UPTEC X07 032

Examensarbete 20 p
Mars 2007

The origin of the Adhesion family of G-protein coupled receptors – an evolutionary study

Linn Wallér

UPPSALA
UNIVERSITET

Bioinformatics Programme

Uppsala University School of Engineering

UPTEC X 07 032	**Date of issue 2007-03**

Author

Linn Wallér

Title (English)

The origin of the *Adhesion* family of G-protein coupled receptors – an evolutionary study

Title (Swedish)

Abstract

The G-protein coupled receptors (GPCRs) involve five families according to the GRAFS-classification system in which the *Adhesion* family was separated into a group of their own. Potential *Adhesion* sequences were found and assembled in several evolutionary distant species with the use of BLAST and BLAT. The sequences were subjected to subsequent phylogenetic studies with the intention of elucidating the history of the family. This included neighbor-joining, maximum parsimony and minimum evolution tree construction.

Keywords

Adhesion, GPCR, G-protein coupled receptors, phylogeny, *Tetraodon nigroviridis, Dictyostelium discoideum*, neighbour-joining, maximum parsimony, minimum evolution, MEGA, PHYLIP

Supervisors

Malin Lagerström and Helgi Schiöth
Department of Pharmacology, Uppsala University

Scientific reviewer

Mikael Thollesson
Department of Molecular Evolution Uppsala University

Project name	Sponsors
Language **English**	Security Secret until March 2008
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **45**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

The origin of the Adhesion family of G protein-coupled receptors – An evolutionary study

Linn Wallér

Sammanfattning

Adhesion-familjen är medlem av superfamiljen av membranbundna proteiner kallad G-protein kopplade receptorer (GPCRer). GPCRer har ett brett spann av såväl funktioner som ligander och är en av de mest studerade proteinfamiljerna inom läkemedelsforskningen. *Adhesion*-familjen särskiljs från övriga medlemmar av att de har exeptionellt långa aminosyra-sekvenser, som sträcker sig ut från cellmembranet, innehållandes en mängd domäner. GPCR proteolytic site (GPS) är en typisk domän för familjen liksom domäner som har med möjlig vidhäftning till andra celler eller proteiner att göra.

I den här studien försökte ursprunget till *Adhesion*-familjen urskiljas genom att först hitta och därefter studera repertoaren av *Adhesion*-sekvenser i en mängd arter, av olika evolutionär ursprung, med hjälp av fylogenetiska metoder. Metoder som användes var bland annat konstruktion av träd med hjälp av neighbor-joining, maximum parsimony och minimum evolution.

Examensarbete 20 p i Bioinformatikprogrammet
Uppsala universitet Mars 2007

Contents

1. Introduction

It has previously been proven that G-protein coupled receptors (GPCRs) are of ancient origin with members in both plants [2] and animals [3, 8-10]. This would mean that they evolved prior to the split leading to these two lineages, estimated to have occurred for about 850 million years ago [5]. In this study we aim at revealing a common history of the *Adhesion*-family, which is part of the GPCR superfamily. To find sequences with clear *Adhesion* affiliation, blastdatabases with members from all relevant GPCR families represented have been established and only sequences fulfilling certain criteria, explained in materials and methods, were selected for further analysis.

With the intention of revealing the history of the *Adhesion*-family, sequences from *Homo sapiens*, *Mus musculus, Gallus gallus, Tetraodon nigroviridis*, *Drosophila melanogaster, Caenorhabditis elegans, Dictyostelium discoideum, Monosiga brevicollis* and *Arabidopsis thaliana* were included. New sequences were found in *Tetraodon nigroviridis*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum.* The subsequent analyses were based on the seven transmembrane regions present in all GPCRs and trees were constructed with the methods neighbor-joining, maximum parsimony and minimum evolution in Phylip 3.65 and Mega 3.1. In order to keep the families and the receptors with the same name such as the *Secretin* family and the secretin receptor all families are denoted in italic and beginning with capital letters.

1.1 Background

1.1.1 The superfamily of G-protein coupled receptors

Guanine nucleotide-binding protein-coupled receptors or G-protein-coupled receptors (GPCRs) comprise one of the largest protein families in human [11], and new members are continuously being found [12, 13]. They are recognized by their seven hydrophobic α-helical transmembrane regions (7TM) with an extracellular N-terminal and an intracellular C-terminal [14]. The 7TMs are organised in a counterclockwise manner within the cellmembrane and has three loops on either side of it (fig 1) [15].

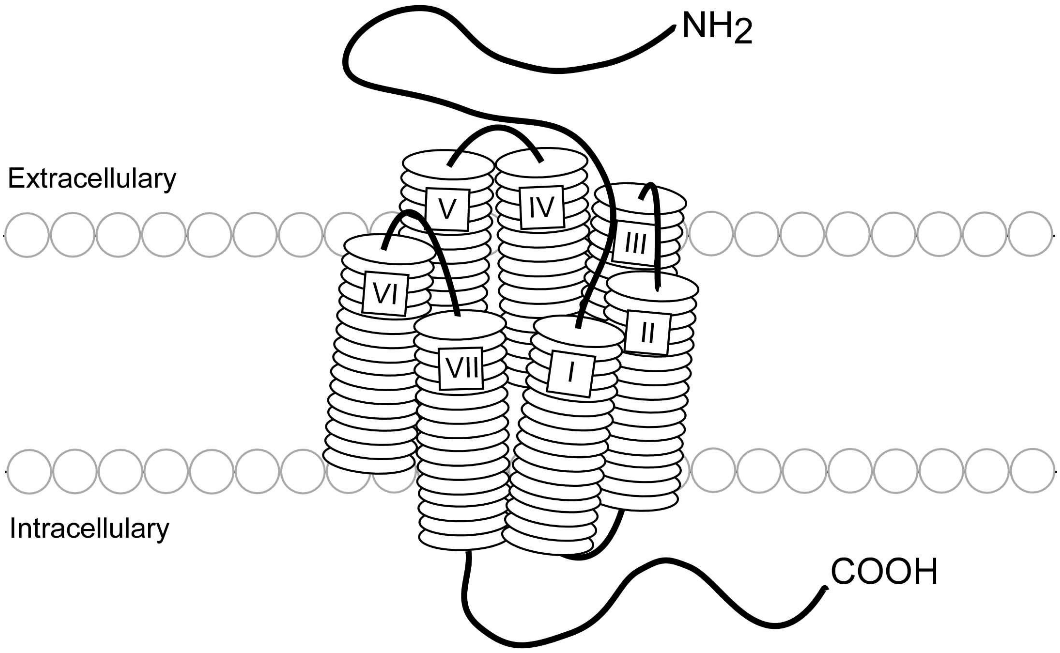


Figure 1. Schematic picture over the 7TM arranged in a counter-clockwise manner with C-terminal (COOH) intracellulary and N-terminal (NH_2) extracellulary.

The name GPCRs is a consequence of the coupling with G-proteins apparent for most of the members. Since not all GPCRs' intracellular response is obviously mediated through G-proteins other names such as serpentine-like receptors, 7-transmembrane receptors or heptahelical receptors have been used [16].

As sequences from several genomes are made publicly available and updated, GPCRs have been discovered in a variety of species ranging from mammals like human [12, 17] and mouse [3] to plants such as *Arabidopsis thaliana* [2]. This inclines that the GPCR superfamily is of ancient origin and since especially the 7TM domains are an overall present trait, has essential functions [12]. These functions are immensely diverse spanning from cellproliferation, brain angiogenesis and immune response to the ability to discern different flavours. In addition to the functionality, the GPCRs are capable of interacting with an immense span of ligands including; nucleosides, nucleotides, peptides, amines, amino acids, Ca^{2+}-ions, glycoproteins, phospholipids, prostanoids, fatty acids, bitter and sweet tastants, photons of light, pheromones and odorants

[18]. The vast functionality, the large number of ligands and the connection between human disease and dysfunctional GPCRs [18] contribute to the pharmaceutical interest in the family and is most likely the reason why the GPCR family is so well studied. Several of the present drugs target this family and others will come. The difficulty lies, amongst other things, in the ambiguity of the structure since merely one of the GPCR's crystal structure has been disclosed, the bovine rhodopsin [19].

Propositions have been made that the vast number of family members for GPCRs have evolved as a result of whole genome duplications (tetraploidizations) [20], but providing no common ancestor is revealed. Alternatively the family can likewise be a result of evolutionary convergence.

Previous attempts have been made with the prospect of unfolding a common evolutionary ancestor to the GPCR family [2, 14]. Parallel to these, studies with the intention of revealing internal relations and species-specific expansions, like *Methuselah* in insects [9] and pheromone receptors in rodents [3], have been made but the potential ancestor is still concealed. For the characterizations, methods like Psiblast [2], phylogenetic studies [12] and clustering [21] have been used, as well as similarities in receptor size taken together with the ligands interaction points [22] and high sequence similarity (>20%) within the TM regions [23]. These different courses of action have resulted in a few classification systems, the A to F system [23] the 1 to 5 system [22] and the GRAFS system [12]. The systems resemble each other but categorize potential GPCRs slightly different due to the different classification methods used and the available data at the time of the organization. In this study the GRAFS system will primarily be used given that it incorporate recently discovered GPCRs and has a relation to this study due to the fact that it is the first classification parting *Secretin* and *Adhesion* GPCRs.

The GRAFS system is constituted of five families *Glutamate* (G), *Rhodopsin* (R), *Adhesion* (A), *Frizzled/Taste2* (F) and *Secretin* (S) divided on the base of sequence similarity in the transmembrane regions [12, 14, 24]. The system is derived from human GPCRs but gives a good estimation of the possible separation in other species as well.

1.1.2 *Glutamate* (G)

The family is also called C [23] or 3 [22] and according to the GRAFS-system it is comprised of 22 receptors involving gamma aminobutyric acid receptors (GABA), taste receptors, metabotropic glutamate receptors, the calcium sensing receptor and a few orphan receptors [25]. The G-family also includes a vomeronasal receptor (V2R) which is especially apparent in rodents, where they have expanded the family with over 140 members [26]. This branch is probably left out from the human related GRAFS classification since mostly pseudogenes of the V2R [25] seems to remain.

Another group within the *Glutamate* family is the major excitatory glutamate neurotransmitter receptors in the central nervous system [18]. The G family binds its ligands in a cleavage between two lobes in the extracellular N-terminal, which by the time of binding undergo conformational changes leading to the enclosure of the ligand [27]. A parallel response to the conformational change is the consequential exposure of amino acid sequences, that can possibly act as a ligand, and thereby be able to interact with the extracellular loops of the 7TMs [18]. This in turn initiates subsequent alterations in the TM conformation and activates the receptor [27, 28].

1.1.3 *Rhodopsin* (R)

The *Rhodopsin* family is the largest of the GRAFS families with as many as 659 components in human [12, 29]. It is also referred to as family A [23] or 1 [22]. Since the relations of such an enormous group is difficult to disclose, further subdivision has been made into 13 divisions collected into four groups; α, β, γ and δ. 388 out of the 659 Rhodopsins are olfactory receptors [30] and are enclosed in the δ–group [29]. They are separated mostly due to the fact that they show extremely high similarity and the noticeable lack of introns [12, 30, 31]. The other three groups have been contrived through phylogenetic studies. The α-group is the largest and contains amongst other the amine receptors, the melatonin receptors, the prostaglandin receptors and the melanocortin/endoglin/cannabinoid/adenosine (MECA) receptor cluster. Group β includes 36 receptors which all have peptides as ligands and the γ–group involve the melanocyte-concentrating hormone (MCH) receptors, the somatostatin/opioid/galanin (SOG) receptor cluster and the chemochine receptor cluster [32].

The *Rhodopsin* receptors differ mostly from the other families in that most of them have short N-terminals and preferably bind their ligands within the 7TMs. They also have the ability to be activated through other approaches such as N-terminal binding domains, cleavage of the N-terminal with the remaining part bound to domains in the extracellular loops, or absorption of light [18].

The *Rhodopsins* are the only family with a crystallised structure of a member, the bovine rhodopsin [19]. It is also the most studied since most of the present drugs target the biogenic amine receptors within this family. Several diseases for instance Parkinson’s disease, dystonias, schizophrenia, drug addiction and mood disorders is connected to the signalling of monoamines through these receptors [33, 34].

1.1.4 *Frizzled/Taste2* (F)

In Kowakalskis characterization the *Frizzled* receptors were referred to the O-family (Other-family) but was rewarded a group of their own when one receptor was proven to couple with a G-protein [35]. They were discovered in *Drosophila melanogaster,* when searching for the responsible mutations for the disruption of polarity in epidermal cells [36, 37]. In mammals there exists 10 *Frizzled* and 1 *Smoothened* receptors [38, 39] which slightly resemble sequences from family B [22] consisting of *Secretin* and *Adhesion* [21, 40]. The N-terminal is cystein-rich forming disulfide bridges, shown to be important for the binding of their endogenous ligand Wnts [41]. Recently another ligand for the *Frizzled* family was revealed, indicating that Norrin, a secreted protein, is able to interact with the mouse FZD4 [42].

At present only inhibitors to the SMO receptor has been publicized [43]. It has however been shown that SMO has the ability to interact with $G_i\alpha$ in *Xenopus* melanophores [44]. The *Frizzled* receptors are seemingly well conserved between species [15], which is likely a result of their functions such as proliferation, control of cell fate and polarity [45].

1.1.5 *Secretin* (S)

The *Secretin* family was previously included in the 2- and B-family [22, 23], but was divided into a group of its own with the publication of the GRAFS system. As mentioned above the members are rich in cysteins in the N-terminal and the only group without any orphans [12]. They bind rather large peptide-ligands which interact with both the secondary structures in the N-terminal formed by the the cysteinbridges as well as the extracellular loops [46]. The interaction causes modifications in the intracellular regions and as a consequence the receptors are activated.

Within the family they share a highly conserved aspartic acid, situated in the connection with the second TM which is crucial for the recognition of its ligand and activation of the receptor [47]. That the *Secretin* family is of ancient origin is accentuated by the presence of the members in various species like *Takifugu rubripes*, *Danio rerio*, *Caenorhabditis elegans*, *Drosophila melanogaster* and even *Ciona intestinalis* [15].

1.1.6 *Adhesion* (A)

The *Adhesion* family is the second largest group of GPCRs with 33 human members and was recently separated from the family B/2 into their own group according to the GRAFS classification [12]. Prior, they have been shown individuality within the B/2 clade by the allotting of various names describing their peculiar topology. EGF-TM7 was used since EGF-module-containing mucin like hormone receptor 1 (Emr1), F4/80 and Cd97 was the first sequences of this family to be cloned and shared constituents for epidermal growth factor (EGF) and 7TM. Another name, LN-TM7 stressed the existence of the large N-terminal (LN) and the expansion to LNB-TM7, the connection to the *Secretin* family [48]. The *Adhesion* family also has some members which demonstrate a hormone binding domain that is also present in all *Secretin* receptors, and has conserved cystein residues in the first and second extracellular loops in common with several other GPCR families [9]. The long N-terminal forms a rigid structure sprawling out from the cell due to a number of mucin-like regions rich in serin and threonin [49]. This and the several domains in the extracellular terminal with connection to adhesion-like functions indicate that the function of the *Adhesion* family member might be to communicate with other cells, membrane proteins on other cells or proteins in the extracellular matrix [39, 48, 50]. The domains include among others epidermal growth factor (EGF), lectin, cadherin, olfactomedin, thrombospondin or immunoglobulin and are unique for the *Adhesion* [51]. The domains previously confirmed to be involved in cell communications are EGF which has one of the widest expression patterns in animals [11, 52, 53]. The protein module is involved in a range of physiological processes such as fibrinolysis, blood coagulation, neural development and cell adhesion [54]. The EGF-domain in Cd97 aids the protein in the binding process of CD55/DAF (Decay accelerating factor) [55] which is expressed on most leucocytes. An EGF-domain shared by both Cd97 and Emr2 has the ability to bind chondroitin sulphate, a glycosaminoglycan which is abundant on cellmembranes and in the extracellular matrix and most often involved in cell-interactions [56]. Possible ligands binding to EGF-domains in Emr2,3 and mouse Emr4 have also shown possible cell-to-cell communication [57]. The Ca^{2+}-dependent cell to cell adhesion domains, cadherines present in Celsr1-3, have been proven to have adhesion-functionality in epithelial cells [58]. The cadherines form cis-dimers on the own cell which are then combined with similar dimers from other cells in a trans-dimer mode [18]. The ligands mentioned previously together with transglutaminase2 (TG2) for Gpr56 are the only ligands found for the *Adhesion* family, the other receptors remain orphans.

Nonetheless various promising functions have been revealed; control of angiogenesis in the brain (Bai1-3), synaptic exocytose (Lec1-3), regulation of immune system (Cd97), definition of cell polarity and synaptogenesis (Celsr1-3) [38, 39]. In the Bai receptors, expressed in both brain and other tissues [59-62], motives with possible ability to act together with thrombospondin type

1 (TSP1) repeats and integrins have been found. Several proteins in the process of guidance cues directing neuronal axons during neuronal development, hold TSP1 [48]. The *Adhesions* are expressed in numerous tissues and cells in the immune system, in smooth muscle cells, hematopoietic cells, lymphocytes, myeloid cells etc [63]. The group was first believed to be involved in the immune system due to the vast expression in cells connected to the immune response. Cd97 is also most likely part of the immune system since activation of the receptor take place in inflammatory sites where it releases its N-terminal [39]. The cleavage is probably mediated by the presence of the GPS located extracellulary in the near proximity of the first TM [3]. The GPS is a trait characteristic for the *Adhesion* family members, although there are exceptions. The functionality of the GPS is still not totally clear but Krasnoperov and colleagues have shown that it is intracellulary cleaved in the primary parts of the golgi apparatus or in the endoplasmic reticulum, resulting in a separation of the N-terminal (NT) from the rest of the receptor (TMC). They argue that this may be a natural step in order to correctly fold the protein or with the purpose of accurately transport the protein to the membrane [64]. The N-terminal is then non-covalently bound to the TM regions [48, 65] but can be released as for Cd97 [39] or be used as a autocrine/paracrine regulator like for Gpr116/Ig-hepta which releases part of the N-terminal to control lung, kidney and heart [66]. Volynski and colleagues mean that the NT and TMC act independently on the plasma membrane where they individually function in signalling and cell-surface reception. They further claim that both parts can re-unite and bind ligands to the NT and thereby transduce signals via the TMC [67]. Even though the *Adhesion* family differ quite remarkably from the rest of the GPCRs it has been revealed that overexpressed Cd97, Emr1 and Gpr64 in *Xenopus* melanphores interact with G-proteins (G_s/G_q) (Jayawickreme C., through [39]). Lec1 also mediates signals through G-proteins ($G_o\alpha$) when bound with α-latrotoxin protein which is a constituent of the venom from the black widow spider [68].

The *Adhesion* family has previously been parted in eight groups (I-VIII) on the basis of the similarity in their 7TM-regions [3]. Group I – Lec1-3 and Etl, group II – Emr1-4 and Cd97, group III – Gpr123,124 and 125, group IV – Celsr1-3, group V – Gpr133 and 144, group VI – Gpr110, 111, 113, 115 and 116, group VII – Bai1-3 and group VIII Gpr56, 64/He6, 97, 112, 114, 126 and 128 (Bjarnadottir et al, 2004). Despite the fact that the division was based on the 7TM regions the receptors show common features in the N-terminals for each group.

The *Adhesion* family is a complex group of sequences which is hard to study as a result of their size and high number of exons. The complex processing steps, including the intracellular cleavage at the GPS, are also contributing factors to their complexity.

1.2 Species

1.2.1 *Tetraodon nigroviridis* (Tn)

Tn is a small freshwater, green spotted puffer fish of the teleost lineage which presently holds one of the smallest sequenced genomes for vertebrates. Even so it contains roughtly almost as many genes as the human genome and is a great model for the vertebrate system [69]. Metpally and colleagues have previously stated that the *Tetraodon nigroviridis* genome incorporates receptors from the *Glutamate, Rhodopsin, Frizzled, Secretin and Adhesion* families. 29 potential *Adhesion* genes have been found in *Tetraodon nigroviridis* under the criteria that they showed specific GPCR patterns and had a 7TM domain [69].

The teleost lineage sprung from the tree or life for 450 Million years ago (Mya) according to molecular studies [70, 71] but fossil records roughly estimates the divergence to have occurred for 410 Mya ago [1], see figure 3.

1.2.2 *Drosophila melanogaster* (Dm)

The fruitfly *Drosophila melanogaster*'s genome contains about 120 million base pair of which it is estimated that 98% have been covered according to Flybase (www.flybase.org). *Drosophila melanogaster* has evolved independently for 993 Mya according to molecular studies [70, 71] but the fossil records only show divergence of 530 Mya [1].

The *Drosophila melanogaster* genome is known to have at least four *Adhesion*-like genes with similarities to the Celsr-family, Gpr56 and Vlgr1 respectively [9]. In the same study nine putative *Methuselah* genes were discovered showing sequence similarities within the 7tm to both *Secretins* and *Adhesion* [9].

1.2.3 *Caenorhabditis elegans* (Ce)

The genome of the nematode *Caenorhabditis elegans* encloses approximately 100 million base pairs and has been assembled by the Wormbase project (www.wormbase.org). According to molecular studies the nematode lineage branched of for about 1177 Mya [70, 71]. Fossil records show a reduced number with 760 Mya [1].

Harmar claim that *Caenorhabditis elegans* has three potential *Adhesion* members which show resemblance to Celsr, Gpr56 and the groups I, II and VIII respectively according to phylogeny [9].

1.2.4 *Dictyostelium discoideum* (Dd)

Dictyostelium discoideum is a social amoebae with a AT-rich genome predicted to incorporate 12500 proteins [8, 72]. It has the ability to function in both unicellular and multicellular forms [8] and has become a superior model for cellular and developmental studies [72, 73].

Eichinger and colleagues have recently found 55 GPCRs in the *Dictyostelium discoideum* genome [72] of which one was a *Secretin*-like receptor, lacking a GPS but with 7TM regions most closely resembling that of the *Secretins*. This inclines that the *Secretin* and possible also the *Adhesion* family predates the divergence of animal and fungi [8].

Dictyostelium discoideum is a species that diverged before the divergence of animals, nonetheless it has been reported to have more than two EGF-repeats in a single gene. Up to 61 predicted genes have been found with EGF/Laminin domains [8]. The divergence of *Dictyostelium discoideum* is estimated to have occurred approximately the same time as the divergence of plants and before the split linking fungi and animals. However *Dictyostelium discoideum* show less of a evolutionary distance to human than human to yeast, partially due to the yeasts higher evolutionary rate [73].

1.2.5 Historical interesting sequences

In order to follow the assumption that GPCRs have a common ancestor, sequences from basal species were included in the study. A sequence from *Monosiga brevicollis* (Mb) was chosen due to its apparent connections to the *Adhesion* family [10] and the species position in the evolutionary tree as a choanoflagellate, likely to be an outgroup to the animal kingdom [74]. With the intention of covering the split linking opisthokonts and plants [75] a sequence from *Arabidopsis thaliana* (At), associated to GPCRs, was incorporated as well [2].

2. Materials and Methods

2.1 Sequence retrieval/assembly

In order to retrieve the most complete set possible, species specific methods were used and multiple verifications were conducted to ensure affiliation to the *Adhesion* family.

2.1.1 Human, mouse and chicken data retrieval

Sequences from human, mouse and chicken were downloaded from previously published articles [3, 24].Global RPS-blast at www.ncbi.nlm.nih.gov/BLAST/ was used against the conserved domain database (CDD) to identify the 7TM regions. Since there is no unique match for the *Adhesion* genes, TM regions for the *Secretin* family (7TM_2) were used to give guidance as to where to cut and additional alignments with ClustalW, were performed to confirm that the entire 7TM region had been collected. The full-length 7TM regions were later used as baits in the assembly of genes in the remaining species.

Human sequences from the other GPCR families were also collected, in the same manner as described previously, and used as a reference group to rule out false positives.

2.1.2 Identification and assembly of *Adhesion* genes in *Tetraodon nigroviridis*

The 33 human *Adhesion* GPCR sequences were used as baits. BLAT (BLAST local alignment tool) was used globally at http://genome.ucsc.edu/cgi-bin/hgBlat and the best hit from each region in the *Tetraodon nigroviridis* genome assembly Feb. 2004, was regarded as a potential *Adhesion* gene. To proceed with the unique hit there had to be an at least 80 bases long hit with sequence identity above 60%. The genomic sequence with additional 10000 bases down- and upstream the actual hit was collected and the gene was manually assembled. This was done by the usage of Editseq, a program part in the DNA Star package version 5.07 (DNASTAR, Madison, Wisconsin, United States) and an alignment search with the collected sequence and its bait using bl2seq with program tblastn at www.ncbi.nlm.nih.gov. The alternative matrices BLOSUM45, 62, 80 and PAM30, 70 where all used depending on which that gave the best coverage of the exons for the gene.

A manual inspection was then performed and emphasized on correctly spliced exons, that is the genomic sequence is manually searched for AG/GT directly up- and downstream the respective exons. When satisfactory boundaries had been found the exons also had to show a continuous sequence which preserved the primary structure of the protein; the frames had to be correct and not shifted. Occasionally exons were not discovered by bl2seq. Complementary searches were then made with the multiple alignment program, ClustalW version 1.8 at www.ebi.ac.uk/clustalw. The genomic part of interest was subsequently translated into the three possible frames and aligned to the exon of interest. The alignments were examined and the most satisfactory one, if any, was investigated to see if it held with correct reading frames and intron-exon boundaries.

The complete 7TM's were assembled and the corresponding protein was compared to the original human bait so that eventual false boundaries could be detected and corrected in as great extent as possible.

Complementary BLAST searches were executed locally. The unmasked *Tetraodon nigroviridis* genome version 7.42 was downloaded chromosome-wise from www.ensembl.org/info/data/download.html. An in-house program in Python translated the sequences into all six reading frames and used BLAST to perform a search with the tblastn method, which utilize a protein sequence against a translated database. Tblastn is used since the human baits are in protein form. The result was ridded of redundancy meaning multiple hits from the same genomic location, and areas corresponding to previously found genes were discarded. This gave additional hits of less obvious matches from new areas of the genome. They where handled and assembled in the same manner as mentioned above. Finally global BLAST searches were made at www.ensembl.org/Tetraodon_nigroviridis/blastview/BLA_XESEl8aNn, with matrix BLOSUM62 and otherwise default settings, revealing additional hits from unlocalized areas of the genome.

To rule out all possibilities that some genes had been overlooked, an additional scan in BLAT and BLAST was done with in-house data from the close relative *Takifugu rubripes* (Fugu). Comparisons to putative *Adhesion* genes mentioned in Metpally’s article [69] were also carried out to ensure complete coverage. When all possible methods to find putative *Adhesion*-genes had been exhausted a complementary hmm study was carried out. An hmm-model was built based on the 7TM of *Adhesion*-sequences from *Homo Sapiens, Mus musculus* and *Gallus gallus.* The successive hmm-searches confirmed presence of 7TM in all sequences.

2.1.3 Identification of *Adhesion* genes in *Drosophila melanogaster* and *Caenorhabditis elegans*

Human *Adhesion* genes were used as baits against the proteome of each species since both of them are well studied model organisms and a number of gene-predictions are available. For this purpose global BLAST was used at www.ncbi.nlm.nih.gov with blastp and target species set to *Caenorhabditis elegans* and *Drosophila melanogaster* respectively. All hits regardless of E-value were collected and only obvious non-*Adhesion* targets were removed, i.e. those annotated as a proteins not corresponding to GPCRs. The same was applied for hits with rps-blast results, against the CDD, with high scores for completely GPCR-unrelated domains. The remaining were cut according to the 7TM_2 in rps-BLAST run against the CDD. If no domain was found the protein was shortened afterwards, according to the alignment of all species-specific genes that had been found, including only the 7TM_2 in the alignment.

To rule out all non-*Adhesion* genes, an alignment with all human *Adhesion* genes as well as a few *Secretin* members, as an outgroup, was conducted. All putative hits that did not cluster within the *Adhesion* group were ignored during further studies. The remaining hits were manually inspected, with focus on splice sites, in the same manner as described previously. Obtained hits were then used as baits against the respective species' genomes.

In *Drosophila melanogaster* a complementary study was carried out with focus on *Methuselah* genes. As baits sequences from Harmar [9] were used and the same methods described previously were performed.

2.1.4 Identification of *Adhesion* genes in *Dictyostelium discoideum*

For *Dictyostelium discoideum* the same approach as for fruitfly and nematode was used with the exception that all hits were kept until further trials with stricter criteria described hereafter. The *Dictyostelium discoideum* sequences is of great interest since they represent the most basal species involved in this study and giving the fact that it has the most divergent genome the criteria for several method has been more or less compromised. Whenever the criteria have been meddled with, it is mentioned in the method in question. For instance in the in-house program where the potential *Adhesion* sequences has to have the first three hits as *Adhesion* as well as an overall of five out of ten *Adhesion* hits.

2.2 Purge of initial set with further verification tests

All previously inspected sequences were collected into a file (initial set) with the entire GPCR repertoire from all other GPCR families in human (in-house dataset). A temporary neighbor-joining (NJ) tree was constructed using the PHYLIP software version 3.65 with a bootstrap of 100, methods described later. All sequences with an obvious relation to another family than *Adhesion* were removed. Thereafter the family relations were further scrutinized with an in-house program described below. All human GPCRs together with additional sequences from *Drosophila melanogaster* and *Dictyostelium discoideum*, covering the *Methuselah* [9] and the *cAMP* family [8] were gathered and used as input together with the initial set. The program transformed the human GPCRs, *Methuselah* and *cAMP* into a blastdatabase with the formatdb command from the blast package and the initial set was searched against it. All sequence names were converted into the family to which it belonged according to the GRAFS classification. This was done with the intention of getting a clearer overview of the belonging of each putative *Adhesion* sequence. In order for a sequence to be kept regarded as a potential *Adhesion* sequence the first 3 and an overall of 5 out of the first ten hits had to be *Adhesion*.

The sequences from the initial set that satisfied this criteria were then subjected to phylogenetic analysis. Since the set did not comprise a stable set, the phylogenetic trees did not display a consistent topology. To cope with this a supplementary clustering analysis was performed, described hereafter.

2.3 KalleClust

With the aim of moderating a stable set, an in-house program (KalleClust) using an ISOdata method of clustering was applied. The clustering is dependent of sequence similarities between all possible combinations of genes. Sequence similarities were calculated with respect to pairwise global alignments using the Needleman-Wunsch algorithm and then normalized according to length. If sequences demonstrated similar distance behaviour to all others they were clustered. The clustering was done 1000 times with the intention of revealing the consistency of each group. Clusters were considered to be stable if its members belonged to it in 75% of the cases.

Promiscuous sequences that did not fulfil the criteria, that is, kept changing clusters, were removed from the set and phylogenetic studies were initiated.

2.4 Phylogenetic analysis

The set that was produced as described earlier was put through a quantity of phylogenetic scrutiny. With ClustalW version 1.83 the sequences were aligned after which the Phylip package version 3.65 and MEGA version 3.1 were used.

2.4.1 ClustalW

The multiple alignments produced by ClustalW are based on a distance matrix formed by pairs of aligned sequences. The matrix is then used to form an initial neighbor-joining tree on which the subsequent multiple alignment rely. The multiple alignments are hereafter constructed by aligning the closest sequences which are then treated as one when the remaining sequences are added one by one. This also happens to be the downside of ClustalW since an initial error in the neighbor-joining tree will propagate in the entire alignment. Otherwise the program has several methods to construct the best alignment possible. Different weight matrices including both PAM and BLOSUM variants are used depending on how closely the sequences are related. This is intended to avoid dominance of strongly related sequences. The gap penalties also differ depending on sequence and position giving a lower penalty for areas where gaps are prominent and for hydrophilic areas, which tend to be loops. Sequences of different lengths and similarities are hereby aligned in an advantageous mode [76]. To get an adequate format to continue with the phylip format had to be activated under output format. The multiple alignments were conducted in a slow/accurate manner [77].

2.4.2 Phylip 3.65

To get a desired number of replicas the Phylip program seqboot was used and subsequently protdist and parsprot were used respectively to produce neighbor joining and maximum parsimony trees.

2.4.3 MEGA 3.1

MEGA functions in a similar way as Phylip with a separate alignmentmodule using, amongst other alignment methods, clustalw. Then either direct phylogenetic trees or bootstrap can be chosen. The file returned from ClustalW to MEGA might have to be controlled for the success of subsequent analysis. All sequences are then of equal length and any gaps, that is space or similar, will result in error. To cope with this the gaps have to be replaced with (-) and the file re-saved.

2.4.4 Seqboot

The .phy file was conveyed to the seqboot program in Phylip which resamples the input data set into multiple data sets. In my study I used molecular sequences which were bootstrapped 100 times by regular sampling fraction. The other settings were left default, meaning that no weights of characters or categories of sites were set.

Seqboot produces replicas of the initial dataset by small alterations of the initial set. Assuming that the characters evolve independently possible alterations are deletions and duplications which finally forms sequences of equally lengths as the originals [78]. These sets can then be used to calculate bootstrap values for the branches in the tree, which is a measure of the support for each branch or clade.

2.4.5 Protdist/Neighbor-joining

The output from seqboot was then passed on to protdist which uses the sequences to calculate a distance matrix [78]. In our case multiple datasets were analyzed resulting in 100 distance matrices calculated using the Jones-Taylor-Thornton (JTT) matrix for amino acid replacement. The JTT is a revised version of the Dayhoff PAM matrix based on a larger dataset than the one used by Dayhoff [79].

In the protdist program in Phylip one category of substitution rates were used and since we did not have the alpha-parameter which is needed to calculate the coefficient of variation of substitution rate among positions, the gamma distribution rates among positions were not selected. Otherwise the remaining settings were kept default with the same values as for seqboot.

The distance matrices were then used to construct neighbor-joining trees with the neigbor program in Phylip. The only setting changed was: analyze multiple data sets, which was set to 100 corresponding to the number of replicas chosen in seqboot. The remaining settings were kept as default. The neighbor-joining method starts with all nodes sprung from one node. Internal branches are then introduced between pairs with the shortest distance. In each step the tree length is recalculated and finally the pairs are connected and the tree with the minimum length of internal branches is presented as an output tree [76].

In MEGA 3.1 the analysis was executed with phylogeny reconstruction, all substitutions were included and the JTT substitution model was utilized. Equally to the Phylip trees a bootstrap value of 100 was applied.

2.4.6 Protpars/Maximum parsimony

The seqboot file was also used as input in the program protpars in Phylip, which calculates maximum parsimony trees with a method that is a compromise between the methods used by Eck and Dayhoff, 1966 [80] and Fitch, 1971 [81]. The first method permits all amino acids to be replaced by all others counting the amount of changes needed, which however is not possible with regard to the genetic code. The other method constructed by Fitch counts the number of nucleotide substitutions needed, including substitutions which do not change the aminoacid. Protpars resembles Fitch method in that it is consistent with the genetic code, however it also allows intermediate steps required to attain a specific amino acid. The program assumes that separate sites and lineages are independent, that changes between branches are relatively global throughout the entire tree and that synonymous changes have a higher probability than nonsynonymous [78].

Maximum parsimony is a discrete character method that rearranges an initial tree to the tree which requires the minimum amount of mutations. This is done repeatedly with different initial trees and the tree with the least amount of mutations is finally chosen [76].

Protpars was set to search for the best tree with ordinary parsimony, the genetic code was maintained Universal and the remaining settings kept default.

MEGA has developed an algorithm of their own for maximum parsimony which uses a heuristic approach for large number of sequences and otherwise uses branch and bound. Branch and bound (BaB) investigate all possible trees but instantly rejects trees with clearly longer lengths. The initial state of this method is a three leaves tree conformed by the sequences with highest diversity. A tree with three leaves can only be combined in one topology. The additional leaves are then adjoined under the criteria that the minimum-length tree is aquired. The heuristic approach resembles BaB but inspect fewer trees [82].

The analysis was conducted with the same settings as for NJ. The MP search options were set to one level of close-neighbor-interchange (CNI) and the initial trees to random addition trees with 10 replications.

2.4.7 Minimum Evolution

Minimum evolution uses pair-wise distances to calculate scores between sequences. The method assumes that all possible pairs are possible and calculate the branchlength for all of them. The lengths of the branches can be approximated by different methods; one that has been used is the Fitch and Margoliash’s method [83]. ME resembles maximum parsimony in that it creates a number of initial trees and swaps the branches to get the shortest distance. The returned tree is then the one with the minimum sum of branch length [82].

In this study the minimum evolution was used with a bootstrap of 100 and with phylogeny reconstruction. The initial tree was constructed by neighbor-joining, maximum number of trees set to one and for the consistency of the study; JTT was used as substitution model.

2.4.8 Consensus

For each method in Phylip that is protpars and protdist the outfile gave 100 trees since seqboot had been set to produce 100 replicas. In order to combine these and to get bootstrap values of the branches the consensus program was used, applying the majority rule of consensus. The trees were treated as unrooted and no specific outgroup was selected. The resulting trees were depicted in TreeView (Win32) version 1.6.6 where the internal edge labels were set to be shown. The trees were then arranged using Canvas 8.0.2 to make the view clearer.

2.5 Re-insertion of sequences

To progress, the sequences removed from the stable set according to KalleClust, were re-installed one by one, on condition that no major rearrangement of the subsequent neighbor-joining tree occurred. When a stable tree involving the most sequences had been established further inspections were performed. In Phylip both a neighbor-joining and a maximum parsimony tree was produced in the same manner as mentioned above and in MEGA version 3.1 one of each was also created see figure 7-10. An additional tree with the method minimum evolution (ME) was also created (Fig 6).

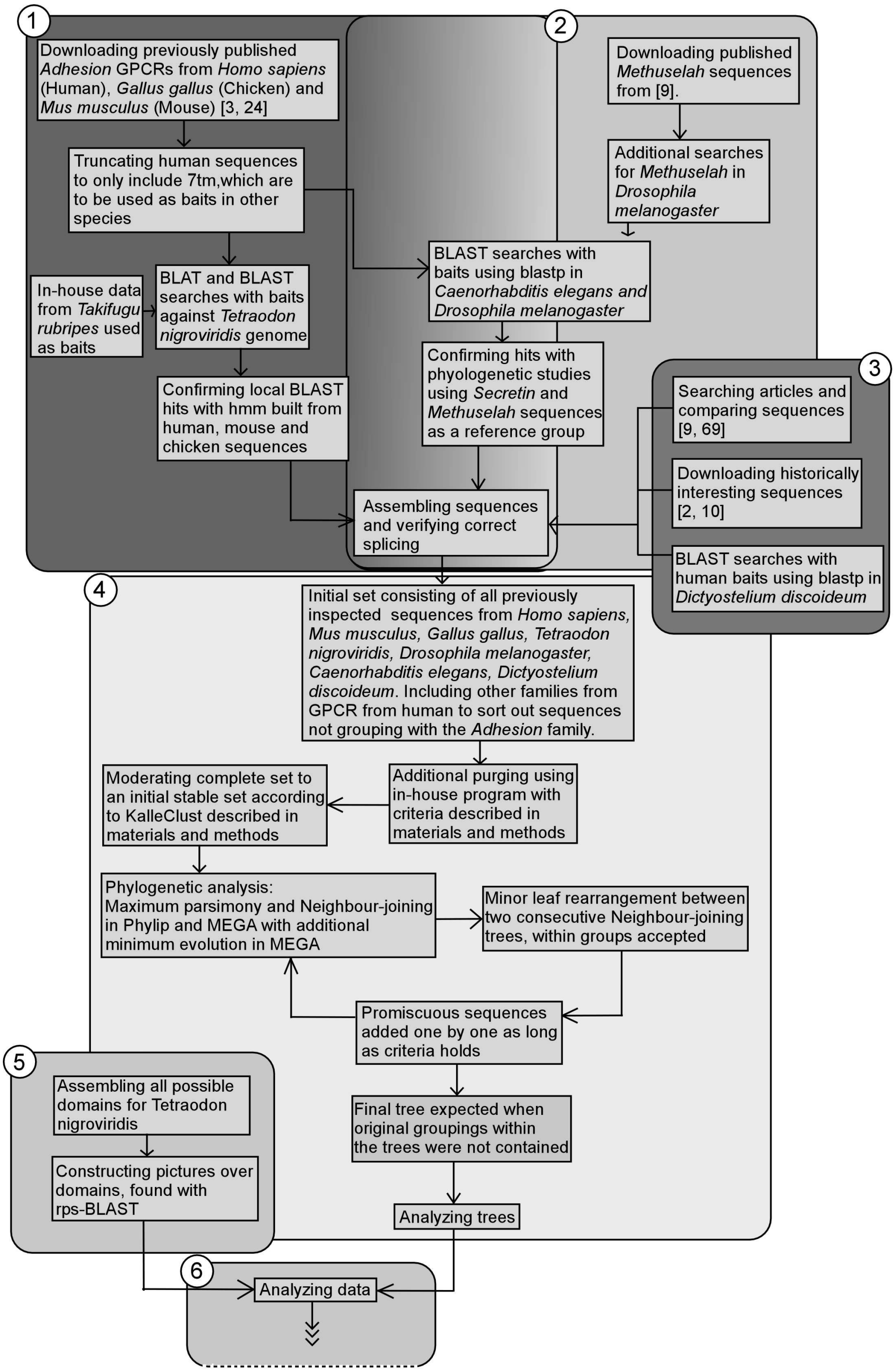
2.6 Domain search

All sequences included in the consensus trees (fig 5-9) were searched for domains in their N-terminal with rps-BLAST with CDD – 12589 PSSMs, at www.ncbi.nlm.nih.gov/BLAST. For *Tetraodon nigroviridis* the domains had to be assembled in the same manner as for the 7TMs but with use of the corresponding full-length human sequence. When no continuous extracellular terminal could be found, another human sequence related to the original bait was used. For the remaining species, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum* the full-length genepredictions acquired through the search with the 7TM were inspected for correct splicesites. With the intention of revealing distantly related domains a high cutoff value, 0.1, for accepting domains was set. An additional control of the domains was conducted using InterProScan at www.ebi.ac.uk/InterProScan. The advantage of InterProScan is its combination of several signature-databases to a nonredundant characterization of family relations, protein domains and functional sites. Integrated databases are PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR superfamily, SUPERFAMILY, Gene3D and Panther [85]. The domains for relevant species are depicted using an alteration of the figure made by Bjarnadottir [3] (fig 11). The figure only displays the domains found with rps-blast since

InterproScan is based on hmm-searches [85] and thereby not as stringent as rps-blast. Rps-blast is calibrated against the number of domain presently in the database and the amount is constantly growing with the downside that previous found domains can be lost in later versions of the database [86].

3 RESULTS

Previous studies have been made to characterize all *Adhesion* GPCR in human, mouse and chicken resulting in a set of 33 human, 31 mouse and 22 chicken *Adhesion*-genes [3, 24]. Since human enclose the most extensive set, these were used to explore the *Adhesion* repertoire in *Tetraodon nigroviridis* (Tn), *Drosophila melanogaster* (Dm), *Caenorhabditis elegans* (Ce) and *Dictyostelium discoideum* (Dd).



1
Downloading previously published Adhesion GPCRs from Homo sapiens (Human), Gallus gallus (Chicken) and Mus musculus (Mouse) [3, 24]
Truncating human sequences to only include 7tm,which are to be used as baits in other species
In-house data from Takifugu rubripes used as baits
BLAT and BLAST searches with baits against Tetraodon nigroviridis genome
Confirming local BLAST hits with hmm built from human, mouse and chicken sequences
2
Downloading published Methuselah sequences from [9].
Additional searches for Methuselah in Drosophila melanogaster
BLAST searches with baits using blastp in Caenorhabditis elegans and Drosophila melanogaster
Confirming hits with phyologenetic studies using Secretin and Methuselah sequences as a reference group
Assembling sequences and verifying correct splicing
3
Searching articles and comparing sequences [9, 69]
Downloading historically interesting sequences [2, 10]
BLAST searches with human baits using blastp in Dictyostelium discoideum
4
Initial set consisting of all previously inspected sequences from Homo sapiens, Mus musculus, Gallus gallus, Tetraodon nigroviridis, Drosophila melanogaster, Caenorhabditis elegans, Dictyostelium discoideum. Including other families from GPCR from human to sort out sequences not grouping with the Adhesion family.
Additional purging using in-house program with criteria described in materials and methods
Moderating complete set to an initial stable set according to KalleClust described in materials and methods
Phylogenetic analysis: Maximum parsimony and Neighbour-joining in Phylip and MEGA with additional minimum evolution in MEGA
Minor leaf rearrangement between two consecutive Neighbour-joining trees, within groups accepted
Promiscuous sequences added one by one as long as criteria holds
Final tree expected when original groupings within the trees were not contained
Analyzing trees
5
Assembling all possible domains for Tetraodon nigroviridis
Constructing pictures over domains, found with rps-BLAST
6
Analyzing data

Figure 2. Flowchart of the analysis routine used in this study. In common for (1) and (2) are the assembling of sequences and the verification of correct splicing.

All of the included species are included in an evolutionary tree depicted in figure 3 and all steps performed during the study are represented in figure 2 in a flowchart.

The *Adhesion* sequences are known to have bulky N-termini with various domains and lengths. They also display the characteristic 7TM which is apparent throughout the entire superfamily of GPCRs. As a result of this the human sequences were truncated to only involve the 7TM which were then used as baits. Different approaches were taken in order to find the most complete set in each species. As seen in the flowchart (Fig 2 box 1), in *Tetraodon nigroviridis* the entire genome was screened with the human baits using BLAT at UCSC (http://genome.ucsc.edu) and both local, using the BLAST package, and global BLAST at ensembl’s homepage (http://www.ensembl.org). All sequences were named after the bait which had discovered them. The BLAT search resulted in 24 potential hits including 7 possible pseudo genes with missing exons or interrupted sequences. The local BLAST gave an additional 11 of which 7 were potential pseudo genes and the final global BLAST gave 1 further hit and 3 possible pseudo genes. The data set was then matched with the gene predictions from Metpally [69] giving two more potential sequences resulting in a total of 24 possible *Adhesion* genes (fig 2 box 3), see all potential *Tetraodon nigroviridis* genes, including pseudo genes in table I. These genes have a percent sequence identity with their human counterpart ranging from 25% between TnGPR56-1 and HsGpr112 to 92.2% between TnBai3-1 and HsBai3. (Table II) *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum* were all searched against their proteome (fig 2 box 2) with global BLAST searches (http://www.ncbi.nlm.nih.gov/BLAST) resulting in a first set of 14 *Drosophila melanogaster*, 43 *Caenorhabditis elegans* and 63 *Dictyostelium discoideum*. The sequences from *Drosophila*

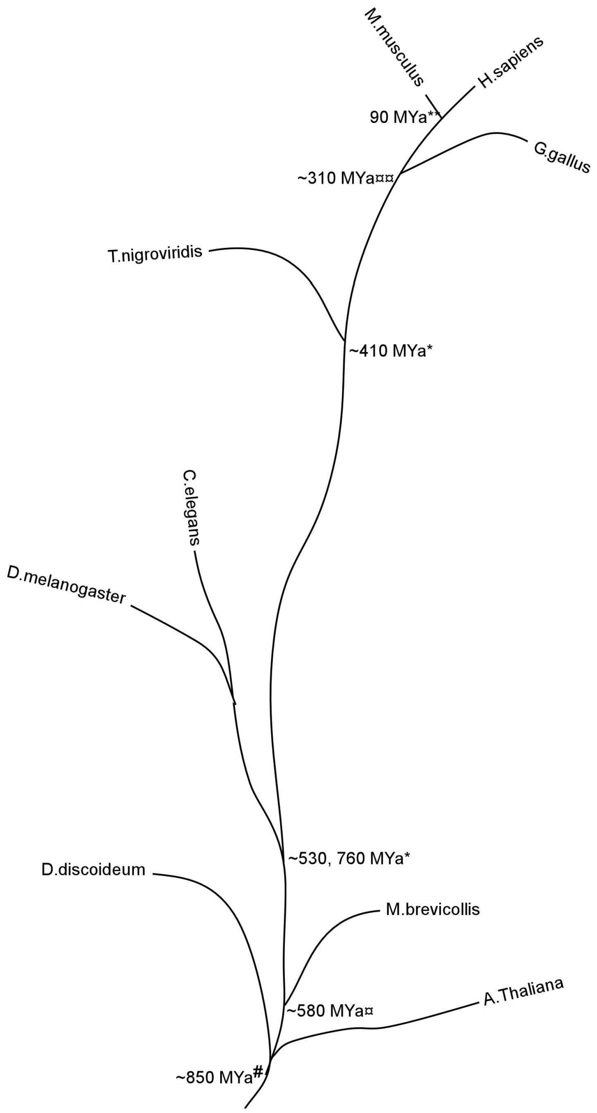


Figure 3. Evolutionary tree of the species included in the study. The numbers at the nodes are potential divergence of the different lineages according to [1]* , [4]**, [5]#, [6]¤ (unicellular choanoflagellates, to which *Monosiga brevicollis* belong, are thought to have evolved during the Ediacaran era) and [7]¤¤.

melanogaster are represented as sDm in order to separate them from the *Methuselah* sequences found in *Drosophila melanogaster*.

Table I: All potential *Adhesion* family genes in *Tetraodon nigroviridis* with unique position in the genome, comment on the quality of the sequences and method with which the sequence was found. BLAT – global BLAT at genome.ucsc.edu/cgi-bin/hgBlat, L BLAST – Local BLAST, LsBLAST – Local BLAST with stricter criteria for position, G BLAST – Global BLAST at www.ncbi.nlm.nih.gov and Metpally – Additional sequences from Metpally and colleagues [69], which did not correspond to any previously found sequences or genome positions

Name	Chr	Start	End	Strand	Method	Full 7tm	Comment
Lec3-1	Un_random	111730955	111748909	+	BLAT	Yes	
Lec3-2	Un_random	56103152	56110343	-	BLAT	No	Missing one aminoacid between exon 3 and 4 leading to break in readingframe
Lec3-3	15_random	412641	417917	-	BLAT	Yes	
Lec2-1	3	11035306	11036255	-	BLAT	No	Stopcodon in one of the last exons
Lec1-1	1	12771231	12779118	+	BLAT	Yes	Gives alternativ pseudogene supported by Halibut – *Hippoglossus hippoglossus*
Bai2-1	21	3520150	3529594	+	BLAT	Yes	
Bai3-1	17	4009410	4017472	-	BLAT	Yes	
Bai3-2	Un_random	44010925	44018913	-	BLAT	No	First exon missing
Bai3-3	21_random	2504735	2505956	-	BLAT	Yes	Should be named BAI1
Celsr1-1	Un_random	55665384	55666281	+	BLAT	Yes	
Celsr1-2	9	735327	736543	+	BLAT	No	4th and last exon missing and readingframe abrupted between 2 and 3 exon
Celsr2-1	11	5867648	5868423	+	BLAT	Yes	
Etl-1	1	12622653	12623894	-	BLAT	Yes	
Vlgr1-1	12	1220514	1222476	+	BLAT	No	Missing first two exons
Tr32-1	3	11146974	11148311	+	BLAT	Yes	Found with *Takifugu rubripes* sequence similar to the EGF-like group
Gpr112-1	Un_random	106929525	106929794	-	BLAT	No	Abrupted frame before the last exon
Tr3-1	Un_random	106928857	106930026	-	BLAT	No	*Takifugu rubripes* sequences similar to human used since they are more closely related, abrupted frame
Gpr112-2	7	9449462	9449731	-	BLAT	Yes	Missing exon 3 but that is consistent with the complementary sequence in *Takifugu rubripes*
Tr3-2	7	9448792	9449963	+	BLAT	No	Abrupted frame and missing third exon
Gpr112-3	1	8843075	8843344	-	BLAT	Yes	
Tr3-3	1	8843078	8843344	-	BLAT	No	Abrupted reading frame
Gpr123-1	Un_random	138354745	138378097	-	BLAT	Yes	
Gpr123-2	Un_random	36439731	36443805	+	BLAT	Yes	Two possible endings one with extra exon
Gpr125-1	Un_random	85083358	85084807	-	BLAT	Yes	
Gpr126-1	Un_random	125937925	125938906	-	BLAT	Yes	
Gpr126-3	6	5895824	5896634	+	BLAT	Yes	
Gpr64-1	7	4123467	4123949	-	BLAT	No	Abrupted reading frame before last exon
Gpr133-1	1	14136298	14136002	-	L BLAST	No	Only two exons found out of eight
Gpr133-2	15	2234757	2234912	+	L BLAST	No	Only one exon found
Gpr133-3	15	773411	773563	+	L BLAST	No	Only one exon found
Gpr133-4	2	15734263	15734418	+	L BLAST	No	Only one exon found
Gpr133-5	2	16656252	16655956	-	L BLAST	No	Only one exon found
Gpr116-1	14	822960	823733	+	L BLAST	Yes	
Gpr116-2	14	901466	901849	+	L BLAST	Yes	Also gives an alternative hit but with abrupted reading frame
Lec1-2	18	1607940	1608167	+	L BLAST	Yes	
Cd97-1	3	8243660	8243887	+	L BLAST	Yes	
Gpr123-3	3	9018722	9018862	+	L BLAST	No	Only one exon found
Mus_Gpr133-1	6	3065054	3065209	+	L BLAST	No	Only one exon found, Mm as bait
GgCelsr3	9	737088	737210	+	LsBLAST	No	Repeted abrupted readingframe, Gallus gallus as bait
GgGpr144	1	14133667	14133548	-	LsBLAST	No	Missing exons and abrupted reading frame. Gallus gallus as bait
Gpr113-1	14	908919	909674	+	LsBLAST	No	Abrupted readingframe
Gpr126-4	Un_random	63631975	63632331	+	G BLAST	No	Only two exons found
Gpr133-6	Un_random	101328430	101647563	-	G BLAST	No	Missing exons
Gpr144-1	Un_random	19867525	19867725	-	G BLAST	Yes	
Lec3-4	Un_random	69157968	69358039	-	G BLAST	No	Only two exons found
Gpr124-1	Un_random	85742430	85749972	+	Metpally	Yes	Manual inspection of geneprediction
Gpr56-1	Un_random	15593733	15600794	-	Metpally	Yes	Manual inspection of geneprediction
Gpr97-1	Un_random	15571743	15573940	-	Metpally	Yes	Manual inspection of geneprediction

The sequences from *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum* were reduced to 5, 5 and 21 respectively by removing:

1. Full-length sequences that obviously did not confirm any domains within the *Adhesion*, *Secretin* or *Methuselah* families
2. Sequences that joined other GPCR families, than the ones just mentioned, in temporary NJ-trees were a few members from all GPCR families were represented (fig 2 box 4).

The remaining putative family B sequences were merged into a large startfile for further confirmation of *Adhesion* affiliations. Alongside the investigation of the *Adhesion* repertoire in *Drosophila melanogaster* a complementary study of the *Methuselah* repertoire took place revealing 7 additional sequences to Harmars formerly found 9 [9] which were also included in the startfile for further confirmation (fig 2 box 2). The methods used were identical to those previously used to find potential *Adhesion* sequences.

In the pursue of the Adhesion family's origin two sequences found in *Monosiga brevicollis* (Mb) and *Arabidopsis thaliana* (At) discovered by King and colleagues [10] and Josefsson and colleagues [2] respectively, were included in the study (fig 2 box 3). Both had previously shown resemblance to the *Secretin* family, or adhesion like functions and were therefore considered to be of most importance. The startfile was subjected to an in-house program removing sequences not fulfilling the specific criteria described in materials and methods. The outcome of the program showed that all of the *Tetraodon nigroviridis*, 5 of 5 *Drosophila melanogaster* and 4 of 5 *Caenorhabditis elegans* were indeed *Adhesion*-like (table II).

Dictyostelium discoideum was to be handled slightly different while two of the sequences Dd16 and Dd38 displayed a strong connection to the *Adhesion* but without fulfilling the criteria. These genes only satisfied one of the criteria each and were chosen under slightly less strict criteria; that the first two hits had to be *Adhesion* and that they demonstrated the typical 7tm_2 domain. Dd1, which fulfilled the original criteria, Dd16 and Dd38 were included in the startset based on their historical importance. The sequence from *Monosiga brevicollis*[10] revealed a high similarity to the *Adhesions* whereas the sequence from *Arabidopsis thaliana*[2] rather grouped with the *cAMP* receptors and was therefore removed from further studies. Verification on the potential Methuselah sequences was also given see table II.

Table II: The ten first hit from in-house program targeting potential *Adhesion* genes against BLAST database containing members from all Human GPCR families as well as *Methuselah* sequences from Harmer [9] and *cAMP* sequences from Eichinger and colleagues [8, 72, 91]. Sequences marked in bold and italic fulfilled established criteria (see materials and methods for further information) and were selected for further investigations. The abbreviations stand for: Mth – *Methuselah*, Adh – *Adhesion*, Rho – *Rhodopsin*, Fz – *Frizzled*, Tas – *Taste2*, Sec – *Secretin*, cA – *cAMP* and Oth – *Other*.

In-house search against BLAST database

Name	1st hit	2nd hit	3rd hit	4th hit	5th hit	6th hit	7th hit	8th hit	9th hit	10th hit
Dm_mth1	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh	Adh
Dm_mth2	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh	Mth	Adh
Dm_mth3	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh
Dm_mth4	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Rho	Rho
Dm_mth5	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Rho
Dm_mth6	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh
Dm_mth7	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh
Dm_mth8	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh
Dm_mth9	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh
Dm_10	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Adh
Dm_11	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Mth	Mth	Adh
Dm_12	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Mth	Adh	cA
Dm_13	Mth	Mth	Mth	Mth	Mth	Mth	Adh	Adh	Adh	Adh
Dm_14	Mth	Mth	Mth	Mth	Mth	Adh	Mth	Adh	Mth	Adh
Dm_15	Mth	Mth	Mth	Mth	Mth	Mth	Mth	Fz	Fz	Fz
Dm_16	Mth	Mth	Mth	Mth	Adh	cA	Mth	Adh	Adh	Adh
sDm_1	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
sDm_2	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
sDm_6	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Rho***	***Adh***
sDm_8	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
sDm_9	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
Ce1	Rho	Mth	Adh	Adh	Adh	Rho	Adh	Tas	Adh	Rho
Ce2	Adh	Adh	Adh	Mth	Adh	Mth	Mth	Adh	Mth	Adh
Ce3	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
Ce4	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***
Ce5	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Oth***	***Mth***	***Adh***	***Sec***
Dd_1	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***cA***	***cA***	***Adh***	***cA***
Dd_6	Adh	cA	Fz	Fz	Adh	Adh	Adh	Adh	Adh	Fz
Dd_7	Adh	Fz	Mth	Adh	Adh	Mth	Mth	cA	Adh	Mth
Dd_8	cA	cA	Adh	Mth	Adh	Rho	Mth	Mth	Oth	Sec
Dd_10	cA	Adh	Fz	Rho	Rho	Rho	Adh	Rho	Fz	Fz
Dd_11	cA	cA	Rho	Fz	Rho	Rho	Rho	Rho	Rho	Fz
Dd_16	***Adh***	***Adh***	***Adh***	***Rho***	***Rho***	***Fz***	***cA***	***Rho***	***Oth***	***Adh***
Dd_17	cA	cA	cA	cA	cA	cA	cA	Adh	Sec	cA
Dd_18	cA	cA	cA	cA	cA	cA	Adh	Oth	cA	cA
Dd_19	cA	cA	cA	cA	Sec	Adh	Adh	cA	Sec	cA
Dd_20	cA	cA	cA	cA	cA	cA	cA	Oth	Adh	Mth
Dd_22	cA	cA	cA	cA	cA	cA	Oth	cA	cA	Rho
Dd_23	Adh	Sec	cA	Adh	Rho	Adh	Rho	Adh	Rho	Rho
Dd_25	cA	cA	cA	cA	cA	cA	cA	Oth	Oth	Adh
Dd_33	Fz	Rho	Rho	Adh	Adh	Rho	Adh	Rho	cA	cA
Dd_34	Fz	Fz	Fz	Fz	Fz	Adh	Adh	Adh	Adh	Adh
Dd_38	***Adh***	***Adh***	***Rho***	***Adh***	***Adh***	***Rho***	***Adh***	***Adh***	***Rho***	***Rho***
Dd_46	Oth	Fz	cA	Adh	cA	Rho	Sec	cA	cA	cA
Dd_49	Sec	Sec	Adh	Mth	Adh	Fz	Fz	Rho	Rho	Rho
Dd_51	cA	Fz	Fz	cA	Sec	Rho	cA	Adh	Mth	Adh
Dd_53	Oth	Fz	cA	Adh	cA	cA	cA	Rho	cA	Rho
At_Rask	cA	cA	Oth	cA	cA	cA	cA	cA	cA	Mth
Mb_King	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***	***Adh***

The resulting set did not however comprise a stable set and a clustering method (KalleClust) (fig 2 box 4) (explained in materials and methods) reduced the set to an initial set for phylogeny from which Ce2, Ce5, Dd1, Dd16, Dd38, Dm_mth1,2,4,5 GgVlgr1, HsVlgr1, MbKing, MmGpr144, TnGpr56-1 and sDm 2,6,8 and 9 were excluded. The sequences removed from the initial set were then reinserted under criteria described in materials and methods. The sequences that were never reinstalled were Ce2 and all of the *Methuselah* sequences previously removed, since these sequences resulted in an inconsistent topology of the trees. However a tree including all sequences from *Caenorhabditis elegans* , *Dictyostelium discoideum* and *Methuselah* was constructed since all of them are of great interest, especially the *Caenorhabditis elegans* sequences with the presence of GPS in their N-terminal. The tree was calculated using a bootstrap value of 1000 and the neighbor-joining method from Phylip 3.65, (described in materials and methods). Ce1 was excluded already with the execution of the in-house program since it had *Rhodopsin* (Rho) as best hit (Table II) and Ce2 did not group with the *Adhesion* group, see the result in figure 4.

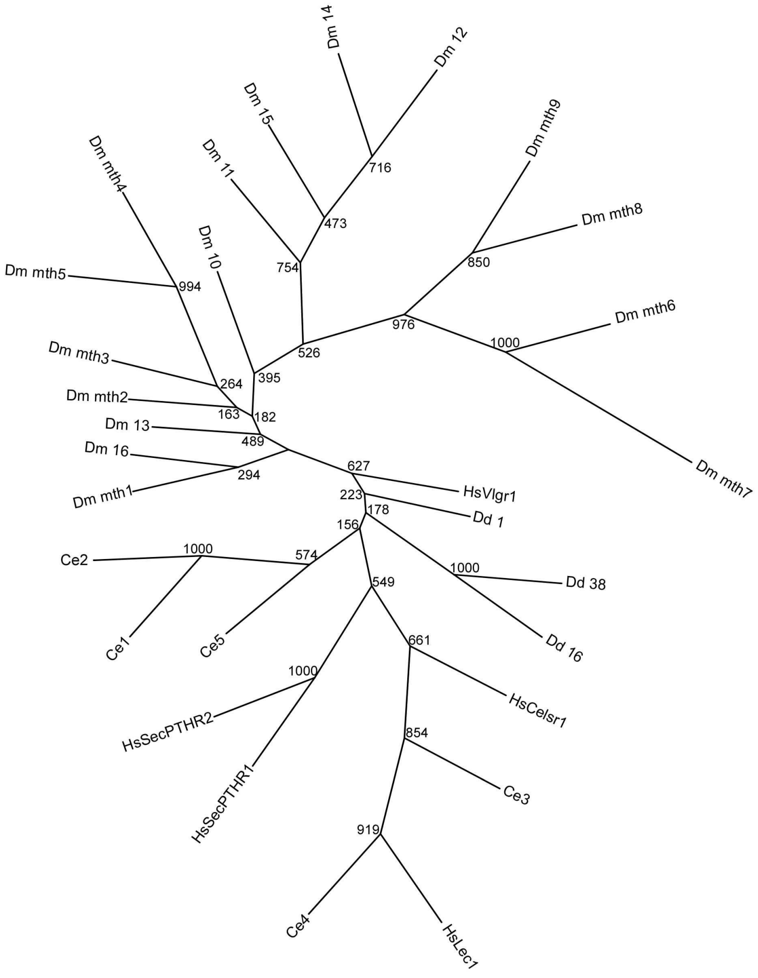


Figure 4. Consensus of Neighbor-joining trees with bootstrap set to 1000, of all *Methuselah*, *Caenorhabditis elegans* and *Dictyostelium discoideum*sequences with additional *Adhesion* (HsVlgr1 and HsLec1) and *Secretin* sequences (HsSecPTHR2 and HsSecPTHR1). Branches with bootstrap values below 50% do not give any information and should be collapsed.

Simultaneously manual assembling of the *Tetraodon nigroviridis*'s N-terminals was conducted in the same manner as for the 7TM but using the full-length human sequences as baits. Subsequently the domains of all exciting sequences were found with rps-blast as well as with interproScan, depicted in figure 11.

The distinct groups I-VIII proposed by Bjarnadottir and collegues [3] were apparent in all trees with the exception of group V and a few sequences that tended to belong to different groups depending on the method used to construct the tree, figure 5-9. Group V kept splitting up but the sequences within the group had mostly an adjacent placement in the trees. When further analysing the trees, sequence homology could be seen with potential sequences from *Tetraodon nigroviridis* following their human baits. A clear orthologous relationship could not be pinned down for all *Tetraodon nigroviridis*-sequences mostly by reason of basal arrangement in the trees. The ortholouges that clearly could be assessed according to the neighbor-joining tree in figure 10 were; TnGpr116-2 to HsGpr113, TnGpr144-1 to HsGpr144, TnGpr125-1 to HsGpr125, TnGpr124-1 to HsGpr124, TnGpr123-1,2 to HsGpr123, TnGpr112-2,3 to HsGpr112, TnGpr126-1 to HsGpr126, TnGpr97-1 to HsGpr97, TnLec3-1 to HsLec3, TnEtl-1 to HsEtl, TnTr32-1 to HsCd97, TnCelsr2-1 to HsCelsr2, TnCelsr1-1 to HsCelsr1, TnBai3-1 to HsBai3, TnBai2-1 to HsBai2 and TnBai3-3 to HsBai1. TnCd97-1 shows relations to group II, sDm6 to group III, sDm1 to group IV, TnLec3-3 to group I, TnGpr116-1 to group VI and TnGpr56-1 and TnGpr126-3 to group VIII (fig 9).

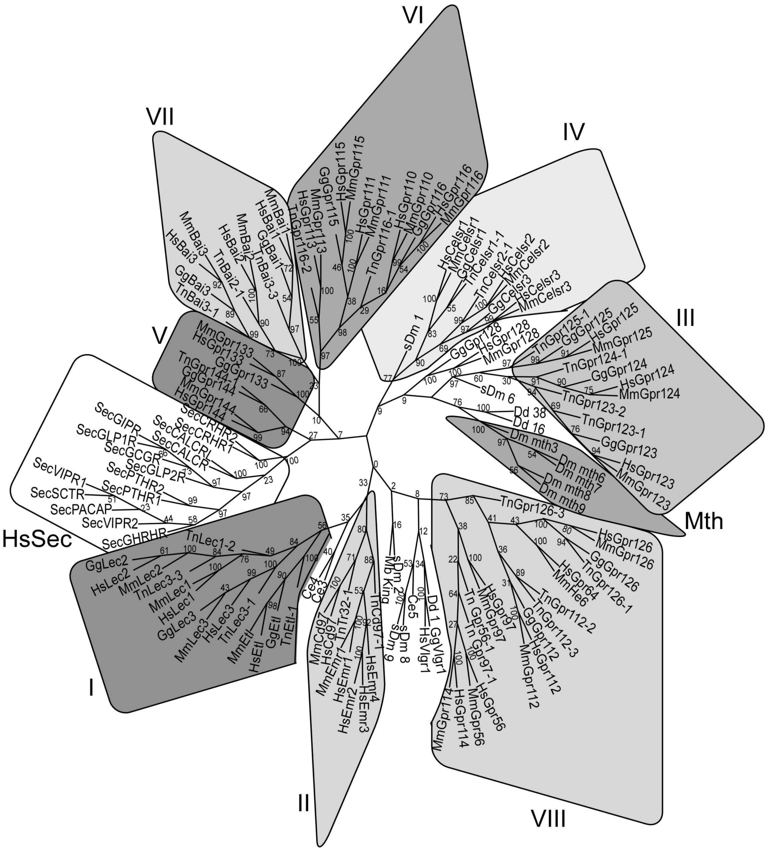
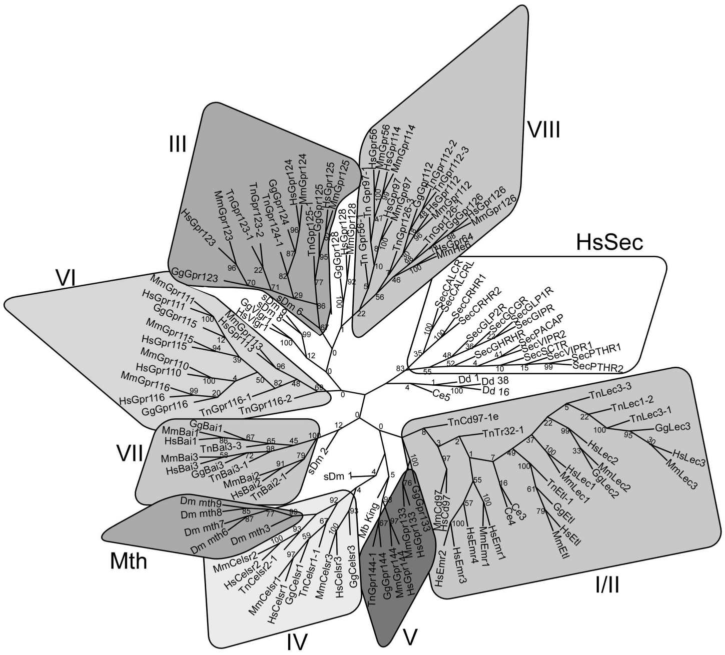


Figure 5. Consensus of minimum evolution (ME) trees with 100 bootstrap constructed in Mega 3.1. Numbers at the internal nodes represent bootstrap values and compartmentalization into group I-VIII is according to Bjarnadottir and colleagues previous division [3]. The abbreviations in fig 5-9 stand for: Hs – *Homo sapiens*, Mm – *Mus musculus*, Gg – *Gallus gallus*, Tn – *Tetraodon nigroviridis*, Tr – *Takifugu rubripes*, Dm – *Drosophila melanogaster*, Ce – *Caenorhabditis elegans*, Dd – *Dictyostelium discoideum*, HsSec – The *Secretin* family from Homo sapiens and Mth – *Methuselah.* The branches with bootstrap below 50% should be collapsed in all trees in fig 5-9.



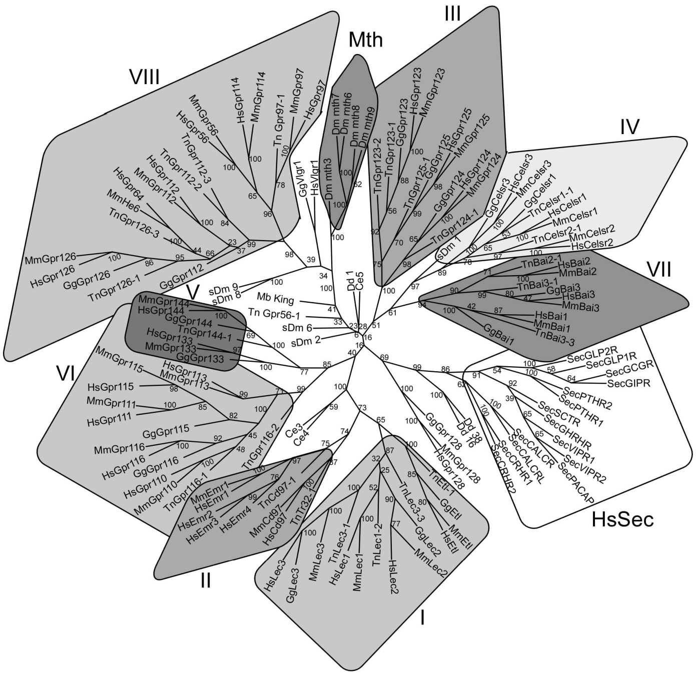


Figure 6. Consensus of maximum parsimony (MP) tree with 100 bootstrap constructed in Mega 3.1. Numbers at the internal nodes represent bootstrap values and compartmentalization into group I-VIII is according to Bjarnadottir and colleagues previous division [3]. Mth represent *Methuselah* and HsSec *Secretin* sequences from human.

Figure 7. Consensus of maximum parsimony (MP) tree with 100 bootstrap constructed in Phylip version 3.65. Numbers at the internal nodes represent bootstrap values and compartmentalization into group I-VIII is according to Bjarnadottir and colleagues previous division [3]. Mth represent *Methuselah* and HsSec *Secretin* sequences from human.

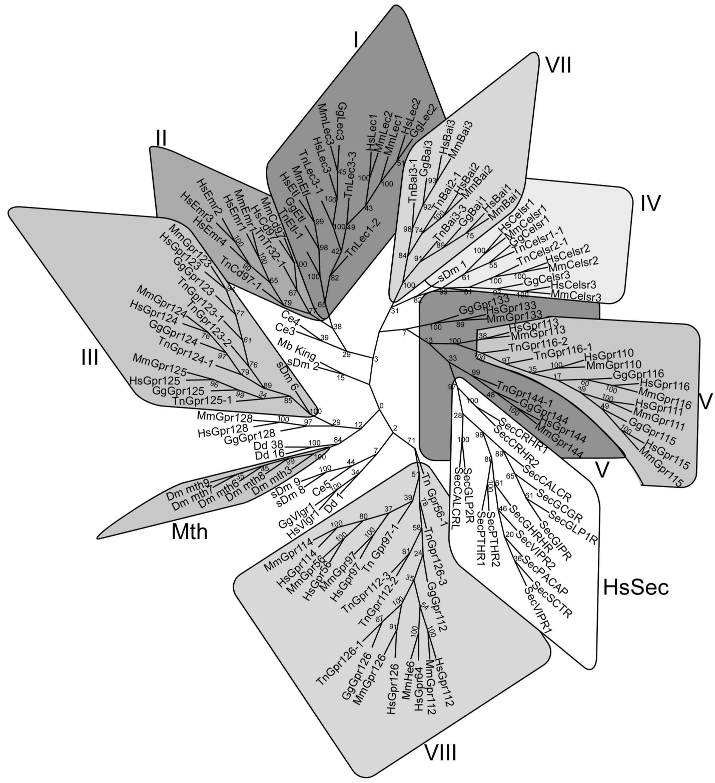


Figure 8. Consensus of neighbor-joining (NJ) tree with 100 bootstrap constructed in Mega 3.1. Numbers at the internal nodes represent bootstrap values and compartmentalization into group I-VIII is according to Bjarnadottir and colleagues previous division [3]. Mth represent *Methuselah* and HsSec *Secretin* sequences from human.

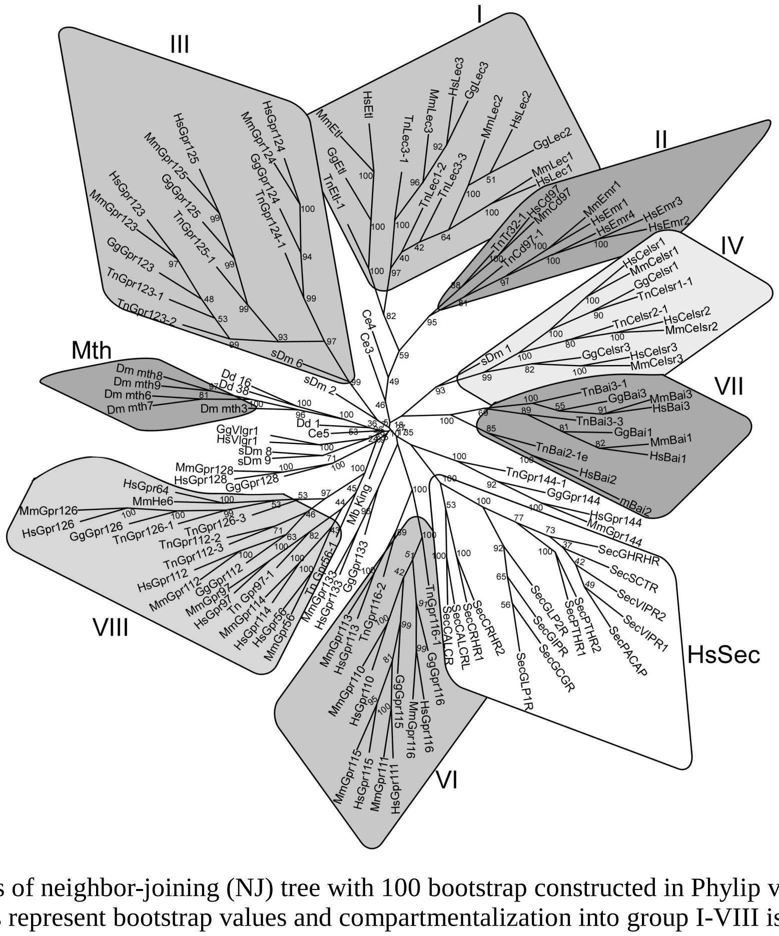


Figure 9. Consensus of neighbor-joining (NJ) tree with 100 bootstrap constructed in Phylip version 3.65. Numbers at the internal nodes represent bootstrap values and compartmentalization into group I-VIII is according to Bjarnadottir and colleagues previous division [3]. Mth represent *Methuselah* and HsSec *Secretin* sequences from human.

The relations described above agree with the sequence identity in most cases, see table III calculated in MegAlign from the DNA Star-package version 5.07 (DNASTAR,Madison, Wisconsin, United States). However some of the sequences show high sequence similarities with sequences from the same species such as TnGPR123-2 to TnGPR123-1. This further strengthens the probable gene duplication which is thought to have occurred in the Teleost lineage to which *Tetraodon nigroviridis* belong [87]. Since the sequence distance to the human sequence, to which they both are most similar is 18.4% one can argue that one of the sequences, probably TnGPR123-2, is about to loose function or evolve to aquire some other functionality. However it is enticing to question the gene duplication given that far from all *Adhesion* GPCRs in *Tetraodon nigroviridis* show copies. If whole genome duplication had occurred it would be most likely that such an essential gene family would show duplications of their sequences. However given that the doubling-up occurred between 300 and 350 million years ago [88, 89] copies might have lost functionality or evolved to oblivion.

Table III: Percent sequence identity between *Tetraodon nigroviridis* sequences and closest human relative. Sequences enclosed in parentheses show higher identity to sequence in their own species

Sequence identity between Hs and Tn		
Tn	**Closest Hs**	**%seq id**
Bai2-1	Bai2	80.1
Bai3-1	Bai3	92.2
Bai3-3	Bai1	91.1
CD97-1	EMR1	39.5
Celsr1-1	Celsr1	65.6
Celsr2-1	Celsr2	65.6
ETL-1	ETL	78.9
Gpr112-2	Gpr112	52.8
Gpr112-3	Gpr112	52.4
Gpr116-1	Gpr116	52.1
Gpr116-2	Gpr115	32.1
Gpr123-1	Gpr123	73.2
Gpr123-2	Gpr123	54.8
(Gpr123-2	TnGpr123-1	56.3)
Gpr124-1	Gpr124	54.7
Gpr125-1	Gpr125	68.5
Gpr126-1	Gpr126	75.5
Gpr126-3	Gpr64	55.0
Gpr144-1	Gpr144	57.5
Gpr56-1	Gpr112	25.0
(Gpr56-1	TnGpr126-1	29.6)
Gpr97-1	Gpr97	35.2
Lec1-2	Lec2	54.0
Lec3-1	Lec3	89.9
Lec3-3	Lec1	74.2
Tr32-1	CD97	52.4

All of the Gpr128 sequences took a more basal position and were hence removed from the previous alliance with group VIII but kept a close proximity to it. Some of the sequences not part of any group, repeatedly clustered together forming something of a group of their own. The group enclose; HsVlgr1, GgVlgr1, Ce5, Dd1, sDm8 and sDm9. The identity within the group was calculated in MegAlign from the DNA Star-package version 5.07, see table IV and V. This disclosed a sequence identity within the group ranging from 10.6%, between GgVlgr1 and Ce5, to 82.4% between GgVlgr1 and HsVlgr1.

Table IV: Pairwise distances between sDm8,9, Dd1, Ce5, Vlgr1 from human and chicken and Gpr128 from human, chicken and mouse. The distances were calculated using ClustalW in the MegAlign program from the DNA Star-package version 5.07 (DNASTAR, Madison, Wisconsin, United States)

Pair Distances of Untitled ClustalW (Slow/Accurate, Gonnet)									
	sDm9	**Ce5**	**Dd1**	**GgGpr128**	**GgVlgr1**	**HsGpr128**	**HsVlgr1**	**MmGpr128**	**sDm8**
sDm9	***	12.5	17.6	21.5	22.7	27.7	25.0	24.6	50.4
Ce5		***	13.1	15.7	10.6	17.8	12.3	18.6	13.1
Dd1			***	25.5	16.7	26.3	16.3	29.5	20.7
GgGpr128				***	13.7	52.2	14.3	53.2	18.4
GgVlgr1					***	16.3	82.4	16.7	23.7
HsGpr128						***	15.8	77.8	21.2
HsVlgr1							***	19.0	22.9
MmGpr128								***	22.0
sDm8									***

Table V: Clearer overview of best and second best hit from table IV

Name	BestHit	%	nBest	%
sDm9	sDm8	50.4	HsGpr128	27.7
Ce5	MmGpr128	18.6	HsGpr128	17.8
Dd1	MmGpr128	29.5	HsGpr128	26.3
GgGpr128	MmGpr128	53.2	HsGpr128	52.2
GgVlgr1	HsVlgr1	82.4	sDm8	23.7
HsGpr128	MmGpr128	77.8	GgGpr128	52.2
HsVlgr1	GgVlgr1	82.4	sDm9	25.0
MmGpr128	HsGpr128	77.8	GgGpr128	53.2
sDm8	sDm9	50.4	GgVlgr1	23.7

Another connection that was acknowledged was Dd16 and Dd38 to the *Methuselah* group. In order to clarify the relations a complementary tree with all *Methuselah* and *Caenorhabditis elegans* sequences together with the *Dictyostelium discoideum* sequences of interest was made, see figure 4. This revealed a middle part made up of the *Dictyostelium discoideum* and some of the *Caenorhabditis elegans* sequences but grouping together with respectively species and *Methuselah* and *Adhesion*/*Secretin* on either side. Due to this the hitlist for selected sequences from the in-house program targeting a BLAST-database with family members from most GPCRs was re-inspected, part of interest shown in table II. The table reveals that the sequences hit a range of different GPCRs, potential *Adhesion* sequences have hits from *Methuselah*, *Secretin*, *cAMP*, *Frizzled*, *Other* and *Rhodopsin* as well, inclining that at least the 7tm-regions might be reasonable similar. Since the *Dictyostelium discoideum* sequences are of ancient origin the Blast hits were collected in a table with the aim of exposing the relations between the GPCR families. As seen in table II the *Dictyostelium discoideum* sequences show resemblance to a range of GPCRs and interestingly not in a constant order. The first hit can be *Adhesion*, the following ones from some other GPCR family and then *Adhesion* hits again.

Dictyostelium discoideum also appears to have high substitution rates and several sequences had to be removed since they compromised the distance matrices and made tree construction impossible (data not shown). For the sake of the study they were kept until discarded by the in-house program targeting them against multiple families in the BLAST-search against the BLAST-database with most GPCR families represented (see materials and methods). Another interesting feature of the *Dictyostelium discoideum*-sequences are that Dd16 hits a 7TM domain with an E-value of 0.33 and Dd38 with an E-value of 0.04 while most other GPCR have a far better E-value in the range around 1e-10 down to 1e-60 (data not shown). Even the *Methuselah* sequences hit their 7TM with better E-values. This could be a result of the large evolutionary distance or perhaps since a *Secretin*-like 7TM domain is the closest hit. It is also enticing to look at the N-terminals of the sequences. Characteristic for the *Adhesion* family are the long N-terminals and the presence of multiple domains [3, 12] but when looking at the more ancient sequences most of them do not render any of those traits, see fig 11.

As a complementary to fig 11 all sequences from *Tetraodon nigroviridis*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Dictyostelium discoideum* and the two sequences from *Monosiga brevicollis* [10] and *Arabidopsis thaliana* [2] were searched against the nr database with blastp at www.ncbi.nlm.nih.gov giving accession numbers for the best match as well as genome localisation

Table VI: Number of exons for every gene in each species. Hs – *Homo sapiens*, Mm – *Mus musculus*, Tn – *Tetraodon nigroviridis*, Dm – *Drosophila melanogaster* and Ce – *Caenorhabditis elegans*. Orthologues are represented with the same name and related sequences are located in the proximity of each other. Exon numbers followed by (7tm) are not fullength and mainly comprise the 7tm domain

Number of exons in the species						
Prot	**Hs**	**Mm**	**Gg**	**Tn**	**Dm**	**Ce**
Lec1	19	20		20		
Lec1-2				5 (7tm)		
Lec2	20	20	19			
Lec3	22	21	19	16		
Lec3-3				19		
Emr1	19	20				
Emr2	18					
Emr3	13					
Emr4	10					
Etl	14	15	13	14		
Cd97	18	18		14		
Tr32				15		
Ce3						18
Ce4						8
Bai1	28	30	28			
Bai2	29	28	7 (7tm)	24		
Bai3	30	29	27	31		
Bai3-3				31		
Celsr1	34	34	23	37		
Celsr2	33	34		34		
Celsr3	34	34	31			
sDm1					4	
Ce5						25
Vlgr1	89	88	27			
sDm8					2	
sDm9					2	
Gpr56	13	13		14		
Gpr64	21	24	11			
Gpr97	12	10	2 (7tm)	9		
Gpr110	10	13				
Gpr111	6	3				
Gpr112	15	17	11	14		
Gpr112-3				14		
Gpr113	13	9				
Gpr114	12	12	3 (7tm)			
Gpr115	8	4	4			
Gpr116	18	17	16	7		
Gpr116-2				10		
Gpr123	16	5 (7tm)	5 (7tm)	5 (7tm)		
Gpr124	19	19	14	20		
Gpr125	16	14	13	18		
sDm6					3	
sDm2					10	
Gpr126	25	11	11	25		
Gpr126-3				5 (7tm)		
Gpr128	16	16	5 (7tm)			
Gpr133	12	10	8 (7tm)			
Gpr144	20	20	7 (7tm)	18		

and potential domains (Table VII). Table VII and fig 10 correspond seemingly well but the sequences in fig 10 have the advantage of being manually inspected and investigated for correct splicing whereas the other mostly are gene-predictions which are correctly spliced in 19% of the cases [90]. Figure 11 also gives a probable grouping for Ce5 since it displays typical domains for group IV. This is also true for sDm2 which have a galactose-binding lectin domain which is only present in group I.

The last investigation performed considered the number of exons in each species, gene and genegroup according to the I-VIII division [3], showing that a change in exon number occurred somewhere between the *Drosophila melanogaster* and *Caenorhabditis elegans* split and the *Tetraodon nigroviridis* (Table VI). The increase in exon number can also be a result of the evolvement of the long N-terminal which is absent in most of the sequences from the species located basally in the evolutionary tree (fig 3). The *Tetraodon nigroviridis* sequences follow the human, chicken and mouse exon number whereas *Drosophila melanogaster, Caenorhabditis elegans* and *Dictyostelium discoideum*have far less but also shorter N-terminals (fig 10).

Figure 10. (See next page) Representation of the *Adhesion* GPCRs N-terminals and their incorporated domains located with rps-blast at www.ncbi.nlm.nih.gov with a cutoff value of 0.1. The sequences are grouped according to Bjarnadottir and colleagues division in group I-VIII [3]. Represented sequences are Hs – *Homo sapiens*, Tn – *Tetraodon nigroviridis*, Dm – *Drosophila melanogaster*, Ce – *Caenorhabditis elegans* and Dd – *Dictyostelium discoideum*. Possible domains are GPS – GPCR proteolytic site, HBD – hormone binding domain, OLF – olfactomedin, GBL – galactose-binding lectin domain, EGF – epidermal growth factor, 7TM – seven transmembrane domain, LRR – leucine rich repeats, Ig – immunoglobulin, SEA – sperm protein, enterokinase and agrin, CA – cadherin domains, TSP1 – thrombospondin 1, PTX – pentraxin domain, LamG – laminin, Calx_beta – Domains in Na-Ca exchangers and integrin-beta4 , OPF – oligoendopeptidase F, TNFR – tumor necrosis factor receptor domain, HprK – Serine kinase of the HPr protein, CUB domain , Urease_beta – Urease beta subunit, Metallothio_PEC – , CLECT – C-type lectin like domain, EPTP – epitempin domain , SIN3 – Histone deacetylase complex and Herpes_gp2 – equine herpesvirus glycoprotein gp2.

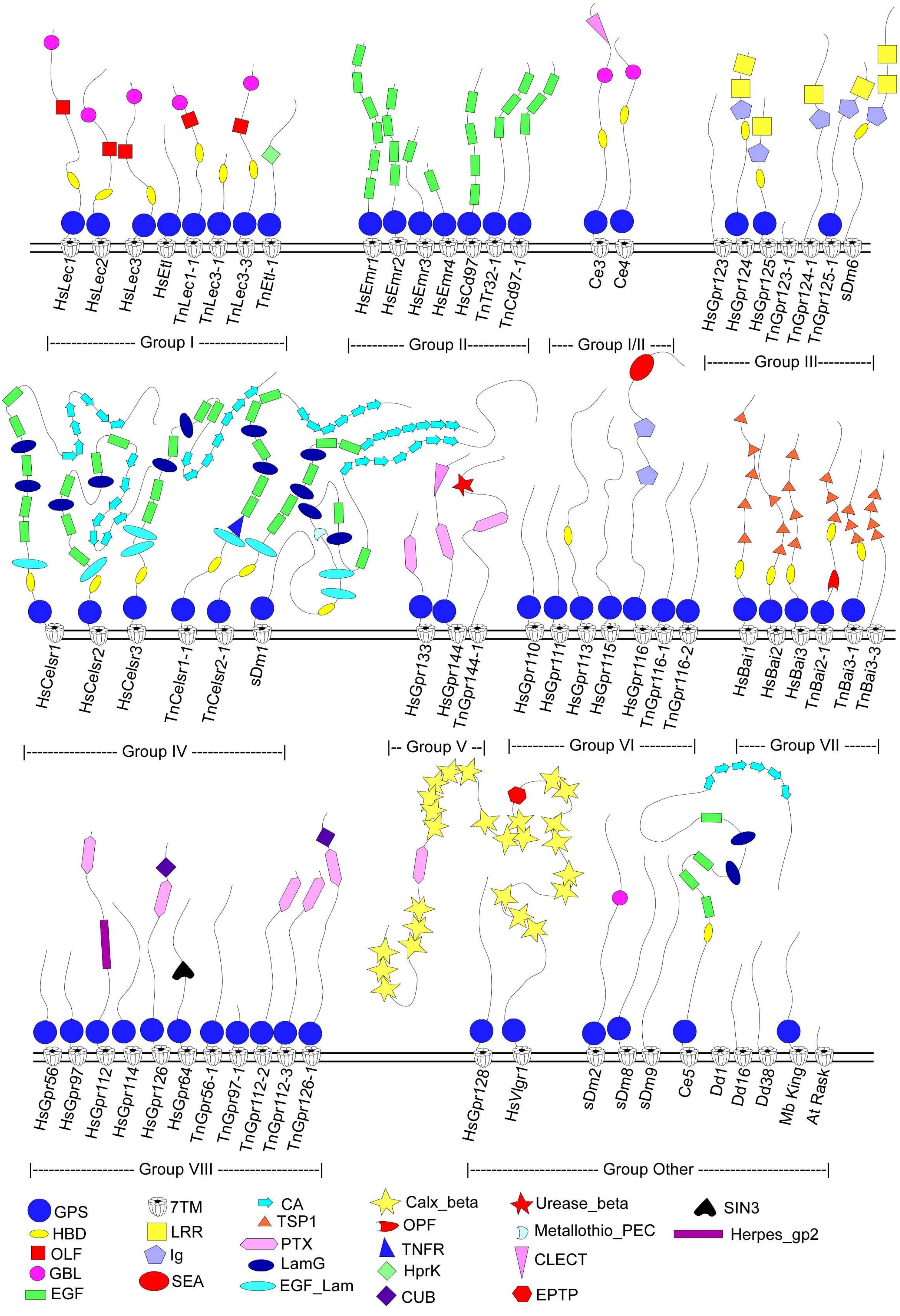

HsLec1
HsLec2
HsLec3
HsEtl
TnLec1-1
TnLec3-1
TnLec3-3
TnEtl-1
|---------------- Group I ----------------|
HsEmr1
HsEmr2
HsEmr3
HsEmr4
HsCd97
TnTr32-1
TnCd97-1
|---------- Group II-----------|
Ce3
Ce4
|---- Group I/II ----|
HsGpr123
HsGpr124
HsGpr125
TnGpr123-1
TnGpr124-1
TnGpr125-1
sDm6
|-------- Group III----------|
HsCelsr1
HsCelsr2
HsCelsr3
TnCelsr1-1
TnCelsr2-1
sDm1
|----------------- Group IV -----------------|
HsGpr133
HsGpr144
TnGpr144-1
|-- Group V --|
HsGpr110
HsGpr111
HsGpr113
HsGpr115
HsGpr116
TnGpr116-1
TnGpr116-2
|---------- Group VI ----------|
HsBai1
HsBai2
HsBai3
TnBai2-1
TnBai3-1
TnBai3-3
|----- Group VII ------|
HsGpr56
HsGpr97
HsGpr112
HsGpr114
HsGpr126
HsGpr64
TnGpr56-1
TnGpr97-1
TnGpr112-2
TnGpr112-3
TnGpr126-1
|------------------- Group VIII -------------------|
HsGpr128
HsVlgr1
sDm2
sDm8
sDm9
Ce5
Dd1
Dd16
Dd38
Mb King
At Rask
|------------------------ Group Other ------------------------|
GPS
HBD
OLF
GBL
EGF
7TM
LRR
Ig
SEA
CA
TSP1
PTX
LamG
EGF_Lam
Calx_beta
OPF
TNFR
HprK
CUB
Urease_beta
Metallothio_PEC
CLECT
EPTP
SIN3
Herpes_gp2

Table VII: Potential *Adhesion* sequences from each species with the accession number from the best hit at www.ncbi.nlm.nih.gov with blastp, chromosomal localisation and a short summary of the blastp hit. Tn – *Tetraodon nigroviridis*, Dm – *Drosophila melanogaster*, Ce - *Caenorhabditis elegans*, Dd – *Dictyostelium discoideum*, Mb – *Monosiga brevicollis* and At – *Arabidopsis thaliana.* For domain names the same abbreviations as for figure 11 have been used

temp name	AC of closest hit	Chr	BestHit blastp
Tetraodon nigroviridis			
Lec1-2	CAG06092	18	Unnamed protein product Tn, GBL, OLF, HormR, 7tm_2
Lec3-1	CAF93446	un	Unnamed protein product Tn, HormR, GPS, 7tm_2
Lec3-3	CAF98480	15	Unnamed protein product Tn, GBL, OLF, HormR, GPS, 7tm_2
Bai2-1	CAF99436	21	Unnamed protein product Tn, TSP1, GPS, 7tm_2
Bai3-1	CAG09441	17	Unnamed protein product Tn, 7tm_2
Bai3-3	CAG10859	21	Unnamed protein product Tn, TSP1, HormR, 7tm_2
Celsr1-1	CAG01167	un	Unnamed protein product Tn, EGF, LamG, EGF_Lam, HormR, GPS, 7tm_2
Celsr2-1	CAG06842	11	Unnamed protein product Tn, CA, EGF, LamG, EGF_Lam, HormR, GPS, 7tm_2
Cd97-1	CAG04105	3	Unnamed protein product Tn, GPS, 7tm_2
Tr32-1	CAG12162	3	Unnamed protein product Tn, GPS, 7tm_2
Etl-1	CAG01189	1	Unnamed protein product Tn, GPS, 7tm_2
Gpr56-1	CAG00694	Un	Unnamed protein product Tn, GPS, 7tm_2
Gpr97-1	CAG00691	Un	Unnamed protein product Tn, GPS, 7tm_2
Gpr112-2	CAG11729	7	Unnamed protein product Tn, 7tm_2
Gpr112-3	CAG09829	1	Unnamed protein product Tn, GPS, 7tm_2
Gpr123-1	CAG10157	Un	Unnamed protein product Tn, 7tm_2
Gpr123-2	CAG00851	Un	Unnamed protein product Tn, 7tm_2
Gpr124-1	CAF90106	Un	Unnamed protein product Tn, LRR, Ig, 7tm_2
Gpr125-1	CAF90021	Un	Unnamed protein product Tn, LRR, Ig, GPS, 7tm_2
Gpr126-1	CAF94956	Un	Unnamed protein product Tn, CUB, PTX, GPS, 7tm_2
Gpr126-3	CAG05443	6	Unnamed protein product Tn, GPS, 7tm_2
Gpr116-1	BAF32963	14	fIg-Hepta *Takifugu rubripes*, SEA, GPS, 7tm_2
Gpr116-2	CAG13000	14	Unnamed protein product Tn, GPS, 7tm_2
Gpr144-1	CAG01306	Un	Unnamed protein product Tn, GPS, 7tm_2
Drosophila melanogaster			
sDm1	BAA84069	2R	Flamingo Dm, CA, EGF, LamG, EGF_Lam, HormR, GPS, 7tm_2
sDm2	AAT47768	2R	RE2528p Dm, GBL, GPS, 7tm_2, cirl
sDm6	NP_572870	X	CG15744-PA Dm, LRR, 7tm_2
sDm8	NP_651842	3R	CG11318-PA Dm, 7tm_2
sDm9	NP_651845	3R	CG15556-PA Dm, 7tm_2
Caenorhabditis elegans			
Ce1	T34248	II	Hypothetical protein, GPS - Present in latrophilin/CL-1, 7tm_2
Ce2	NP_494739	II	F31D5.5, GPS – Present in latrophilin/CL-1, 7tm_2
Ce3	NP_001040724	II	Latrophilin receptor family member (lat-2) Ce, CLECT, GBL, HormR, GPS, 7tm_2
Ce4	NP_495894	II	Latrophilin receptor family member (lat-1) Ce, GBL, HormR, GPS, 7tm_2
Ce5	AAQ84880	V	flamingo-like protein FMI-1 Ce, CA, EGF, LamG, EGF_Lam, HormR, GPS, 7tm_2
Monosiga brevicollis			
King	AAP78684	Not available	MB7TM1 Mb, GPS, 7tm *Secretin*-like
Dictyostelium discoideum			
Dd1	XP_637908	4	GPCR family protein Dd, *Secretin*-like receptor; latrophilin receptor-like, 7tm similar to *Secretin*
Dd16	XP_636816	5	GPCR family protein Dd, *Frizzled* and *Smoothened*-like protein and 7TM similar to *Secretin*
Dd38	XP_636809	5	GPCR *Frizzled* and *Smoothened*-like protein
Arabidopsis thaliana			
Josefsson	CAA72145	1	GPCR At, slime mold *cAMP* receptor

4 Discussion

Adhesion sequences of the GPCR-family have been collected from *Homo sapiens* (human) [3], *Gallus gallus* (chicken) [24], *Mus musculus* (mouse) [3], *Tetraodon nigrovididis* (puffer fish), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (Nematode) and *Dictyostelium discoideum* (amoeba). This resulted in 24 potential *Adhesion* sequences from *Tetraodon nigroviridis* with complete 7TM and for 21 of them the N-terminal could be assembled (fig 11). For *Drosophila melanogaster*, *Caenorhabditis elegans* and *Dictyostelium discoideum*, 5,3 and 3 respectively potential *Adhesion* sequences could be located. This comprises one more *Drosophila melanogaster*-sequence than previously reported by Harmar [9] and two more *Dictyostelium discoideum*-sequences than formerly reported by Eichinger and colleagues [8, 72, 91]. However the most interesting sequences, taken the *Adhesion* perspective, are the two *Caenorhabditis elegan*-sequences excluded from the Adhesion set, one due to questionable positions in the phylogenetic trees (Ce2) and one not fulfilling the criteria for the in-house program (Ce1), see materials and methods and table II. The reason of the interest is the presence of a GPS domain in both sequences, which is characteristic for the *Adhesion*-family of GPCRs [3, 12, 29]. A possible explanation could be that the sequences can not be compared according to the 7TM alone since a prominent feature of the *Adhesion*-family is the long N-terminal. Contradicting this is the apparent consistency of similar domains in the I-VIII groups where the groups have been divided on the phylogenetic relationship of the 7TM [3], which also is evident for the *Tetraodon nigroviridis*-sequences, see fig 11. Taking both of these previous statements in consideration an additional explanation might be that the 7TMs can be used single-handedly for sequences derived from species with a closer evolutionary distance. As seen in table II sequences from species situated basally in the evolutionary tree (fig 3) have a higher probability to hit sequences from multiple GPCR families. This also stresses the possibility for a common ancestor for both the *Adhesion* family and the entire GPCR family which is also supported by Prabhu and Eichinger [91]. They argue that the presence of a member of family 2 (*Secretin/Adhesion*) in *Dictyostelium discoideum* emphasizes the appearance of a common origin before the split of animals and fungi but since no members have been found in fungi the sequence must have been lost in this lineage [91].

A theory I would like to introduce in this study is the prospect of the *Frizzled* and *Smoothened* receptors as predecessor of the *Adhesion*–receptors. The assumption is based on the resemblance of *Frizzled* and *Smoothened* to the B-family containing both *Secretin* and *Adhesion* receptors [22] and the annotated likeness of Dd16 to the *Frizzled* and *Smoothened*-like protein but with *Secretin*-like 7TM and Dd38 as *Frizzled* and *Smoothened* protein (Table VII). Both of these sequences group with the B-family in this study (fig 5-9) which further accentuate the similarity between the families in basally located species. An interesting thought is then that a common ancestor to the *Secretin*, *Adhesion* and *Methuselah* receptor families resides before the divergence of *Dictyostelium discoideum*, which according to Harmar are the three components of the B-family [9]. The presence of 25 *Frizzled/Smoothened*-like receptors in *Dictyostelium discoideum* [91] presents the possible interpretation that some of these might have evolved to *Adhesion* or *Secretin*-like receptors in higher species since at most ten *Frizzled* and one *Smoothened* receptor can be found in those species [38, 39]. This could however also be a result of a species-specific expansion of the *Frizzled/Smoothened*-family in *Dictyostelium discoideum*. The lack of the *Adhesion*-specific domain GPS in the *Dictyostelium discoideum*-sequences could also incline that the specificity of the *Adhesion*-family evolved later but before the upcoming of *Monosiga brevicollis*since the sequence derived from King and colleagues [10] possesses a GPS-domain (fig 11).

The discrepancy throughout the different trees seen in fig 5-9 for the sequences not part of any of the eight groups is a consequence of the different methods used for the different trees. Maximum parsimony (MP) trees risk getting caught in local minima meaning that the best tree might not be the one returned. Further is MP not preferable when the lineages have different evolutionary rates [92] and should not be used when the sequence distances exceeds 10%. This is also true for constant evolutionary rates if the tree presents short internal branches [93, 94]. Previous evaluations of the phylogenetic methods put maximum likelihood (ML) and neighbor-joining (NJ) as superior methods to MP [84, 95] in fact there are very few cases when MP is to prefer. This is when having; constant nucleotide substitution rates, no strong transition/transversion or GC-biases and when analysing a few, similar, long sequences including over 1000 bases [96, 97]. ML has not been used in this study since the method is immensely time consuming and would take up to much time.

Comparisons of Minimum evolution and NJ show that the trees demonstrate similar topology for small numbers of sequences but may differ remarkably if the number increases [98]. Minimum evolution has however the same disadvantage as maximum parsimony since it might enter local minima and will then not return the best ME tree. In this study the ME and NJ tree from MEGA 3.1 correspond really well with only a few differences, such as different placements for sDm2 and Mb King.

A feature of most phylogenetic trees and also neighbor-joining trees is the formation of groups due to long-branch attraction. This happens when sequences not strongly related to any other participants in the tree group together. When looking at the low percentage of sequence identity within the group containing HsVlgr1, GgVlgr1, Ce5, Dd1, sDm8 and sDm9 the group could be the result of long-branch attraction.

This arouses suspicion that additional cluster analysis, structure studies and functionality and ligand studies have to be performed. It is obvious that the structure ought to have a great impact on the protein which perhaps overshadows the sequence similarity. This has already been proven for RDC1 receptor which is a GPCR which shows high sequence similarity to the adrenomedullin receptor (ADMR) [12, 99] but shows structural similarity with the chemokine receptor CXCR4 according to a threading assembly refinement (TASSER) methodology. Strengthening this assumption is the previous discovery that both RDC1 and CXCR4 bind the same ligand, CXCL12 [100].

To conclude, 24 *Tetraodon nigroviridis* sequences potentially belonging to the *Adhesion* family were found and were elongated with their 7TM when possible. One *Drosophila melanogaster* sequence additional to Harmar previously found sequences was discovered, as well as two further *Caenorhabditis elegans* [9] but with questionable membership in the *Adhesion* family. In *Dictyostelium discoideum* two complementing sequences to Eichinger and colleagues first family B-like sequence [8, 72, 91] were found but a clear evolutionary history has not been discerned. Continued investigation of the *Adhesion*-family is of most interest since they seem to have an impact in tumour repression and are involved in several developmental stages [59-62, 101].

5 References

1. Lee, M.S., *Molecular clock calibrations and metazoan divergence dates.* J Mol Evol, 1999. **49**(3): p. 385-91.
2. Josefsson, L.G., *Evidence for kinship between diverse G-protein coupled receptors.* Gene, 1999. **239**(2): p. 333-40.
3. Bjarnadottir, T.K., et al., *The human and mouse repertoire of the adhesion family of G-protein-coupled receptors.* Genomics, 2004. **84**(1): p. 23-33.
4. Dehal, P., et al., *The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins.* Science, 2002. **298**(5601): p. 2157-67.
5. Cavalier-Smith, T., *Only six kingdoms of life.* Proc Biol Sci, 2004. **271**(1545): p. 1251-62.
6. Knoll, A.H., *The early evolution of eukaryotes: a geological perspective.* Science, 1992. **256**(5057): p. 622-7.
7. Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**: p. 695-716.
8. Eichinger, L., et al., *The genome of the social amoeba Dictyostelium discoideum.* Nature, 2005. **435**(7038): p. 43-57.
9. Harmar, A.J., *Family-B G-protein-coupled receptors.* Genome Biol, 2001. **2**(12): p. REVIEWS3013.
10. King, N., C.T. Hittinger, and S.B. Carroll, *Evolution of key cell signaling and adhesion protein families predates animal origins.* Science, 2003. **301**(5631): p. 361-3.
11. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
12. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
13. Fredriksson, R., et al., *Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions.* FEBS Lett, 2002. **531**(3): p. 407-14.
14. Fredriksson, R. and H.B. Schioth, *The repertoire of G-protein-coupled receptors in fully sequenced genomes.* Mol Pharmacol, 2005. **67**(5): p. 1414-25.
15. Schioth, H.B. and R. Fredriksson, *The GRAFS classification system of G-protein coupled receptors in comparative perspective.* Gen Comp Endocrinol, 2005. **142**(1-2): p. 94-101.
16. Pierce, K.L., R.T. Premont, and R.J. Lefkowitz, *Seven-transmembrane receptors.* Nat Rev Mol Cell Biol, 2002. **3**(9): p. 639-50.
17. Larhammar, D., A.G. Blomqvist, and C. Wahlestedt, *The receptor revolution--multiplicity of G-protein-coupled receptors.* Drug Des Discov, 1993. **9**(3-4): p. 179-88.
18. Kristiansen, K., *Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function.* Pharmacol Ther, 2004. **103**(1): p. 21-80.
19. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
20. Ohno, S., *Evolution by Gene Duplication*. 1970, Berlin, Heidelberg, New York: Springer-Verlag.
21. Graul, R.C. and W. Sadee, *Evolutionary relationships among G protein-coupled receptors using a clustered database approach.* AAPS PharmSci, 2001. **3**(2): p. E12.
22. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success.* Embo J, 1999. **18**(7): p. 1723-9.
23. Kolakowski, L.F., Jr., *GCRDb: a G-protein-coupled receptor database.* Receptors Channels, 1994. **2**(1): p. 1-7.
24. Lagerstrom, M.C., et al., *The G protein-coupled receptor subset of the chicken genome.*

PLoS Comput Biol, 2006. **2**(6): p. e54.

25. Bjarnadottir, T.K., R. Fredriksson, and H.B. Schioth, *The gene repertoire and the common evolutionary history of glutamate, pheromone (V2R), taste(1) and other related G protein-coupled receptors.* Gene, 2005. **362**: p. 70-84.
26. Keverne, E.B., *Pheromones, vomeronasal function, and gender-specific behavior.* Cell, 2002. **108**(6): p. 735-8.
27. Pin, J.P., T. Galvez, and L. Prezeau, *Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors.* Pharmacol Ther, 2003. **98**(3): p. 325-54.
28. Tateyama, M., et al., *Ligand-induced rearrangement of the dimeric metabotropic glutamate receptor 1alpha.* Nat Struct Mol Biol, 2004. **11**(7): p. 637-42.
29. Bjarnadottir, T.K., et al., *Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse.* Genomics, 2006. **88**(3): p. 263-73.
30. Niimura, Y. and M. Nei, *Evolution of olfactory receptor genes in the human genome.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12235-40.
31. Glusman, G., et al., *The olfactory receptor gene superfamily: data mining, classification, and nomenclature.* Mamm Genome, 2000. **11**(11): p. 1016-23.
32. Fredriksson, R., et al., *Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives.* FEBS Lett, 2003. **554**(3): p. 381-8.
33. Jacoby, E., et al., *The 7 TM G-protein-coupled receptor target family.* ChemMedChem, 2006. **1**(8): p. 761-82.
34. Nemeroff, C.B. and M.J. Owens, *Treatment of mood disorders.* Nat Neurosci, 2002. **5 Suppl**: p. 1068-70.
35. Slusarski, D.C., V.G. Corces, and R.T. Moon, *Interaction of Wnt and a Frizzled homologue triggers G-protein-linked phosphatidylinositol signalling.* Nature, 1997. **390**(6658): p. 410-3.
36. Adler, P.N., *Planar signaling and morphogenesis in Drosophila.* Dev Cell, 2002. **2**(5): p. 525-35.
37. Vinson, C.R., S. Conover, and P.N. Adler, *A Drosophila tissue polarity locus encodes a protein containing seven potential transmembrane domains.* Nature, 1989. **338**(6212): p. 263-4.
38. Foord, S.M., *Receptor classification: post genome.* Curr Opin Pharmacol, 2002. **2**(5): p. 561-6.
39. Foord, S.M., S. Jupe, and J. Holbrook, *Bioinformatics and type II G-protein-coupled receptors.* Biochem Soc Trans, 2002. **30**(4): p. 473-9.
40. Barnes, M.R., D.M. Duckworth, and L.J. Beeley, *Frizzled proteins constitute a novel family of G protein-coupled receptors, most closely related to the secretin family.* Trends Pharmacol Sci, 1998. **19**(10): p. 399-400.
41. Miller, J.R., *The Wnts.* Genome Biol, 2002. **3**(1): p. REVIEWS3001.
42. Xu, Q., et al., *Vascular development in the retina and inner ear: control by Norrin and Frizzled-4, a high-affinity ligand-receptor pair.* Cell, 2004. **116**(6): p. 883-95.
43. Chen, J.K., et al., *Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened.* Genes Dev, 2002. **16**(21): p. 2743-8.
44. DeCamp, D.L., et al., *Smoothened activates Galphai-mediated signaling in frog melanophores.* J Biol Chem, 2000. **275**(34): p. 26322-7.
45. Gho, M. and F. Schweisguth, *Frizzled signalling controls orientation of asymmetric sense organ precursor cell divisions in Drosophila.* Nature, 1998. **393**(6681): p. 178-81.
46. Unson, C.G., et al., *Roles of specific extracellular domains of the glucagon receptor in ligand binding and signaling.* Biochemistry, 2002. **41**(39): p. 11795-803.
47. Langer, I., et al., *Lysine 195 and aspartate 196 in the first extracellular loop of the VPAC1 receptor are essential for high affinity binding of agonists but not of antagonists.* Neuropharmacology, 2003. **44**(1): p. 125-31.
48. Stacey, M., et al., *LNB-TM7, a group of seven-transmembrane proteins related to family-B G-protein-coupled receptors.* Trends Biochem Sci, 2000. **25**(6): p. 284-9.

49. Baud, V., et al., *EMR1, an unusual member in the family of hormone receptors with seven transmembrane segments.* Genomics, 1995. **26**(2): p. 334-44.
50. McKnight, A.J. and S. Gordon, *The EGF-TM7 family: unusual structures at the leukocyte surface.* J Leukoc Biol, 1998. **63**(3): p. 271-80.
51. Fredriksson, R., et al., *There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini.* Biochem Biophys Res Commun, 2003. **301**(3): p. 725-34.
52. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
53. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
54. Campbell, I., Bork, P, *Epidermal growth factor-like domains.* Curr Opin Struct Biol, 1993. **3**: p. 385-392.
55. Hamann, J., et al., *The seven-span transmembrane receptor CD97 has a cellular ligand (CD55, DAF).* J Exp Med, 1996. **184**(3): p. 1185-9.
56. Stacey, M., et al., *The epidermal growth factor-like domains of the human EMR2 receptor mediate cell attachment through chondroitin sulfate glycosaminoglycans.* Blood, 2003. **102**(8): p. 2916-24.
57. Stacey, M., et al., *EMR4, a novel epidermal growth factor (EGF)-TM7 molecule up-regulated in activated mouse macrophages, binds to a putative cellular ligand on B lymphoma cell line A20.* J Biol Chem, 2002. **277**(32): p. 29283-93.
58. Takai, Y., et al., *Nectins and nectin-like molecules: roles in cell adhesion, migration, and polarization.* Cancer Sci, 2003. **94**(8): p. 655-67.
59. Kaur, B., et al., *Brain angiogenesis inhibitor 1 is differentially expressed in normal brain and glioblastoma independently of p53 expression.* Am J Pathol, 2003. **162**(1): p. 19-27.
60. Kaur, B., et al., *Vasculostatin, a proteolytic fragment of brain angiogenesis inhibitor 1, is an antiangiogenic and antitumorigenic factor.* Oncogene, 2005. **24**(22): p. 3632-42.
61. Kee, H.J., et al., *Expression of brain-specific angiogenesis inhibitor 3 (BAI3) in normal brain and implications for BAI3 in ischemia-induced brain angiogenesis and malignant glioma.* FEBS Lett, 2004. **569**(1-3): p. 307-16.
62. Kee, H.J., et al., *Expression of brain-specific angiogenesis inhibitor 2 (BAI2) in normal and ischemic brain: involvement of BAI2 in the ischemia-induced brain angiogenesis.* J Cereb Blood Flow Metab, 2002. **22**(9): p. 1054-67.
63. Kwakkenbos, M.J., et al., *The EGF-TM7 family: a postgenomic view.* Immunogenetics, 2004. **55**(10): p. 655-66.
64. Krasnoperov, V., et al., *Post-translational proteolytic processing of the calcium-independent receptor of alpha-latrotoxin (CIRL), a natural chimera of the cell adhesion protein and the G protein-coupled receptor. Role of the G protein-coupled receptor proteolysis site (GPS) motif.* J Biol Chem, 2002. **277**(48): p. 46518-26.
65. Krasnoperov, V.G., et al., *alpha-Latrotoxin stimulates exocytosis by the interaction with a neuronal G-protein-coupled receptor.* Neuron, 1997. **18**(6): p. 925-37.
66. Fukuzawa, T. and S. Hirose, *Multiple processing of Ig-Hepta/GPR116, a G protein-coupled receptor with immunoglobulin (Ig)-like repeats, and generation of EGF2-like fragment.* J Biochem (Tokyo), 2006. **140**(3): p. 445-52.
67. Volynski, K.E., et al., *Latrophilin fragments behave as independent proteins that associate and signal on binding of LTX(N4C).* Embo J, 2004. **23**(22): p. 4423-33.
68. Lelianova, V.G., et al., *Alpha-latrotoxin receptor, latrophilin, is a novel member of the secretin family of G protein-coupled receptors.* J Biol Chem, 1997. **272**(34): p. 21504-8.
69. Metpally, R.P. and R. Sowdhamini, *Genome wide survey of G protein-coupled receptors in Tetraodon nigroviridis.* BMC Evol Biol, 2005. **5**: p. 41.
70. Hedges, S.B. and S. Kumar, *Genomics. Vertebrate genomes compared.* Science, 2002. **297**(5585): p. 1283-5.
71. Hedges, S.B. and S. Kumar, *Precision of molecular time estimates.* Trends Genet, 2004. **20**(5): p. 242-7.

72. Williams, J.G., A.A. Noegel, and L. Eichinger, *Manifestations of multicellularity: Dictyostelium reports in.* Trends Genet, 2005. **21**(7): p. 392-8.
73. Insall, R., *The Dictyostelium genome: the private life of a social model revealed?* Genome Biol, 2005. **6**(6): p. 222.
74. Burger, G., et al., *Unique mitochondrial genome architecture in unicellular relatives of animals.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 892-7.
75. Patterson D.J., S.M.L. *Eukaryotes. Eukaryota, Organisms with nucleated cells.* 2000 [cited 2007 23 jan]; Version 08:[Available from: http://tolweb.org/Eukaryotes/3/2000.09.08 in The Tree of Life Web Project, http://tolweb.org/.
76. Salemi, M., Vandamme, A.-M., *The Phylogenetidc Handbook: A Practical Approach to DNA and Protein Phylogeny*. 2003, Cambridge: Cambridge University Press. 406.
77. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
78. Felsenstein, J., *PHYLIP (Phylogeny Inference Package)*. 2004, Distributed by the author. Department of Genome Sciences, University of Washington: Seattle.
79. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.
80. Eck, R.V., Dayhoff, M.O., *Atlas of Protein Sequence and Structure*. 1966, Silver Springs, Maryland: National Biomedical Research Foundation.
81. Fitch, W.M., *Towards defining the course of evolution: Minimum change for a specific tree topology.* Systematic Zoology, 1971. **20**: p. 406-416.
82. Kumar, S., K. Tamura, and M. Nei, *MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.* Brief Bioinform, 2004. **5**(2): p. 150-63.
83. Fitch, W.M. and E. Margoliash, *Construction of phylogenetic trees.* Science, 1967. **155**(760): p. 279-84.
84. Saitou, N., Nei, M., *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
85. Zdobnov, E.M. and R. Apweiler, *InterProScan--an integration platform for the signature-recognition methods in InterPro.* Bioinformatics, 2001. **17**(9): p. 847-8.
86. Marchler-Bauer, A. and S.H. Bryant, *CD-Search: protein domain annotations on the fly.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W327-31.
87. Jaillon, O., et al., *Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.* Nature, 2004. **431**(7011): p. 946-57.
88. Christoffels, A., et al., *Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes.* Mol Biol Evol, 2004. **21**(6): p. 1146-51.
89. Vandepoele, K., et al., *Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1638-43.
90. Rogic, S., A.K. Mackworth, and F.B. Ouellette, *Evaluation of gene-finding programs on mammalian sequences.* Genome Res, 2001. **11**(5): p. 817-32.
91. Prabhu, Y. and L. Eichinger, *The Dictyostelium repertoire of seven transmembrane domain receptors.* Eur J Cell Biol, 2006. **85**(9-10): p. 937-46.
92. Felsenstein, J., *Cases in which parsimony of compatibility methods will be positively misleading.* Systematic Zoology, 1978. **27**: p. 401-410.
93. DeBry, R.W., *The consistency of several phylogeny-inference methods under varying evolutionary rates.* Mol Biol Evol, 1992. **9**(3): p. 537-51.
94. Hendy, M.D.a.P.D., *A framework for the quantitative study of evolutionary trees.* Systematic Zoology, 1989. **38**: p. 297-309.
95. Tateno, Y., N. Takezaki, and M. Nei, *Relative efficiencies of the maximum-likelihood,*

neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol, 1994. **11**(2): p. 261-77.

96. Lin, J. and M. Nei, *Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data.* Mol Biol Evol, 1991. **8**(3): p. 356-65.

97. Sourdis, J. and M. Nei, *Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree.* Mol Biol Evol, 1988. **5**(3): p. 298-311.

98. Rzhetsky, A. and M. Nei, *Theoretical foundation of the minimum-evolution method of phylogenetic inference.* Mol Biol Evol, 1993. **10**(5): p. 1073-95.

99. Ladoux, A. and C. Frelin, *Coordinated Up-regulation by hypoxia of adrenomedullin and one of its putative receptors (RDC-1) in cells of the rat blood-brain barrier.* J Biol Chem, 2000. **275**(51): p. 39914-9.

100. Zhang, Y., M.E. Devries, and J. Skolnick, *Structure modeling of all identified G protein-coupled receptors in the human genome.* PLoS Comput Biol, 2006. **2**(2): p. e13.

101. Chang, G.W., et al., *CD312, the human adhesion-GPCR EMR2, is differentially expressed during differentiation, maturation, and activation of myeloid cells.* Biochem Biophys Res Commun, 2007. **353**(1): p. 133-8.