G Protein-Coupled Receptors; Discovery of New Human Members and Analyses of the Entire Repertoires in Human, Mouse and Rat

DAVID E. GLORIAM
Dissertation presented at Uppsala University to be publicly examined in Room B21, BMC, Uppsala, Friday, April 28, 2006 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

G protein-coupled receptors (GPCRs) are signal mediators that have a prominent role in the regulation of physiological processes and they make up the targets for 30-45% of all drugs.

Papers I and II describe the discovery of new human GPCRs belonging to the Rhodopsin family, a family which contains many common drug targets. The new receptors have only weak relationships to previously known GPCRs. However, they have been evolutionary conserved in several species and most of them display distinct expression patterns.

In paper III we identified new human GPCRs belonging to the Adhesion family, which is characterised by very long N-termini containing conserved domains. The different compositions of conserved domains as well as the expression patterns suggest that the Adhesions can have several different functions.

In paper IV we revealed remarkable species variations in the repertoires of Trace Amine-Associated Receptors (TAARs), which are relatives of the biogenic amine receptors. The human, mouse and rat TAAR genes are located in only one locus and are therefore most likely the result of gene tandem duplications. 47 of the 57 zebrafish TAARs were mapped to nine different loci on six chromosomes containing from 1 to 27 genes each. This study suggests that the TAARs arose through several different mechanisms involving tetraploidisation, block duplications, and local duplication events.

Papers V and VI are overall analyses of the repertoires of GPCRs in humans, mice and rats; which contain approximately 800, 1800 and 1900 members, respectively. The repertoires were compared to distinguish between species-specific and common (orthologous) members, something which is important for example when predicting drug effects from experiments in rodents. The Glutamate, Adhesion, Frizzled and Secretin families show no or very little variation between human and rodents, whereas the repertoires of olfactory, vomeronasal and Taste2 receptors display large differences between all three species.

Keywords: G protein-coupled receptor, GPCR, Bioinformatics, Evolution, Orphan, Rhodopsin, Adhesion, Phylogeny

David E. Gloriam, Department of Neuroscience, Pharmacology, Box 593, Uppsala University, SE-75124 Uppsala

© David E. Gloriam 2006

ISSN 1651-6206
ISBN 91-554-6522-6
urn:nbn:se:uu:diva-6745 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-6745)
To my beloved wife, Anna
LIST OF PUBLICATIONS

I. Gloriam DE, Schiöth HB, Fredriksson R.
Nine new human Rhodopsin family G-protein coupled receptors: identification, sequence characterisation and evolutionary relationship.

II. Fredriksson R, Höglund PJ, Gloriam DE, Lagerström MC, Schiöth HB.
Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives.

III. Fredriksson R, Gloriam DE, Höglund PJ, Lagerström MC, Schiöth HB.
There exist at least 30 human G protein-coupled receptors with long Ser/Thr-rich N-termini.
Biochem Biophys Res Commun. 2003 Feb 14; 301(3):725-34.

IV. Gloriam DE, Bjarnadottir TK, Yan YL, Postlethwait JH, Schiöth HB, Fredriksson R.
The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish.

V. Bjarnadóttir TK, Gloriam DE, Hellstrand SH, Kristianssen H, Fredriksson R, Schiöth HB.
Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse
Submitted

VI. Gloriam DE, Fredriksson R, Schiöth HB.
The overall repertoire of G protein-coupled receptors in rats
Manuscript
INTRODUCTION ........................................................................................11
The physiologic and therapeutic significance of GPCRs .................11
Molecular structure, activation and signalling of GPCRs ...............11
The GRAFS nomenclature for the human GPCRs..........................12
Evolution of the GPCRs ..................................................................13
The GPCRs superfamily; dynamic and expanding .......................13
High species variation within the repertoires of Trace Amine
Associated Receptors (TAARs).....................................................14
The Rhodopsin family of GPCRs ...................................................15
Receptor ligands and their binding ..............................................15
Rhodopsin family subgroups ......................................................15
Orphan Rhodopsins ..................................................................16
The Adhesion family of GPCRs ...................................................17

AIMS OF STUDIES .............................................................................22

METHODS ..........................................................................................23
Searching for GPCRs ......................................................................23
Sources of sequence databases ..................................................23
Sequence similarity based search strategies .............................23
Filtering the large search results ..............................................24
Removing GPCRs from other families/groups and non-GPCRs ......24
Removing duplicate hits ............................................................25
Manual curation of predicted coding sequences ......................25
Phylogeny - Determining sequence relationships ....................26
Expression profiles from EST data ..........................................27
Submitting new sequences to a public database .......................28

RESULTS AND DISCUSSION ...........................................................29
Paper I - Nine new human Rhodopsins ....................................29
Paper II - Seven new human Rhodopsins ..................................31
Paper III - Six new human Adhesions .......................................32
Paper IV - The evolution of the Trace Amine-Associated Receptors ..34
Papers V-VI – The GPCR repertoires in human, mice and rats ......36

MAIN CONCLUSIONS .....................................................................38
### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7TM</td>
<td>Seven Trans Membrane</td>
</tr>
<tr>
<td>Adhesion(s)</td>
<td>GPCR(s) belonging to the Adhesion family</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>BLAT</td>
<td>BLAST-Like Alignment Tool</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding Domain Sequence</td>
</tr>
<tr>
<td>CNS</td>
<td>Central Nervous System</td>
</tr>
<tr>
<td>EGF</td>
<td>Epidermal Growth Factor</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag</td>
</tr>
<tr>
<td>E-value</td>
<td>Expectation value</td>
</tr>
<tr>
<td>G protein</td>
<td>Guanine nucleotide binding protein</td>
</tr>
<tr>
<td>GPCR</td>
<td>G Protein-Coupled Receptor</td>
</tr>
<tr>
<td>GPS</td>
<td>GPCR-Proteolytic Site</td>
</tr>
<tr>
<td>GRAFS</td>
<td>The human GPCR families; Glutamate, Rhodopsin, Adhesion, Frizzled/ Taste2 and Secretin</td>
</tr>
<tr>
<td>HBD</td>
<td>Hormone-Binding Domain</td>
</tr>
<tr>
<td>HGNC</td>
<td>HUGO Gene Nomenclature Committee</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HUGO</td>
<td>HUman Genome Organisation</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MP</td>
<td>Maximum Parsimony</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>NCBI (US)</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbor Joining</td>
</tr>
<tr>
<td>nr</td>
<td>NBCIs non-redundant database</td>
</tr>
<tr>
<td>OR</td>
<td>Olfactory Receptor</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>Rhodopsin(s)</td>
<td>GPCR(s) belonging to the Rhodopsin family</td>
</tr>
<tr>
<td>TA</td>
<td>Trace Amine</td>
</tr>
<tr>
<td>TAAAR</td>
<td>Trace Amine-Associated Receptor</td>
</tr>
<tr>
<td>TAS1R</td>
<td>Taste Receptor type 1</td>
</tr>
<tr>
<td>Taste2(s)</td>
<td>GPCR(s) belonging to the Taste2 (Taste Receptor type 2) family</td>
</tr>
<tr>
<td>Vomeronasal1</td>
<td>GPCR(s) belonging to the <em>V1R</em> (Vomeronasal Receptor type 1) family</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>V2R</td>
<td>Vomeronasal Receptor type 2</td>
</tr>
<tr>
<td>VR</td>
<td>Vomeronasal Receptor</td>
</tr>
</tbody>
</table>
INTRODUCTION

The physiologic and therapeutic significance of GPCRs

The super-family of G protein-coupled receptors (GPCRs) is one of the largest protein families in humans (Lander et al., 2001; Venter et al., 2001). GPCRs are signal mediators that have a prominent role in most major physiological processes at both the central and peripheral level. Several human diseases are caused by their dysfunction and many pharmaceuticals and drugs of abuse convey their effects by modulating them. It has been estimated that GPCRs represent 30-45% of current drug targets (Drews, 2000; Hopkins and Groom, 2002). Drugs have only been developed for a small number of the GPCRs and the potential for further drug discovery within this field is very large.

Molecular structure, activation and signalling of GPCRs

The key structural component of GPCRs is seven transmembrane (7TM) α-helices that span the cell membrane. Every helix is constituted by 25-35 amino acid residues of relatively high hydrophobicity. The 7TM helices are preceded by an extracellular N-terminus, connected by three loops on each side of the cell membrane and succeeded by an intracellular C-terminus.

*Figure 1.* A schematic view of the GPCR structure. The seven transmembrane helices span the cell membrane and are connected by three extracellular and three intracellular loops. The N-terminus is outside and the C-terminus inside of the cell.
Many large ligands activate the GPCRs by binding to the extracellular N-terminus and/or loops, whereas most small ligands bind within a pocket between the 7TM helices. GPCRs respond to large variety of signals including endogenous ligands such as amines, peptides, amino acids, glycoproteins, prostanoids, phospholipids, fatty acids, nucleosides, nucleotides, Ca\(^{2+}\) ions as well as many exogenous sensory messages such as light (photons), odors, bitter and sweet taste, pain and pheromones (Bockaert and Pin, 1999). Upon activation, the 7TM helices change configuration to uncover sites within the second and third intracellular loops and/or the C-terminus, which can interact with G proteins or other intracellular signal mediators (Wong, 2003). G proteins are heterotrimers composed of one \(\alpha\), one \(\beta\) and one \(\gamma\)-subunit. Upon activation by a ligand-bound receptor, the \(G_{\alpha\beta\gamma}\) complex dissociates into the active \(G_{\alpha}\) and \(G_{\beta\gamma}\). These can initiate cellular response by altering the activity of specific effector molecules inside the cell (mostly enzymes and ion channels). Eventually, \(G_{\alpha}\) re-associates with the \(G_{\beta\gamma}\) dimer, regenerating the inactive \(G_{\alpha\beta\gamma}\) heterotrimer. A single activated GPCR, G protein or the G protein-activated enzyme/ion channel can activate several downstream effectors leading to signal amplification. Regulation of the receptor activity occurs through several cellular mechanisms, including receptor desensitation (fast), internalisation and down-regulation (slow).

The GRAFS nomenclature for the human GPCRs

The most widely used nomenclature for GPCRs hold families designated A-F and was introduced in the first version of the GPCR database, GPCRdb (Kolakowski, 1994). These families, although initially not called A-F, were derived from the use of characteristic “fingerprints”, unique compositions of conserved regions within a multiple alignment (Attwood and Findlay, 1993; Attwood and Findlay, 1994). By another nomenclature the GPCRs are divided into “clans” 1-5 using a combination of physiologic (e.g. location of ligand binding site) and structural features (including phylogenetic analyses) (Bockaert and Pin, 1999). These systems attempt to cover the entire GPCR repertoire in several developmental lineages but do not include some of the more recently discovered families.

More recently, the human GPCRs were divided into five main families strictly based on large scale phylogenetic analyses (Fredriksson et al., 2003). These families were named Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2, and Secretin; and are together abbreviated GRAFS.
Table 1. The human GPCR families according to the GRAFS nomenclature, designations in other classification systems and the number of full-length and interrupted members. The Taste2 and Frizzled families group together as one in phylogenetic analyses of only the human members, but separately if including the rat Taste2s. The Vomeronasal1s and V2Rs are excluded because they are non-functional in human.

<table>
<thead>
<tr>
<th>GRAFS</th>
<th>Designation in the A-F and 1-5 Systems</th>
<th>Number of full-length members</th>
<th>Number of interrupted members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutamate (incl. TAS1Rs)</td>
<td>C, 3</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Rhodopsin (incl. ORs)</td>
<td>A, 1</td>
<td>286 + 388 OR</td>
<td>26 + 479 OR</td>
</tr>
<tr>
<td>Adhesion</td>
<td>B (B2), 2</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>Frizzled/ Taste2</td>
<td>F, 5</td>
<td>11/25</td>
<td>0/10</td>
</tr>
<tr>
<td>Secretin</td>
<td>B, 2</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

The two largest families in human, the Rhodopsin (including the ORs) and Adhesion families are described later in this thesis. The Glutamate family includes the receptors that bind glutamate, GABA and calcium as well as the sweet taste receptors (TAS1R) and vomeronasal type 2 receptors (V2Rs). The Secretins bind the large peptides such as secretin, parathyroid hormone, glucagon, glucagon-like peptide, calcitonin, vasoactive intestinal peptide, growth hormone releasing hormone and pituitary adenyl cyclase activating protein. The Frizzleds bind among others the Wnt ligands and play an important role in the embryonic development. The Taste2s are involved in bitter taste sensing. Apart from the GRAFS GPCRs, a few (8 in human) receptors exist that do not show sequence similarity to any of the main families.

Evolution of the GPCRs

The GPCRs superfamily; dynamic and expanding

It is uncertain whether all GPCRs share a common ancestor. GPCRs in fungi, plants and animals have no sequence similarity, except for one Adhesion-like GPCR found in thale cress (Arabidopsis thaliana) (Josefsson and Rask, 1997). For example, the MLO GPCRs exist only in plants and the STE2 and STE3 GPCR families are only found in yeast. The GPCR families
in human, the GRAFS families, arose before the chordate lineage diverged from the lineage leading to nematodes because *Caenorhabditis elegans* has more than 100 receptors belonging to these families (Fredriksson and Schioth, 2005). In parallel to these, other GPCR families have arisen or evolved in specific lineages/species, like the nematode chemosensory receptors (Robertson, 1998), insect gustatory receptors (Hill et al., 2002) and mammalian vomeronasal receptors involved in pheromone recognition (Kouros-Mehr et al., 2001).

High species variation within the repertoires of Trace Amine Associated Receptors (TAARs)

Trace amines (TAs), substances closely related to the classical biogenic amines, have been known to exist in mammalian brains since the seventies (Boulton, 1976) but it was not until recently receptors that bind TAs were discovered (Borowsky et al., 2001; Bunzow et al., 2001; Lindemann et al., 2005). These receptors, referred to as Trace Amine-Associated Receptors (TAARs), according to a new nomenclature (Lindemann et al., 2005), have now been described in several species. The number of full-length TAARs varies remarkably between primates (human 6 and chimpanzee 3), rodents (rat 17 and mouse 15) and fishes (zebrafish 57 and pufferfish 8) (Gloriam et al., 2005; Lindemann et al., 2005). The primate and rodent TAARs are located in tandem in one locus. Phylogenetic analyses indicate that 9 TAARs were present before the lineages leading to primates and rodents split (Lindemann et al., 2005). Fishes and mammals however, only share one orthologous receptor, TAAR1, whereof there are multiple copies in zebrafish and fugu (Gloriam et al., 2005).

Ligands have only been identified for TAAR1 (tyramine, phenylethylamine, amphetamines, ergoline derivates and adrenergic agents) and TAAR4 (phenylethylamine) (Borowsky et al., 2001; Bunzow et al., 2001; Hart et al., 2006; Lindemann et al., 2005). It has been proposed that the other TAARs could bind non-TA substances (Lindemann and Hoener, 2005). Likewise, it has been suggested that the role of non-TAAR-binding TAs is to modulate the level of other neurotransmitters or that they can activate other receptors than the TAARs. Elucidating the role of the TAs is of great importance because they have been indicated to be involved in several pathological conditions, such as schizophrenia, depression, attention deficit hyperactivity disorder, substance abuse and eating disorders (Branchek and Blackburn, 2003; Davenport, 2003; Premont et al., 2001).
The *Rhodopsin* family of GPCRs

Receptor ligands and their binding

The *Rhodopsin* family is the largest of the human GPCR families. There are at least 286 human non-olfactory *Rhodopsins* whereof the majority binds peptides, biogenic amines or lipid-like substances. The receptors binding endogenous peptides have an import role in mediating the effects of a wide variety of neurotransmitters, hormones and paracrine signals. The receptors that bind biogenic amines, e.g. norepinephrine, dopamine, and serotonin, are very commonly modulated by drugs. Pathological conditions, including Parkinson's disease, schizophrenia, drug addiction, and mood disorders are examples of where imbalances in the levels of biogenic amines cause altered brain functions.

Unlike the receptors within other GPCR families, most *Rhodopsins* have very short N-termini and bind their ligands within the 7TM region. However there are also other ways of activating GPCRs within this family. The large glycoprotein hormones (follicle-stimulating hormone, luteinising hormone and thyroid-stimulating hormone) have long N-termini which bind their ligands. The protease-activated receptors (F2R, F2RL3, F2RL1 and F2RL2) have their N-termini cut by a protease and the shortened N-termini then bind to domains in the extracellular loop and activate the receptor. Furthermore, the opsins (Rho and OPNs) hold a covalently bound ligand (11-cis retinal) which upon the absorption of a photon undergo a change in conformation which activates the receptor.

*Rhodopsin* family subgroups

The relationships of many single, pairs and groups of *Rhodopsins* can not be resolved in phylogenetic analyses because they have too low or ambiguous sequence similarity to the other family members. Examples of such *Rhodopsins* are DUFFY (single), GPR153-GPR162 (pair) and the prostaglandin-binding receptors (group). Resolved relationships are obtained mainly for the biogenic amine-, MECA (defined in paper VI), purine- and most peptide-binding receptors. Several clusters of receptors binding to different ligand types, such as peptides and lipids, group adjacently in phylogenetic analysis showing that the ability to bind a certain ligand type could have evolved many times during evolution.
Orphan Rhodopsins

Receptors for which the ligand is still undiscovered are called orphan receptors. We studied the literature and found that 70 of the 284 human Rhodopsins are still orphan (Paper VI). Because of the high abundance of drug targets within the Rhodopsin family the finding and characterisation of orphan receptors is important (Hopkins and Groom, 2002).

GPCRs that bind the same ligands generally have the highest sequence similarity, but the orphans are not closely related to any characterised receptors and no specific ligands can be inferred. However, some orphans have a low sequence similarity to larger groups of Rhodopsins that bind the same type of ligand such as peptides, amines or lipids. It is likely that these orphans share the type of ligand although the specific ligand is unknown. Most of the orphans are so atypical that they are not part of any of the groups of Rhodopsins that share ligand types. For these the sequence similarity is so low that it is very difficult to subtract any information about the type of ligand.
The *Adhesion* family of GPCRs

The GRAFS classification system is the only system in which the *Adhesions* are clearly recognised as a family. For example the A-F classification system (Kolakowski, 1994) and the conserved domain search at the NCBI website, identify the *Adhesion* and *Secretin* families as one. The *Adhesions* are equipped with a wide diversity of conserved domains in their N-termini, whereas the secretin receptors only have a HBD (hormone-binding domain). Most of the human domains are present in mice (Bjarnadottir et al., 2004) and also in more distant species (unpublished data).

The conserved domain that exists in most of the *Adhesions* is the GPCR proteolytic site (GPS), which all family members except GPR123 have. The GPS domain is located in the extracellular N-terminus not far from the first transmembrane helix. Following translation, the receptor protein can be cleaved within the GPS (Abe et al., 2002; Gray et al., 1996; Ichtchenko et al., 1999; Krasnoperov et al., 1997; Moriguchi et al., 2004; Nechiporuk et al., 2001; Obermann et al., 2003; Ponting et al., 1999; Stacey et al., 2002; Usui et al., 1999). For LEC2 it has been shown that the N-terminus and the 7TM region can function independently at different locations in the cell surface (Volynski et al., 2004). In the example of LEC1 and GPR126 (Ig-Hepta) the 7TM part link non-covalently to the extra-cellular unit forming a heterodimer (Abe et al., 1999, (Kwakkenbos et al., 2004). Only a few *Adhesions*; CD97, EMR1, GPR64; have been associated with G protein-coupled signalling (Foord et al., 2002).

The N-termini of the *Adhesions* are very rich in glycosylation sites (NXS or NXT) and for GPR126 it has been shown that the glycosylation is responsible for 59% of its molecular weight (Stehlik et al., 2004). The glycosylation makes the N-terminus more rigid so that it forms a stalk-like structure that erects from the cell surface. In this way the conserved domains in the N-termini can reach and interact with adjacent cells or the extracellular matrix. The composition of domains (types and numbers) can often be altered by alternative splicing resulting in different interaction properties for the same receptor. Several *Adhesions* have a similar composition of conserved domains in the N-termini and are therefore though to be functionally grouped. As can be seen in Figure 2, this grouping correlates well with phylogenetic grouping based on the 7TM domain.
One such group is made up of the **BAIs** (brain-specific angiogenesis inhibitor receptors) that have one HBD and several (3-5) thrombospondin 1 (TSP1) domains in their N-termini. TSP1 domains have been found in proteins that help navigate neuronal axons during neuronal development and are involved in cell-cell interactions and inhibition of angiogenesis (Stacey et al., 2000). The BAIs have been suggested to participate in the regulation of ischemia-induced brain angiogenesis (Kee et al., 2002) and in the progression of gliomas (Kaur et al., 2003; Kaur et al., 2005).

The **CELSRs** (cadherin EGF LAG seven-pass G-type receptors) also have a HBD and in addition multiple cadherin, epidermal growth factor (EGF) and laminin domains. Cadherins are involved in Ca\textsuperscript{2+}-mediated cell-cell adhesion in epithelial cells, whereas laminin domains can have binding sites for various molecules such as steroids, heparin and sulfatides (Marchler-Bauer et al., 2005). Mutations in mouse CELSR1 has been shown to cause neural tube defects (Curtin et al., 2003). The fruitfly paralogue “flamingo/Starry night” and the zebrafish and chicken orthologues have been shown to have an essential role within the nervous system and for planar cell polarity (Formstone and Mason, 2005a; Formstone and Mason, 2005b; Usui et al., 1999). Recently, it was also discovered that the CELSRs are required for spermato genesis (Beall et al., 2005).

The **EMRs** (egf-like module containing receptors; EMR1-4, CD97 and ETL) contain multiple EGF domains which can bind to decay-accelerating factor and chondroitin sulfate. This group of receptors seem to have a role in the generation of adhesive contact and/or migration of immune cells as well as binding of antigens (Kop et al., 2005; Kwakkenbos et al., 2005).

The **LECs** (lectomedin receptors), also called CIRLs (calcium-independent receptors for latrotoxin) or LPHNs (lathrophilin receptors), contain N-terminal HBD, galactose-binding lectin and olfactomedin domains. They were identified based on their ability to bind alfa-lathrotoxin, the major component of the black widow spider venom, but neither their endogenous ligand nor function is known. LEC1 and LEC3 are selectively expressed in the brain, whereas LEC2 are uniformly expressed in all tissues.
Figure 2. A Phylogenetic tree of all human Adhesions surrounded by the respective N-termini. The illustrations of the N-termini are from Bjarnadottir et. al (Bjarnadottir et al., 2004) and have been modified using information form additional sources (Bjarnadottir et al., 2004; Kwakkenbos et al., 2004; Stehlik et al., 2004).
Other Adhesions do not form well defined groups with common properties (phylogenetic clustering, function, expression and/or domain content). However, VLGR1 can have up to seven copies of the EPTP (Epitempin) repeat which is found in a heterogeneous group of proteins linked to epileptic diseases (Scheel et al., 2002; Staub et al., 2002). A nonsense mutation causing a deletion of the C-terminal 126 amino acid residues has been identified in one family with febrile and afebrile seizures (Nakayama et al., 2002). Furthermore, mutations in VLGR1 can cause the Usher syndrome which is characterised by hearing impairment and retinitis pigmentosa (Weston et al., 2004). Moreover, GPR126 is expressed within the endothelial cells of certain blood vessels upon lipopolysaccharide or thrombin treatment and has been suggested to function in cell-cell adhesion in inflammation and coagulation (Stehlik et al., 2004). GPR126 contains a CUB domain and other proteins that have been found to contain the CUB domain are involved in blood coagulation, cleavage of collagen, development and immunity (Bork and Beckmann, 1993; Duke-Cohan et al., 2000; Hulmes et al., 1997).

The remaining Adhesions have no known function but some contain conserved domains that can provide clues. GPR116 has a SEA (Sperm protein, Enterokinase, and Agrin) box in the beginning of the N-terminus. Just like the GPS, this domain can serve as a site for proteolytic cleavage (Abe et al., 2002). GPR116, GPR124 and GPR125 all have immunoglobulin (IG) domains. IG domains are likely to be involved in protein-protein and protein-ligand interactions and are found in hundreds of proteins, like antibodies, receptor tyrosine kinases and muscle kinase titins (Marchler-Bauer et al., 2005). GPR124 and GPR125 have a domain termed LRR (leucine rich repeat). LRRs are short motifs present in a number of proteins with diverse functions and cellular locations (Marchler-Bauer et al., 2005). The LRR domains in the two receptors are similar, but are not found in any other GPCRs. GPR144, GPR126, GPR112 contain a pentraxin domain. The pentraxins consist of up to five non-covalently bound identical subunits that are arranged in a flat pentameric disk. Other proteins that contain pentraxin domains have been linked to the activation of immunological response (Goodman et al., 1996).

Ten Adhesions lack conserved domains apart from the GPS. However, the fact that they have glycosylation sites strengthens the likelihood that these receptors have domains that have not been recognised yet. HE6 and GPR56 are two Adhesions for which no N-terminal domains have been identified although they have both been shown to have adhesive properties. The HE6 attachment properties are required for maturing germ cells and mutation of this receptor in mouse has resulted in male infertility (Davies et al., 2004). GPR56 has been shown to be overexpressed in gliomas and function in tumor cell adhesion (Shashidhar et al., 2005). Moreover, mutations in GPR56
have been shown to cause a human brain cortical malformation (Piao et al., 2004).
AIMS OF STUDIES

The Aims of this thesis were to:

1. Identify new human GPCRs belonging to the Rhodopsin and Adhesion families (Papers I-III)

2. Study the evolution of the Trace Amine-Associated Receptors (TAARs) (Paper IV)

3. Compile and analyse the overall repertoires of GPCRs in human, mice and rats (Papers V and VI)
METHODS

This section describes some of the practical aspects of the methodology used in this thesis. More detailed method descriptions can be found in the material and method sections of the papers.

Searching for GPCRs

Sources of sequence databases
The sources of the nucleotide and protein sequence databases used in this thesis; the US National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/), Ensembl (http://www.ensembl.org/index.html) and Celera (www.celera.com); contain complementary information. NCBI and Ensembl, as well as the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp), are primary annotation centrals for researchers world-wide. These exchange data frequently and therefore, in most cases, hold identical information. However differences are seen between the Ensembl and Refseq (from NCBI) protein databases due to different annotation criteria (Larsson et al., 2005). Moreover, the Ensembl and NCBI online services hold different sets of downloadable databases, search options/parameters and genome browsers. Celera is a commercial enterprise that focuses on the discovery of therapeutic targets and the discovery and development of new drugs. We used the Celera Genomics human genome assembly as a complement before the public version of the human genome was completed. Its databases can not be downloaded (except for the first human genome draft) and the online interface is slow, making it difficult to use for large scale searches. Thus, we now only use Celeras databases when searching for small sets of sequences such as missing rat and mouse orthologues.

Sequence similarity based search strategies
The easiest way of performing a sequence similarity based search is perhaps to use the NCBI online BLAST (Basic Local Alignment Search Tool) search programs and databases (http://www.ncbi.nlm.nih.gov/blast). The hits are
ordered in decreasing similarity and are linked to a vast amount of information, e.g. to publications, facilitating their recognition and collection of relevant information.

For extensive studies however there is a need to perform multiple simultaneous searches (batch searches). Batch BLAST can be performed online at Ensembl, but only for a limited number of query sequences and options. NCBI distributes a free BLAST program package which makes it possible to query an online database from a local computer or to download and search databases locally (ftp://ftp.ncbi.nih.gov/blast/executables).

BLAST (Altschul et al., 1997) rapidly identifies sequences of high similarity, but when searching for distantly related homologues, such as atypical human or evolutionary distant family members, a higher sensitivity is often obtained using Hidden Markov Models (HMMs) (Eddy, 1998). These two methods work in different ways and can thus give different (complementary) results. BLAST programs identify the local alignments sharing the highest number of identical or similar residues (hits), whereas the HMM algorithm makes a statistical model from a multiple alignment of a training sequence set that is used to scan a database for sequences belonging to the set.

Besides the choice of search method, the completeness (degree of annotation) of the sequence databases will affect the outcome. The new human GPCRs presented in this thesis were mainly found as putative proteins predicted from the genome sequence by Genscan (Burge and Karlin, 1997) or as full-length cDNAs deposited in the NCBI non redundant (nr) database (http://www.ncbi.nlm.nih.gov/About/index.html), two types of data that are not dependent on human annotation and therefore more extensive. The genomes are even more complete sources, especially for gene families that have proven difficult to annotate. Paper VI in this thesis have utilised TBLASTN against the human, mouse and rat genomes to mine large and diverse GPCR families. Individual or small groups of sequences can also be searched in a genome using BLAT (Kent, 2002). BLAT, which has a relatively low sensitivity but is quick, is appropriate for instance when searching for missing orthologues (if the evolutionary distance is not too long).

Filtering the large search results

Removing GPCRs from other families/groups and non-GPCRs
Sequence searches yield many GPCRs belonging to other families than the currently studied as well as proteins that are not GPCRs. The desired/correct
hits can be defined as hits having GPCRs from the right family as first hits in a BLAST search against a database containing all protein families. Undesired/false hits will have other GPCR- or protein families as a first hits. When using this method, it is important that the validation database has a good representation of all major protein families and that its sequences have been names that describe their identity. One good choice is Refseq which contains most annotated proteins.

Removing duplicate hits

Duplicate hits that are identical to sequences in the start/reference dataset can easily be identified using BLASTCLUST, which can list identical sequences (threshold of minimum identity set to 100 percent). Hits that are non-identical duplicates, such as splice variants and incorrectly predicted proteins, can be identified by using local BLAST against the reference dataset but require individual inspection of their alignments. Hits that are sequence duplicates having only a few scattered sequence differences, such as polymorphic genes or genes containing sequencing errors, can be difficult to distinguish from unique genes by visual inspection of alignments. For the tough cases BLAT can be used to distinguish between pairs of unique genes (have two unique genomic locations) and sequence duplicates (map to the same place). This works well for relatively complete genome assemblies, such as that of human, mouse and rat, but can give false results if a relevant part of the genome assembly is missing.

Manual curation of predicted coding sequences

Genscan predicts 80% of exons correctly (Burge and Karlin, 1997) and sometimes misses and/or include false exons in a protein. This kind of errors can be identified by comparing it to other members of the same family with for example BLAST or alignment program. Falsely included exons can often be removed directly when identified after defining their borders using BLAT, whereas missing exons have to be identified from physically expressed sequences such as full-length cDNAs or ESTs, which can be found in the nr and dbEST (http://www.ncbi.nlm.nih.gov/About/index.html) respectively. These ESTs, full-length cDNAs and the gene sequence is assembled and the missing exon is filled in. This procedure can also be used to identify splice variants.

Figure 3. A part of the assembly of one GPCR and three EST sequences. The GPCR is missing one exon, which can be added using the ESTs as template.
Phylogeny - Determining sequence relationships

The *Rhodopsin* family of GPCRs contains some members which have so atypical sequences that their relationships can not be determined by phylogenetic analyses. These will group ambiguously when performing phylogenetic analysis using different methods or versions of the dataset and should therefore be excluded from phylogenetic analyses. Such atypical family members can be distinguished from the rest of the dataset using BLAST-CLUST to group sequences with a minimum threshold of sequence similarity.

When aligning GPCR protein sequences in order to make phylogenetic calculations it is appropriate to use only the 7TM region, which is generally under stronger evolutionary pressure and therefore more conserved than the N- and C-termini. The latter parts are usually so divergent that they can only be aligned for receptor subtypes (GPCRs sharing ligands). The borders of the 7TM regions of *Rhodopsins* can be defined by aligning them to the sequence of bovine rhodopsin for which the crystal structure has been determined (Palczewski et al., 2000). Some GPCRs, e.g. NPY5R, also have very long loops that, if not removed, split up the alignment.

We have used bootstrapping, to obtain multiple datasets (e.g. 100) from the alignment of the 7TM regions. Bootstrapping makes the multiple datasets by duplicating and deleting random columns of the original alignment. Each and every dataset is analysed individually and generates one tree. The correctness of the outcome can be estimated from the variation of the results, i.e. the number of identical nodes among all trees generated.

The phylogenetic analyses in this thesis have been performed with different methods included in the PHYLIP software package (Felsenstein, 1993). Maximum Likelihood (ML) is regarded as the most accurate method. However this method is very time-consuming for large datasets. The NJ (Neighbor Joining) method is more suited for large datasets. Maximum Parsimony (MP) has intermediate time consumption and has in several analyses in this thesis given more resolved trees. However, MP can falsely group sequences that do not have a common descent but have similarity at random or because of convergent evolution. This bias is called long-branch attraction.

In several of the papers in this thesis, sequence similarities have been measured using BLAST searches. A more visual way to display the distance of sequence relationships is to calculate branch lengths for the phylogenetic trees. BLAST can also be used for validation the relationships of sequences.
in a phylogenetic tree. As a rule of thumb, a phylogenetic tree is more reliable if the order of BLAST hits agree with the topology of the tree.

Expression profiles from EST data
In the four first publications of this thesis ESTs identified from BLAST searches with a crude cut-off (e-value = $1e^{-12}$), which can miss a few true hits but generate very few false hits (obtained in rare cases when there are other genes that have very high sequence similarity to the queried gene). In a refined approach we then BLAST searched all ESTs identified against all known GPCRs in order to match them correctly (paper V). Recently, we mapped all human and mouse ESTs and genes to their corresponding genomes assemblies enabling a nearly perfect matching defined from genomic overlap (unpublished).

Estimates of expression are more likely to be accurate if applying a threshold for the number of ESTs (e.g. demanding 4 ESTs in a certain tissue/organ). The ESTs databases have grown very fast and several species now have a considerable number of annotated ESTs (Table 2), but the annotation is biased towards certain diseases and developmental stages and many tissues and genes are not represented in the EST libraries. However, although often incomplete and/or unspecific (e.g. brain/CNS), the expression patterns derived from ESTs give the possibility to choose genes or tissues for more detailed analyses using for instance RT-PCR or In situ hybridisation. It is important to remember that EST analysis is not quantitative and expression levels can not be directly inferred from the numbers of ESTs within a certain tissue (Gupta, Zink et al. 2004).

Table 2. The ten organisms having the highest number of EST entries (January 20, 2006).

<table>
<thead>
<tr>
<th>Organism</th>
<th>EST Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em> (human)</td>
<td>7,596,977</td>
</tr>
<tr>
<td><em>Mus musculus + domesticus</em> (mouse)</td>
<td>4,690,536</td>
</tr>
<tr>
<td><em>Xenopus tropicalis</em> (pipid frog)</td>
<td>1,038,272</td>
</tr>
<tr>
<td><em>Bos taurus</em> (cattle)</td>
<td>837,648</td>
</tr>
<tr>
<td><em>Rattus sp.</em> (rat)</td>
<td>812,662</td>
</tr>
<tr>
<td><em>Danio rerio</em> (zebrafish)</td>
<td>689,613</td>
</tr>
<tr>
<td><em>Ciona intestinalis</em> (sea squirt)</td>
<td>686,395</td>
</tr>
<tr>
<td><em>Zea mays</em> (maize)</td>
<td>662,884</td>
</tr>
<tr>
<td><em>Triticum aestivum</em> (wheat)</td>
<td>600,205</td>
</tr>
<tr>
<td><em>Gallus gallus</em> (chicken)</td>
<td>588,739</td>
</tr>
</tbody>
</table>
Submitting new sequences to a public database

After identifying new human GPCRs we have sent them to the human nomenclature committee (Wain et al., 2004) that confirm whether they are new or not, and if they are, assign them official names (GPR-“number”). Subsequently, new human GPCRs (papers II and III) have been submitted to Genbank making them publicly available in nr. However, the most recent human GPCRs (paper I) could not be included in nr because the Genbank administration changed its policy to only approve sequences obtained from physical experiments. When other variants of the new GPCR sequences, such as Gen-scan predictions or full-length cDNAs, exist in Genbank they adopt the new GPCR names their records are not linked to our publications and are often hold incomplete variants of the sequence.
RESULTS AND DISCUSSION

Paper I - Nine new human *Rhodopsins*

In this publication we identified nine new human *Rhodopsins*. Human GPR139, GPR146, GPR148, GPR150, GPR152, GPR153, GPR154, GPR162 and GPR165P (names provided by the HUGO Gene Nomenclature Committee; HGNC) were found from BLASTP searches in the nr database and HMM search of the human Genscan dataset.

The coding regions of all new GPCRs were manually curated by assembling the gene sequences to human and mouse ESTs/cDNA. As we obtained coverage of all internal splice sites we could verify all exon-intron boundaries and we also found a total of seven alternative splice variants for three receptors. The mouse GPR153 was found to have one splice variant with a shortened C-terminus, whereas all other alternative splice variants do not code for proteins with full-length 7TM regions. Interestingly, some orthologues have longer C-terminus or third intracellular loop, -the parts involved in G protein binding.

We identified 37 orthologues (Table 1 in paper I) and could trace the origin of the new *Rhodopsins* to before the split of the lineage leading to humans and those leading to insects (GPR154), bony fishes (GPR139, GPR146, GPR148, GPR153), chickens (GPR152, GPR162) and rodents (GPR150, GPR165). Only two species-specific duplicates were found (GPR148 in zebrafish and GPR154 in mosquito). Low conservation was seen for the orthologues of GPR148 (from 23% in chicken to 29% in fugu and zebrafish), whereas very high conservation was observed for GPR139 (71% in fugu to 96% in mouse and rat) and GPR162 (96% in mouse).

GPCRs often exist in multi-gene (4-8 members) families, but phylogenetic analyses and BLAST searches show that the new receptors are atypical (Figure 1 in paper I). Four receptors show weak relationships to previously characterised GPCRs: GPR150 and GPR154 to oxytocin/vasopressin receptors, GPR152 to CRTH2/FPRs and GPR165 to GPR72/NPYR. Since this sequence similarity of the new receptors to the characterised receptors is relatively low it is likely that they are activated by other ligands. Four of the
receptors GPR139, GPR146, GPR153 and GPR162 are related to one other orphan GPCR, whereas GPR148 has no closely related receptor.

Moreover, we made expression profiles based on human and mouse expressed sequence tags (ESTs) records (Figure 2 in paper I). GPR139, GPR150, GPR154 and GPR165 were almost exclusively found in the CNS. For GPR146, GPR153 and GPR162 the majority of the ESTs were derived from the CNS but ESTs were also found from several peripheral tissues. For GPR152 we did not find more than two ESTs in the same any tissues (lung and eye) and for GPR148 we found no ESTs at all.
In this work we identified seven new human *Rhodopsins*. Human GPR100, GPR119, GPR120, GPR135, GPR136, GPR141 and GPR142 (names provided by HGNC) were found from BLASTP searches in the Celera database and HMM search of the human Genscan dataset.

The predicted coding regions were manually curated by assembling the GPCR gene sequences to a combination of vertebrate ESTs and/or cDNAs and also from alignment with machine predicted mouse orthologues (GPR120, GPR136 and GPR142 contain internal splice sites not covered by ESTs). No splice variants were identified, probably because of the relatively low number of ESTs found.

A total of 16 orthologues could be identified (Table 1 in paper II). Five receptors had orthologues in fish, whereas GPR120 and GPR141 were only found in rodents. The only species-specific duplicate found was for GPR142 in fugu but the extra copy was later found to be another GPCR subtype (GPR139 in paper I). Some orthologues display an unusually high conservation. The TM regions of human, mouse and rat GPR141 are over 95% identical and fugu GPR135 and GPR136 sequences are 69% and 74% identical to their human counterpart.

Only GPR100 has a close structural relative in the human genome, SALPR, and the other six new GPCRs make up their own "single member subgroups" (Table 2 in paper II).

The expression patterns derived from EST data gave the following information: GPR100 (marrow), GPR119 (pancreas), GPR120 (stomach*2), GPR135 (eye*2, cervix, stomach, brain, testis), GPR136 (testis, unknown), GPR141 (marrow, unknown*5) and GPR142 (none). The relatively low number of ESTs can be due to low expression of the genes or the under-representation of many tissues and genes in the EST database.
Paper III - Six new human Adhesions

In this work, our group identified the sequences of six new human Adhesions. Human GPR123, GPR124, GPR125, GPR126, GPR1127 and GPR128 (names provided by the HGNC) were found from BLASTP searches in the Celera database.

As Adhesions have so many exons, it is extra important to manually curate their predicted coding sequence. The coding regions of GPR123, GPR125, GPR126 and GPR128 could all be fully verified as complete coverage was obtained when assembling them to vertebrate ESTs. GPR125 was identified in three (splice) variants, one of them introducing a leucin rich domain in the N-terminus. The CDS coverage was not complete for GPR124 (splice sites 5-9 covered) and GPR127 (no ESTs) and machine predicted mouse orthologous from nr were used as a complement to verify their exon-intron boundaries.

In the phylogenetic analysis (Figure 2 in paper III) three of the new receptors; GPR123, GPR124 and GPR125; form a new group. Two of receptors, GPR126 and GPR128, show phylogenetic relationship to HE6 and GPR56, which have been known for a few years, and three receptors that our group described recently (Fredriksson et al., 2002). The sixth receptor, GPR127, shows clear phylogenetic relationship to EMR1-3.

Several types of conserved domains were identified in the N-termini of the new receptors (Table 1 and Figure 3 in paper III) using conserved domain search at NCBI (Marchler-Bauer and Bryant, 2004). These were CA (Caderhin), EGF (Epidermal Growth Factor), GBL (Galactose Binding Lectin domain), GPS (Gpcr Proteolytic Site), HBD (Hormone-Binding Domain), Ig (immunoglobulin), Lam (Laminin), LRR (Leucin Rich Repeat), Olf, (olfactomedin), PTX (pentraxin), SEA (Sperm protein, Enterokinase, and Agrin) and TS (thrombospondin). Most of these domains are also found in other proteins than the GPCRs. Surprisingly, no GPS domain was found in GPR123 (all other known Adhesions have this domain). Other domains spread into some of the six Adhesion family subgroups (e.g. HBD & EGF) or occur specifically in a certain phylogenetic cluster (e.g. TS-BAI; CAD and Lam-CELSR; and GBL and OLF-LEC), suggesting that some physiological functions are shared whereas others are unique for subfamilies.

The tissue distribution derived from the EST data shows highly individual patterns for each of the receptors (Table 3). Unexpectedly not a single EST was found belonging to the immune system, the first suggested place of action of the Adhesions.
Table 3. The ESTs of the GPCRs in paper III divided into three main categories.

<table>
<thead>
<tr>
<th>GPR123</th>
<th>CNS</th>
<th>Reproductive</th>
<th>Periphera/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>brain</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPR124</td>
<td>dorsal root ganglion</td>
<td>uterus, placenta, ovary</td>
<td>lung</td>
</tr>
<tr>
<td>GPR125</td>
<td>brain</td>
<td>ovary, cervix, testis, uterus</td>
<td>kidney, colon, skin, liver, lung, breast, heart, stomach, eye, NT2, head, nervous</td>
</tr>
<tr>
<td>GPR126</td>
<td>-</td>
<td>placenta, germ cell, testis</td>
<td>pancreas, muscle, liver, lung/spleen, small intestine, skin, total fetus</td>
</tr>
<tr>
<td>GPR127</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPR128</td>
<td>-</td>
<td>Cervix</td>
<td>skin</td>
</tr>
</tbody>
</table>
Paper IV - The evolution of the Trace Amine-Associated Receptors

In this study we revealed remarkable species variations between the repertoires of Trace Amine-Associated Receptors (TAARs) among mammals (human 8, mouse 15 and rat 17 members) and fishes (zebrafish 57 and pufferfish 8). Lindemann et al. simultaneously described the variations between the TAAR repertoires in mammals and showed that differences are also seen for primates (human 6 and chimpanzee 3 full-length members) (Lindemann et al., 2005).

The TAARs have among the shortest (316-384 residues) GPCR protein sequences and are intronless (TAAR2/GPR58 has a transcript variant with two exons). The mammalian TAARs are located in only one locus (Figure 1 in paper IV). Their order is conserved and the rodent duplicates are positioned in direct adjacency showing that they are most likely the result of gene tandem duplications. 47 of the 57 zebrafish TAAR genes were mapped nine different loci (on six chromosomes) containing from 1 to 27 genes, indicating that they arose through several different mechanisms involving tetraploidisation, block duplications, and local duplication events.

In the tree calculated using maximum likelihood (ML) (Figure 2 in paper IV) two zebrafish and five fugu TAARs group with the mammalian TAAR1/TA1, however none of them do so in the maximum parsimony (MP) and neighbor joining (NJ) analysis and thus it is uncertain whether there are any orthologous fish and mammalian TAARs. The TAARs that are located adjacent in the genome most often group together in the phylogenetic analysis (Figure 2 in paper IV). Several TAARs seem to be very recent duplicates and have so high sequence similarity that their relationships can not be resolved in the phylogenetic analysis using any of the methods; ML, MP or NJ. In group VII 32 such zebrafish receptors spring from only one node and in group VIII several unresolved rodent TAARs are found, most of them species-specific.

Borowsky et al. showed that the invertebrate receptors that bind trace amines group with the other amine receptors (Borowsky et al., 2001) and thus they are not TAARs (according to phylogenetic grouping). We can confirm this (phylogenetic tree not displayed) and find it likely that the ancient amine binding receptors have at least twice developed specific preference for trace amines. We searched for, but did not find, invertebrate TAARs in the genome assemblies of the fruitfly, the sea squirt Ciona intestinalis and the nematode Caenorhabditis elegans.
We found very few ESTs for the TAARs, which is in accordance with earlier studies that have reported a low or intermediate expression levels of these GPCRs (Borowsky et al., 2001; Bunzow et al., 2001; Lee et al., 2000; Zeng et al., 1998).
Papers V-VI – The GPCR repertoires in human, mice and rats

In the two final papers of this thesis we analysed 793 human (+582 pseudogenes), 1859 rat (+737 pseudogenes) and 1776 mouse (+702 pseudogenes) GPCR gene sequences found by sequence mining using BLAST, BLAT, Hidden Markov Models and searched other GPCR datasets and literature data.

There are some large differences between the rodent and human GPCR repertoires (Table 1 in paper VI). We could confirm that mice have more than twice as many olfactory receptors (ORs) as humans (Malnic et al., 2004; Niimura and Nei, 2003; Zhang and Firestein, 2002) and also showed that the rat has even more ORs than the mouse (1235 compared to 1081). The vomeronasal receptors (VRs), which respond to pheromones, are non-functional in human whereas both the rodent species have over 100 members in both the Vomeronasal1 family and the large V2R subgroup of Glutamates. The human and rodent Taste2s are about as many, 35 in rodent and 25 in human, but make up only nine orthologous pairs (18%). Moreover, the human-mouse Taste2 orthologues have a lower conservation than the orthologues in other GPCR families (paper V supplementary material 10). This is likely to be an important contribution to the different senses of taste in humans and rodent. The Frizzled, Secretin and Glutamate (if not considering the V2Rs) families are the same (all members are orthologues) and the Adhesions are slightly fewer in rat than human (30 of the 33 human Adhesions are found in rat). Most of the Rhodopsins make up one-to-one orthologous pairs, except for the Olfactory Receptors (ORs), Trace Amine-Associated Receptors (TAARs), Formyl Peptide Receptor Like receptors (FPRLs) and Mas-Related G protein-coupled receptors (MRGs) that all display very large species differences, mainly resulting from species-specific expansions.

The extensive diversification of the Taste2s, VRs, ORs, TAARs and MRGs has been very fast as their evolutionary history is shorter than for the other mammalian GPCRs (Fredriksson and Schioth, 2005). They, as well as the nematode chemosensory receptors, bind relatively small molecular ligands requiring few interaction points to bind to the receptor. It can thus be speculated that the capacity of these GPCRs to evolve rapidly comes from a lower number of evolutionary constraints on the 7TM sequence.

In Paper VI we investigated the relationships between the GPCR families (Figure 1 in paper VI). Based on sequence similarity measurements we stated earlier that the Taste2s make up a family of their own (Schioth and Fredriksson, 2005) although phylogenetic analyses, with both the NJ and MP
methods, on the human dataset clustered them together with the *Frizzled* family (Fredriksson et al., 2003). In paper VI the rat GPCRs were included in the analysis. Now the *Taste2* family grouped apart from the *Frizzled* family and closer to the *Rhodopsin* and *Vomeronasal1* families, a grouping which is in accordance with sequence similarity measurements. The *Adhesion* and *Secretin* families group together, which is accordance with previous studies (Fredriksson et al., 2003). Furthermore we obtained a stable relationship (same in both NJ and MP) of all GPCR families, however many GPCR families have fairly low sequence similarity and the relationships in addition to those described above need to be confirmed by more detailed studies. In addition to the above GPCR families there are 8 GPCRs defined by us as “others” that form their own one or two member families. Two of these, GPR157 and Noname1, have a sequence similarity to other GPCR families (although low) and were included in the phylogenetic analysis of the relationship of the GPCR families. However, their relationship to other GPCR families could not be determined as NJ and MP display non-concordant relations.

Paper VI we also made a new investigation of the relationship of the *Rhodopsins* (Figure 2 in paper VI) based on the updated human dataset (45 receptors added since our publication in 2003) and including the rat orthologues. The new analysis gave similar, but not identical, results as the previous. Extensive reading of Pubmed literature showed that among the characterised human non-olfactory *Rhodopsins* there are 128 peptide-, 42 lipid-, 40 biogenic amine- and 9 purine (nucleosides and nucleotides) binding receptors (color coded in Figure 2 in paper VI). We found that there are 70 orphan non-olfactory *Rhodopsins* (red in Figure 2 in paper VI) spread into many subgroups. 61 of the human orphans have rat orthologue and 59 of them display 1:1 orthologous relationships.

Paper V presents EST charts (Figures 6 and 7 in paper VI) showing which mouse GPCRs that are broadly expressed and which that are specific for a certain tissue/organ.
MAIN CONCLUSIONS

Papers I-II - New human Rhodopsins
- The new Rhodopsins have low sequence similarity to characterised receptors, suggesting that they could have unique interactions with, or even new types, of ligands and/or G proteins.
- The new GPCRs have been indicated to be physiologically relevant as they have been evolutionary conserved and their genes are expressed (note: this was not shown for GPR142 and GPR148).

Paper III - New human Adhesions
- The conserved domains in the N-termini as well as the broad expression suggest that the Adhesion family is involved in several different physiologic processes.

Paper IV - The evolution of the Trace Amine-Associated Receptors
- The common ancestor of the vertebrate TAARs arose before the split between the lineages leading to the ray-finned and lobe-finned fishes.
- This group of receptors has undergone many and different duplication events leading to remarkable species differences.

Papers V-VI – The GPCR repertoires in human, mice and rats
- Rats have more olfactory receptors than mice (1235 compared to 1081).
- The human-rodent Taste2 orthologues are few and are less conserved than GPCR orthologues belonging to other families. This is likely to be an important contribution to the different senses of taste in humans and rodent.
- The Frizzled and Secretin families are the same (all members are orthologues) in human and rodents and 30 of the 33 human Adhesions are found in rat.
- GPCR families and groups that bind small molecular ligands requiring few interaction points to bind to the receptor seem to evolve more rapidly.
- There are currently 70 orphan non-olfactory Rhodopsins. 61 of the human orphans have orthologues in rat and of the orthologues 59 of these display one-to-one orthologous relationships.
FUTURE PERSPECTIVES

The human repertoire of GPCRs is most likely very near complete and the number of new human receptors reported is decreasing (Wistrand et al., 2006). The sequences of the new receptors are often very atypical, which is probably one of the reasons that it took long time to identify them. If more human GPCRs are found, it is likely that also they too have very atypical sequences. Thus it may prove difficult to determine if they have a common descent with other GPCRs. There are also many other groups of membrane proteins that have not been carefully mined, such as the solute carriers. We find it likely that the annotation process of membrane proteins will continue for many more years. This is in particular true for genes that have very complicated genomic structures, do not show classical protein motifs and are not abundantly expressed. It is, at this stage, difficult to know whether membrane proteins with multiple TM regions have a common origin.

There are many genomes that have not been studied as carefully as the human, mouse, rat and chicken (Lagerström et. al, submitted) genomes. Mining of the GPCR repertoires in bony fishes, cartilaginous fishes, lamprey, hagfish, amphioxus and sea squirts (e.g. *Ciona intestinalis*) could provide information about how the GPCRs were formed and have evolved. Much work is also needed to investigate if all the animal GPCRs have common origin or if they may have arisen many times during the evolution.

One of the important biological challenges is to find the ligands and functions for the orphan *Rhodopsins*. Three of the *Rhodopsins* identified in this thesis; GPR100, GPR119 and GPR120; have been deorphanised after we published them (Hirasawa et al., 2005; Liu et al., 2005; Soga et al., 2005). GPR119 and GPR120 have now received tremendous interest from the pharmaceutical industry because they have been indicated to be involved in the regulation of insulin release and could prove potential drug targets.

It is likely that many of the other orphan *Rhodopsins* will be deorphanised during the coming years because of the large capacity that the pharmaceutical industry has for searching for new ligands. Ligands are often the key to determining the functional role but much can also be learnt about their possible functions through detailed anatomical expression studies.
SAMMANFATTNING PÅ SVENSKA

G protein kopplade receptorer (GPCRer) utgör en av de största proteinfamiljerna i människa. Deras funktion är att förmedla signaler från cellernas utsida till insida och de har en framträdande roll i regleringen av de flesta fysiologiska processer i kroppen. Om GPCRerna inte fungerar kan sjukdomar uppkomma och många läkemedel (30-45%), och även droger som missbrukas, har sin effekt genom att påverka dem.

Artiklarna I-III i denna avhandling beskriver upptäckten av nya humana GPCRer. Vår grupp har kunnit visa att de nya GPCRerna troligen har en viktig funktion eftersom deras gener är uttryckta och har bevarats under en lång evolutionär tid. Vi har sökt ledtrådar om deras funktioner genom att studera i vilka vävnader och organ som de uttryckts, vilka tidigare kända receptorer de liknar och, för en del, även genom att identifiera konserverade domäner med känd funktion. Tre av de nya GPCRerna har fått sin ligand (aktiverande signal) bestämd av andra forskargrupper.

Människans och maskars GPCRer härstammar från en gemensam förfader, dvs. de är av mycket gammalt ursprung. Vissa familjer och grupper av GPCRer har bevarats mycket väl under evolutionen medan andra har förlorat eller erhållit nya medlemmar i olika arter vilket har gjort det möjligt att möta olika behov. Artikel IV beskriver en grupp av GPCRer som skiljer sig väldigt mycket mellan människa, gnagare (råtta och mus) och fiskar. Denna grupp är besläktad med kända amin-bindande receptorer (tex. de för serotonin, adrenalin och dopamin) och kallas ”Trace Amine-Associerade Receptorer”.

Artiklarna V-VI analyserar hela GPCR uppsättningarna i mus och rätta och jämför dem med människans. Denna sorts studier är viktiga eftersom rätter och möss är vanlig försökdjur och används för att förutsäga tex. ett läkemedels effekt på människor.
ACKNOWLEDGEMENTS

First, I would like to thank everyone in the groups of Helgi Schiöth, Dan Larhammar and Lars Orlæand at the Department of Neuroscience for friendship, inspiration and shared knowledge.

Especially, I thank:

Robert Fredriksson, for the uncountable excellent answers and ideas.

Helgi Schiöth, for the expertise and dynamic working environment which have given me the means to grow as a researcher.

Tobias Hill, for friendship and the much lively exchange of ideas.

Malin Lagerström, for the frequent exchange of opinions and tips.

Thora Bjarnadottir, for the much productive co-work.

Pär Höglund, for the many tips, including the tip to apply for conference funding from Smålands nation.

Jenny Tigerholm, Majd Mirza and Karl (Kalle) Nordström; for ambitious co-work, creative feedback and a lot of fun.

Tomas Larsson, for valuable discussions about phylogeny and evolution.

Anna Gloriam, for all the love and tenderness. With you I have discovered more of myself both as a researcher and person.
REFERENCES


Shashidhar S, et al. (2005) GPR56 is a GPCR that is overexpressed in gliomas and functions in tumor cell adhesion. *Oncogene* 24:1673-82.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 133

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”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-6745