamate (G) family of GPCRs ................................................................. 20
  The Rhodopsin (R) family of GPCRs ................................................................. 21
  The Adhesion (A) family of GPCRs ................................................................... 23
  The Frizzled/Taste2 family of GPCRs .................................................................. 23
  The Secretin family of GPCRs ............................................................................. 24
The mouse as a genetic model for the human ......................................................... 26
Research aims .......................................................................................................... 27

Results ....................................................................................................................... 28
  Paper I .................................................................................................................... 28
  Paper II ................................................................................................................... 28
  Paper III .................................................................................................................. 29
  Paper IV ............................................................................................................... 30
  Paper V .................................................................................................................. 30

Discussion ................................................................................................................ 31
  The Adhesion family forms several clans ........................................................... 31
  Ligand binding of Adhesion GPCRs ................................................................. 31
  Functional domains of the Adhesion family ..................................................... 33
    The Adhesion family GPS domain ................................................................. 36
  Tissue distribution and function of the Adhesion GPCRs ................................... 37
    The EGF-clan .................................................................................................... 37
    The BAI-clan .................................................................................................... 39
The CELSR-clan .......................................................................................... 40
The LEC-clan ............................................................................................... 41
The remaining Adhesion GPCRs ................................................................ 42
Conserved motifs of the Glutamate family ............................................. 45
The GPCR superfamily in human and mouse ........................................... 47

Conclusions ................................................................................................. 50
Future perspectives ..................................................................................... 52
Acknowledgements ...................................................................................... 54
References .................................................................................................... 55
Abbreviations

BAI Brain specific angiogenesis inhibitor
BLAST Basic Local Alignment Search Tool
BLAT BLAST like alignment tool
CA Cadherin repeats
CD Cell differentiating antigen
CELSR Cadherin EGF LAG seven-pass G- type receptor
CRD Cysteine-rich domain
CS Chondroitin sulphate
cAMP Cyclic adenosine monophosphate
EGF Epidermal growth factor domain
EGF-Lam Laminin type epidermal growth factor
EGF-TM7 Epidermal growth factor–seven transmembrane receptors
EMBL European Molecular Biology Laboratory
EMR EGF-module containing mucin-like hormone receptor
EST Expressed sequence tags
GABA γ-aminobutyric acid
G-protein Guanine nucleotide binding protein
GDP Guanosine 5´-diphosphate
GPCR G protein coupled receptor
GPS GPCR proteolytic domain
GRM Glutamate receptor, metabotropic
GTP Guanosine 5´-triphosphate
HBD Hormone binding domain
HE6 Human epididymal gene product 6
HMM Hidden Markov Model
LamG Laminin domain
LEC Lectomedin receptor
LNB-TM7 Long N-terminal B-family seven helical transmembrane receptor
ML Maximum Likelihood
MP Maximum Parsimony
mRNA Messenger ribonucleic acid
NCBI The National Center for Biotechnology Information
NJ Neighbor Joining
OLF Olfactomedin domain
RPS-BLAST Reversed position specific BLAST
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT-PCR</td>
<td>Real-time polymerase chain reaction</td>
</tr>
<tr>
<td>T2R</td>
<td>Taste receptors type 2</td>
</tr>
<tr>
<td>TA</td>
<td>Trace amine</td>
</tr>
<tr>
<td>TM</td>
<td>Transmembrane</td>
</tr>
<tr>
<td>7TM</td>
<td>Seven helical transmembrane</td>
</tr>
<tr>
<td>V1R</td>
<td>Pheromone receptor type 1</td>
</tr>
<tr>
<td>V2R</td>
<td>Pheromone receptor type 2</td>
</tr>
<tr>
<td>VF</td>
<td>Venus flytrap</td>
</tr>
<tr>
<td>Wnt</td>
<td>Wingless protein</td>
</tr>
</tbody>
</table>
Introduction

During the past decades there have been major advances in the field of molecular genetics. The wide availability of methods for DNA sequencing in the early 1980s and the sequencing of the first microbial genome in 1995 (Fleischmann et al., 1995) represented significant steps along the way. Now, ten years later, the genome sequencing of 176 eukaryotic organisms is in progress, 19 are considered complete and 85 are already assembled to some extent (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi), including the human (Lander et al., 2001; Venter et al., 2001) and mouse genomes (Gregory et al., 2002; Waterston et al., 2002). Not only are genomic sequences created at an exponential rate, but we are also gaining better understanding of the complexity of eukaryotic genomes. The existence of introns and exons have been described (Gilbert, 1978) as well as the importance of non-coding DNA sequences (Nei, 1969). All this has contributed to a very rapid accumulation of biological information, which has created a need for efficient ways to store biological data as well as practical tools to view and analyse it. Thus, the momentum has increasingly been shifted towards computational science, and created a foundation for a new field, which we now know as bioinformatics. This field of expertise strives to organise and bring biological information together using computers (in silico) and furthermore to extract meaningful knowledge from this information, which will lead to a better understanding of the biological system.

The methods of bioinformatics

Data banks

The rapid creation of biological information has prompted creation of bioinformatic resources in form of databases that store vast amount of sequence data. Examples include NCBI’s GenBank (Bilofsky et al., 1986), the EMBL data Library (Hamm & Cameron, 1986) and later the Celera Discovery Systems database (Kerlavage et al., 2002). GenBank and EMBL along with the DNA DataBank of Japan (DDBJ) are part of the International Nucleotide Sequence Database Collaboration. They are all publicly available, free of charge and since they exchange data on a daily basis they should contain equivalent information. The Celera database on the other hand is a private...
database, which can be accessed only on subscription. To give an estimation
of the size of these databases GenBank currently contains about 47 million
sequence records (http://www.ncbi.nlm.nih.gov/Genbank/index.html). Read-
ily accessible genome browsers also soon became available, for example
UCSC’s Human Genome browser, a web tool for rapid and reliable display
of any requested portion of the genome at any scale (Kent et al., 2002).

Search tools and sequence alignments
Development of search tools for the databases followed, with one of the
major breakthroughs being the development of the Basic Local Alignment
Search Tool (BLAST) (Altschul et al., 1990). BLAST breaks the sequences
of a given dataset into short fragments and makes use of a similarity score
matrix to look for an identical or close match between those fragments. Once
such a hit is encountered the hit is extended in both directions to generate a
local alignment segment. Identical and conserved residues between segments
get positive scores, while unlikely replacements get negative scores. The
scores are summed up to find sequence segments with the highest identity,
defined as maximal segment pairs (MSP). BLAST can search all local MSPs
resulting in relatively conserved subsequences within two sequences. Since
the method does not require the sequence similarity to be global (consistent
throughout the whole sequences), the method is able to detect weak but bio-
logically significant sequence similarities. This sensitivity enables compari-
son of partially sequenced genes and distantly related proteins, which share
only isolated regions of similarity. A similar BLAST-like alignment tool
(BLAT) exists for searches in UCSC’s genome browser (Kent, 2002). The
application of profile Hidden Markov Models (HMMs) has also proven im-
mensely useful in detecting sequence identity. By creating a profile HMM
from related sequences, it is possible to define conserved motifs of the data-
set. A consensus sequence for the desired motif can be build and further used
in multiple types of searches (Baldi et al., 1994; Krogh et al., 1994).

The basic search tools and databases can facilitate the rapid gathering of
sequence data for gene or protein families to be further analysed. Some of
the most widely used tools in bioinformatical analyses include multiple se-
quence alignment tools such as CLUSTALW (Thompson et al., 1994). The
multiple alignment is built up progressively by a series of pair-wise align-
ments used to generate a distance matrix and subsequently a phylogenetic
tree. The multiple alignment is then built following the branching order in a
phylogenetic tree. Calculations of scores for all possible pairs of aligned
residues, with gap penalties taken into account, are used to align the two
closest sequences first. For further alignment the two sequences are treated
as one, so that any gaps created between the two cannot be moved. Again,
two of the closest related sequences are aligned and so on, gradually adding
in the more distant ones. For more accurate alignments, gap penalties are
reduced in short stretches of hydrophilic residues (usually indicating loop or random coil regions) or positions, where there are already many gaps. A series of four score-matrixes (for instance BloSum62) are available, as different matrices will be optimal at different evolutionary distances, or for different classes of proteins.

Expressed sequence tags

Expressed sequence tags (ESTs) are short sequence reads of cDNA, typically about 300-700 nucleotides long (Adams et al., 1991). They are produced from cloned mRNAs derived from certain cells, tissues or organs. The mRNA is converted to cDNA (a much more stable compound) using reverse transcriptase and thereafter sequenced by one-shot sequencing from either end to produce 5’ESTs or 3’ESTs. The 5’ESTs usually code for proteins whereas the 3’ESTs are likely to fall in non-coding or untranslated regions (http://www.ncbi.nlm.nih.gov/About/primer/est.html).

Nowadays, ESTs are generated on a massive scale relatively inexpensively. The sequence data is stored in databases such as the NCBI dbEST, which currently contains over 7.5 million human and 4.5 million mouse ESTs (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). EST data have proven extremely useful in for example helping to determine the coding sequence of genes, creating expression charts, predicting intron-exon boundaries, determining alternative splicing, and finding single nucleotide polymorphisms.

However, it must be taken into consideration that EST data are in some ways incomplete. Because they are unedited one-shot sequencing reads, EST are prone to errors, attaining at best 97% accuracy rate (Hillier et al., 1996). Moreover, ESTs generated are in proportion to the abundance of the mRNAs in the tissues. Thus genes expressed at very low levels are not likely to be found within EST datasets, while abundantly expressed genes can be over represented. Normalisation and subtractive methods have been used to compensate for this bias (Marra et al., 1998). Other shortcomings of ESTs include vector and genomic contamination, premature mRNA, and intronic contamination (Murray et al., 2005).

Despite these shortcomings ESTs have proven extremely valuable in characterising the human genome (Marra et al., 1998). In the future, with EST databases growing and representation of different sequences and tissues increasing, ESTs will probably continue to be very important for genomic research.
Phylogenetic methods

Other examples of methods widely used for sequence data and evolutionary relationship analyses include algorithms for calculation of phylogenetic trees such as Neighbor-Joining (NJ), Maximum Parsimony (MP), and Maximum Likelihood (ML) seen in the commonly used PHYLIP package (Felsenstein, 1989; Felsenstein, 2005; Vandamme, 2003). The three methods are based on different principles. In NJ (Figure 1a), the sequences are initially arranged in a star shaped tree. A pairwise distance matrix is then used to link the least evolutionary distant pair of sequences. Distances between sequences are expressed in fraction of sites that differ. The least different pairs are connected by an ancestral node. Then the distance matrix is modified and the calculations repeated regarding the two sequences, previously joined by a node, as one. Another most related pair is then found in the revised matrix and the calculations continue in this way until all pairs have been given a score.

In theory, MP uses an alignment for a given dataset, evaluates all possible tree topologies and aims to find the tree topology that can be explained with the smallest number of evolutionary changes (i.e. mutations). Not all sites will be informative in such an alignment. For example characters conserved in the same position throughout all the sequences are non-informative. The MP algorithm evaluates all possible trees for each informative site and the tree with fewest evolutionary changes is chosen (see Figure 1b).

In theory, ML also uses an alignment, considers every reasonable tree topology and evaluates the support for every topology by calculating the probability of observing the sequence found in leaf nodes. The probabilities are dependent on which evolutionary model is used. The tree most likely to explain the given dataset is eventually chosen (see Figure 1c). As described above, both MP and ML in theory score all possible trees aiming to find the one considered to best fit the sequence alignment. However, as the size of the dataset increases, so do the possibilities for different tree topologies and it soon becomes unrealistic to perform an exhaustive search. Thus, for datasets exceeding the size of ten sequences, only a subset of possible trees is examined using algorithms, faster than exact searches but with no guarantees that the best possible tree is examined. As for now, all the bioinformatic tools and methods mentioned above are widely used for genetic analyses as scientists try to figure out the various complex aspects of genomes.
Figure 1: a) NJ: In this case B and C are the least different sequences according to a pairwise distance matrix (not shown). They are paired and joined by an ancestral node where after they are treated as one and the distance matrix modified. Again the least different sequences are paired, in this case A and B/C are joined by an ancestral node etc. b) MP: In theory, all tree topologies are tested. The tree with fewest evolutionary substitutions (ES) is then chosen, in this case tree one with ES = 10. Figure redrawn from (http://www.icp.ucl.ac.be/~opperd/private/parsimony.html) c) ML: Likewise all tree topologies are in theory considered, looking at each amino acid position and calculating the probability of the expected amino acid in an ancestral node. In this example there are four possible nucleotides for node x and node y (left tree). Thus there are 16 different possible trees; one of these, and how the probability is calculated, is shown (right tree). After calculating P1-P16, the probabilities are added to get the probability of the tree to the left and then a tree with another topology is evaluated. The tree with the highest probability is considered the most likely tree.
The complexity of the human genome

The completion of human genome sequencing was awaited with great anticipation, and in the hope that it would shed light on unanswered questions, such as how many proteins the human proteome contains? That question still remains unanswered, but recent estimations for the number of protein coding genes in human lie around 25,000 (International-Human-Genome-Sequencing-Consortium, 2004). This is thought to constitute only around 1.5% of the whole genome material (Brosius, 2003). The existence of exons and introns lead to the discovery of alternative splicing mechanisms, including exon skipping, alternative exon insertions, use of alternative 5’ or 3’ splice sites, and intron retention. Each of these mechanisms is now recognised as an important contributing part to the complexity of eukaryotic proteomes. Present predictions state that an average of three human protein products can possibly result from each gene (Humphery-Smith, 2004).

The remaining 98.5% of the genome was for a long time regarded as merely non-sense RNA (Nei, 1969) or “junk” RNA (Brosius, 2003). However, it now seems that at least 30-50% of the genome is transcribed (Mattick, 2003). Whereof around 29-49% are considered to represent untranslated or non-coding RNA (ncRNA) (Brosius, 2003) including small nucleolar RNAs (snoRNAs), micro RNAs (miRNAs), short interfering RNAs (siRNAs), and other tiny RNAs, in addition to larger untranslated RNAs (Brosius, 2005). Some of these sequences lie in introns, which account for at least 30% of the human genome (Mattick, 2005) and over time evidence has been gathering supporting functional roles for some of the ncRNAs for example snoRNAs direct modification of ribosomal RNAs, miRNAs are involved in gene expression, and siRNAs mediate down regulation of gene expression (Szymanski et al., 2003). The larger ncRNAs resemble mRNA in that they are polyadenylated, often spliced but lack substantial open reading frames and could very well have cellular function, including regulatory roles (Brosius, 2005).

Thus even after the complete sequencing of the human genome we are still only in the early stages of identifying all the functional components of the human genome and its products. Sequence analyses and genome comparison between species is however a very important step and can provide significant leads for further studies such as determination of functional roles, interaction with other proteins, signal pathways, and drug targets.
Evolution

Various mechanisms are considered to underlie the complicated process of evolution. Not only can new genes arise but existing genes can also undergo functional changes or even be silenced. New genes can be generated for example through whole genome duplications, duplications of individual genes or chromosomal segments. The 2R hypothesis theory for whole genome duplication proposes that two rounds of large-scale genomic duplications (tetraploidisations) occurred in early vertebrate ancestry, more than 400 MYA, resulting in up to four copies of each gene originated from invertebrates, such as *Drosophila* (Lundin, 1993; Ohno, 1970; Wolfe, 2001). Observations from the genomic databases for several eukaryotic species suggest that duplicate genes arise at a very high rate, on average ~0.01 per gene per million years (Lynch & Conery, 2000). Far from all of those genes remain permanently in the genome since they can undertake changes such as neofunctionalisation (the gene copy acquiring a novel function that becomes preserved by natural selection), subfunctionalisation (both copies become partially compromised by mutation accumulation to the level of the ancestral gene) or the most common, nonfunctionalisation (one copy is simply silenced within a few million years of the duplication) (Force *et al.*, 1999). Together these above mentioned mechanisms result in complex evolutionary relationships between species. Many genes can be found conserved as orthologues (a homologous sequence found in different species and derived from a common ancestral gene); other may have undergone expansions or deletions within a certain species. In the past, we have mainly relied on analyses of fossils for interpreting our evolutionary history. Now with the material for genomic comparison accumulating, it should be possible to use it to find the ancestral ties between diverse organisms. Thus combining research on fossils with comparison of genomic material may enable us to uncover evolutionary relationships between the different forms of life.

The superfamily of G protein-coupled receptors

G protein-coupled receptors (GPCRs) form one of the largest superfamilies of cell-surface receptors. It constitutes around 800 human genes, which accounts for about 2-3% of all human genes (Venter *et al.*, 2001). Members of the superfamily are situated transmembranally in cells where they recognise endogenous ligands (such as hormones, neurotransmitters, growth, and developmental factors), or sensory messages (such as light, odors, vision, and pain). The role of GPCRs is to transduce a signal over the membrane to a G-protein (Bockaert & Pin, 1999). GPCRs are expressed virtually in all types of tissues in the body (Fredriksson & Schioth, 2005). They are involved in most types of physiological and pathological processes. However, they are
often expressed at low levels and in specific cells types, which contributes to the fact that they are the most important family of proteins serving as targets in drug discovery. At present, approximately 50% of all newly introduced drugs are targeted at GPCRs and 25% of the 100 top-selling drugs are targeted at members of this protein family (most to GPCRs that bind amines). From the several hundred members of the GPCRs family only around 30 representative targets of currently marketed drugs have been revealed. There are natural ligands still to be found for all the so-called orphan receptors (where neither ligand nor physiological function is known) that have been identified within the human genome (Klabunde & Hessler, 2002).

Molecular structure and function of GPCRs

GPCRs consist of a polypeptide chain of variable length (from about 300-1000 amino acids) that passes repeatedly through the cellular membrane, making up the distinctive feature characterising all GPCRs, the seven α-helical transmembrane (7TM) regions (Ulloa-Aguirre et al., 1999). So far, only one GPCR, the bovine-rhodopsin, has been structurally determined by crystallisation and thus provides the most accurate information on GPCR structure (Palczewski et al., 2000). The common 7TM helices are of unequal length ranging from 20-27 amino acids with diverse degrees of hydrophobicity (Bockaert & Pin, 1999). The helices are also irregular in orientation and in some cases steric hindrances between amino acid side chains can elicit their shape to be slightly bent. In general, however, they are thought to form a barrel-shape perpendicular to the plane of the membrane with TMIII in the center, as has been shown for the bovine rhodopsin (Stenkamp et al., 2002).

The helices are kept in close proximity of one another and hydrogen bonding, among other things, helps to maintain the core tightly packed in an inactive state. Three intracellular loops (IC) and three extracellular loops (EC) connect the 7TM helices. They are usually predicted to be about 10-40 amino acids in length, except for IC3, which may be as long as 150 amino acids. IC2 and 3 are the two main loops engaged in G-protein recognition and activation. EC1, EC2, and EC3 are considered to play an important part in structure stabilisation and the binding of ligands (Ulloa-Aguirre et al., 1999). An N-terminus protrudes from TMI at the extracellular side and at the intracellular side a C-terminus connected to TMVII. Both termini are highly variable in length, and the N-termini can comprise different functional domains each of which is able to provide specific properties to the relevant receptor (Bockaert & Pin, 1999).

Binding of a ligand at the extracellular side activates GPCRs. The ligand binding varies depending on the particular subfamily of GPCRs in question as well as on the size and structure of the ligand. For instance, subfamilies containing members with short, or almost non-existing N-termini, which
bind relatively small ligands most often bind at the upper part of the 7TM regions. Subfamilies with long N-termini, which bind larger ligands, tend to use the N-terminus, its functional domains, EC loops and sometimes also the 7TM regions for binding. GPCRs undergo conformational changes upon ligand binding (Kristiansen, 2004) and the orientation of TMIII and TMVI is considered to unmask the GPCRs binding sites for various G-proteins on the intracellular side which can transduce a signal to a range of intracellular effector molecules (Boekaert & Pin, 1999). Most GPCRs activate a chain of events that alters the concentration of one or more small intracellular signalling molecules through complex pathways (Neves et al., 2002).

G-proteins

G-proteins are named so because of their interaction with the guanine nucleotides, GTP and GDP. Classically, G-proteins are heterotrimers made up of α-, β-, and γ-subunits. The α-subunit binds GDP or GTP and has slow GTPase activity. The β- and γ-subunit form a tightly associated complex, which is anchored to the intracellular side of the plasma membrane by a lipid chain covalently attached to the γ-subunit. Upon activation, a conformational change occurs allowing the G_{αγ} complex to displace GDP with GTP. Subsequently this leads to activation of the α-subunit as well as dissociation and activation of the βγ-subunit. The GTP-bound form of the G-protein, the α-subunit, and in some cases the free βγ-subunits, initiate cellular responses by altering the activity of specific effector molecules. Gradually, GTP is hydrolysed to GDP, leading to dissociation of G_α from the effector and reassociation with the G_βγ dimer, regenerating the inactive G_{αγ} heterotrimeric complex (Radhika & Dhanasekaran, 2001). To date, 28 α-subunits (formed from 16 genes), 5 β-subunits, and 12 γ-subunits have been cloned and identified (Cabrera-Vera et al., 2003). G-proteins can be divided into four different families according to their α-subunits sequence similarity: G_α_s, G_α_i/o, G_α_q/11 and G_α_{12/13} (Neer, 1995; Rens-Domiano & Hamm, 1995). The family of G_α_s and G_α_i/o produce stimulation and inhibition of the enzyme adenylyl cyclase, which in turn affects the production of cAMP within the cell. G_α_{q/11} family members activate phospholipase C, generating both inositol trisphosphate (IP_3) and diacylglycerol (DAG). IP_3, is a soluble molecule, which can diffuse through the cytosol and bind to receptors on the endoplasmic reticulum causing the release of Ca^{2+} ions into the cytosol. DAG on the other hand remains in the cell membrane where it recruits protein kinase C (PKC), which is able to phosphorylate different proteins leading to their activation or inactivation. Proteins from the G_α_{12/13} family are implicated in the regulation of small GTP binding proteins, such as Rho, which can further activate phospholipase D (Plonk et al., 1998; Yuan et al., 2001). The βγ-dimer has been shown to
regulate inward rectifier G-protein gated potassium channels (GIRKs), adenylyl cyclase and phospholipase Cβ (PLCβ).

Classification of GPCRs in mammalian species

There are different approaches for classifying the GPCRs. One of the most frequently used methods is to divide them into clans (families A-F) and then further separating them into sub-clans. The well known A-F system is designed for both vertebrate and invertebrate GPCRs. Family A contains receptors similar to rhodopsin and biogenic amine receptors, family B secretin- and calcitonin related receptors, and family C holds the metabotropic glutamate receptors. However, some families of the A-F system do not exist in humans (e.g. clan D and E, which represent fungal pheromone receptors and cAMP receptors) (Kolakowski, 1994). Therefore, another system has been suggested for classifying mammalian GPCRs, namely the GRAFS classification system. The receptors are grouped into five major families, according to phylogenetic analyses, named Glutamate (G, with 15 members), Rhodopsin (R, 701 members), Adhesion (A, 30 members), Frizzled/Taste2 (F, 24 members), and Secretin (S, 15 members). Twenty-three protein sequences, which could not be designated to any of the five families, were categorised as "other 7TM receptors" (Fredriksson et al., 2003c). Here, we opt to use the GRAFS classification system.

The Glutamate (G) family of GPCRs

The Glutamate family (also termed family C) has previously been described to contain receptors for the main neurotransmitters, glutamate, and GABA (γ-aminobutyric acid), one receptor for Ca²⁺ and Mg²⁺ binding (CASR), three type 1 taste receptors (T1R1-3), pheromone receptors type 2 (V2Rs), and a few orphan receptors. These clans agree with the results of phylogenetic analysis. The eight metabotropic glutamate binding receptors (GRMs) can be further divided into three clans according to their sequence similarity, transduction pathways, and pharmacology. Group I includes GRM1 and GRM5, Group II: GRM2 and GRM3, and Group III: GRM4, GRM6, GRM7, and GRM8 (Hermans & Challiss, 2001; Pin et al., 2004). To date, most knowledge has been gained on the GRM, GABA, CASR, and T1Rs clans of the Glutamate family. In general, they all contain very long N-termini, which make up a ligand-binding site, similar to what is seen for the bacterial periplasmic amino acid-binding proteins. The mechanism of the binding site is most often referred to as the Venus flytrap (VF) module (O’Hara et al., 1993) and the details of its functions are addressed in the Discussion chapter.
The *Rhodopsin* (R) family of GPCRs

The family of *Rhodopsin* (also termed family A) is by far the largest of the GPCR families. The vast number of members and their extremely diverse ligands have made this receptor family the most studied from both structural and functional point of view over the past years. According to the GRAFS classification system it is subdivided into α-, β-, γ-, and δ-Rhodopsin (Fredriksson *et al.*, 2003c). The α-Rhodopsin is the largest subfamily out of the four, containing the only GPCR crystallised to date, the bovine rhodopsin. The family makes up several clear phylogenetic branches, which in most cases are consistent with their ligand binding profiles and/or pharmacological properties.

Many α-Rhodopsin family members studied to date bind biogenic amines (Kroeze *et al.*, 2002; Neve *et al.*, 2004; Strosberg, 1993). Figure 2a provides a simplified schematic illustration of one such amine binding receptor, the adrenergic β3 receptor (ADRB3) and its ligand norepinephrine (Nagatomo *et al.*, 2001; Strosberg, 1997). In general, the biogenic amines are small compounds, and as exemplified by ADRB3, they are able to bind within a small hydrophobic pocket made up from, in this case, four of the 7TM regions (Strosberg, 1997). Thus the N-termini of amine-binding receptors show no obvious role in ligand binding. The β-Rhodopsin is the smallest subfamily out of the four built up of smaller branches than the α-Rhodopsin. Most β-Rhodopsin members bind peptide ligands including neuropeptide Y, neuropeptide FF, and neuromedin U (Brighton *et al.*, 2004; Cabrele & Beck-Sickinger, 2000; Mollereau *et al.*, 2002). Binding of most peptides differs from the binding of amines. While most amines are prone to bind in hydrophobic pockets created by the several helixes of the TM regions, peptide receptors seem to rely on the upper parts of the TM helices, extracellular loops, and even N-termini to bind a peptide as shown for the Tachykinin 1 receptor (TAC1R) and its ligand Substance P (SP) (Elling *et al.*, 2000; Turcatti *et al.*, 1997). Figure 2b illustrates a simplified version of TAC1R and the amino acids, which have been found important in SP binding. The largest clan of γ-Rhodopsins, the chemokine receptors, contains numerous receptors that bind to chemotactic cytokines or chemokines (Onuffer & Horuk, 2002). However, members of this subfamily also bind to opiates and peptides such as formyl peptide. Finally, the δ-Rhodopsin subfamily comprises the fewest clans, the largest clan binding to purines (ADP, ATP, UDP, and UTP) (von Kugelgen & Wetter, 2000). All the above-mentioned subfamilies contain very interesting orphan GPCRs where neither ligand nor physical function is yet known.
Figure 2: a) A simplified illustration of a biogenic amine binding receptor. The ligand, norepinephrine, is shown in purple and the TM regions of the receptor, adrenergic β3 (ADRB3) in orange, extracellular loops are marked e1, e2, e3, and intracellular loops i1, i2, i3. Amino acids, which have been found to be involved in binding of the ligand by mutagenesis and photoaffinity labelling, include the following (shown in red): D117 (likely forms a salt bridge), S169, S209 and S112 (are all thought to form hydrogen bonds), and F309 (involved in hydrophobic interactions). D83 and Y336, are thought to be involved in G-protein activation along with the i2, i3 and the cytoplasmic tail. Redrawn from (Nagatomo et al., 2001; Strosberg, 1997).

b) A simplified illustration of a peptide binding GPCR represented by the Tachykinin 1 receptor (TAC1R) and its ligand, Substance P (SP). It is suggested that the ligand does not penetrate deeply below the extracellular face of TAC1R (as in the case of ADRB3) since the residues most important for binding lie at the upper parts of the TM helices, in the N-terminal, or extracellular loops. Redrawn from (Axelrod & Saavedra, 1977; Elling et al., 2000; Turcatti et al., 1997).
The Adhesion (A) family of GPCRs

The Adhesion family (also termed family B2 since they are a distinct branch within family B), has also been referred to as EGF-TM7 (epidermal growth factor-seven span transmembrane receptors), or called the LNB-TM7 family (long N-terminal seven transmembrane receptors related to family B). The main characteristic of the family is their relatively long N-termini, distinguishing them from for example the Rhodopsin GPCRs. The long N-termini can extend up to a few thousand amino acids and each typically exhibits one or more functional domains, many of them with adhesive properties. These functional domains are generally unique for the Adhesion members and not found within other GPCR families (Foord et al., 2002; Harmar, 2001). The possible roles of the N-termini are discussed in detail in the Discussion chapter. In contrast to the Rhodopsin family, the Adhesions are coded for by many exons and their genomic structure is in general very complex, which is one of the reasons that most of these genes have been difficult to study and were described only relatively recently. Thus the majority of Adhesion GPCRs are still orphans, where neither ligand nor function is known, making them a challenging group to study.

The Frizzled/Taste2 family of GPCRs

The family of Frizzled/Taste2 contains receptors from two receptor clans, the Frizzled clan and the Taste2 receptor clan, that are weakly similar in sequence (Fredriksson et al., 2003c). The Frizzled clan consists of ten Frizzled (FZD) and one smoothened (SMOH) receptor. In general, the FZD receptors bind to secreted wingless (Wnt) proteins of approximately 350 amino acids. The binding site is situated within a cysteine rich domain (CRD) in the N-termini (Dann et al., 2001), consisting of 120-125 residues with ten conserved cysteines forming disulphide bonds (Huang & Klein, 2004). Figure 3c shows a simplified illustration of the CRD domain and TM regions of a FZD receptor. Both FZD1 and FZD2 have been shown to couple to G-proteins (Malbon et al., 2001) through two different pathways. FZD1 by the canonical Wnt/β-catenin pathway, resulting in stabilisation of β-catenin when activated by the ligand, and FZD2 by the Wnt/calcium pathway, which can result in increased intracellular calcium (Huang & Klein, 2004). Overall the role of the FZD receptors include generation of cell polarity, embryonic induction, and specification of cell fate (Cadigan & Nusse, 1997; Moon et al., 1997). On the other hand, the SMOH bind Hedgehog proteins connected to a signalling pathway effecting cell growth and differentiation and pathological conditions such as growth of tumours (Lum & Beachy, 2004). As their name implies, the Taste2 receptors are involved in
mediating bitter taste perception (Adler et al., 2000; Chandrashekar et al., 2000; Nelson et al., 2001).

The Secretin family of GPCRs

The Secretin family contains 15 receptors, which interact with large glycoprotein hormones (30-140 amino acid residues in length). The family includes receptors for: secretin (SCTR), calcitonin (CALCR, CALCRL), corticotrophin-releasing factor and urocortin (CRHR1, CRHR2), glucose-dependent insulintropic peptide (GIPR), glucagon or glucagon like peptides (GCGR, GLP1R, GLP2R), growth hormone releasing hormone (GHRHR), parathyroid hormone (PTHR1, PTHR2), pituitary adenylate cyclase-activating polypeptide (PACAP), and vasoactive intestinal peptide (VIPR1, VIPR2) (Martin et al., 2005). In general, all Secretin receptors comprise a moderately long N-terminus (120-140 amino acids). Each receptor contains six conserved cysteine residues in the N-terminus. They are connected by three disulphide bonds supposedly forming a distinct tertiary structure as has been demonstrated for CRHR1 (Perrin et al., 2001), PTHR1 (Grauschopf et al., 2000), and GLP1R (Bazarsuren et al., 2002).

A two-domain model for the mechanism of Secretin ligand-receptor interaction has been proposed in which the C-termini part of the ligand binds to the N-termini of the receptor creating an affinity trap, whereby the N-termini part of the ligand can bind to the TM region of the receptors leading to activation (Gardella & Juppner, 2001; Hoare, 2005). Figure 3a and b show a schematic illustration of the two-domain model and how PTHR1 is thought to bind its ligand. Overall, family members of Secretin receptors are involved in various physiological processes such as: regulation of bone turnover and calcium homeostasis, thus linked to osteoporosis (CALCR and PTHR) (Body, 2002; Hodsmann et al., 2005), insulin release and glucose homeostasis (GLP1R and GIPR) (Nauck et al., 2004), stress response (CRHR1) (Bale & Vale, 2004; Pellemounter et al., 2002), and vasorelaxation (VIPR) to name some (Sherwood et al., 2000).
Figure 3: a) The two-domain ligand-binding model for the Secretin family. The C-terminus of the ligand has high affinity for the N-terminus of the receptor, leaving the ligand N-terminus in close proximity to the receptor TM region inducing a low affinity binding. Once bound, the ligand activates the receptor and a G-protein can attach at the intracellular site. Redrawn from (Hoare, 2005). b) A schematic figure of the ligand binding of human parathyroid hormone receptor 1 (PTHR1). Cysteine residues (black), conserved among the Secretin family members, fix the receptor N-terminus in a tertiary structure. Cross-linking between the ligand and receptor residues are shown in dotted lines. Redrawn from (Gardella & Juppner, 2001; Hoare, 2005; Hoare & Usdin, 2001). c) A simplified figure of a Frizzled receptor and the cysteine rich domain (CRD) (Sagara et al., 1998). According to the crystal structure of mouse Fzd8 CRD it is held together in a tertiary structure by cysteine disulphide (light grey lines; C3-C64; C11-C57; C48-C87; C76-C115; C80-C104) (Dann et al., 2001). The CRD is necessary, and alone sufficient, for binding of Wnt (Huang & Klein, 2004) giving the TM regions no obvious role in ligand binding. Redrawn from (Dann et al., 2001; Sagara et al., 1998).
The mouse as a genetic model for the human

GPCRs can be found in almost all eukaryotic organisms, including insects (Hill et al., 2002) and plants (Josefsson, 1999). Nevertheless, like for most other research on human disease and development, the mouse serves as one of the premiere genetic models for GPCRs. Firstly, it is a mammal and detailed analyses of organs, tissues, and cells have revealed many physiological, anatomical and metabolic parallels with humans, including whole organ system reproduction and behaviour (Bradley, 2002).

Secondly, the mouse has the closest to ideal genetic tractability. The genomes of human and mouse are approximately the same size, containing around $3 \times 10^9$ base pairs. In the mouse, these are distributed on 20 chromosomes compared to 23 chromosomes for the human. A counterpart for virtually every gene in the human genome can readily be identified in mouse and both genomes contain large segments of synteny (10-20 mega base segments containing dozens to hundreds of genes that have the same gene order and similar intergenic distances between the two species) (Perkins, 2002).

Thirdly, genetic manipulation within the living mouse has become routine and can these days be done with extraordinary precision. The ability to engineer mutations in specific genes, and to generate mice with induced mutations, facilitates great possibilities for identification of genetic variants of biological interest.

In summary the sequenced genomes of human and mouse and the bioinformatic tools available can be used to analyse for example the family of GPCRs to give meaningful knowledge regarding orthologous relationships and receptor conservation between the species. Such information would be useful for studying of the orphan receptors as the mouse orthologues can serve as a genetic model for leads regarding tissue distribution, ligand binding, and function in the human.
Research aims

The overall aim was to investigate the genetic repertoire of the family of G protein-coupled receptors (GPCRs), in human and mouse in particular. The specific aims were:

- To search the human and mouse genomes for novel genes belonging to the relatively newly recognised Adhesion family of GPCRs. Moreover, to determine the orthologous relationships between the receptor of human and mouse as well as to gather expressed sequence tag (EST) data to get an overview of the tissue expression of members of this family.

- To identify novel splice variants for human members of the Adhesion family based on EST and mRNA database searches. Furthermore, to classify the splice variants into functional and non-functional as well as determine, which functional domains are present or absent in the N-termini of the splice variants compared with the wild type receptors.

- To collect a representative and up to date dataset of human and mouse GPCRs. Further, to determine their evolutionary relationships using phylogeny as well as to study ESTs to establishing expression charts of GPCRs in these species.
Results

Paper I
Thorough searches in NCBI’s human and mouse genome databases (http://www.ncbi.nlm.nih.gov) as well as the Celera genome database (http://www.celera.com) led to the finding of two new human genes and seventeen mouse genes belonging to the *Adhesion* family of GPCRs. Coding regions for each of the original findings were verified using mRNA and EST data. The two new human sequences were confirmed unique and provided with GPR numbers, GPR133 and GPR144. The mouse genes were named according to their closest human orthologue (Gpr110, Gpr111, Gpr112, Gpr113, Gpr114, Gpr115, Gpr116, Gpr123, Gpr124, Gpr125, Gpr126, Gpr128, Lec1, Lec2, Lec3, Gpr133, and Gpr144).

A phylogenetic analysis of the 7TM regions of the entire set of *Adhesion* GPCRs was carried out (excluding VLGR1; very large G protein-coupled receptor 1). It showed that there exist eight clusters of *Adhesion* GPCRs, whereof the human GPR133 and GPR144 receptors, together with their orthologues, make up a distinct cluster. Each of the human receptors group together with a mouse receptor in a one-to-one orthologous pair. The only exception is EMR2 and EMR3, which do not seem to have any orthologues in mouse. Finally, alignments of the consensus sequence of each phylogenetic group showed conserved motifs of the family. Database searches resulted in findings of functional domains in the N-termini of the *Adhesion* GPCRs, which were illustrated, as well as the expression patterns in several major organs.

Paper II
In this study, we used mRNA and ESTs sequences to identify splice variants in all human *Adhesion* family GPCRs. A total of 239 mRNA-sequences and 1218 EST-sequences were gathered with uneven distribution between the receptors. In general the majority (over 70%) had between three and eight mRNA-sequences and over nine EST-sequences. These sequences gave support for 53 unique splice variants subsequently categorised into functional and non-functional variants. A variant was only considered functional if it
contained the whole seven-transmembrane region \( \text{RPS-BLAST model 7tm}_2 \) for Adhesion GPCRs) and otherwise classified as non-functional.

In total, we identified 29 functional splice variants for the following nineteen receptors (number of variants in parenthesis): CD97 (2), CELSR3 (2), EMR2 (1), EMR3 (1), GPR56 (2), GPR110 (1), GPR112 (1), GPR113 (1), GPR114 (1), GPR116 (2), GPR123 (1), GPR124 (1), GPR125 (1), VIGR/GPR126 (2), GPR133 (2), HE6 (5), LEC1 (1), LEC2 (1), and LEC3 (1). The splice variants for GPR116, GPR125, GPR126 and HE6 were the only ones found conserved in other species. The majority of the splice variants proved to differ in the N-terminus or contained extended or truncated extracellular or intracellular loops. In about half of these, the alterations do not affect the number of conserved domains. However, we found ten cases where alternative splicing resulted in the variant being deprived of one or more functional domains. The alternative splice variants never showed more domains than the recognised wild type. Changes in parts, other than the N-termini, were rarely found.

**Paper III**

Mining of NCBI’s human, mouse, Fugu and zebrafish genome databases (http://www.ncbi.nlm.nih.gov) with the addition of BLAST searches in both the Celera genome (http://www.celera.com) and Ensembl databases (http://www.ensembl.org/) were carried out in order to identify new genes belonging to Glutamate GPCR family. The findings resulted in 163 sequences considered functional genes. Thereof, 73 mouse, Fugu, and zebrafish pheromone sequences (V2Rs) were verified, using genome alignments to related sequences and named according to their chromosomal positioning.

Comparison of the repertoires between the species confirmed that they vary greatly. The mouse has by far most members (79) compared with the others: Fugu (30), zebrafish (32), and human (22). Phylogenetic trees for the Glutamate GPCRs dataset were created and confirmed that the sequences cluster in five major groups. All the human proteins have a mouse orthologue, apart from the V2Rs (which do not seem to be functional in human). However, the overall phylogenetic relationship of the four species is complex with evidence of expansions or deletions within the genomes of the different organisms. Conserved elements in the N-termini, 7TM regions, and C-termini of all the Glutamate GPCRs were investigated, using alignments of consensus sequences of different phylogenetic groups as well as an overview of which receptors can be found in numerous major organs.
Paper IV

Search for trace amine (TA) GPCR receptors using various methods and databases (NCBI’s, Ensembl’s, Celera genomics, and JGIs-Joint Genome Institute databases) yielded in a dataset of 37 mammalian receptors whereof four rat (PNR, GPR57, GPR58, and TaA) and 14 mouse (TA2-TA4, TAA, TAH, PNR; GPR57, GPR58) sequences represented new receptors. Intron-exon boundaries of those were manually verified and they were named according to their human orthologues (given the name “TA” A–H if a human orthologue was missing). Additionally 57 previously unknown zebrafish genes coding for TA receptors and eight Fugu genes (as well as seven pseudogenes) were found.

The evolutionary relationship of the protein family was investigated using phylogenetic analysis and showed that the TA receptors grouped into seven groups, three of which contain solely mammalian genes and three only fish genes. Syntenic regions identified by chromosomal mapping at UCSC showed well conserved genomic organisation between human, mouse, and rat TA receptors, all tightly arranged on HSA6 (6q23.2), mouse chromosome 10, and rat chromosome 1 (1p12). Receptors of the syntenic regions clustered in the same phylogenetic groups.

Paper V

Systematic searches applying several search tools such as BLAST, BLAT, Hidden Markov Models (HMMs), and searches in literature data, in both the human and mouse genomes, resulted in a comprehensive dataset of a full-length version of 495 mouse and 400 human functional non-olfactory GPCRs. The dataset was classified into families according to the GRAFS classification system; Glutamate (G), Rhodopsin (R) (subdivided into α-, β-, γ-, and δ-Rhodopsin), Adhesion (A), Frizzled/Taste2 (F), and Secretin (S).

A detailed phylogenetic analysis of the transmembrane regions of each family was carried out in order to establish accurate orthologous pairs. Overall, 329 of the receptors were found in one to one orthologous pairs, while 119 mouse and 31 human receptors seem to have originated from species-specific expansions or deletions. The average percentage similarity of the orthologue pairs is 85% while it varies between the main GRAFS families from an average of 59% to 94%. The orthologues pairs for the lipid binding GPCRs had the lowest levels of conservation while the biogenic amines had highest level of conservation. More than 17,000 ESTs matching the GPCRs in mouse and human were collected, providing information about their expression patterns.
Discussion

The *Adhesion* family forms several clans

Overall, the structure of *Adhesion* family members is complex. Their long N-termini are coded for by many exons, which make them quite difficult to study. The first *Adhesion* receptor identified, the Epidermal growth factor module-containing receptor 1 (EMR1), was described relatively recently (Baud *et al.*, 1995). Only eight years after the cloning of EMR1, the number of *Adhesion* GPCRs in humans was up to thirty (Fredriksson *et al.*, 2003c). The *Adhesion* family can be divided into several clans according to phylogenetic analysis. Interestingly, even though the phylogenetic analysis is based on 7TM regions alone (Paper I) the clans are in agreement with the functional domains contained in the N-termini. For example the receptors containing epidermal growth factor (EGF) domains form one phylogenetic clan.

Ligand binding of *Adhesion* GPCRs

The EGF-clan is probably the most studied of the *Adhesion* clans considering both ligand binding and tissue expression. Overall, the human and mouse receptors have between two and six EGF domains in their N-termini (Figure 5), most of which contain Ca$^{2+}$ binding sites (McKnight & Gordon, 1998) and have the following consensus sequence: (D/N)X(D/N)(E/Q)X(D/N)X(Y/F), where X are variable (Downing *et al.*, 1996). Ca$^{2+}$ sites within EGF domains are a requirement for stabilisation of protein-protein interactions for some non-GPCR proteins (Selander-Sunnerhagen *et al.*, 1992) and supposedly they play a similar role in EGF containing GPCRs. Thus, by participating in binding to other cell surface molecules, or extracellular matrix components, the EGF-domains could provide cell-adhesion (McKnight & Gordon, 1998). In fact, this theory has been strengthened by results for the CD97 receptor, which splice variants contain between three to five EGF domains. Two of these domains require Ca$^{2+}$ for interdomain stabilisation and together with hydrophobic interactions
Figure 4: A simplified illustration of an Adhesion receptor, cell differentiating receptor 97 (CD97) shown in purple. This particular CD97 splice variant contains three epidermal growth factor domains (EGF) and a GPCR proteolytic (GPS) domain in its N-terminus. The GPS domain is thought to enable proteolytic cleavage, in this case between Leu and Ser (shown in read), of the N-terminus. EGF1 and EGF2 contain cysteines (yellow) that make up disulphide bridges (grey lines). Furthermore they require Ca\(^{2+}\) for interdomain stabilisation and together with hydrophobic interactions they mediate binding to a short consensus repeat (SCR) on decay accelerating factor (DAF/CD55). Interestingly, splice variants of CD97 containing more EGF domains have been shown to bind to a different ligand, chondroitin sulphate glycosaminoglycans. Redrawn from (Kwakkenbos et al., 2004; Lin et al., 2001).
they mediate binding to a short consensus repeat (SCR) on decay acceler-
ating factor (DAF/CD55) (Downing et al., 1996; Hamann et al., 1996; Lin et al., 2001; Stacey et al., 2003). DAF is found for example on surfaces of lymphocytes and erythrocytes. The binding of DAF to CD97 can thus func-
tion to recruit immune cells by cell-adhesion. Similarly Ca\(^{2+}\)-dependent EGF domains are involved in CD97 and EMR2 binding to chondroitin sulphate (CS) glycosaminoglycans. Glycosaminoglycans are present on cell mem-
branes, and in the extracellular matrix, and have been implicated in biologi-
cal processes such as cell-adhesion, proliferation, tissue repair, and immune responses (Stacey et al., 2003).

Figure 4 shows a simplified illustration of CD97, depicting the domains and residues important in binding of the ligand. Interestingly the binding profile of CD97 differs greatly for the two ligands (DAF or CS). The first two EGF domains are involved in DAF binding, whereas it is the fourth EGF domain that is required for the binding of CS (the same EGF domain as for CS binding of EMR2) (Stacey et al., 2003). Previously, three splice variants of CD97 have been reported, which contain three (EGF1,2,5), four (EGF1,2,3,5), and five (EGF1,2,3,4,5) EGF domains (Gray et al., 1996). In our searches for splice variants of the Adhesion family, using EST and mRNA data, we found five CD97 variants, these three previously recognised and additionally two unique variants both containing three EGF domains (EGF1,4,5 and EGF3,4,5) (Paper II). The splice variants are shown in Figure 5. Since EGF1 and 2 mediate DAF binding it is likely that those two splice variants will be unable to bind DAF to the same extent as the EGF1 and EGF2 containing variants. Moreover, two of the previously recognised vari-
ants (EGF1,2,5 and EGF1,2,3,5) lack the fourth EGF domain, which makes them unable to bind CS. Only one of the splice variants (EGF1,2,3,4,5) con-
tains all five EGF-domains enabling binding of both ligands (Kwakkenbos et al., 2005; Stacey et al., 2003). From these findings it seems apparent that alternative splicing can influence how Adhesion receptors interact with other proteins.

Functional domains of the Adhesion family

Apparently, the Ca\(^{2+}\)-dependent EGF domains play an important role in ligand-binding and cell-adhesion but how about the other functional do-
mains? Figure 5 provides an overview of the different functional domains according to RPS-blast searches at NCBI’s conserved domain database. The Cadherin EGF LAG seven-pass G type receptor (CELSR) clan is made up of proteins containing both EGF domains, cadherin repeats (CA), and laminin G domains (LamG). The CA found in GPCRs make up one subgroup of a large CA superfamily, which contain motifs with the following conserved sequences: DRE, DXNDNAPXF, and DXD. In general, the role of CA
Figure 5: An overview of the 33 Adhesion family members. The figure illustrates ESTs and mRNA data. The following functional domains can be seen: GPS (GPCR type 1), PTX (pentraxins), EGF (epidermal growth factor) domains, OLF (olfacto-Immunoglobulin-like) domains, SEA (sea urchin sperm protein), Calx-beta (leucine-rich repeats) and glycosylation sites (consensus sequences: NXT or NXS,
the functional domains according to RPS-blast and splice variants supported by proteolytic) domains, HBD (hormone binding domains), TSP1 (thrombospondin medin) domains, GBL (galactose binding lectin) domains, CA (cadherin repeats), Ig domain), LamG (laminin G) domains, EGF-Lam (laminin type EGF) domains, LRR where X can represent any amino acid).
involves cell-cell interactions. Just as in the case of the EGF-domains, Ca\(^{2+}\) is important for CA adhesive function as they ensure proper folding of the domain and rigidity in the N-termini (Wheelock & Johnson, 2003). Approximately 80 LamG domains have been identified in diverse families of extracellular and transmembrane proteins, where they occur as single modules or pairs (Timpl et al., 2000). These domains are known to mediate cell-adhesion. Thus, possession of these domains undeniably makes the CELSRs candidates for mediating cell-adhesion. We found one splice variant, CELSR3, missing five CA and four EGF domains, which could certainly affect ligand-binding properties of the receptor (Paper II).

The lectomedin receptor (LEC) clan consists of proteins containing hormone binding domains (HBD), olfactory domains (OLF), and galactose binding lectin domains (GBL). The HBD is known to have four conserved cysteines, which probably form disulfide bridges. The domain is found in many hormone-binding receptors, including most of the *Secretin* receptors (Fredriksson et al., 2003a). GBL have been postulated as modulators of cell-cell and cell-extracellular matrix interactions (Rabinovich, 1999). Altogether members of the LEC clan are thought to have a function in synaptic cell-adhesion even though their natural ligands are still unknown, as are their roles in vivo (Sudhof, 2001). Thus, in general it seems that many members of the *Adhesion* family are involved in cell-adhesion even though some of them have domains, of which the functional role is still not known.

The *Adhesion* family GPS domain

Together with the sequence similarity of the TM regions, the GPCR proteolytic (GPS) domain is common to almost all *Adhesion* GPCRs. The GPS domain is found in a single copy adjacent to the TM region and contains four conserved cysteines, one glycine, and two conserved tryptophan residues (Stacey et al., 2002). This domain has proven to be essential in the proteolytic cleavage of N-termini for a number of *Adhesions* such as CD97, ETL, EMR4, EMR2, and LEC1 (Chang et al., 2003; Krasnoperov et al., 1999; Kwakkenbos et al., 2002; Wang & Roehrl, 2002). Basically, a peptide bond within the GPS domain (often between Leu/Ser or Leu/Thr) is cleaved in a still unknown mechanism. After the cleavage, the N-terminus is non-covalently linked to the TM region, making up a two subunit structure (Chang et al., 2003). This proteolytic process has been shown to be essential for surface expression of LEC1. Since heterodimer structure exist for numerous other GPS containing *Adhesions*, it has been suggested that the mechanism applies to all members of the family (Krasnoperov et al., 2002). In our searches for *Adhesion* splice variants, we found two GPS deprived variants, GPR56 and GPR124 (Paper II). Both are still orphans. However, as shown for LEC1, it is possible that the alternative splicing mechanism is used to regulate the surface expression of GPCRs, including GPR56 and
GPR124, and may be such mechanisms could even take part in regulating tissue specific surface expression.

Tissue distribution and function of the Adhesion GPCRs
Since most of the Adhesion members are still orphans, we felt it would be interesting to get an overview of the tissues distribution of the family members. Extensive human and mouse EST and mRNA sequence (hereafter simply referred to as ESTs) searches were carried out at NCBI’s dbEST and UCSC homepage. The results, taken together with the previous laboratory results from the literature, show some interesting and distinct expression patterns for different clans and members.

The EGF-clan
As previously mentioned, the EGF-clan is the probably the best studied group of Adhesion GPCRs. There are several studies that look at the tissue expression pattern of these receptors (see Table 1). Most were carried out on human tissue using Northern blotting (Eichler et al., 1994; Jaspars et al., 2001; Lin et al., 2000; Stacey et al., 2001), in situ hybridisation (Kwakkenbos et al., 2002) or by RT-PCR (Baud et al., 1995; Stacey et al., 2001). Others used rat or mouse tissues and one of the methods mentioned above (Kwakkenbos et al., 2002; Lin et al., 1997; Stacey et al., 2002). Our results showed that 80% (144 of 179) of the mouse EMR1 ESTs are found in leukocytes, and 25% (7 of 28) of the human EMR2 ESTs are also found in leukocytes. The highest number of ESTs were found for CD97, 22% (41 of 183) come from immune-system related tissues, such as bone marrow, leukocytes, spleen, and stem cells (Paper I and II). Only a few ESTs were found for EMR3, and EMR4. However, some of them where also found in immune system related tissues, including spleen, bone marrow, and haematopoietic stem cell (for EMR3), spleen and stem cell (for EMR4). The ETL receptor expression profile was more spread between different tissues according to the EST data.

Taking together our results and previous knowledge from literature, the EGF-clan members are in general clearly expressed by cells of the immune system and smooth muscle cells. Accordingly, CD97 has been shown to have a physiological function within the immune system, inducing localised inflammatory responses (Gray et al., 1996). Furthermore, chondroitin sulphate (CS), ligand to CD97 and EMR2, has been implicated in the pathogenesis of rheumatoid arthritis raising the question of whether CD97 could be involved in the disease.
Table 1: EGF-clan tissue expression according to the available literature.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Tissue expression</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD97</td>
<td>Adrenal gland, bladder, brain, dendritic cells, heart, intestine, kidney, leukocytes (granulocytes, monocytes, T-cells, B-cells), liver, lung, lymph node, pancreas, placenta, prostate, skeletal muscle, skin, spleen, stomach, testis, thymus, tonsils, uterus</td>
<td>(Aust et al., 2002; Eichler et al., 1994; Jaspars et al., 2001; Kwakkenbos et al., 2002; Lin et al., 2000; McKnight &amp; Gordon, 1998)</td>
</tr>
<tr>
<td>EMR1</td>
<td>Haematopoietic cells (monocytes, macrophages, myeloid cells)</td>
<td>(Baud et al., 1995; Lin et al., 1997)</td>
</tr>
<tr>
<td>EMR2</td>
<td>Bone marrow, leukocytes (monocytes, granulocytes), liver, lymph nodes, placenta, skin, spleen, thymus, tonsils</td>
<td>(Kwakkenbos et al., 2002; Lin et al., 2000)</td>
</tr>
<tr>
<td>EMR3</td>
<td>Bone marrow, heart, kidney, leukocytes (neutrophiles, monocytes), liver, lung, lymph node, macrophages, pancreas, placenta, skeletal muscle, spleen, thymus, tonsils</td>
<td>(Stacey et al., 2003)</td>
</tr>
<tr>
<td>EMR4</td>
<td>Kidney, liver, lung, macrophages, spleen, thymus</td>
<td>(Stacey et al., 2002)</td>
</tr>
<tr>
<td>ETL</td>
<td>Brain, heart (cardiomyocytes), liver, lung (bronchiolar smooth muscle cells), muscle (vascular smooth muscle cells), kidney, spleen</td>
<td>(Nechiporuk et al., 2001)</td>
</tr>
</tbody>
</table>

These speculations were supported by localisation of the ligands, DAF and dermatan sulfate (DS), as well as their receptors, CD97 and EMR2, in rheumatoid synovial tissue. Supposedly, CD97 (present on all leukocytes) functions as a primary DS receptor, whereas EMR2 (expressed on macrophages and dendritic cells) may serve as a second DS receptor contributing to the increase of macrophages and dendritic cells seen in rheumatoid synovial tissue (Kop et al., 2005). Considering the expression patterns and preliminary results for the other EGF-members, it seems likely that they play a role in mediating immune responses (except for ETL). EMR3 recognises a ligand on macrophages and neutrophils. It has been suggested that by modulating myeloid-myeloid interactions the receptor can amplify or reduce immune and inflammatory responses (Stacey et al., 2001). Similarly mouse EMR4 is the first receptor known to mediate the cellular interaction between myeloid cells and B-cells, giving it potential for a specific role in recruiting B-cells (Stacey et al., 2002). Quite on the contrary to the general expression of EGF-members, ETL is primarily found in smooth muscle cells. Its expression is
developmentally regulated in rat and human heart suggesting it plays an important role in the growth phase and maturation of cardiac muscle (Nechiporuk et al., 2001). Thus, even thought the EGF-clan members predominantly seem to carry out functions in the immune-system, there is a possibility that their roles are not restricted to only one function. For example, CD97 is also found in smooth muscle as well as various peripheral tissues such as kidney, stomach, and pancreas, giving a possibility for cell-adhesion roles else where than the immune system.

The BAI-clan

Experiments for localising tissue expression of the BAI-clan have also mostly been carried out using Northern blot analysis and human tissue (Shiratsuchi et al., 1998; Shiratsuchi et al., 1997) and the results are listed in Table 2. Our EST findings are in accordance with these results, since 35-70% of the human and mouse BAI1 ESTs, 57-65% of human and mouse BAI ESTs, and 39-53% human and BAI3 ESTs were found in brain. Non-GPCR genes containing thrombospondin type 1 repeats (TSP, see Figure 5), such as found in the BAI receptors, are thought to inhibit angiogenesis (Bornstein, 1992). Further support for BAI1 as a candidate for tumour suppression has been gathered since BAI1 transcripts are shown induced by p53, the most frequently altered gene in human glioblastoma and expression of BAI1 proved absent or significantly reduced in seven out of nine glioblastoma cell lines tested (Koh et al., 2001; Nishimori et al., 1997). BAI3 has also been found absent or significantly reduced in five of nine glioblastoma cell lines examined. However, despite that and the similar tissue specificity among the three BAI receptors, only BAI1 is transcriptionally regulated by p53. This suggest that the BAI receptors share similar roles in inhibiting angiogenesis in glioblastomas, even though BAI1 seems transcriptionally regulated in a different manner to the other two (Shiratsuchi et al., 1997). Although all three members of the BAI-clan are found in many various subregions of the brain, they are also found in tissues of the periphery. In fact, the BAI2 transcripts expressed in heart and skeletal muscle were longer than that found in brain (Shiratsuchi et al., 1997). Thus, it is quite possible that these longer transcripts are splice variants, and it cannot be excluded that different splice variants are expressed in different tissues where they may regulate angiogenesis.
Table 2: Tissue expression of the BAI-clan according to the available literature.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Tissue expression</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAI1</td>
<td>Brain (amygdala, caudate nucleus, cerebral cortex, corpus callosum, frontal lobe, hippocampus, medulla, occipital pole, putamen, substantia nigra, subthalamic nucleus, temporal lobe, thalamus), colon, heart, kidney, leukocyte, liver, lung, ovary, pancreas, placenta, prostate, skeletal muscle, small intestine, spinal cord, spleen, testis, thymus</td>
<td>(Koh et al., 2001; Nishimori et al., 1997; Shiratsuchi et al., 1998; Shiratsuchi et al., 1997)</td>
</tr>
<tr>
<td>BAI2</td>
<td>Brain (amygdala, caudate nucleus, cerebral cortex, corpus callosum, frontal lobe, hippocampus, medulla, occipital pole, putamen, substantia nigra, subthalamic nucleus, temporal lobe, thalamus), heart, skeletal muscle, spinal cord, thymus</td>
<td>(Kee et al., 2002; Shiratsuchi et al., 1997)</td>
</tr>
<tr>
<td>BAI3</td>
<td>Brain (amygdala, caudate nucleus, cerebral cortex, corpus callosum, frontal lobe, hippocampus, medulla, occipital pole, putamen, substantia nigra, subthalamic nucleus, temporal lobe, thalamus), heart, spinal chord</td>
<td>(Kee et al., 2004; Shiratsuchi et al., 1997)</td>
</tr>
</tbody>
</table>

The CELSR-clan

The CELSR-clan tissue expression has mainly been studied in mouse and rat, using Northern blots, in situ or RT-PCR (Formstone et al., 2000; Formstone & Little, 2001; Hadjantonakis et al., 1997; Nakayama et al., 1998). Comparison of our EST results with literature findings (shown in Table 3) strengthens the conclusion that these receptors have important expression in the brain. A total of 57% (25 of 44) of CELSR1, 43% (42 of 97) CELSR2, and 58% (18 of 31) CELSR3 mouse ESTs were found in brain. During mouse embryonic development, the CELSR family exhibits restricted but distinct patterns of expression suggestive of functional roles in a number of developmental processes such as the early development of the hindbrain as well as the early peripheral nervous system (Formstone & Little, 2001). In adults, CELSR1 could possibly be involved in cell-cell signalling in brain (Hadjantonakis et al., 1997).
Table 3: Tissue expression of the CELSR-clan according to the available literature.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Tissue expression</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELSR1</td>
<td>Brain (<em>area postrema</em>, <em>choroids plexus</em>, lateral, third, and fourth ventricle), eye, kidney, lung, spinal cord, spleen</td>
<td>(Formstone <em>et al.</em>, 2000; Hadjantonakis <em>et al.</em>, 1997)</td>
</tr>
<tr>
<td>CELSR2</td>
<td>Brain (cerebral cortex, cerebellum, hippocampus, brain stem, olfactory bulb), eye, kidney, lung, spinal cord, spleen</td>
<td>(Formstone <em>et al.</em>, 2000; Nakayama <em>et al.</em>, 1998)</td>
</tr>
<tr>
<td>CELSR3</td>
<td>Brain (cerebellum, olfactory bulb), eye, spinal cord</td>
<td>(Formstone <em>et al.</em>, 2000; Formstone &amp; Little, 2001; Nakayama <em>et al.</em>, 1998)</td>
</tr>
</tbody>
</table>

The LEC-clan

Results from the literature regarding tissue expression of the LEC-clan can be seen in Table 4. According to these, LEC1 is primarily expressed in brain but is also present in peripheral tissues, LEC3 is also found highly enriched in brain whereas LEC2 shows no preferential expression in brain but is instead uniformly present in peripheral tissues. LEC2 expression was found in liver, placenta, heart, lung, kidney, pancreas, spleen, and ovary (Matsushita *et al.*, 1999; Sugita *et al.*, 1998). Looking at our EST data, 14-28% of the human and mouse EST for LEC1 are found in brain, as are 25-45% of human and mouse LEC2 ESTs, and 8-22% of human and mouse LEC3 ESTs. In addition, peripheral expression of LEC2 was supported by nine human and one mouse EST found in lung, one from each species in kidney, and a single human EST in placenta. However, no LEC2 ESTs were found in heart, ovary, or spleen. The discrepancy between findings can be explained by the fact that even thought EST data are accumulating rapidly, the EST databases still contain large variability in sequence quality as well as variations in the representation of sequences, tissues and cell states (Murray *et al.*, 2005) limiting the accuracy of expression patterns.

Thus, even though we did not find any ESTs for LEC2 in heart, ovary, or spleen, this does not exclude its expression in these regions. LEC1 is at least 50-fold more abundant in brain than in any other tissue (Matsushita *et al.*, 1999). Additionally, LEC1 can bind $\alpha$-Latrotoxin, which serves a purpose in recruiting the toxin to synapses and stimulating exocytosis. Thus the restricted localisation of LEC1 in nerve terminals suggests it carries out highly specialised functions in synaptic cell-adhesion in the brain (Sudhof, 2001), which could couple through a G-protein to exocytosis of synaptic vesicles.
Table 4: The LEC-clan tissue expression according to the available literature.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Tissue expression</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEC1</td>
<td>Brain (cortex, hippocampus, striatum), fibroblasts, heart, kidney, liver, lung, pancreas, placenta, skeletal muscle, spleen</td>
<td>(Krasnoperov et al., 1997; Lelianova et al., 1997; Matsushita et al., 1999; Sugita et al., 1998; White et al., 1998)</td>
</tr>
<tr>
<td>LEC2</td>
<td>Brain, colon, heart, kidney, liver, lung, ovary, pancreas, placenta, prostate, skeletal muscle, spleen, testis</td>
<td>(Ichtchenko et al., 1999; Matsushita et al., 1999; Sugita et al., 1998)</td>
</tr>
<tr>
<td>LEC3</td>
<td>Brain, heart, kidney, lung, pancreas, placenta, spleen, testis</td>
<td>(Ichtchenko et al., 1999; Matsushita et al., 1999; Sugita et al., 1998)</td>
</tr>
</tbody>
</table>

LEC2 on the other hand, has been shown to be unable to function as a α-Latrotoxin receptor when expressed in liver membranes and similar effects have been described for LEC3 suggesting that the two in deed have distinct functions from those of LEC1 (Matsushita et al., 1999). With its widespread expression, LEC2 could be a part of the peripheral nervous system, whereas LEC3, found enriched in brain, seems involved in effects other than exocytosis.

The remaining Adhesion GPCRs

Much work needs to be done on the remaining 18 receptors of the Adhesion family. So far, expression studies have been described for GPR56 (also known as TM7XN1), GPR113, GPR116 (also known as Ig-hepta), GPR126, and HE6. GPR56, GPR126, and HE6 group together according to phylogeny (Paper I, see Figure 4) as do GPR113 and GPR116. GPR56 expression has been studied in human and rat using Northern blotting, RT-PCR and in situ approaches (Liu et al., 1999; Zendman et al., 1999). The highest level of mRNA was detected in thyroid gland but it was also present in specific brain areas and a wide variety of peripheral tissues (Liu et al., 1999). Neither our human nor mouse searches yielded any thyroid gland ESTs. However, 15-17% of the mouse and human ESTs for GPR56 were found in brain, and the rest widespread in the periphery. HE6 expression has been studied by Northern blotting in human and rat tissue. The receptor was found expressed in epididymis tissue (Osterhoff et al., 1997).
Table 5: The tissue expression of GPR113, GPR126, and HE6 according to the available literature.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Tissue expression</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPR56 (TM7XN1)</td>
<td>Bladder, brain (amygdala, hippocampus, hypothalamic nucleus, caudate nucleus, corpus callosum, cortex, paraventricular nucleus, substantia nigra, subthalamic nucleus, thalamus), heart, kidney, lung, pancreas, prostate, skeletal muscle, testis, thyroid gland, uterus</td>
<td>(Liu et al., 1999; Zendman et al., 1999)</td>
</tr>
<tr>
<td>GPR113</td>
<td>Circumvallate papillae, testis</td>
<td>(LopezJimenez et al., 2005)</td>
</tr>
<tr>
<td>GPR116 (Ig-hepta)</td>
<td>Lung (alveolar walls), heart, kidney (intercalated cells in the collection duct of kidney)</td>
<td>(Abe et al., 1999)</td>
</tr>
<tr>
<td>GPR126</td>
<td>Endothelial cells, heart, lung</td>
<td>(Moriguchi et al., 2004; Stehlik et al., 2004)</td>
</tr>
<tr>
<td>HE6</td>
<td>Epididymis tissue</td>
<td>(Osterhoff et al., 1997)</td>
</tr>
</tbody>
</table>

Most of our mouse EST hits come from breast 35% (8 out of 23) and our human ESTs from placenta 35% (6 of 17). Also 29% (5 of 17) of human ESTs come from the male reproductive organs, which further support the expression of HE6 in the reproductive organs. GPR126 (also known as VIGR or DREG) expression, studied by Northern blots in human tissue and RT-PCR in mouse, was found selectively expressed on endothelial cells derived from larger vessels, but not from micro-vessels (Stehlik et al., 2004), and at high levels in the heart during embryogenesis and in the adult lung (Moriguchi et al., 2004). According to our EST data, only two mouse GPR126 ESTs are found in lung and none is detected in heart. However, 52% (15 of 29) of the human ESTs came from placenta. More research will have to be done to determine the expression pattern of GPR126. RT-PCR and in situ experiments showed that GPR113 was found both in testis and taste cells (circumvallate papillae) of mouse (LopezJimenez et al., 2005). All four human ESTs found in our studies came from male reproductive system, which is consistent with the finding in testis. High expression levels of GPR116 (Ig-hepta) were found in rat lung using Northern blot (Abe et al., 1999). Lower expression levels were detected in kidney and heart. ESTs for all three tissues can be found in our results even though only 12% (6 of 49) of the total human ESTs came from lung, 8% from kidney, and 4% from heart. Other ESTs come from various peripheral and CNS tissues, suggesting a very widespread distribution of GPR116. Unfortunately, our searches for
other members of the two groups discussed above resulted in relatively few EST hits (1-6) making the results difficult to interpret. However, an exception is the very long GPCR1 (VLGR1) with 166 human ESTs and 51 mouse ESTs. A majority (27%) of the mouse ESTs came from brain, whereas the human ESTs seem widespread in tissues such as adrenal gland, brain, breast, eye, gastrointestinal, kidney, lung, and pancreas. Even though our searches for alternative splice variants did not yield in any VLGR1 splice variant, it cannot be excluded that they exists, especially since it is the longest of the Adhesions and is coded for by over 80 exons. It would thus be possible that there existed tissue-specific splice variants, in various tissues.

The only clan of receptors where laboratory data could not be found for any of the members contains GPR123-125 (see Figure 4). Between 7 and 238 ESTs were found for these receptors during our searches. Fewest ESTs were found for GPR123. Out of the human ESTs, 57% (3 of 7) were found in brain, as were 100% (11) of the mouse EST. For GPR124, 32% (7 of 22) of human ESTs and 24% (4 of 17) of the mouse ESTs came from breast. It is tempting to suggest that those two have distinct expression patterns, GPR123 in the brain and GPR124 in breast. GPR125 on the other hand was found expressed in a wide variety of tissues. Out of the 238 human ESTs, 13% are found in brain, 11% in female reproductive system, 10% in gastrointestinal tissues and 7% in the eye. Other tissues include: liver, placenta, male reproductive system, lung, heart, kidney, breast, pancreas and skin. The mouse EST data for GPR125 also shows many (16%, 25 of 177) ESTs, supporting expression in brain as well as in the eye (12%). Thus GPR125 seems expressed both centrally and in the periphery.

All the EST results presented here are interesting leads for expression patterns of Adhesion receptors, especially for the many receptors where tissue expression has not been studied specifically. A vast amount of ESTs could be found for some of these receptors but, unfortunately, searches for others only yielded few EST hits. Even so, each EST hit alone should be considered important. When the Rhodopsin GPR119 receptor was first discovered (Fredriksson et al., 2003b), EST searches resulted in only one single EST from pancreas. Today, there are evidence that GPR119 is predominantly expressed in pancreatic β-cells and plays a pivotal role in inducing insulin secretion from pancreatic β-cells as well as the intact pancreas (Soga et al., 2005). Thus the relevance of few EST hits should not be underestimated.
Conserved motifs of the *Glutamate* family

Just like the *Adhesion* family, most *Glutamate* GPCRs constitute long N-termini (around 600 amino acids). The *Glutamate* N-termini are divided into a cysteine rich domain (CRD) and a ligand binding domain (Hermans & Challiss, 2001; Pin et al., 2004). The sequence similarity between the ligand binding domain and bacterial periplasmic amino acid-binding proteins supported the existence of a Venus flytrap (VF) module among the *Glutamate* GPCRs (O'Hara et al., 1993).

Two thirds of the extracellular domain participates in the two lobe VF mechanism, which has proven critical for recognising and binding ligands for the metabotropic glutamate receptors (GRM), γ-aminobutyric acid (GABA), and Ca\(^{2+}\) sensing (CASR) receptor clans. The structure of the mouse GRM1 VF has been solved by X-ray crystallography in the absence and presence of both agonist and antagonist (Kunishima et al., 2000; Tsuchiya et al., 2002). Figure 6 shows a simplified illustration of mouse GRM1, highlighting amino acids, which have been found conserved within the extracellular domain of the GRM subgroup (Hermans & Challiss, 2001; Jingami et al., 2003; Pin & Duvoisin, 1995).

In order to search for specific conserved elements in the N-termini of all the *Glutamate* GPCRs, we created consensus sequences of all the different phylogenetic clans (CASR, GRM, taste receptors type 1 (TAS1R), GPRC6A (orphan GPCR), and GABA) and aligned them (Paper III). Alignment comparison shows that Ser165 in LB1 is conserved even in the subgroups of CASR, and TAS1R, GPRC6A, and GABA receptors. Moreover, Asp208 and Tyr236 in LB2 are extremely well conserved in CASR, V2Rs, TAS1R, and GPRC6A receptors. It is also noteworthy that Thr188 in LB1 has been replaced with a Ser188 in all subgroups, except for the GRM and GPRC6A receptors. Thus, residue conservation supports a common mechanism of ligand binding that is general conserved among the *Glutamate* family members, excluding the orphan GPRC5(A-D), which do not contain long N-termini.

Since GPRC5s (A-D) lack a VF module, their ligand binding domain, giving they have one, has to be found within the TM-regions, as is the case for most *Rhodopsin* GPCRs. Since the ligand activated VF domain of *Glutamate* GPCRs has been suggested to interact with the 7TM to activate the receptor (Parnot & Kobilka, 2004) it seems possible that the ligand for the GPRC5 receptors could be structurally similar to the N-terminus of the other *Glutamate* GPCRs. It is hence possible that a soluble ligand similar to a ligand activated VF domain can bind and activate *Glutamate* GPCRs of the GPCR5 clan. However, the ligand binding and function of GPCR has yet to be determined and they could very well lack ligand binding capabilities of their own and function only as partners in heterodimers, paring with other *Glutamate* receptors that are able to bind ligands (Kunishima et al., 2000).
Figure 6: a) A schematic figure of the mouse metabotropic glutamate receptor 1 (GRM1) (purple) binding to its agonist (glutamate, shown in purple) via the Venus flytrap (VF) module made up of ligand binding domain 1 (LB1) and ligand binding domain 2 (LB2). Residues, which contribute to the binding of glutamate, are shown in red. Connection of the VF to the 7TM regions is established through a cysteine rich domain (CRD) predicted to possess four $\beta$-strands and three disulphide bridges (grey lines) between cysteines (yellow) (Liu et al., 2004). Residues considered important for binding of the G-protein at the intracellular site (IS) are: P698, C694, T695, K697, S702, R775, and F781 (coloured in orange) (Hermans & Challiss, 2001). Redrawn from (Hermans & Challiss, 2001; Kunishima et al., 2000; Liu et al., 2004).
A cysteine rich domain (CRD) containing nine conserved cysteine residues connects the VF to the 7TM region (except in the GABA receptors, which lack the CRD) (Kaupmann et al., 1997). We checked for conservation of the CRD for the different Glutamate clans by aligning consensus sequences for the different clans (Paper III). The CRD is clearly seen since nine cysteine residues are found within the consensus sequences of the CASR, V2R, TAS1R, GRM, and even the GPRC6A receptors. Four of the cysteines align perfectly indicating that they are very well conserved and probably serve an important function in receptor mechanism. The domain appears necessary for the activation of GRM and CASR receptor (Pin et al., 2005). However, its exact role is not known. It may simply function as a rigid spacer between the 7TM region and the ligand-binding domain. The GPR158 receptor clan, which to our knowledge had not previously been described as a Glutamate GPCR, also has long N-termini. However, they are highly divergent from the N-termini of other Glutamate GPCRs, and in fact the only common feature seems to be the CRD, although it appears to be somewhat shorter in GPR158.

The GPCR superfamily in human and mouse

According to our results the repertoire of the human and mouse superfamily of GPCR contains over 800 non-olfactory GPCRs (Paper V), which could be classified into the five GRAFS families (Fredriksson et al., 2003c). Overall, the majority of the receptors within the different families form well conserved orthologous pairs, and a clear majority of these pair with highest possible bootstrap values. All the human receptors in the families of Frizzled and Secretin possess an orthologue in mouse, while the other families show more variation.

There are also some important differences in the level of residue conservation between the different families. The family that has the highest conservation in the TM regions between orthologues is the Frizzled GPCRs. On average, the Frizzled human-mouse pairs has 94% identical amino acids (See Figure 7). This can be contrasted with the Taste2 receptors, which have the lowest conservation between the orthologues pairs of 59%. This is interesting, considering the fact that they group together in phylogeny and this reflects the common evolutionary history of these groups. The Frizzled receptors are highly conserved in vertebrates, and even in invertebrates, while the Taste2 seem to be rapidly evolving and expanding in numbers (Fredriksson & Schioth, 2005).
Figure 7: Overview of orthologue relationships of the GPCR human and mouse superfamily. All the human receptors in the families of Frizzled and Secretin possess an orthologue in mouse while the other families show more variation. Orthologue conservation is listed in boxes. On average, the Frizzled human-mouse pairs has 94% identical amino acids. On the contrary the Taste2 receptors have the lowest conservation, 59%, between the orthologous pairs.
The Taste2 family appears to have undergone a complex combination of local duplications, inter-chromosomal duplications, and deletions (Conte et al., 2003).

The pattern for the Taste2 family is similar to those seen for distinct clusters of the Rhodopsin family: olfactory, trace amine, and chemosensory receptors, as well as pheromone receptors (Gloriam et al., 2005; Robertson, 1998; Rodriguez et al., 2002; Young et al., 2002). These rapid species-specific expansions coincide with the fact that they all bind relatively small molecular ligands. It can be speculated that there are only relatively few receptor residues that are involved in binding which could contribute to the three-dimensional structure not being as fixed as in receptors binding larger complex molecules, such as the case of Wnt protein bound by the Frizzled receptors. It is thus possible that this type of ligand binding mode contributes to the “dynamic” gene repertoire.

The Rhodopsin family holds many interesting differences between the human and mouse repertoire. Single receptor gene deletions, or duplications, are not uncommon in the Rhodopsin family and it also contains clusters of receptors originating from large expansions. Each of the Rhodopsin α-, β-, γ- and δ-subgroups have receptors that bind different kinds of ligands while all of these four groups contain receptors that bind peptide ligands. The receptors that bind lipids are the only group that has the exact equal number of receptors in each subfamily in human and mouse. This might suggest that the lipid binding GPCRs are more highly conserved between the two species than the receptors that bind other ligands. However, calculation of the percentage of identical and positively matching amino acids, in the TM regions shows that this is not the case. The receptor pairs that bind lipids have averages of 86% identical amino acids, and 91% positive matching amino acids, which are lower than for each of the other ligand groups (Paper V). It seems thus that the presence of lipid binding GPCRs is well conserved between the mouse and human genomes even though their amino acid identities is somewhat lower than for the other Rhodopsins. This may reflect that the lipid ligands in general do not provide the same evolutionary constraints on the primary structure of their receptors, while this does not suggest less important functional role for these receptors, which could be implicated for the receptors that are missing orthologues in one of the species.
Conclusions

We thoroughly searched the human and mouse genomes for new members belonging to the Adhesion family of GPCRs (Paper I). Our searches resulted in two novel human GPCRs and 17 mouse GPCRs. Thus the Adhesion family comprises 33 human and 31 mouse members. Each of the human receptors group together with a mouse orthologue, except for EMR2 and EMR3, which do not seem to exist in mouse. A phylogenetic analysis of the TM regions of the entire set of Adhesion GPCRs shows that there exist eight clans of receptors, whereof the novel GPR133 and GPR144 together with their orthologues, make up one. These clans, based on the TM regions, are in good accordance with the functional domains comprised in the N-termini. The expression patterns, charted from EST data, are highly variable between the different receptors suggesting they participate in a number of physiological processes.

We continued to study the complex family of Adhesion GPCRs, making use of extensive mRNA and ESTs search results, to identify human splice variants among the family members (Paper II). These sequences gave support for 53 unique splice variants, which were categorised into functional (containing the whole 7TM region) and non-functional variants (missing parts of the 7TM region). According to our criteria, 29 out of the 53 were considered functional variants. The majority of the splice variants showed variations in the N-termini, intracellular loops or extracellular loops (extended or truncated loops). The alterations do not affect the number of conserved domains in about half of the functional variants. However, in ten cases alternative splicing results in the variant being deprived of one or more functional domains. This is quite intriguing considering that the functional domains of this family are involved in processes including ligand binding and cell-surface expression of the Adhesion GPCRs. Thus it seems possible to control factors such as these by alternative splicing excluding the relevant functional domain.

Thorough searches in various databases were carried out in order to identify new genes belonging to Glutamate GPCR family of human, mouse, Fugu, and zebrafish (Paper III). The searches resulted in 73 putative mouse, Fugu, and zebrafish receptors belonging to the pheromone receptor clan (V2Rs). We verified the intron-exon boundaries of these, using genome alignments and related sequences. Phylogenetic trees for Glutamate GPCRs dataset were created and confirmed that the sequences cluster in five major
groups (the large V2Rs clan shown in a separate tree). All the human proteins have a mouse orthologue, with the exception of V2Rs (which do not seem to be functional in human). Even though the phylogenetic relationships of the four species is complex with evidence of expansions/deletions within the genomes of the different organisms we found elements conserved in the N-termini for the majority of Glutamate receptors clans supporting a common mechanism of ligand binding among the Glutamate family members, excluding the orphan GPRC5(A-D), which do not contain long N-termini.

Our searches for trace amine (TA) GPCR receptors using various databases resulted in a dataset of 37 mammalian receptors whereof four rat and 14 mouse sequences represented new receptors (Paper IV). Intron-exon boundaries of these were manually verified. Additionally 57 previously unknown zebrafish genes coding for TA receptors and eight Fugu genes were found. Considering these numbers the repertoires clearly differ remarkably between the species. The evolutionary relationship of the family was investigated using phylogenetic analysis and it further supported species-specific expansions or deletions among the TA clan such as seen with other GPCRs clans including the olfactory or chemoreceptors. These clans all bind small ligands with a number of similar ligands within the system. It is possible that the rigidity of the structure of these clans is lower compared with receptors, which bind larger ligands. Less rigidity could result in a more dynamic repertoire a possible explanation to the expansions or deletions.

We continued searching for GPCRs in the human and mouse genomes now on even larger scale as we went on to collect a comprehensive dataset for the whole GRAFS superfamily of GPCRs relying on extensive database and literature searches (Paper V). The searches resulted in 495 mouse and 400 human full-length functional non-olfactory GPCRs. The dataset was classified into families according to the GRAFS classification system and phylogenetic trees were calculated for each family. The clear majority of GPCRs are found in one to one orthologous pairs between human and mouse. However, apparently some distinct clans have undergone interesting species-specific expansions or deletions. The average percentage similarity of the orthologue pairs varies considerably between the different families. Interestingly, the human-mouse pairs of the FZD clan bear 94% amino acid identity whereas the most closely related clan Taste2, according to phylogeny, showed the lowest identity with an average of 59%. These rapid species-specific expansions coincide with the fact that they all bind relatively small molecular ligands. It can be speculated that there are only relatively few receptor residues that are involved in binding which could contribute to the structure not being as fixed as in receptors binding larger complex molecules, such as in the case of the FZD receptors. It is thus possible that this type of ligand binding mode contributes to the “dynamic” gene repertoire.
Future perspectives

- Nowadays genomic sequences are being created at an exponential rate resulting in the completion of genome-sequencing projects for increasing number of organisms. We have already mined the mouse and human genomes for an up-to-date dataset of the whole GRAFS family of GPCRs. Recent efforts have also resulted in GPCR repertoires of *Fugu* (*T. nigroviridis*) (Metpally & Sowdhamini, 2005), rat (Gloriam, submitted), and chicken (Lagerström, submitted). It would be interesting to take it a step further and mine genomes from additional sequencing projects that could give more information on the evolution of GPCRs.

  GPCRs can be found in almost any eukaryotic organism including insects and plants indicating that they are of ancient origin. Moreover, members of each individual GRAFS family can already be found in *C. elegans* showing that the families arose before the split of nematodes from the chordate lineage (Fredriksson & Schioth, 2005). Thus it would be interesting to mine the genome of an even earlier animal than *C. elegans*, for example a marine hydrozoan. The sequencing project of the marine hydrozoan, *H. symbiolongicarpus*, is ongoing (http://www.genomesonline.org/) and analysis of the GPCR repertoire of the genome could help get a better idea on when the different GRAFS families originated.

  Furthermore, there are few families that are lineage specific in bilateral and thus do not belong to the GRAFS families. Examples include the chemosensory receptors only found in nematodes, and the gustatory receptors only found in fruit fly and mosquito (Fredriksson & Schioth, 2005). Mining of the genomes of species such as amphioxus, hagfish or lamprey could reveal if the lineage-specific subgroups exist in any of these chordata species.
- The mechanism of alternative splicing seems to be important in determining the amount of functional domains situated in the N-termini of *Adhesion* family members. This is interesting considering that functional domains can have important roles in ligand binding (EGF domains) as well as proteolytic cleavage (GPS domain), which can affect trafficking to the cell membrane. We have already collected 29 functional *Adhesion* splice variants supported by expressed sequence tags (EST) and mRNA data. Now we are studying the tissue expression of these variants. Tissue specific expression of splice variants, especially those that differ in functional domains, could suggest that the alternative splicing plays a role as a regulator in tissue specific ligand binding.

- ESTs have proven interesting leads for discovering expression patterns of GPCRs, as discussed in this thesis. After collecting ESTs for the whole GRAFS superfamily we are eager to use wet laboratory experiments to see how well they correspond to our EST results. Creation of expression charts would be valuable, especially for the orphan GPCR that are found relatively recently and little is known about. Moreover, knockout models in mouse are likely to prove important to uncover the functional roles of many orphans. Once we begin to understand where and how a gene is expressed under normal circumstances, we can study what happens in an altered state, such as in disease and perhaps reveal potential drug targets.
Acknowledgements

I would like to thank all my co-workers and students whom I have collaborated with the last three years at the Department of Neuroscience, Uppsala University.

My sincere gratitude to Helgi B. Schiöth, my supervisor for giving me the opportunity to be a part of his group, for being so enthusiastic, an endless resource of ideas (it is truly inspiring), and for always being ready to lend a helping hand in all situations whether involving work or not.

I would like to thank my second supervisor Robert Fredriksson for all guidance and good advise. I really admire your broad knowledge on, what seems to me, everything science related.

Extra special thanks to: Malin for being my roommate and a very good friend. Thanks for all laughs and lively discussions regarding everything within this universe! Tatjana for being a very good friend and for all shared moments whether in swedish class, the gym, or just hanging out.

Thanks to other fellow Ph.D. students: David for teaching and getting me started with bioinformatics, Tobias for patiently answering questions concerning everything computer related, and Pär for his help along the way. Thanks to past and present members of Helgi group (in particular Tommy and Andreas, for teaching me Swedish), my ex-job students, and all co-others (especially Malena and Kristín).

Thanks to everyone in Lars Orelands group, Dan Larhammars group (especially Tomas, Paula, Torun, Helena, and Görel), and Klas Kulanders group (especially Nadine).

Of course none of this would ever have been possible without my family! Þúsund þakkir Pabbi minn fyrir að vera að vera mér góð fyrirmynd, alltaf til taks, reiðubúinn að hjálp og veita ráð ☺. Þúsund þakkir Mamma minn fyrir að vera svo hvetjandi og upplifandi. Það hefur ekki liðið sá dagur að þú hafir ekki glatt mig með tölvupóstí, símtali eða heimsóknun ☺. Þúsund þakkir Brynja og Andri fyrir samveruna í Svíþjóð. Ótrúlega notalegt að hafa systu svona innan handar ☺. Þúsund þakkir Hildur fyrir allar heimsóknirnar og góðar stundir heima.

Þúsund þakkir elsku besti Hákon minn fyrir að fylgja mér lönd og strönd og styðja mig svo ég gæti látið draum minn rætast ☺.
References


Felsenstein J. (2005). PHYLIP (Phylogeny Inference Package), Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.


http://www.genomesonline.org/.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 121

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)