# A bioinformaticians view on the evolution of smell perception

Patrizia Anders[1§]

[1] School of Humanities and Informatics,
University of Skövde,
Högskolevägen
Skövde
Sweden

[§] Corresponding author

Email address:
Pat.anders@gmx.de

## Abbreviations

MOS – main olfactory system

VNO – vomeronasalorgan

OR – olfactory receptor

V1R, V2R – vomeronasal receptor, types 1 & 2

TM – trans membrane

MSA – multiple sequence alignment

AA – amino acid

DNA – desoxyribonucleic acid

ML – Maximum Likelihood

HMM – hidden Markov model

PIM – percent identity matrix

AOS – average overlap score

LRT – likelihood ratio test

# Abstract

### Background

The origin of vertebrate sensory systems still contains many mysteries and thus challenges to bioinformatics. Especially the evolution of the sense of smell maintains important puzzles, namely the question whether or not the vomeronasal system is older than the main olfactory system. Here I compare receptor sequences of the two distinct systems in a phylogenetic study, to determine their relationships among several different species of the vertebrates.

### Results

Receptors of the two olfactory systems share little sequence similarity and prove to be a challenge in multiple sequence alignment. However, recent dramatical improvements in the area of alignment tools allow for better results and high confidence. Different strategies and tools were employed and compared to derive a high quality alignment that holds information about the evolutionary relationships between the different receptor types. The resulting Maximum-Likelihood tree supports the theory that the vomeronasal system is rather an ancestor of the main olfactory system instead of being an evolutionary novelty of tetrapods.

### Conclusions

The connections between the two systems of smell perception might be much more fundamental than the common architecture of receptors. A better understanding of these parallels is desirable, not only with respect to our view on evolution, but also in the context of the further exploration of the functionality and complexity of odor perception. Along the way, this work offers a practical protocol through the jungle of programs concerned with sequence data and phylogenetic reconstruction.

# Background

Progress in genome sequencing projects of many different species enables us to construct more reliable estimates of evolutionary processes. Given its fundamental role in animal behaviour, the exploration of the sense of smell is of substantial scientific and general interest, not only since the Nobel prize in medicine or physiology 2004 was awarded to Linda Buck and Richard Axel for the discovery of olfactory receptors and the elucidation of the physiology of the sense of smell. A better understanding of the evolution of this sensory apparatus and the relationships of its distinct subsystems is therefore within reach, using advanced bioinformatic methods on the large amounts of chemosensory sequence data. OR proteins are encoded by the largest and most diverse gene family in the vertebrate genome, which consists of more than a thousand different genes in mammals (1). These are distributed in tandem clusters on the vertebrate chromosomes, revealing a record of evolutionary processes that have led to rapid expansion of the olfactory subgenome, gene duplication and conversion, extensive gene loss by pseudogenization, and diversifying selection (2-5). This superfamily is subject to a broad variety of bioinformatics research areas, e.g. genomics, phylogenetic reconstruction and structure prediction. Such challenges make those special GPCRs an interesting case for research.

## Molecular biological background

Olfaction is subserved by two anatomically and functionally distinct systems, the main olfactoy system (MOS) and the accessory olfactory system which includes the vomeronasal organ (VNO) as the sensory surface (6-8). Both systems hold receptors of the superfamily of G-protein-coupled receptors, GPCRs, a vast protein family that encompasses a wide range of functions (including vision, neurotransmission and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be classified into distinct groups. The MOS expresses odorant receptors (1,10) that belong to the family of *rhodopsin* GPCRs, like the V1 receptors (V1R) (11). V2 receptors (V2R)(12-14) of the VNO belong to the *glutamate* family and are predominantly utilized for pheromone-detection (15,16).

Pheromones (17) are semio-chemicals communicated by individuals within one species, which release stereotyped behavioural or physiological responses in the recipient. Pheromones are used by almost all animals. These scents are usually blends of a few types of molecules in species specific ratios (18,19). V1Rs appear to show very high affinity and specificity in ligand binding(15,20). However, pheromone-perception does not seem to be limited to the VNO, as some pheromone-induced behaviours are not perturbed by removal of the VNO (21-24), but are rather mediated by the MOS (25,22). Single V1R genes in humans and goats are even known to be expressed in the MOS (27,28). Conversely, some ‘common’ odorants can also be detected by the VNO (29,30), implying that these two olfactory systems might have functional overlap (31). It should also be mentioned that the complexity of the VNO morphology and the degree of its involvement in communication varies substantially among different mammals (33), and the VNO-mediated functions differ from species to species (34).

In comparison, ORs, V1Rs and V2Rs are coded by three complex gene superfamilies, which lack obvious sequence similarity, and yet seem to be organized in conceptually similar ways. These commonalities not only include their genomic organization (35,36), compact gene structure, and monoallelic expression (37-38) - of one (or a few) genes per neuron (40-44, all reviewed in 45). But also do the three receptor types share certain structural similarities. Until today, the only crystal structure available of a GPCR fold is that of *bovine rhodopsin* (26,32), to which many GPCR family members cannot be modelled with accuracy.

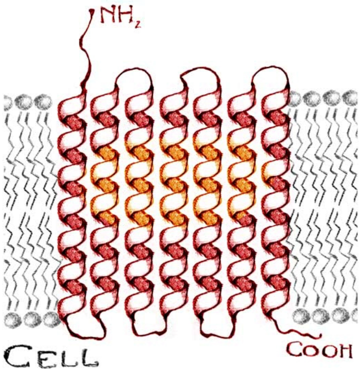


**Fig. 1**: sketch of predicted OR and V1R secondary structure. Yellow marks the region where odorant ligands are likely to bind.

As to be seen in figures 1 and 2, the GPCR fold comprises seven transmembrane (7TM) α-helices, with a cytosolic C-terminus, and an extracellular N-terminus, that can be very long in V2Rs (12-14,46,47). In principle, however, structure-function relationships for *rhodopsin*-like GPCRs are well established: extra cellular ligands are thought to bind within the 7TM region, inducing conformational changes that are linked to the binding of intracellular G-protein heterotrimers (48,32). Ligand binding in V2Rs is assumed to occur within the globular extracellular domain. Nevertheless, the understanding of structure-function relationships for specific GPCRs is far from complete. Furthermore are V1R genes intronless, while the large N-terminus in V2Rs is encoded by multiple exons, with the 7TM region contained within a single exon. Very few receptors have been deorphaned to date (reviewed in 45), but V2-like receptors in goldfish have been shown to bind arginine and lysine (50), suggesting amino acid-binding properties rather than binding of volatile organic compounds. This observation is supported by the homology between V2R extracellular regions and bacterial proteins with amino acid-binding properties (51,52).

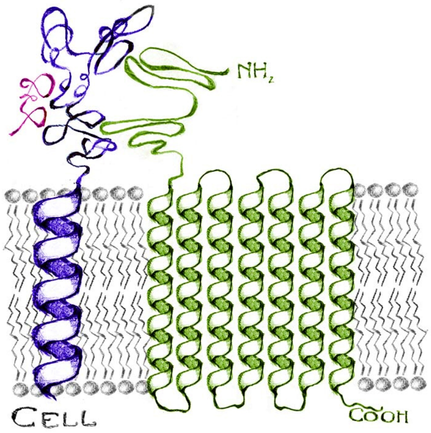


**Fig. 2** sketch of predicted V2R secondary structure in complex with M10 (blue) and β2m (pink).

## Evolutionary background

Evolutionarily, early tetrapods were probably the first animals to have separate olfactory and vomeronasal organs (53). Due to the existence of both systems solely in tetrapods and the aqueous content of the VNO, Broman (54) supposed their VNO to be homologous with the nasal chemosensory system of fishes and that the MOS arose later as an adaption to terrestrial life. In contradiction to that, the olfactory bulb in fishes and the main olfactory bulb of tetrapods reveals very similar axonal projections, so that Bertmar later inverted Broman's hypothesis, suggesting that the VNO arose in tetrapods as an adaptation to terrestrial life (55). This hypothesis remains controversial, since the VNO is generally present in aquatic salamanders, caecilians, and frogs (reviewed in 56, 57,58). The presence of the VNO in both amphibians and amniotes indicates that the system must have been present in the last common ancestor of these two groups, which was probably aquatic (59,60). This work will provide more evidence, on the basis of receptor sequence data, that the VNO really is the evolutionarily older organ.

## Bioinformatics background

During the past years, sequence databases were boosted and expanded exponentially. Genome sequencing projects of more and more species are completed, dramatically increasing the scope in which answers to our scientific questions can be found. Furthermore, bioinformatic tools and methods have been developed and improved tremendously, with increasing accuracy, speed, and tenability. Especially in the field of multiple sequence alignment (MSA), programs became very sophisticated. These achievements are of great value to this work, since multiple alignments of OR and V1R sequences proved to be quite difficult to construct, due to their rather distant relatedness.

On the other hand, these rapid developments in the recent past also mean that there now is a multitude of programs and routines, aggravating the choice of the best suited tools. In most cases, the user has to compare and evaluate many different methods that tackle the same problem. It might be tempting to simply trust the result a program comes up with, but quite frequently another method will find a contradicting solution. It’s one of the aims of this work to find a suitable path through the contemporary jungle of programs in this field.

The approach of this work can roughly be described as follows: In the first step, a data set of annotated amino acid (AA) sequences was established, containing about 1700 OR and 220 V1R sequences. Their separate MSAs were used as a basis to filter out closely related sequences, to reduce redundancy and to find representative subsets. The remaining 100 sequences of each kind were then realigned together, so that regions of positive selection could be identified and excised accordingly. Then a raw tree was constructed upon that final alignment, since more than 200 sequences exceed the computable maximum for the envisaged maximum likelihood approach by far. To obtain a reliable tree, at most eleven sequences may be contained in the dataset, to allow a feasible exhaustive search. The coarse tree helped to make a biologically reasonable choice of those eleven, and maximum likelihood calculation was used to determine the best tree among more than 34 million.

# Results

## Sequence retrieval

To build the fundament for a reliable phylogenetic reconstruction, the input datasets were chosen with great care. For the MSA, AA sequences were preferred over DNA sequences, because AAs allow to survey evolutionary history more deeply than nucleotides due to the fact that they show similarity in addition to identity. Also does it seem reasonable to exclude V2 receptors from the alignment, since they belong to a completely different family of GPCRs due to their long extracellular which most likely is involved in ligand binding. Several thousand sequences of ORs and V1Rs can be found in the gene and genome databanks, and a restriction to vertebrates does not decrease their number prominently. Therefore, to also impose a strict quality measure on the data, the input sequences were limited to those introduced in previously published work. Table 1 gives an overview of the number and sources of those annotated sequences. Of course, pseudogenes and partial sequences were excluded as well, since they are relieved from selective pressures, and would certainly import more noise to the estimation than information content.

| | Canis | Danio | Gallus | Xenopus | Mus. | Bos | Opossum | Primates | Lamprey |
|---|---|---|---|---|---|---|---|---|---|
| V1R | 8 (b) | | | | 104 (h) | 32 (b) | 49 (b) | | |
| V1R-like | | 7 (g) | | | | | | 13 (f) | |
| V3R | | | | | 8 (a) | | | | |
| OR | 90 (e) | 86 (c)<br>9 (d) | 78 (c) | 411 (c) | 1037 (c) | | | | 10 (j) |

**Table 1**: annotated sequences (see references for sources)

Taken together, the starting datasets eventually comprise about 1700 OR- and 220 V1R sequences, which were subject to further selection and preparation in the next stages.

### Eliminating sites of positive selection using PAML

Much care was invested into the question, whether or not, and if yes, how to pre-process data. It is much about the issue, what kind of information can be expected from the sequence data, and how to reduce noise caused by hypervariable regions.

The reason for this deliberation is the way, in which adaptive evolution can be expected to have created the great variety of GPCRs in general, but also the diversity of odorant receptors, that are involved in processes as diverse as foraging, reproductive behaviour, and predator avoidance. In general, most AA sites of a protein are subject to strong functional constraints, which are not equally distributed over the sequences in many cases. According to most studies of molecular evolution, adaptive changes occur at only a subset of sites (61). The diversifying selective pressure, however, causes such regions to evolve more rapidly than the rest of the sequence, usually to expand functionality of the resulting protein. In order to focus on the evolutionary path of whole gene families hidden behind the strong patterns of diversifying evolution of its individual members, it appears reasonable to exclude such sites, that underlie positive selection, from the multiple alignment of – in this case - olfactory and vomeronasal receptors. Ultimately, the inferred phylogeny should not reveal the evolution of discriminatory details of smell perception, but rather the relationships between different fundamental types of receptors, distinct from their odorant specificity.

A rational measure for the kind of selective pressure on a protein-coding gene is the ratio ($\omega$) of non-synonymous (amino acid replacement) versus synonymous (silent) substitution rates (62). Three different categories of selective pressure can be distinguished, by means of $\omega$. Negative, purifying selection reduces the fixation rate of unfavourable non-synonymous mutations, keeping $\omega$ below 1, whereas mutations are called neutral when they are fixed at the same rate as silent substitutions, i.e. when $\omega$ equals 1. Positive, diversifying selection, however, yields $\omega$ values above 1, since non-synonymous mutations tend to be fixed at a higher rate than silent ones. To detect sites under this positive selective pressure, PAML was applied, Phylogenetic Analysis by Maximum Likelihood (63). This software package allows to compare different models, endorsing positive selection if models of neutral and purifying selection can be rejected and if $\omega$ significantly exceeds 1 (64,65). For more detailed descriptions, see the "methods" section.

Besides PAML, several results from previous publications were incorporated to get a broader picture of the approximate distribution of sites under positive selection. A variety of studies (see table 2) was focusing on the question, which residues are involved in odorant binding activity, using different techniques to determine such subsequences. In most cases these regions are under diversifying, positive selection, which explains the great diversity of olfactory receptors and the ability of vertebrates to perceive thousands of different odours. These works were integrated and their results marked on a multiple alignment of representative sequences, together with the results of my own PAML analysis. Incorporating these hits, the information gathered here guided the process of deciding which subsequences to exclude from the final alignment. This way, the results of different studies support the excision of certain regions of the alignment.

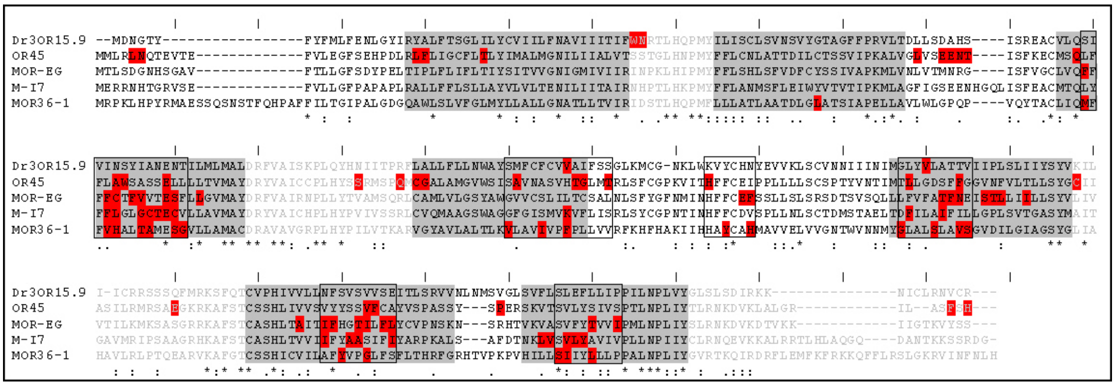


**Fig. 3**: MUSCLE (81) alignment of olfactory receptors previously involved in binding site studies (for details see table 2). Colouring pattern: transmembrane helices highlighted in dark grey, extra cellular loops shown in light grey, positions marked in red are expected to be involved in ligand binding to some extent. Boxed regions were excluded from the final alignment to reduce noise caused by hypervariable subsequences (see text for details). Stars below the alignment point out conserved regions, two dots show nearly conserved regions and one dot marks sites with different residues that share chemical attributes (hydrophobicity et cetera). The positions of TM helices are shown according to the structure of *bovine rhodopsin*.

As illustrated in figure 3, the hits do not show an a random distribution over the sequence alignment, but rather form distinct clusters at certain locations. Even though multiple methods were applied in the publications incorporated, they have one aspect in common: the search for the putative binding site of ORs. Their distinct approaches are listed in table 2, but can be divided into two groups: docking simulation studies (MOR-EG and M-I7) and studies detecting variability (DrOR15.9, OR45 and MOR36-1). However, not all these information are of equal importance to the decision which regions to excise. It may certainly be assumed that not all residues involved in ligand binding are extraordinarily variable, even positions that appear conserved in this alignment were described as part of the binding pocket by the docking simulation studies. Therefore, only non-conserved, hypervariable positions located within hypervariable clusters were chosen for the excision. Restraints appeared necessary in order to limit sequence loss exceeding an approximate one third of the original alignment length, and to exclude false positives, since all applied methods exhibit certain drawbacks. These disadvantages range from *rhodopsin* as a model for ORs in structure-based approaches, up to the point of weaknesses of model assumptions in sequence-based methods. Finally, even two of my own results (shown in the first sequence of figure 3) were spared from excision since this area was not proposed by any other study. The other two statistically significant results of my PAML analysis, however, fall into two of the very obvious clusters that were cut out from the alignment.

| Receptor | Description | Approach | Sites | Reference |
|---|---|---|---|---|
| **Dr3OR15.9**<br>(*danio rerio*) | PAML results inferring positive selected sites | ML calculation of extremely variable sites using PAML (63)<br>See Methods for details. | W46 N47 V154 V196 | *this work* |
| **OR45**<br>(mouse, NP_667174.1) | PAML results of 5 OR subfamilies combined | Extensive ML calculation of extremely variable sites using PAML (63) | D5 N6 D18 G25 M26 I32 A33 V86 R88 E91 R92 G101 A105 W106 T112 Q134 T139 T142 I153 C161 Y162 I164 L165 K177 A200 H220 I206 R233 A255 T269 I256 R308 Y312 | Emes 2004 (64) |
| **MOR-EG**<br>(mouse, AAL61116.1) | residues in the predicted binding pocket | functional analysis of several site-directed mutants, ligand docking simulation studies, rational receptor design with predictable changes in ligand affinity. | F102 F105 C106 F108 V109 E112 S113 L116 E181 F182 T205 F206 N207 S210 T211 L212 I215 A248 I251 F252 L259 T280 I283 | Katada 2005 (65) |
| **M-I7**<br>(mouse, AAD13307.1) | residues within 3.5Å of ligands in final models | 3D prediction of structure and ligand docking studies using MembStruck and HierDock, prediction of responses to 56 odorants and comparison with experimental data | F109 L110 G113 C114 T115 E116 C117 K164 F205 I209 I255 A258 A259 I263 L278 V279 V281 L282 Y283 | Hall 2004 (66) |
| **MOR36-1**<br>(mouse, AAL61401.1) | predicted functional residues | alignment & analysis of 1441 OR sequences, assuming binding-involved residues to be conserved among pairs of orthologous receptors and considerably less conserved among paralogous pairs. | L87 M115 V118 H119 T122 A123 S126 G127 V166 I170 F173 Y192 H195 G213 S217 V220 S221 Y266 G269 S301 I302 L305 | Man 2003 (67) |

**Table 2**: Details about figure 3: Different studies unveiling residues possibly involved in OR ligand binding. Site numbers are according to each sequence and may differ from positions in alignments (fig.3 & fig.4) due to sequence shifts caused by gaps.

Table 2 provides an overview of the literature that was taken into account in this part of the project. Distinct receptors were subject to these different approaches, they are shown in a multiple alignment in figure 3, to get an overview of the suspicious AA sites in relation to each other.

Earlier work regarding the search for the binding pocket (Vaidehi, Floriano, Singer, Ashar, 68-71) was not taken into account, due to the reduced reliability of docking simulation studies using the low resolution structure of *rhodopsin* (7.5 Å). These lines had to be drawn to achieve a justifiable extent of accuracy, - even though it is clear that in the end it will not be possible to eliminate the noise due to hypervariability completely - , but its reduction is sensible matter.

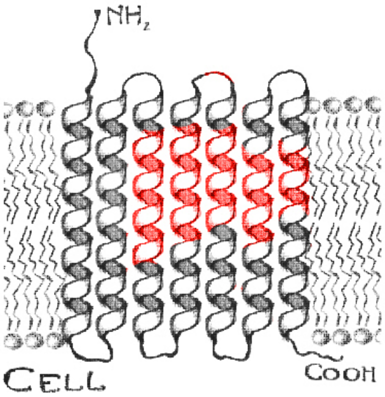


**Fig. 5:** OR topology layout, regions that will be excluded from the sequence alignment marked in red

Figure 5 illustrates the approximate locations of segments within the receptor that will be excised from the sequence data. In general, there are seven helices located in the lipid bilayer, forming a kind of pocket to bind certain ligands. The closer to the intracellular side, the narrower the pocket gets, opening up towards the extra cellular side. This simple sketch of the structure can be assumed in analogy to r*hodopsin* and other GPCRs of this kind. Thus the localization of the chosen regions is biologically plausible located in the upper half or middle of some helices, where ligand interaction most certainly takes place (see fig. 1). Furthermore, most studies do not expect the first two helices to be candidates for domains involved in ligand binding. Marking the proposed candidates of binding-pocket involvement into a schematic presentation of an olfactory receptor, the following details catch the eye:

(1) The amino and carboxy terminals are not suspected to interact with ligands, neither the first TM helix. (2) Other loops, and helices 2 and 7 have no or very few

suspicious amino acids that have been mentioned by no more than one study. (3) Most likely to form the binding pocket are helices 3, 4, 5 and 6, as suggested by the majority of previous studies and my results.  As to be seen in Fig. 5, the aminoacids predicted to form the binding pocket cluster within the  in a two thirds of each of the helices 3, 4, 5 and 6 facing the extracellular side. These clusters that are presumably under positive selection will be excluded from the alignment to prevent misleading conclusions about the phylogeny of the olfactory receptors apart from the developments in smell-detection.

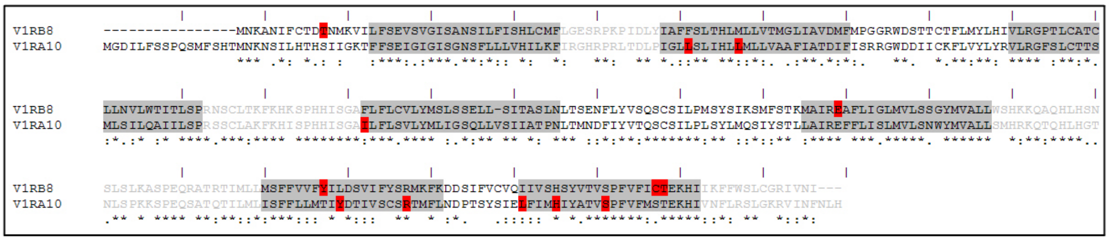
When combining the final alignment, it will be necessary to extract the very same regions from all other aligned ORs, which presents a challenge regarding the V1Rs. Even though these receptors are similar to ORs of the main olfactory system, there is reason to expect differences in ligand interaction:

ORs are known to feature a broad spectrum of ligands that cause different levels of activation (9).  V1Rs, however, are much more specific in ligand affinity, which is biologically sensible. So there is a discrepancy in the way in which the two kinds have a very similar structure and still differ in specificity, which might most likely be due to slightly different ways of ligand binding.

Thus the procedure has to be carried out on V1Rs as well, determining the regions which should be excluded from the alignment. If they differ (and that is likely), they have to be cut out from the ORs of the main olfactory system as well.

## Applying the truncation process to V1Rs

Regarding the vomeronasal receptors, considerably less literature exists exploring the parts of the protein possibly involved in ligand interaction. Fig. 6 illustrates the localization of hypervariable residues according to PAML. Both sequences are representatives for sets of 10-22 sequences involved in the calculation, V1RB8 in my case, and V1RA10 in the work of Emes et al (64). This time the results do not obviously cluster together, possibly due to the distance between the datasets. However, it appears helpful to find support for some biologically plausible situated hits (almost all are found in tm domains) within the noisy regions of olfactory receptors.

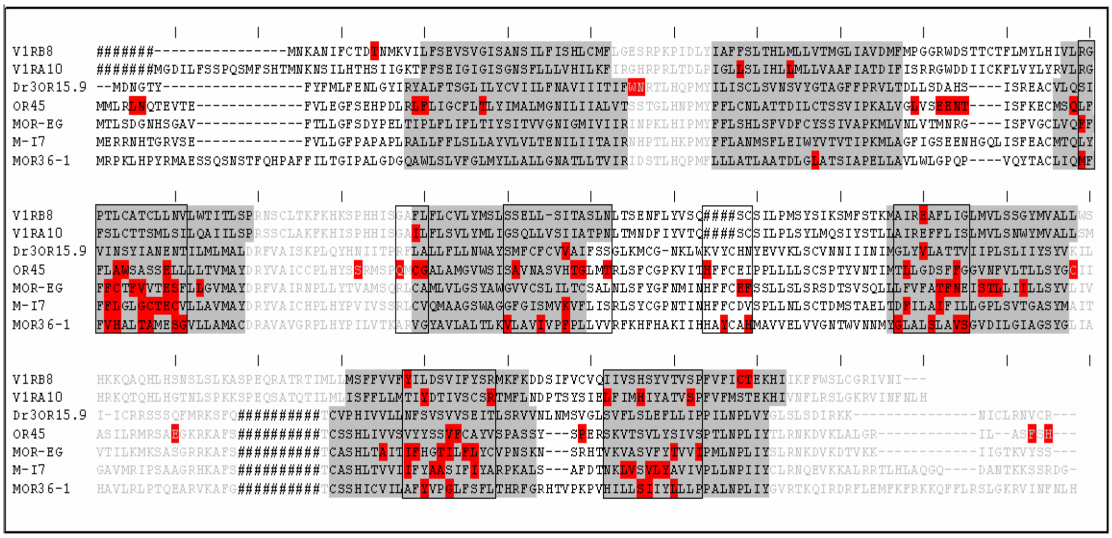


```
V1RB8      ----------------MNKANIFCTDTNMKVILFSEVSVGISANSILFISHLCMFLGESRPKPIDLYIAFFSLTHLMLLVTMGLIAVDMFMPGGRWDSTTCTFLMYLHIVLRGPTLCATC
V1RA10     MGDILFSSPQSMFSHTMNKNSILHTHSIIGKTFFSEIGIGISGNSFLLLVHILKFIRGHRPRLTDLPIGLLSLIHLLMLLVAAFIATDIFISRRGWDDIICKFLVYLYRVLRGFSLCTTS
                           ***  .*: *.: :   :***:.:***.**:*:: *:  *:    **:  ** *.::** **::*:. .:**.*:*:.    **.  *.**:**: **** :**:*.

V1RB8      LLNVLWTITLSPRNSCLTKFKHKSPHHISGAFLFLCVLYMSLSSELL-SITASLNLTSENFLYVSQSCSILPMSYSIKSMFSTKMAIREAFLIGLMVLSSGYMVALLWSHKKQAQHLHSN
V1RA10     MLSILQAIILSPRSSCLAKFKHISPHHISGAILFLSVLYMLIGSQLLVSIIATPNLTMNDFIYVTQSCSILPLSYLMQSIYSTLLAIREFFLISLMVLSNWYMVALLSMHRKQTQHLHGT
           :*.:* :* ****.***:**** ********:***.**** :.*:** ** *: *** ::*:**:*******:** ::*::** :**** ***.*****. ******  *:**:****..

V1RB8      SLSLKASPEQRATRTIMLLMSFFVVFYILDSVIFYSRMKFKDDSIFVCVQIIVSHSYVTVSPFVFICTEKHIIKFFWSLCGRIVNI---
V1RA10     NLSPKKSPEQSATQTILMLISFFLLMTIYDTIVSCSRTMFLNDPTSYSIELFIMHIYATVSPFVFMSTEKHIVNFLRSLGKRVINFNLH
           .** * **** **:**::*:***::: * *:::  **  * :*.   .::::: * *.*******:.*****::*: **  *::*:
```

**Fig. 6**: Alignment of two reference V1Rs to show PAML results of this project (V1RB8) and of Emes et al.'s research (V1RA10). Alignment made using MUSCLE. Positions of TM helices calculated using TMHMM (72) and marked in grey, sites under positive selection, according to PAML, highlighted in red.

The alignment between receptors of the main olfactory and the vomeronasal system is quite a challenge. Achieving a biologically meaningful alignment is very difficult and this high goal is situated much further down the road of this work. However, it is necessary to get some information about the approximate correlations of these two receptor types, in order to exclude the same regions from both kinds of sequences. At this point it is very unlikely to match exact positions with high certainty. However, a rough secondary structure alignment helps to guide the process. It provides sufficient information about the relative localizations of hypervariable regions. The positions of the TM domains were estimated using the online program TMHMM (72), which was tested on the OR sequences as well, to find out how much this prediction method differs from the results of the structural alignment method using *rhodopsin*. It turned out that TMHMM gave a very reliable prediction in this case (data not shown).

Figure 7 illustrates the way in which the distinct types of receptors – ORs and V1Rs were set into approximate correlation based upon their domain structure. Now many of the marked V1R positions (Fig. 3) fall into the obvious clusters already found in ORs. Two of the previously chosen regions to exclude were expanded and one small new one introduced (amino terminus of TM helix 4), since more clues coincided on these areas. Especially the clusters in helices six and seven gained support through the V1R results, which make the extension of excludable areas reasonable. Shown in figure 7 in black-lined boxes, those regions sum up to almost 70 residues. Thus less than one third of the receptor sequence will be cut out from the alignment.



```
V1RB8       #######----------------MNKANIFCTDTNMKVILFSEVSVGISANSILFISHLCMFLGESRPKPIDLYIAFFSLTHLMLLVTMGLIAVDMFMPGGRWDSTTCTFLMYLHIVLRG
V1RA10      #######MGDILFSSPQSMFSHTMNKNSILHTHSIIGKTFFSEIGIGISGNSFLLLVHILKFIRGHRPRLTDLPIGLLSLIHLLMLLVAAFIATDIFISRRGWDDIICKFLVYLYRVLRG
Dr3OR15.9   --MDNGTY-----------------FYFMLFENLGYIRYALFTSGLILYCVIILFNAVIIITIFWNRTLHQPMYILISCLSVNSVYGTAGFFPRVLTDLLSDAHS----ISREACVLQSI
OR45        MMLRLNQTEVTE-------------FVLEGFSEHPDLRLFLIGCFLTLYIMALMGNILIIALVTSSTGLHNPMYFFLCNLATTDILCTSSVIPKALVGLVSEENT----ISFKECMSQLF
MOR-EG      MTLSDGNHSGAV-------------FTLLGFSDYPELTIPLFLIFLTIYSITVVGNIGMIVIIRINPKLHIPMYFFLSHLSFVDFCYSSIVAPKMLVNLVTMNRG----ISFVGCLVQFF
M-I7        MERRNHTGRVSE-------------FVLLGFPAPAPLRALLFFLSLLAYVLVLTENILIITAIRNHPTLHKPMYFFLANMSFLEIWYVTVTIPKMLAGFIGSEENHGQLISFEACMTQLY
MOR36-1     MRPKLHPYRMAESSQSNSTFQHPAFFILTGIPALGDGQAWLSLVFGLMYLLALLGNATLLTVIRIDSTLHQPMFLLLATLAATDLGLATSIAPELLAVLWLGPQP----VQYTACLIQMF

V1RB8       PTLCATCLLNVLWTITLSPRNSCLTKFKHKSPHHISGAFLFLCVLYMSLSSELL-SITASLNLTSENFLYVSQ####SCSILPMSYSIKSMFSTKMAIREAFLIGLMVLSSGYMVALLWS
V1RA10      FSLCTTSMLSILQAIILSPRSSCLAKFKHISPHHISGAILFLSVLYMLIGSQLLVSIIATPNLTMNDFIYVTQ####SCSILPLSYLMQSIYSTLLAIREFFLISLMVLSNWYMVALLSM
Dr3OR15.9   VINSYIANENTILMLMALDRFVAISKPLQYHNIITPRFLALLFLLNWAYSMFCFCVVAIFSSGLKMCG-NKLWKVYCHNYEVVKLSCVNNIIINIMGLYVLATTVIIPLSLIIYSYVKIL
OR45        FLAWSASSELLLLTVMAYDRYVAICCPLHYSSRMSPQMCGALAMGVWSISAVNASVHTGLMTRLSFCGPKVITHFFCEIPPLLLLSCSPTYVNTIMTLLGDSFFGGVNFVLTLLSYGCII
MOR-EG      FFCTFVVTESFLLGVMAYDRFVAIRNPLLYTVAMSQRLCAMLVLGSYAWGVVCSLILTCSALNLSFYGFNMINHFFCEFSSLLSLSRSDTSVSQLLLFVFATFNEISTLLIILLSYVLIV
M-I7        FFLGLGCTECVLLAVMAYDRYVAICHPLHYPVIVSSRLCVQMAAGSWAGGFGISMVKVFLISRLSYCGPNTINHFFCDVSPLLNLSCTDMSTAELTDFILAIFILLGPLSVTGASYMAIT
MOR36-1     FVHALTAMESGVLLAMACDRAVAVGRPLHYPILVTKARVGYAVLALTLKVLAVIVPFPLLVVRFKHFHAKIIHHAYCAHMAVVELVVGNTWVNNMYGLALSLAVSGVDILGIAGSYGLIA

V1RB8       HKKQAQHLHSNSLSLKASPEQRATRTIMLLMSFFVVFYILDSVIFYSRMKFKDDSIFVCVQIIVSHSYVTVSPFVFICTEKHIIKFFWSLCGRIVNI---
V1RA10      HRKQTQHLHGTNLSPKKSPEQSATQTILMLISFFLLMTIYDTIVSCSRTMFLNDPTSYSIELFIMHIYATVSPFVFMSTEKHIVNFLRSLGKRVINFNLH
Dr3OR15.9   I-ICRRSSSQFMRKSFQ##########TCVPHIVVLLNFSVSVVSEITLSRVVNLNMSVGLSVFLSLEFLLIPPILNPLVYGLSLSDIRKK---------------NICLRNVCR---
OR45        ASILRMRSAEGKRKAFS##########TCSSHLIVVSVYYSSVFCAYVSPASSY---SPERSKVTSVLYSIVSPTLNPLIYTLRNKDVKLALGR------------IL--ASFSH---
MOR-EG      VTILKMKSASGRRKAFS##########TCASHLTAITIFHGTILFLYCVPNSKN---SRHTVKVASVFYTVVIPMLNPLIYSLRNKDVKDTVKK------------IIGTKVYSS---
M-I7        GAVMRIPSAAGRHKAFS##########TCASHLTVVIIFYAASIFIYARPKALS---AFDTNKLVSVLYAVIVPLLNPIIYCLRNQEVKKALRRTLHLAQGQ----DANTKKSSRDG-
MOR36-1     HAVLRLPTQEARVKAFG##########TCSSHICVILAFYVPGLFSFLTHRFGRHTVPKPVHILLSIIYLLLPPALNPLIYGVRTKQIRDRFLEMFKFRKKQFFLRSLGKRVINFNLH
```

**Fig. 7**: Domain-based correlation between roughly aligned V1R and OR sequences. Manually inserted gaps marked with '# '. Colouring scheme according to fig. 1. Approximate regions to cut out are shown in boxes, relative to domain structure.

### Reducing redundancy – FILTER

Filter is a custom-made perl program I designed to adjust the level of similarity among sequences within one dataset and exclude redundant data due to highly similar sequences. Based upon a sequence percent identity matrix (PIM) constructed by ClustalX, any degree of similarity can be chosen, and sequences exceeding that threshold will be removed from the dataset. In short, the program evaluates the PIM matrix above the diagonal for scores higher than the threshold. The corresponding sequences are memorized and all others written into a new file, which now contains sequences that are more distantly related than the threshold. This value was initially set to 60, because subfamilies are defined as sequences, that share more than 60% identities on the amino acid level. However, maintaining one member of each subfamily, these reduced datasets still contained too many sequences to be feasible in upcoming MSA calculations. About 100 sequences for each kind – OR and V1R – appeared reasonable. Therefore the threshold was set to 40% in OR sequences, and 45% in V1R sequences, yielding  a little over 100 sequences for both datasets. The source code of filter.pl along with comments is provided as additional file “filter.pdf”.

### The multiple sequence alignment

A lot of attention in the field of bioinformatics has always been focused on the MSA problem, mainly because MSAs play a central role in the annotation of genomes (73). Several new and accurate alignment methods have been introduced (74-76), that are applied and compared in this work.

Among the various different approaches to the MSA problem, at least in theory, hidden Markov models (HMM) present a very sound solution to multiple sequence alignments, using statistical models to describe evolutionary processes. The states of a HMM can be viewed as representing the sequence of columns in a multiple sequence alignment, with provisions for arbitrary position-dependent insertions and deletions in each sequence. The models are trained on a family of protein sequences using an expectation-maximization algorithm and a variety of algorithmic heuristics. A trained model can then be used to generate multiple alignments. There are two popular implementations of HMMs for MSA: SAM and HMMER (77,78), and according to Wistrand and Sonnhammer (49), SAM performs better in several aspects.

Looking at the extreme complexity of the optimality landscape of a MSA-problem containing several hundreds of sequences, it seems advisable to use an existing MSA (derived from heuristic methods), instead of unaligned sequences as a starting point for the HMM training. Then it might be more likely to end up near the global maximum, instead of some local optimum. However, all efforts to change settings and parameters did not yield any improvement of the given MSA, even worse, it obviously decreased the quality of the initial MSAs. Personal communication to Kevin Carplus, who is one of the authors of SAM, finally reassured that it is technically very unlikely to improve an existing MSA with a HMM, due to several reasons. Not only are the programs in the SAM package fine tuned for applications in structure prediction, and not thoroughly tested on generating MSAs, but also have heuristic methods (MUSCLE et cetera) become much more powerful in the recent past, so that they are now the best choice for generating high quality MSAs ( reviewed in 57).

| Program | MSA Strategy |
|---|---|
| Clustal (79) | Profile-based progressive MSA, implementation similar to Feng-Doolittle method |
| Mafft (80) | Tree generation via fast Fourier transform, tree-based iteration to refine the alignment by optimizing the weighted sum of pairs objective function |
| Muscle (81) | Iterative refinement method, initial rough draft of alignment built, using a crude guide tree to construct progressive alignment, profiles aligned upon log expectation score |
| Praline (82) | Homology extended MSA strategy, sequences are psi-blasted against the databases and together with their hits converted into profiles that are progressively aligned |
| Probcons (83) | Initial alignment using a pair-HMM generated with the maximum expected accuracy objective function, guide tree based upon sums of posterior probabilities, subalignments combined via sum of pairs, iterative refinement of the final alignment |

**Table 4**: overview of different popular approaches to the MSA problem.

The computational problem of generating MSAs is to define a model of sequence evolution that assigns probabilities to elementary sequence edits (i.e. character changes) and seeks a most probable directed graph in which edges represent edits and terminal nodes are the observed sequences. Until today, no feasible method for finding such a graph is known. The heuristic approach is to seek an MSA that optimizes the sum of pairs (SP) score, i.e. the sum of pairwise alignment scores. However, this task is NP complete (89) and can be achieved by dynamic programming with time and space complexity O($L^N$) in the sequence length L and number of sequences N (90). There is another alternative, a progressive method (91,92), which first estimates a tree and then constructs a pairwise alignment of the sequences within the subtrees found at each internal node. A subtree is represented by its profile, i.e. a multiple alignment is treated as a sequence by regarding each column as an alignable symbol. Most modern heuristics make use of this successful strategy.

Over the past 15 years, dozens of MSA programs and algorithms have been proposed. In order to compare them and measure accuracy, benchmark test sets have been designed, in the form of databases of precompiled alignments to which the alignments generated by test algorithms are compared. BAliBASE is the first large-scale purpose-built benchmark (93), and the five methods listed above are among the best-scoring ones. With input data as diverse as OR and V1R sequences, it is unsurprising that the output alignments do yield differences. Thus the dilemma, again, is to find the method that yields the most reasonable alignment.

If the best algorithm cannot be selected a priori, then it becomes a viable strategy to employ several alignment algorithms to construct alternative solutions, and subsequently to select the best among these. Fortunately, a recent publication (84) not only discusses these subjects, but also provides a tool (MUMSA) – accessible via online server – that compares several alignments. According to Lassmann and Sonnhammer, it is good to provide as many different alignments for this process, as possible, meaning MSAs of the same data-set, retrieved by different MSA-generating programs. The following methods were applied: CLUSTALX (79), MAFFT (80), MUSCLE (81), PRALINE (82) and PROBCONS (83). Unfortunately the set of sequences was too large for the complete praline process, after nine days the allowed memory consumption exceeded 1.5 GB, thus the praline alignment remains provisory.

### MSA quality assessment

Three different approaches were applied to get impressions of the quality of the alignments: sequence logos (114), MUMSA (84), and information content evaluation via PAUP (88). Sequence logos are a graphical representation of an MSA, which consist of stacks of symbols, one stack for each position in the sequence. While the height of symbols within the stack indicates the relative frequency of each AA at that position, the overall height of the stack represents the sequence conservation at that position. Thus, a sequence logo provides a richer and more precise description of a conserved position, than would a consensus sequence, and is very helpful in getting  a visual impression about the quality of an MSA an the patterns within.

The praline alignment, for example, shows fewer gaps than the ones derived by the other methods, but obviously fails to clearly identify conserved regions, as to be seen in the sequence logos.

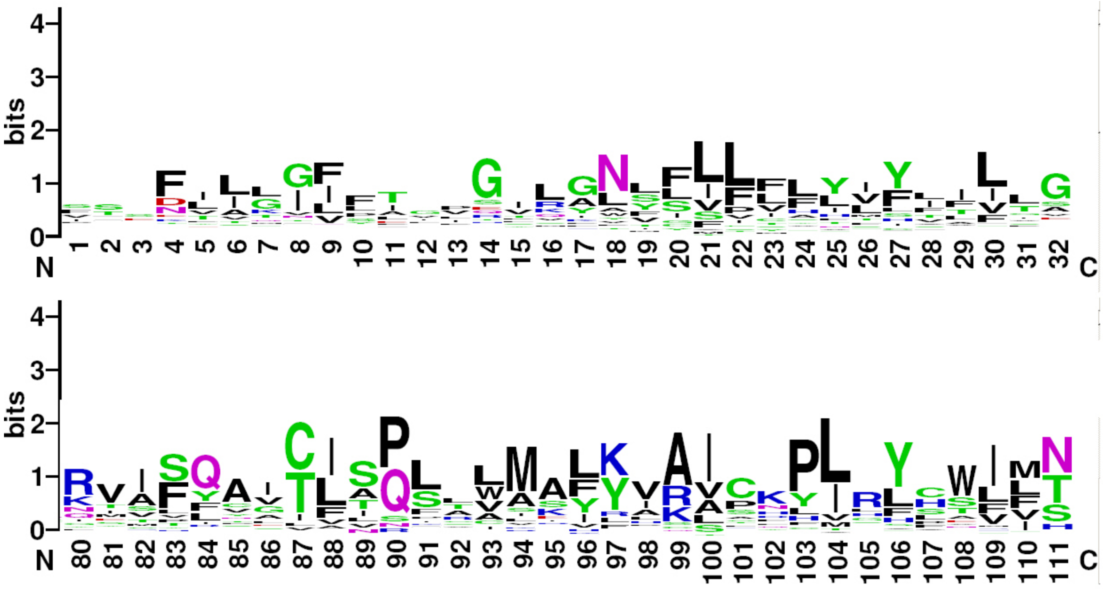


**Fig 8**: Sequence logo of PRALINE MSA (excerpt)

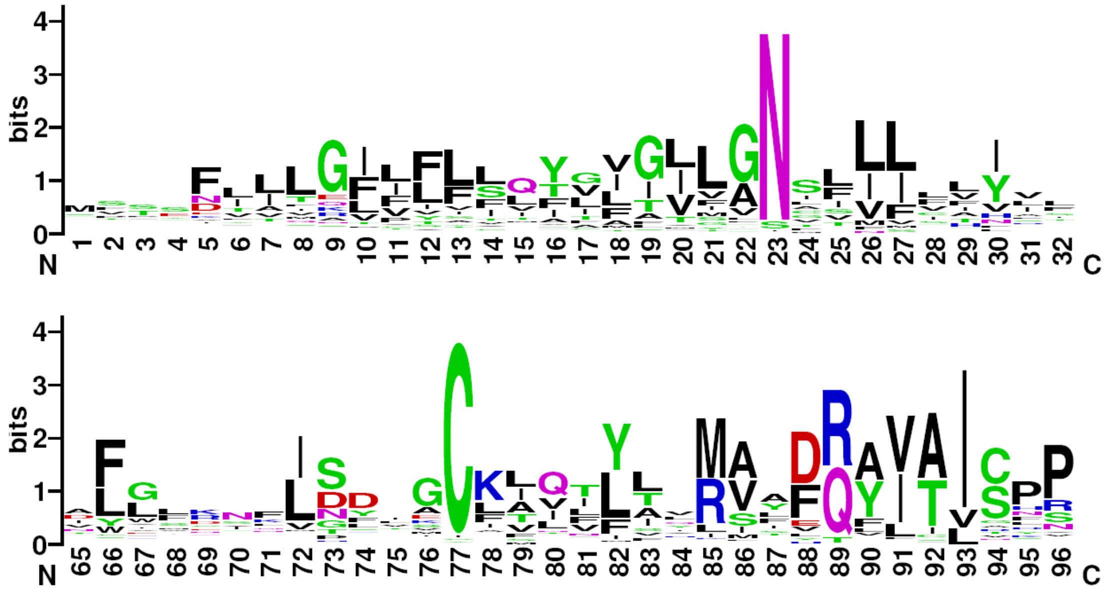


**Fig. 9**: Sequence logo of MUSCLE MSA (excerpt)

Next to each other, the sequence logos of the MSAs by MUSCLE (fig. 9) and PROBCONS (fig. 10) look very similar, in contrast to the rather flat PRALINE MSA (fig. 8). However, PROBCONS introduced a multitude of gaps when using default parameter settings, producing single columns of residues that lack biological sense.

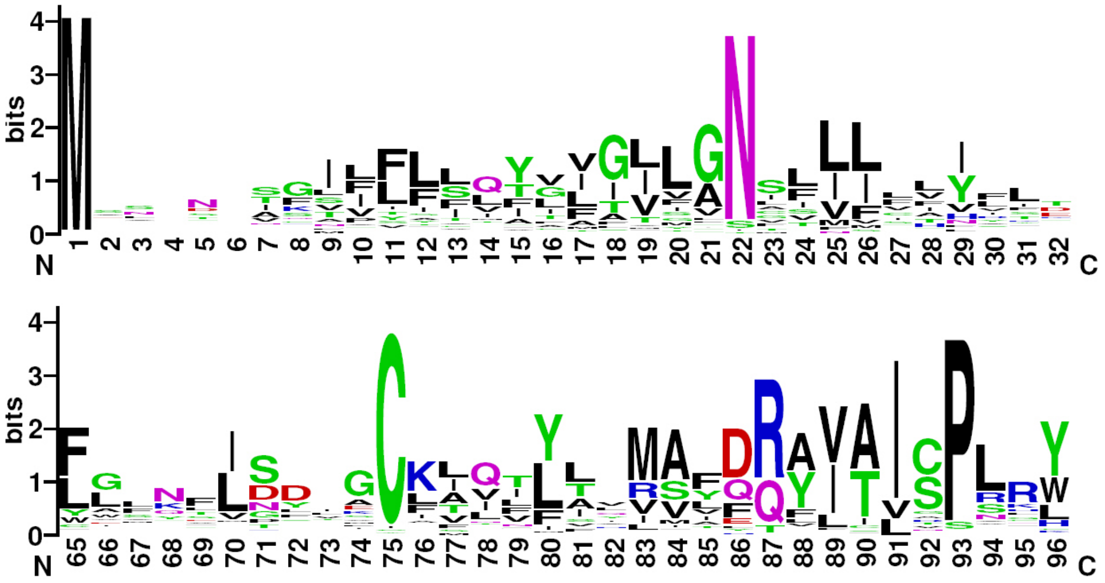


**Fig. 10**: sequence logo of PROBCONS MSA (excerpt)

The MSA by CLUSTAL looks similar to the MUSCLE one, but in the sequence logo presentation the differences become clearer:

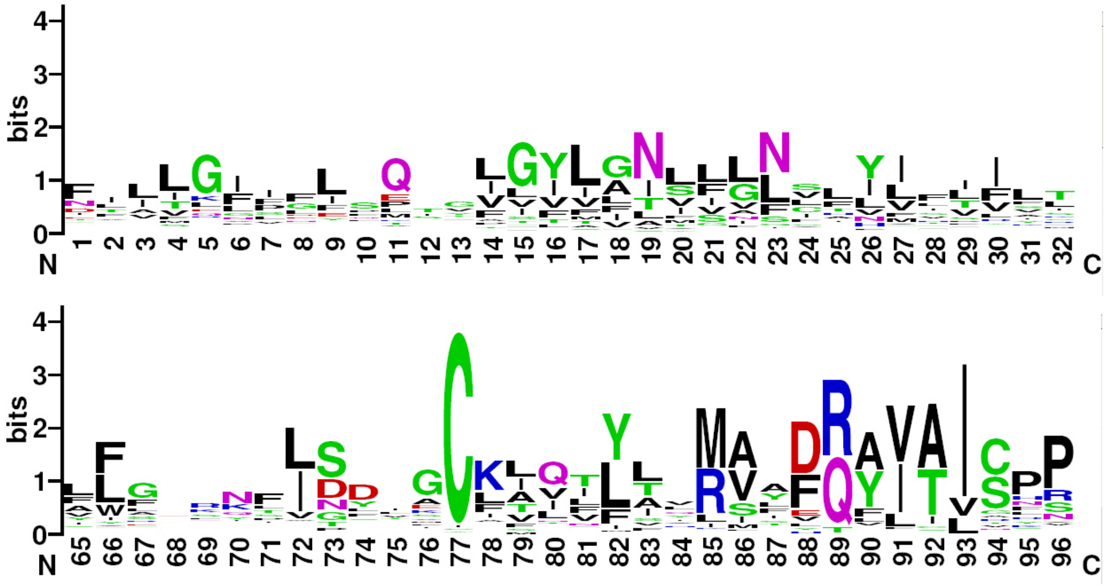


**Fig. 11**: Sequence logo of CLUSTAL MSA (excerpt)

In figure 11 (CLUSTALX), especially the beginning of the MSA does not exhibit as strong patterns as the alignments by MAFFT (fig. 12) or MUSCLE (fig. 9). Therefore, conserved regions were obviously not aligned sufficiently well in CLUSTAL.

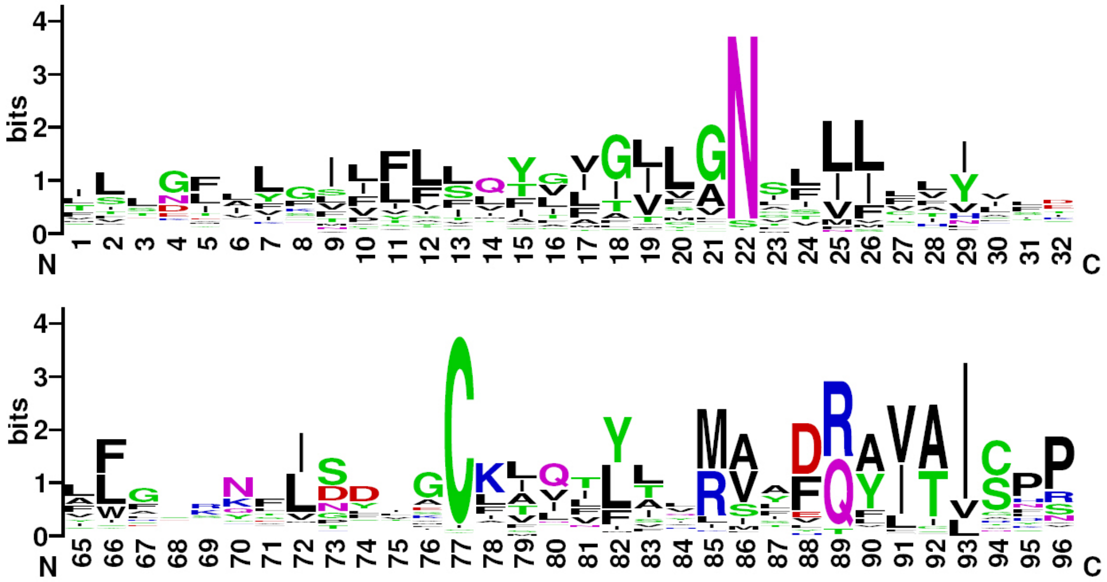


**Fig. 12**: Sequence logo of MAFFT MSA (excerpt)

The sequence logos were made of MSAs in which columns were eliminated containing more than 80% gaps. Here, only the beginning and a middle part of each logo are shown, featuring the MAYDRY motif typical of olfactory receptors.

For complete logos see additional files 2-6.

A more qualitative approach to assess and compare the quality of MSAs is a program called MUMSA (84), that aims at facilitating the decision, which alignment of the same input sequences is the most appropriate for the purpose at hand. It employs solely the criterion of inter-consistency, the similarity between alternate pairwise alignments of the same sequences. Basically, the program searches for regions which are identically aligned in different alignments, assuming that these are more reliable than regions that differ from MSA to MSA. Thereby, the method also estimates the difficulty of the alignment case and assesses the quality of individual alignments. Lassmann et. al. assume that in difficult cases alignment programs will generate more dissimilar alignments than in trivial cases.

It has been shown on the Balibase benchmark set that even methods with a low average accuracy outperform the best methods in many individual cases (81). Thus, as recommended, several different methods were applied; MUSCLE and MAFFT were even run several times. Motivation for this was the fact that the alignments generated prior to the respective iterative refinement are often quite different from the final alignment. Here, MUSCLE was restricted to two iterations of refinement steps, and MAFFT was applied in two different run modes (medium and slow) using two different Matices, PAM30 and PAM45, for rather distantly related sequences. Altogether, six different alignments were generated as input for MUMSA: probcons, muscle_max2, muscle, mafft30, mafft45, and clustal. Figure 11 shows the MUMSA multiple overlap scores, with the values of standard deviation and mean on top, average overlap score (AOS) below. Obviously, none of the alignment scores is significantly below or above the mean, only the score of the MSA by ClustalX clearly appears worse than the rest, however without significant statistical support. Ultimately, the results of MUMSA indicate that the input MSAs are of similar quality, except maybe ClustalX, and the differences between them are too smal to allow a statistically sound preference. The AOS of 0.55, however, clearly states that my alignment case is rather difficult. Basically, this is a crude measure of how dispersed alignments are in the space of all solutions, approaching 1 in simple alignment cases with similar alignments, and 0 in difficult cases.

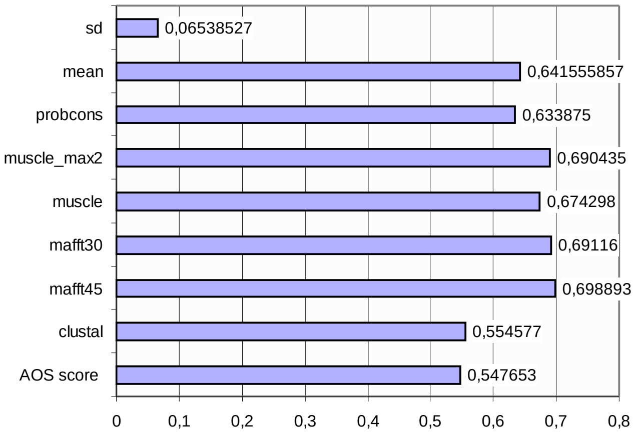


**Fig. 11:** MUMSA multiple overlap scores

The MOS scores, however, appear more encouraging. Lassmann et. al. state that a cutoff of 0.8 MOS is practical for trusting the quality of an alignment, while 0.5 MOS or less implies that the alignment is probably incorrect. Except the scores of ClustalX and PROBCONS, all other scores are just below 0.7, which is not yet completely convincing, but closer to the trustworthy 0.8 cutoff than to the 0.5 threshold of unreliability. In summary, MUMSA failed to clearly differentiate between superior and inferior alignments in this case, which indicates that the best methods produced comparable results. This outcome, however, should not be surprising, because otherwise the choice of program would have been trivial beforehand. After all, there is yet another way to distinguish between different levels of alignment quality.

Phylogenetic Analysis Using Parsimony (PAUP) is one of the most widely used software packages for the inference of evolutionary trees. One of its features is the generation and evaluation of random trees, which in turn provides insight into the information content of an MSA. The tree-lengths are calculated and represented in a diagram, a bell-shaped curve. If the MSA contains information, this curve will reveal a skewness, leaning to one side instead of being perfectly symmetrical. As a negative control, I checked the information content of the sequences after permutation (using PHYLIP). Shown in figure 12, the skewness value remains close to zero in this case. For more details about the calculation of skewness and inference of information content, see “methods”.

As to be seen in figure 12, the skewnessvalues for the different MSAs according to PAUP are also nearly equally distributed, and, except for the negative control (permuted data), no alignment score is significantly above or below the mean. Even though the differences between the scores never exceed two standard deviations, it is

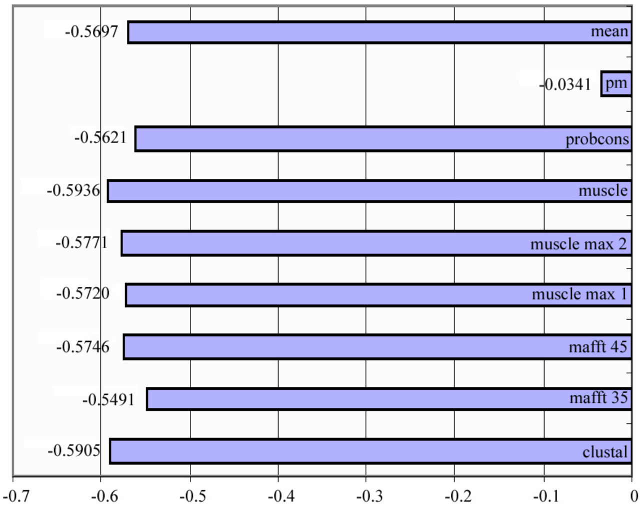
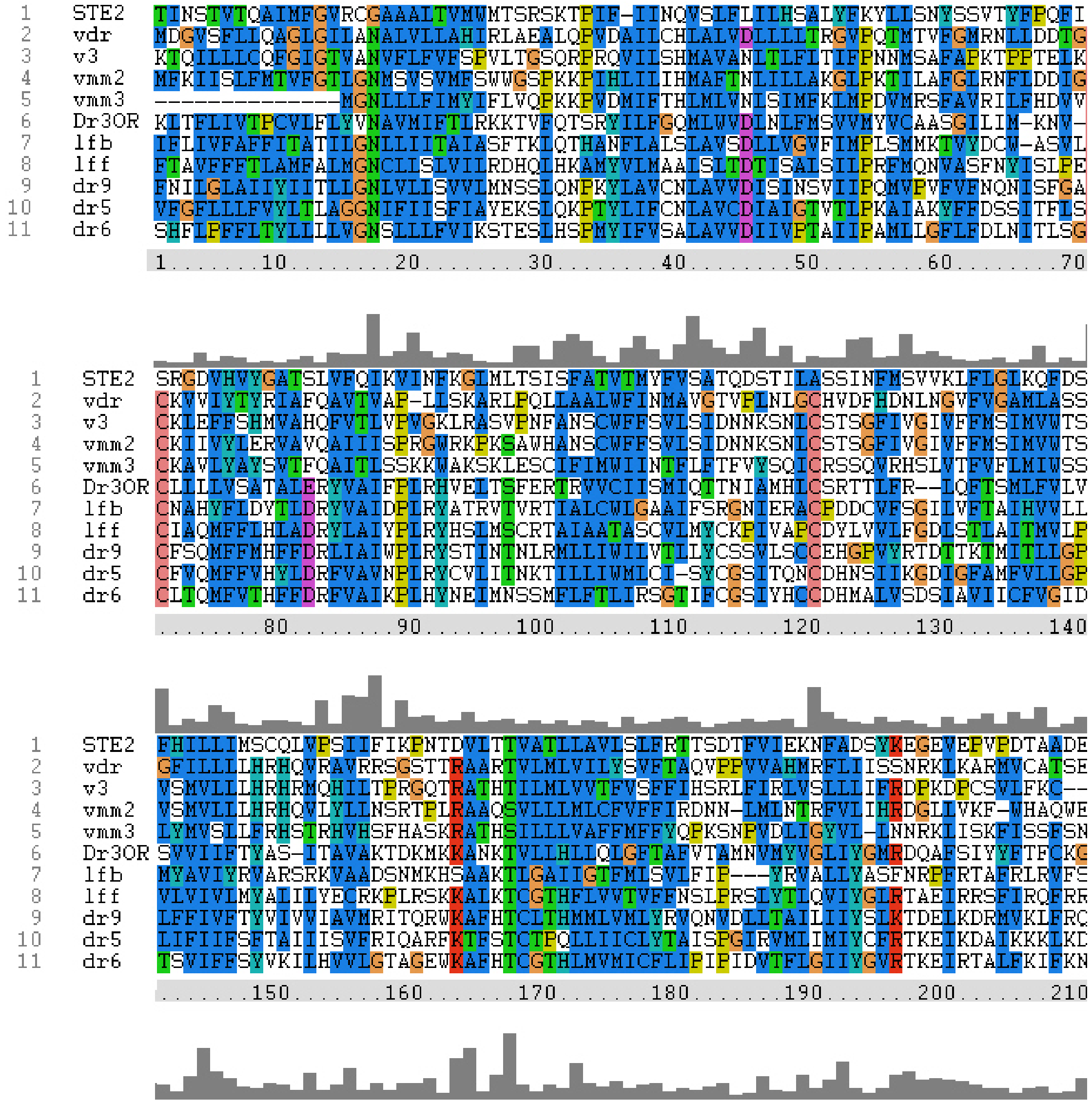


**Fig. 12 :** skewness values calculated by PAUP

obvious that MUSCLE and CLUSTAL received highest scores, which are quite similar. Altogether, neither MUMSA nor PAUP offer substantial statistical support to prefer any of the six alignments. Nevertheless, a reasonable decision needs to be made, even if the choices obviously fail to differ significantly. Out of the three approaches to assess alignment quality, PAUP’s calculation of the information content should be regarded as the most important measure, because the information accumulated in an alignment should certainly be maximized, since this is the desired output of these efforts. Thus the hierarchy of criteria used is the following, in decreasing order: information content, multiple overlap and finally, consistency of sequence logos. Now, according to PAUP, the alignments by MUSCLE and CLUSTAL yield the highest information content. However, CLUSTAL’s alignment scores much lower on MUMSA, and visual inspection of the sequence logo also reveals the drawbacks of this alignment. Thus the final choice is the alignment by MUSCLE, shown in additional file 7. Following guide tree construction, the fundamental step of the program MUSCLE is pairwise profile alignment, which is used first for progressive alignment and then for refinement.

An excerpt of the complete MUSCLE alignment, featuring eleven representative sequences, is shown in figure 13. Here, sites are excluded that reveal gaps in more than 2 columns.



```
 1  STE2   TINSTVTQAIMFGVRCGAAALTVMWMTSRSKTPIF-IINQVSLFLIILHSALYFKYLLSNYSSVTYFPQFI
 2  vdr    MDGVSFLIQAGLGILANALVLLAHIRLAEALQPVDAILCHLALVDLLLLTRGVPQTMTVFGMRNLLIDDTG
 3  v3     KTQILLLCQFGIGTVANVFLFVFSPVLTGSQRPRQVILSHMAVANLTLFLTIFPNNMSAFAPKTPPTELK
 4  vmm2   MFKIISLFMTVFGTLGNMSVSVMFSWWGSPKKPIHLILIHMAFTNLILLAKGLPKTILAFGLRNFIDDIG
 5  vmm3   --------------MGNLLLFIMYIFLVQPKKPVDMIFTHLMLVNLSIMFKLMPDVMRSFAVRILFHDVV
 6  Dr3OR  KLTFLIVTPCVLFLYVNAVMIFTLRKKTVFQTSRYILFGQMLWVDLNLFMSVVMYVCAASGILIM-KNV-
 7  lfb    IFLIVFAFFITATIILGNLLIITAIASFTKLQTHANFLALSLAVSDLLVGVFIMPLSMMKTVYDCW-ASVL
 8  lff    FTAVFFFTLAMFALMGNCLLSLVILRDHQLHKAMYVLMAALSLTDTISALSIILPRFMQNVASFNYISLPE
 9  dr9    FNILGLAIIYIITLLGNLVLLSVVLMNSSLQNPKYLAVCNLAVVDISINSVIIPQMVPVFVFNQNISFGA
10  dr5    VFGFLLLFVYLTLAGGNIFIISFIAYEKSLQKPTYLIFCNLAVCDIAIGTVTLPKAIAKYFFDSSITFIS
11  dr6    SHFLPFFLTYLLLLVGNSLLLFVIKSTESLHSPMYIFVSALAVVDIIVPTAIIPAMLLGFLFDLNITLSG
           1.......10........20........30........40........50........60........70

 1  STE2   SRGDVHVYGATSLVFQIKVINFKGLMLTSISFATVTMYFVSATQDSTILASSINFMSVVKLFLGLKQFDS
 2  vdr    CKVVIYTYRIAFQAVTVAP-LLSKARLPQLLAALWFINMAVGTVPLNLGCHVDFHDNLNGVFVGAMLASS
 3  v3     CKLEFFSHMVAHQFVTLVPVGKLRASVPNFANSCWFFSVLSIDNNKSNLCSTSGFIVGIVFFMSIMVWTS
 4  vmm2   CKIIVYLERVAVQAIIISPRGWRKPKSAWHANSCWFFSVLSIDNNKSNLCSTSGFIVGIVFFMSIMVWTS
 5  vmm3   CKAVLYAYSVTFQAITLSSKKWAKSKLESCIFIMWIINTFLFTFVYSQICRSSQVRHSLVTFVFLMIWSS
 6  Dr3OR  CLLLLVSATALERYVAIFFPLRHVELTSFERTRVVCIISMIQTTNIAMHLCSRTTLFR--LQFTSMLFVLV
 7  lfb    CNAHYFLDYTLDRYVAIDPLRYATRVTVRTIALCWLGAAIFSRGNIERACPDDCVFSGILVFTAIHVVLL
 8  lff    CIAQMFFLHLADRYLAIYPLRYHSLMSCRTAIAATASCVLMYCKPIVAPCDYLVVLFGDLSTLAITMVLP
 9  dr9    CFSQMFFMHFFDRLIAIWPLRYSTINTNLRMLLIWILVTLLYCSSVLSCCEHGPVYRTDTTKTMITLLGP
10  dr5    CFVQMFFVHYLDRFVAVNPLRYCVLITNKTILLIWMLCI-SYCGSITQNCDHNSIIKGDIGFAMFVLLGP
11  dr6    CLTQMFVTHFFDRFVAIKPLHYNEIMNSSMFLFTLIRSGTIFCGSIYHCCDHMALVSDSIAVIICFVGID
           .........80........90.......100.......110.......120.......130.......140

 1  STE2   FHILLIMSCQLVPSIIFIKPNTDVLTTVATLLAVLSLFRTTSDTFVIEKNFADSYKEGEVEPVPDTAADE
 2  vdr    GFILLLLHRHQVRAVRRSGSTTRAARTVLMLVILYSVFTAQVPPVVAHMRFLIISSNRKLKARMVCATSE
 3  v3     VSMVLLLHRHRMQHILTPRGQTRATHTILMLVVTFVSFFIHSRLFIRLVSLLLIFRDPKDPCSVLFKC--
 4  vmm2   VSMVLLLHRHQVLYLLNSRTPLRAAQSVLLLMLCFVFFIRDNN-LMLNTRFVLIHRDGLLVKF-WHAQWE
 5  vmm3   LYMVSLLFRHSTRHVHSFHASKRATHSILLLVAFFMFFYQPKSNPVDLIGYVL-LNNRKLISKFISSFSN
 6  Dr3OR  SVVIIFTYAS-ITAVAKTDKMKKANKTVLLHLLQLGFTAFVTAMNVMYVGLIYGMRDQAFSIYYFTFCKG
 7  lfb    MYAVIYRVARSRKVAADSNMKHSAAKTLGAIIGTFMLSVLFIP---YRVALLYASFNRPFRTAFRLRVFS
 8  lff    VLVIVLMYALIILYECRKPLRSKKALKTCGTHFLVVTVFFNSLPRSLYTLQVIYGLRTAEIRRSFIRQFRR
 9  dr9    LFFIVFTYVIVVIAVMRITQRWKAFHTCLTHMMLVMLYRVQNVDLLTAILIIYSLKTDELKDRMVKLFRQ
10  dr5    LIFIIFSFTAIIISVFRIQARFKTFSTCTPQLLIICLYTAISPGIRVMLIMIYCFRTKEIKDAIKKKLKD
11  dr6    TSVIFFSYVKILHVVLGTAGEWKAFHTCGTHLMVMICFLIPIPIDVTFLGIIYGVRTKEIRTALFKIFKN
           .......150.......160.......170.......180.......190.......200.......210
```

**Fig. 13**: Excerpt of the MUSCLE multiple sequence alignment, viewed in the CLUSTALX editor. Eleven representative sequences are shown, numbers 2-5 of the VNO, 6-11 of the MOS. The first sequence will be used as outgroup in upcoming tree calculations. Residue colouring patterns are those of the CLUSTALX editor standards.

## Tree search

All previous steps aimed at generating an optimal alignment. Finally, this work can now be taken to the next level, the search for the maximum likelihood tree, based upon this alignment. In general, the tree-seeking scientist has the following approaches to choose from, depending on the optimality criterion employed: parsimony, distance data, and maximum likelihood.

The strategies based upon distance data are also called a phenetic approaches (94). Here, data are clustered according to their similarities, minimizing the total distance among taxa in the output tree. There has been a long-standing controversy over whether distance data contain reliable information that can be used to infer phylogenies. After all, distance methods should only be applied if the characters used to compute the distance are unlikely to demonstrate convergence (95,96).

When applying parsimony every character is evaluated and the preferred tree topology is the one that minimizes the total amount of evolutionary change that has occurred. Maximum parsimony is a very simple approach, and is popular for this reason. However, it is not statistically consistent. Particularly it assigns every exchange the same cost without regard of AA similarity, thereby disregarding a lot of information contained in the MSA. Thus, they collapse the information content of every sequence pair to one number, the distance/similarity. That way, it is not guaranteed to produce the true tree with high probability, given sufficient data.

Maximum likelihood is a general methodology in the field of mathematical statistics, which can also be applied to the task of estimating phylogenies. Basically, the idea is to maximize the Bayesian likelihood that a given model generated the observed sequence data. More detailed, the question becomes: Given a formula describing the probability that different types of AA substitutions will occur, and given a particular tree, how likely am I to obtain this particular set of AA sequences? Simulation and other theoretical studies have shown that when rates of change vary among lineages, maximum likelihood can outperform parsimony methods in recovering the correct tree (reviewed in 97).

The reasons to prefer the Maximum Likelihood approach are the following. ML estimation often has lower variance than other methods, ie. it is frequently the estimation method least affected by sampling error, and it is statistically well founded. Also does it tend to be robust to many violations of the assumptions in the evolutionary model, it evaluates different tree topologies and uses all the sequence information. Thus, even with very short sequences, ML methods tend to outperform alternative approaches such as parsimony or distance methods. Basically, merely two serious implementations of ML estimation in tree reconstruction exist, TREE-PUZZLE (101) and MOLPHY (102).

In any case, the search for the optimal tree is an NP-hard problem, because the number of possible trees grows polynomially in the number of taxa. Simply put this means that time - and/or memory resources necessary to solve it will not suffice for more than an extremely limited set of taxa. Therfore, most methods employ heuristics in some step of the calculation.

As the name already states, TREE-PUZZLE makes use of a heuristic method, the quartet-puzzling algorithm (85). Quartet trees are unrooted and feature four taxa. In its first step, all quartet tree topologies are generated from the set of sequences and evaluated to get a set of quartets, which are best supported by the underlying sequence alignment. According to their posterior probability weights in the ML estimation, quartet topologies are chosen and recursively combined to generate intermediate trees. Finally, a ML consensus tree is created based upon those partial trees.

For my 230 sequences, this procedure took several days on a local PC. MOLPHY on the other hand is the implementation of an exhaustive tree search algorithm. Therefore the number of sequences this program mode is still feasible for, is very limited. To optimise the signal-to-noise ratio, the data set should certainly be as large as time and memory resources allow for a sophisticated ML analysis. Since the number of possible tree topologies increases explosively with the number of taxa (86), it is a serious problem to find the best tree among the huge number of alternatives. Until today, this complex calculation cannot be carried out for more than eleven sequences in an exhaustive tree search, since more than 34 million trees need to be evaluated in that case. Thus the rough draft tree estimation by TREE-PUZZLE serves the purpose of providing the information needed to chose eleven representative sequences. This tree is shown in fig. 14 on the next page, with taxa of OR sequences coloured in red, and those of V1R sequences coloured in blue. The eleven sequences chosen for further tree search are highlighted in yellow, featuring six OR sequences of the lamprey and the zebrafish (*danio rerio*), and five V1R sequences of the mouse and the zebrafish. The criterium for selecting those eleven representatives is the intention to use sequences that do not cluster together closely but are rather placed in different branches of the big tree. One of the lamprey receptors was chosen as outgroup in the data set of the eleven representative sequences, but proved unsuited for this task. Presumably it is not older than most other sequences, since it usually clustered together with olfactory sequences in narrow branches. Thus, in the further tree search process, a new outgroup was substituted for one of the mouse receptors.

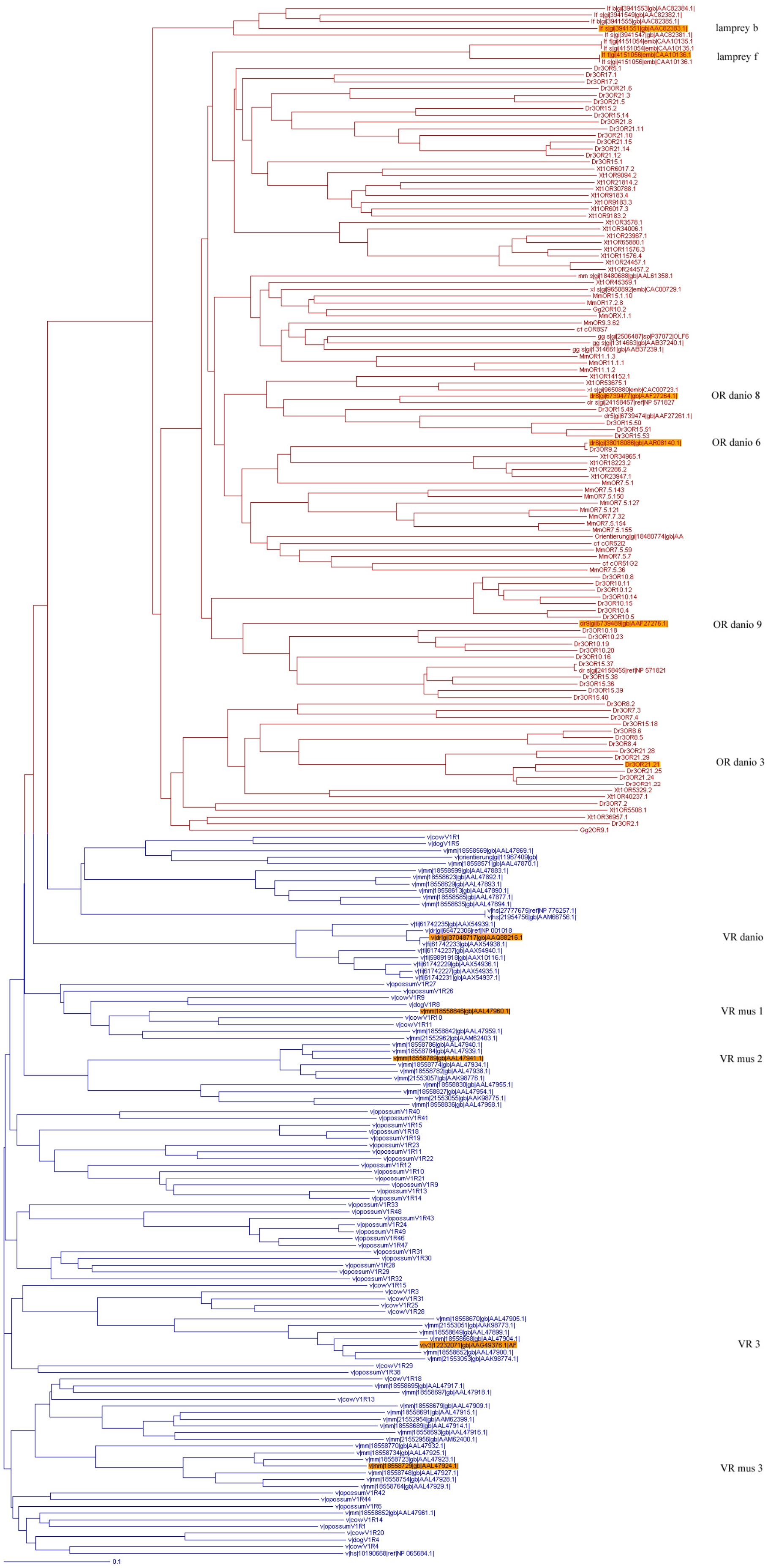

lf b|gi|3941553|gb|AAC82384.1|
lf s|gi|3941549|gb|AAC82382.1|
lf b|gi|3941555|gb|AAC82385.1|
lf s|gi|3941551|gb|AAC82383.1|
lamprey b
lf s|gi|3941547|gb|AAC82381.1|
lf f|gi|4151054|emb|CAA10135.1
lf s|gi|4151054|emb|CAA10135.1
lf f|gi|4151056|emb|CAA10136.1
lamprey f
lf s|gi|4151056|emb|CAA10136.1
Dr3OR5.1
Dr3OR17.1
Dr3OR17.2
Dr3OR21.6
Dr3OR21.3
Dr3OR21.5
Dr3OR15.2
Dr3OR15.14
Dr3OR21.8
Dr3OR21.11
Dr3OR21.10
Dr3OR21.15
Dr3OR21.14
Dr3OR21.12
Dr3OR15.1
Xt1OR6017.2
Xt1OR9094.2
Xt1OR21814.2
Xt1OR30788.1
Xt1OR9183.4
Xt1OR9183.3
Xt1OR6017.3
Xt1OR9183.2
Xt1OR3578.1
Xt1OR34006.1
Xt1OR23967.1
Xt1OR65880.1
Xt1OR11576.3
Xt1OR11576.4
Xt1OR24457.1
Xt1OR24457.2
mm s|gi|18480688|gb|AAL61358.1
Xt1OR45359.1
xl s|gi|9650892|emb|CAC00729.1
MmOR15.1.10
MmOR17.2.8
Gg2OR10.2
MmORX.1.1
MmOR9.3.62
cf cOR8S7
gg s|gi|2506487|sp|P37072|OLF6
gg s|gi|1314663|gb|AAB37240.1|
gg s|gi|1314661|gb|AAB37239.1|
MmOR11.1.3
MmOR11.1.1
MmOR11.1.2
Xt1OR14152.1
Xt1OR53675.1
xl s|gi|9650880|emb|CAC00723.1
dr8|gi|6739477|gb|AAF27264.1|
OR danio 8
dr s|gi|24158457|ref|NP 571827
Dr3OR15.49
dr5|gi|6739474|gb|AAF27261.1|
Dr3OR15.50
Dr3OR15.51
Dr3OR15.53
dr6|gi|38018086|gb|AAR08140.1|
OR danio 6
Dr3OR9.2
Xt1OR34965.1
Xt1OR18223.2
Xt1OR2286.2
Xt1OR23947.1
MmOR7.5.1
MmOR7.5.143
MmOR7.5.150
MmOR7.5.127
MmOR7.5.121
MmOR7.7.32
MmOR7.5.154
MmOR7.5.155
Orientierung|gi|18480774|gb|AA
cf cOR52I2
MmOR7.5.59
MmOR7.5.7
cf cOR51G2
MmOR7.5.36
Dr3OR10.8
Dr3OR10.11
Dr3OR10.12
Dr3OR10.14
Dr3OR10.15
Dr3OR10.4
Dr3OR10.5
dr9|gi|6739489|gb|AAF27276.1|
OR danio 9
Dr3OR10.18
Dr3OR10.23
Dr3OR10.19
Dr3OR10.20
Dr3OR10.16
Dr3OR15.37
dr s|gi|24158455|ref|NP 571821
Dr3OR15.38
Dr3OR15.36
Dr3OR15.39
Dr3OR15.40
Dr3OR8.2
Dr3OR7.3
Dr3OR7.4
Dr3OR15.18
Dr3OR8.6
Dr3OR8.5
Dr3OR8.4
Dr3OR21.28
Dr3OR21.29
Dr3OR21.21
OR danio 3
Dr3OR21.25
Dr3OR21.24
Dr3OR21.22
Xt1OR5329.2
Xt1OR40237.1
Dr3OR7.2
Xt1OR5508.1
Xt1OR36957.1
Dr3OR2.1
Gg2OR9.1
v|cowV1R1
v|dogV1R5
v|mm|18558569|gb|AAL47869.1|
v|orientierung|gi|11967409|gb|
v|mm|18558571|gb|AAL47870.1|
v|mm|18558599|gb|AAL47883.1|
v|mm|18558623|gb|AAL47892.1|
v|mm|18558629|gb|AAL47893.1|
v|mm|18558613|gb|AAL47890.1|
v|mm|18558585|gb|AAL47877.1|
v|mm|18558635|gb|AAL47894.1|
v|hs|27777675|ref|NP 776257.1|
v|hs|21954756|gb|AAM66756.1|
v|fi|61742235|gb|AAX54939.1|
v|dr|gi|66472306|ref|NP 001018
v|dr|gi|37048717|gb|AAQ88216.1
VR danio
v|fi|61742233|gb|AAX54938.1|
v|fi|61742237|gb|AAX54940.1|
v|fi|59891918|gb|AAX10116.1|
v|fi|61742229|gb|AAX54936.1|
v|fi|61742227|gb|AAX54935.1|
v|fi|61742231|gb|AAX54937.1|
v|opossumV1R27
v|opossumV1R26
v|cowV1R9
v|dogV1R8
v|mm|18558846|gb|AAL47960.1|
VR mus 1
v|cowV1R10
v|cowV1R11
v|mm|18558842|gb|AAL47959.1|
v|mm|21552962|gb|AAM62403.1|
v|mm|18558786|gb|AAL47940.1|
v|mm|18558784|gb|AAL47939.1|
v|mm|18558789|gb|AAL47941.1|
VR mus 2
v|mm|18558774|gb|AAL47934.1|
v|mm|18558782|gb|AAL47938.1|
v|mm|21553057|gb|AAK98776.1|
v|mm|18558830|gb|AAL47955.1|
v|mm|18558827|gb|AAL47954.1|
v|mm|21553055|gb|AAK98775.1|
v|mm|18558836|gb|AAL47958.1|
v|opossumV1R40
v|opossumV1R41
v|opossumV1R15
v|opossumV1R18
v|opossumV1R19
v|opossumV1R23
v|opossumV1R11
v|opossumV1R22
v|opossumV1R12
v|opossumV1R10
v|opossumV1R21
v|opossumV1R9
v|opossumV1R13
v|opossumV1R14
v|opossumV1R33
v|opossumV1R48
v|opossumV1R43
v|opossumV1R24
v|opossumV1R49
v|opossumV1R46
v|opossumV1R47
v|opossumV1R31
v|opossumV1R30
v|opossumV1R28
v|opossumV1R29
v|opossumV1R32
v|cowV1R15
v|cowV1R3
v|cowV1R31
v|cowV1R25
v|cowV1R28
v|mm|18558670|gb|AAL47905.1|
v|mm|21553051|gb|AAK98773.1|
v|mm|18558649|gb|AAL47899.1|
v|mm|18558668|gb|AAL47904.1|
v|v3|12232071|gb|AAG49376.1|AF
VR 3
v|mm|18558652|gb|AAL47900.1|
v|mm|21553053|gb|AAK98774.1|
v|cowV1R29
v|opossumV1R38
v|cowV1R18
v|mm|18558695|gb|AAL47917.1|
v|mm|18558697|gb|AAL47918.1|
v|cowV1R13
v|mm|18558679|gb|AAL47909.1|
v|mm|18558691|gb|AAL47915.1|
v|mm|21552954|gb|AAM62399.1|
v|mm|18558689|gb|AAL47914.1|
v|mm|18558693|gb|AAL47916.1|
v|mm|21552956|gb|AAM62400.1|
v|mm|18558770|gb|AAL47932.1|
v|mm|18558734|gb|AAL47925.1|
v|mm|18558723|gb|AAL47923.1|
v|mm|18558729|gb|AAL47924.1|
VR mus 3
v|mm|18558748|gb|AAL47927.1|
v|mm|18558754|gb|AAL47928.1|
v|mm|18558764|gb|AAL47929.1|
v|opossumV1R42
v|opossumV1R44
v|opossumV1R6
v|mm|18558852|gb|AAL47961.1|
v|cowV1R14
v|opossumV1R1
v|cowV1R20
v|dogV1R4
v|cowV1R4
v|hs|10190668|ref|NP 065684.1|
0.1

**Fig. 14 (previous page)**: Preliminary tree of OR (red) and V1R (blue) sequences that are no closer related to each other than 40% and 45%. This phylogram is based upon the alignment by MUSCLE and was calculated by TREE-PUZZLE. The tree visualization program used is TREEVIEW X. Based upon this tree, eleven representative sequences (highlighted in yellow) were chosen for further tree calculations. Next to these, written in black, is the name they are called in further explorations, for the sake of recognition. Taxa names without access numbers are named according to Niimura (103).

Among others, the MOLPHY package features a program called PROTML, that applies the ML estimation to AA sequences. The method is an advanced version of Felsenstein's DNAML. The straightforward approach to inferring a tree would be to evaluate all possible tree topologies one after another and pick the one which gives the highest maximum likelihood. Again, this is computationally prohibitive. Therefore, the exhaustive tree search was performed using approximate likelihood, which basically carries out a tree search applying a criterion that is computationally less expensive than maximum likelihood but chooses similar trees. The top 1000 trees were then evaluated via ML, to choose the best tree. Even thought Molphy is capable of both steps, the evaluation of the 1000 best trees was left to PAML, since the AA substitution matrixes are more up-to date in this program package.

Table 5 shows the top ten trees according to the PAML estimations. The ranking is based upon their ML scores, assigning tree number 394 the highest likelihood to be the correct tree. However, differences between the ML scores are very small. Regarding the statistical support, PAML provides a score called pRELL, bootstrap proportions, calculated via the RELL method (109), as well as the method of Shimodaira and Hasegawa (110) with a correction for multiple comparison. This measure is applied in order to assess the significance of the difference in significance of likelihood scores

| *Tree number* | *ML score* | *pRELL* |
|---|---|---|
| 394 | -5748.017 | 0.081 |
| 428 | -5748.586 | 0.060 |
| 488 | -5748.733 | 0.056 |
| 411 | -5748.964 | 0.055 |
| 979 | -5749.016 | 0.123 |
| 391 | -5749.184 | 0.006 |
| 536 | -5749.236 | 0.051 |
| 251 | -5749.365 | 0.005 |
| 443 | -5749.399 | 0.044 |
| 507 | -5749.532 | 0.044 |

**Table 5:** Top ten trees according to PAML

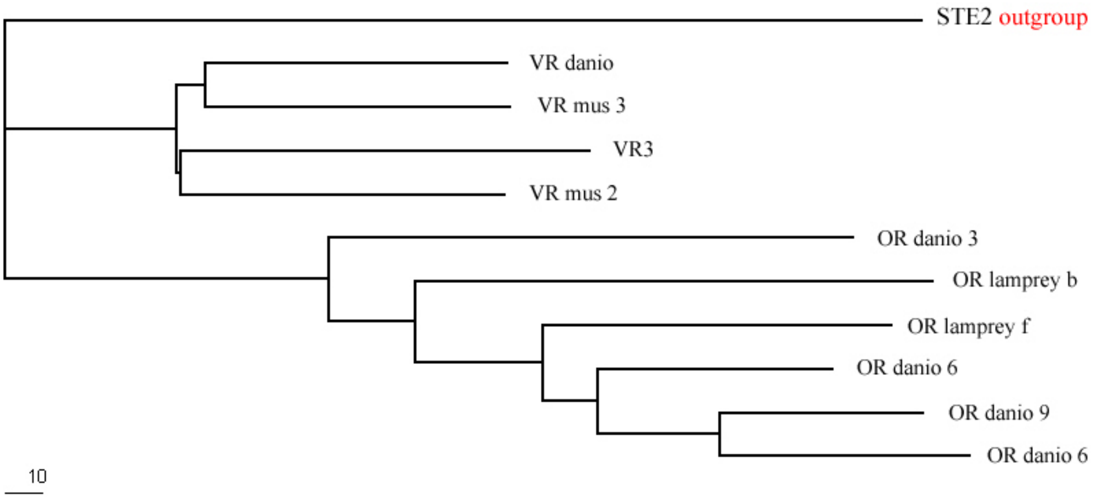


**Fig. 15:** Maximum Likelihood tree of eleven representative sequences of ORs and V1Rs.

of the tree topologies of the 1000 trees evaluated. Tree 394 (fig. 15) features the best likelihood value (-5748.017), and the second best pRELL (0.081). Since the sum of all pRELL values of these trees becomes 1, this is actually a high score. Furthermore, only 121 tree topologies have a pRELL greater than zero. Out of these, 99 trees support the main feature of the V1Rs branching off earlier than ORs and their scores add up to 0.945 pRELL or 95.8%. The top ten trees listed in table 5 also share this feature, additional file 8 shows their topologies in detail.

The above phylogram can be regarded as the final result of this work. It presents evolutionary relationships between receptors of the main olfactory system and the vomeronasal system. The new outgroup was chosen with some consideration, since it should be related to all other sequences without being to closely related or biased towards one of the two groups. STE2 is an evolutionary ancient receptor found in yeast, which clearly belongs to the GPCR superfamily (112), and according to INTERPRO (99) the STE2 receptor bears its own unique '7TM' signature, without obvious sequence similarity to other GPCR families. For all other chordate GPCRs it is a priori unclear whether they are ancestors or descendants of chemosensory receptors. A protist GPCR however cannot be derived from chordate chemoreceptors,

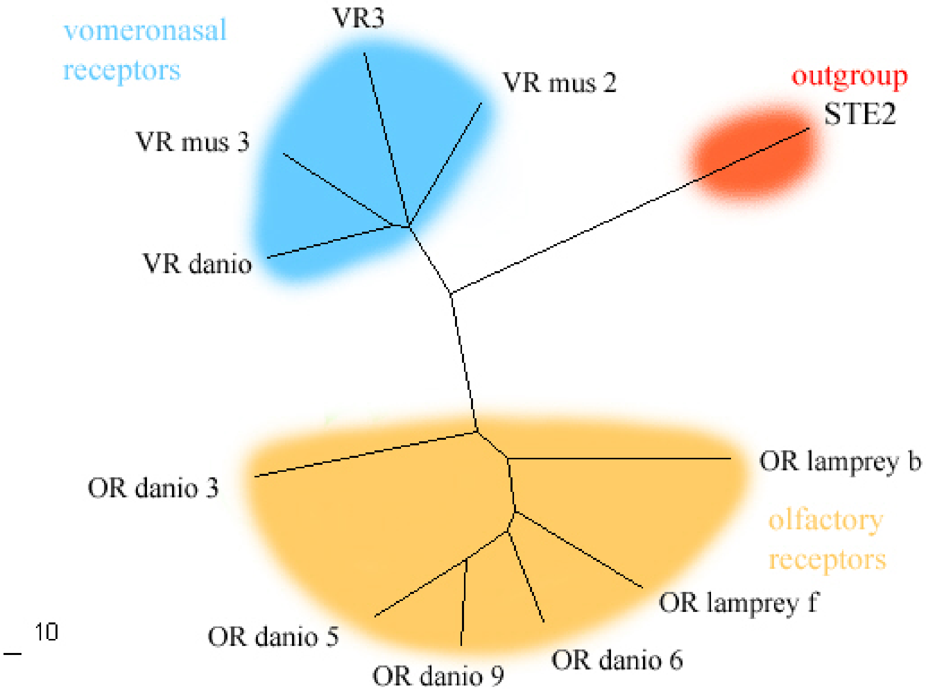


**Fig. 16:** Radial tree representation of the ML tree with receptor groups marked in different colours (outgroup – red, ORS – yellow, V1Rs – blue)

which makes it an acceptable outgroup of an early ancestor. Fig. 15 and 16 also reveal the way in which the two major groups, ORs and V1Rs are separated. Interestingly, the branch of the VRs diverges evolutionarily earlier, indicating the ancestral character of the vomeronasal system. Nevertheless, the exact ranking of sequences within the two major subtrees is of negligible importance to the meaning of this work.

# Discussion

In this project I sought the answer to a scientific question, assessing and applying the best bioinformatic methods available today. This involved sequence retrieval from databanks, exploration and extraction of regions under positive selection from the sequence material, reduction of the number of sequences according to their similarity, multiple sequence alignment and maximum likelihood estimation to find the tree that supports the data best. Since this work is based on sequence data and programs, and conducted solely on the computer, careful estimation and interpretation of the results at every step was of great importance. Embedded in a life science environment it is easier to identify wrong results than with calculations based upon models that represent nature always as an approximation. But methods of statistical evaluation were improved along with bioinformatics programs. Therefore, the answer to the question, which system of smell is the evolutionarily older one, can be provided with a certain amount of reliability.

Initially, a basis of information was generated, and sequence data of both kinds of receptors (olfactory and vomeronasal) were retrieved from the databanks. Here, the quality criteria were the following: Pseudogenes and partial sequences were not considered, since they evolve at different rates due to the lack of selective pressure. Furthermore, only sequences were gathered, that have been annotated in previously published papers, to make sure the input sequence data is as reliable as possible. The source organisms were chosen as representatives of the chordates. Few genomes have been sequenced yet, and of those, e.g. frog, mouse, dog, the complete olfactory subgenome was taken into the further calculations, to maximize the spectrum of sequences.

To avoid possible false interpretations of the phylogenetic reconstruction, this sequence data was carefully pre-processed.  Presenting the largest gene family in vertebrates, the sequences of olfactory receptors feature a lot of variation, in analogy of the multitude of ligands their receptors recognize. This variation is usually restricted to the locations of binding sites, since the selective pressure is greatest upon them. Therefore, it seemed advisable to exclude such regions under positive, diversifying selection pressure from the alignment. Otherwise the resulting evolutionary tree would rather describe the evolution of smell detection, instead of the divergence of different receptor families. Here, a maximum likelihood approach (PAML) was used to identify sites within the sequences that underlie positive selection. To not completely rely on the program's results, several other publications about the location of binding sites and regions under diversifying selective pressure were integrated to back up each other. Thus, several subsequences could be excised with relatively high confidence. When mapping these clusters of sites proposed to be under positive selection onto the secondary structure presentation of such a GPCR, it seems biologically plausible that these regions are likely to be involved in ligand interaction.

Another step of preparing the dataset for the generation of an MSA was the reduction of the large number of sequences to feasible datasets suited for further calculations. To avoid the unnecessary loss of information, only sequences were discarded that revealed a certain level of similarity to others. Therefore, I wrote a program in PERL, to define the level of similarity and automatically exclude those sequences that are above this threshold. In the case of ORs, the level of sequence similarity is now <= 40%, and among the V1Rs, <= 45%.

The next challenge was the generation of a multiple sequence alignment. ORs and V1Rs are known to be very difficult to align, due to their rather distant relatedness. Furthermore, several different MSA programs of similarly good performance and reliability are available. Here, the top five MSA methods were applied, and the resulting MSAs compared using three different approaches. Their quality was assessed regarding the information content, the interconsistency, and the impression of their sequence logos. In spite of the fact that several different means of quality evaluation were applied, it turned out that four of the five MSAs were similarly good. Thus the alignment chosen for further calculations (by MUSCLE) lacks statistical support, even though it was assigned high scores.

Finally, based upon this MSA, the phylogenetic tree was reconstructed. Here, Maximum likelihood estimations were applied, implemented in the programs MOLPHY and PAML. Presumably, this is the most accurate way to do so, regarding the drawbacks of distance based methods and the parsimony approach. For this task the sequences had to be reduced to a maximum of eleven, to keep the calculation feasible. MOLPHY evaluated all >34 million possible trees according to approximate likelihood, and PAML then evaluated the best 1000 trees among those, using maximum likelihood. Reason for the two programs to share the task is the fact that MOLPHY is the only program that performs an exhaustive tree search employing ML, and PAML, on the other hand, is better in evaluating the best trees, using more up-to-date substitution matrixes. Even though a best-scoring tree was proposed, the differences in ML scores among the top ten trees are very small, and the bootstrap values acceptable. Therefore, it is difficult to choose a single tree to be the optimal representation of the data. However, it is not necessary to define this best tree, since the highest scoring trees share the main feature: V1Rs branch off earlier than ORs from the common root, implying that vomeronasal receptors are actually older than receptors of the MOS. Thus, the details of the internal branches of subtrees are of minor importance to this general message, which many of the best trees have in common, providing sufficient statistical support.

This answer to my scientific question is to be seen within the context of the sequence data used, and the limitations of the programs applied. When looking for such weak evolutionary signals, one has to carefully evaluate, at what point the price for more sensitivity is too much loss of reliability. By applying and comparing several different programs at each step of this work, I tried to minimise this effect. Ultimately, this is about the best possible way to solve such problems today.

## Conclusion

It should be mentioned that some researchers in different fields earlier also suggested reconsideration of this aspect of our understanding of evolution. The morphologist H. Eisthen, for example, argues for the idea of the MOS of teleost fishes to actually be more like a combination of an olfactory and vomeronasal system (100). To make a final judgement whether or not the VNO is older than the MOS, clearly goes beyond the scope of this project, which can only provide inspiration and support for this thesis. However, from the viewpoint of a bioinformatician, it certainly seems like the vomeronasal receptors evolved before those of the main olfactory system.

## Methods

In the following the programs are listed that were employed using the standard parameter settings and/or were described sufficiently in the “results” section:

- CLUSTALX version 8.3 (79)
- MAFFT version 5 (75)
- PRALINE (82)
- PROBCONS (74)
- TREEVIEW version 1.6.6
- TREE PUZZLE version 5.1 (101)

- PHYLIP version 3.65
- MOLPHY version 2.3 (102)
- SAM version 3.2 (113)
- TMHMM version 2.0 (72)
- WEBLOGO (87)

## Detection of positive selection sites using PAML version 3.15

To argue for positive selection, models of neutral evolution and purifying selection must be rejected, including ω to be significantly greater than 1 (104,105).

$\omega = \mathbf{d_N/d_S}$ = nonsynonymous/ silent substitution

|  |  |
|---|---|
| ω < 1 | negative (purifying) selection |
| ω = 1 | neutral selection |
| ω > 1 | positive (diversifying) selection |

The software package PAML by Z. Yang (106,107) offers many possibilities to assess such information and compare different models. A *Markov* process is used to describe substitutions between 61 non-stop codons. The PAML subprogram CODEML evaluates the relative fit of codon substitution models with actual data by likelihood ratio (LR) statistics, which are assumed to be $\chi^2$ distributed with degrees of freedom equal to the difference in the number of parameters between models. LR tests for positive selection compare a model in which there is a class of sites with ω > 1 against a model that does not allow for this class.

The olfactory sequences employed to mark positions of positive selection belong to the genome of the zebrafish (*Danio rerio*), and present a subset of 10 sequences that cluster together with high bootstrap values. The representative sequences were derived from the data sets provided by Niimura & Nei. It may be assumed that regions exhibiting positive selection are similar among olfactory receptors, since those share a common tertiary structure and functional domains for ligand binding.

The analysis was performed based upon several site-specific models in which selective pressure varies among different sites but the site-specificity is identical across all lineages:

| **Model** | **Description** |
|---|---|
| M0 | null model with a single ω ratio among all sites |
| M1 | "nearly neutral" model, with two categories of site with fixed ω ratios of 0 and 1 |
| M2 | "selection" model, with three categories of site, two with fixed ω ratios of 0 and 1, a third estimated ω ratio |
| M3 | "discrete" model, three categories of site with the ω ratio free to vary for each site |
| M7 | "beta" model, eight categories of site, with eight ω ratios in the range of 0-1 taken from a discrete approximation of the beta distribution |
| M8 | "beta plus ω" model, expanding model 7 by an additional category of site with an ω ratio that is free to vary from 0 to greater than 1 |

**Table 6**: Overview of models used in PAML

PAML calculates the Maximum likelihood estimates for all parameters and the ω ratios that are free to vary under these models, as well as the proportion of sites with each ratio, as shown in table 7 (values in bold indicate positive selection.). Then an empirical Bayes approach is used to predict the most likely ω class for each site, with the posterior probability providing a measure of reliability.

| Model | Parameter estimates | ln L | Positively selected sites |
|---|---|---|---|
| One ratio (M0) | ω = | -4035.79 | None |
| Nearly neutral (M1) | $\omega_1 = 0.13$ $\omega_2 = 1.00$<br>$p_1 = 0.67$ $p_2 = 0.33$ | -3986.60 | not allowed |
| Selection (M2) | $\omega_1 = 0.13$ $\omega_2 = 1.00$<br>$\omega_2 = \mathbf{5.69}$<br>$p_1 = 0.67$ $p_2 = 0.32$<br>$p_3 = 0.00$ | -3986.24 | 1 site |
| Discrete (M3) | $\omega_1 = 0.05$ $\omega_2 = 0.40$ | -3984.79 | 33 sites |

| | | | |
|---|---|---|---|
| | $\omega_2$ = **1.51** | | |
| | $p_1$ = 0.37 $p_2$ = 0.48 | | |
| | $p_3$ = 0.14 | | |
| Beta (M7) | p = 0.34 | -3988.08 | not allowed |
| | q = 0.65 | | |
| Beta + ω (M8) | $p_0$ = 0.89 p = 0.72 | -3984.95 | 16 sites |
| | q = 1.89 | | |
| | ($p_1$= 0.12) | | |
| | **ω = 1.56** | | |

**Table 7:** Maximum likelihood estimates of $d_N/d_S$ ratios (ω) using PAML: site-specific models over all 10 lineages Values in bold indicate positive Selection, that subsequently needs to be tested for significance using LR statistics (table 8).

The likelihood ratio tests (LRT) listed in table 7 help determine whether particular models provide a significantly better fit to the data than other related models. They are compared to critical values of the Chi square distribution with the appropriate degrees of freedom (108). P values for sites potentially under positive selection were obtained using a Bayesian approach in PAML.

LR- comparing the one-ratio model M1 with the discrete model M3 is a test of variable selective pressures among sites.

| Models compared | Likelihood ratio statistic: -2($L_1$-$L_2$) | P value | df |
|---|---|---|---|
| M1 versus M2 | 0.71 | 0.70 | 2 |
| M1 versus M3 | 3.61 | 0.46 | 4 |
| M7 versus M8 | 6.26 | 0.04 | 2 |

**Table 8** Likelihood ratio tests of models presented in table 6

To specifically test for a portion of sites evolving by positive selection, LRTs are conducted to compare M1 with M2 and M7 with M8. Positive selection is indicated when a freely estimated ω parameter is >1 and the LRT is significant. Therefore, the model suggested by the data is M8, as to be seen in tables 7 and 8.

Under this model, the following sites are estimated to significantly ($P>0.95$) be subject to positive selective pressure: W46, N47, V154, V196 (using Dr3OR15.9 as reference). These results were derived using 10 closely related *Danio rerio* ORs as input. Examined in context with previous results, it appears likely that W46 and N47, situated in the first TM helix, are false positives, possibly indirectly influenced by other positively selected amino acids. None of the many experimental studies ever considered this helix as part of the pocket. V154 and V196 are found in helix 5, which is certainly part of the binding site, according to most studies in this area.

A second run of PAML using more sequences (26) yielded no significant hits, indicating that further distance within the data introduces too much noise to draw reliable conclusions. It has to be pointed out though, that these results were derived by applying and comparing site-specific models only.

| *Model type* | ω | *Description* |
|---|---|---|
| Lineage specific | variable on lineages,<br>fixed on sites | Adapive evolution during a certain time interval is found, but it detects positive selection for a lineage only if the average $d_N$ over all sites is higher than the average $d_S$ , thus it assumes that all sites were subject to identical selective pressure. |
| Codon<br>/site specific | Variable on sites, fixed on lineages | Detects positive selection at individual sites only if the average $d_N$ over all lineages is higher than the average $d_S$, assuming constant selection pressure over time / along the tree. |
| Branch-site | Sites & lineages | useful if adaptive evolution occurred at a few time points and affected a few sites. |

**Table 9**: overview of model types implemented in PAML.

In this case, the lineage-specific models are not very useful, since we expected the selective pressure to vary equally in all lineages. Therefore, the branch-site model does not seem reasonable in this case. (111).

The final decision, which subsequences to exclude from the alignment, was based on the combination of my results and those of the previous experimental and theoretical studies.

### The multiple sequence alignment – MUSCLE version 3.6

In summary, the following steps were carried out to derive an OR-V1R multiple sequence alignment: Annotated sequences from previous studies were collected and combined. Based upon their a priori multiple alignment (MUSCLE (76)) a percent identity matrix was generated (CLUSTALX), which in turn served as input for my self-written program FILTER. It allows to exclude sequences from the data set that exhibit more sequence similarity to another sequence, than a certain threshold. To retain equal numbers of OR and V1R sequences, OR-sequences are now no closer related than 45%, and V1Rs 40%.  After the "redundant" data was eliminated, the sequences left were realigned, and truncated according to the estimations of sites under positive selection (PAML). Deciding, which regions to exclude, was a rather complicated undertaking and is described separately. Finally, the prepared input data was aligned employing several different multiple alignment programs: CLUSTALX, MAFFT, MUSCLE, PRALINE and PROBCONS. The best alignment among their outcomes was chosen, based upon the three criteria information content (PAUP), inter consistency (MUMSA), and appearance of sequence logos.

### PAUP version 4.0  – skewness as measure of information content

The word "skewness" addresses the shape of the frequency distribution of tree lengths a set (100000) of randomly generated trees evaluated for a given alignment. A biologically senseless multiple alignment holding no information would yield a perfect bell-shaped curve, more or less symmetrical. The presence of  information content, however, has a skewing effect upon that curve, as to be seen in figure 17:

**Fig. 17 (next page):** tree length frequency distribution based upon the alignment by muscle. 100000 random trees generated and evaluated by PAUP (88) Version 4.0b10 for Macintosh.

```
3127.00000 /----------------------------------------------------------------------
3133.06000 | (1)
3139.12000 | (2)
3145.18000 | (0)
3151.24000 | (0)
3157.30000 | (0)
3163.36000 | (1)
3169.42000 | (0)
3175.48000 | (1)
3181.54000 | (0)
3187.60000 | (0)
3193.66000 | (0)
3199.72000 | (1)
3205.78000 | (6)
3211.84000 | (0)
3217.90000 | (3)
3223.96000 | (4)
3230.02000 | (5)
3236.08000 | (4)
3242.14000 | (11)
3248.20000 | (5)
3254.26000 | (16)
3260.32000 | (14)
3266.38000 | (11)
3272.44000 | (8)
3278.50000 | (19)
3284.56000 | (18)
3290.62000 |# (30)
3296.68000 |# (31)
3302.74000 |# (44)
3308.80000 |# (43)
3314.86000 |# (57)
3320.92000 |# (65)
3326.98000 |## (82)
3333.04000 |## (126)
3339.10000 |## (122)
3345.16000 |## (127)
3351.22000 |### (135)
3357.28000 |### (174)
3363.34000 |#### (202)
3369.40000 |#### (231)
3375.46000 |##### (237)
3381.52000 |###### (302)
3387.58000 |####### (365)
3393.64000 |######## (401)
3399.70000 |######## (419)
3405.76000 |########## (498)
3411.82000 |########### (576)
3417.88000 |############# (650)
3423.94000 |############## (742)
3430.00000 |############### (802)
3436.06000 |##################### (1074)
3442.12000 |################### (1001)
3448.18000 |##################### (1086)
3454.24000 |######################## (1261)
3460.30000 |########################### (1376)
3466.36000 |############################## (1541)
3472.42000 |################################ (1684)
3478.48000 |#################################### (1864)
3484.54000 |##################################### (1915)
3490.60000 |######################################## (2070)
3496.66000 |########################################### (2220)
3502.72000 |############################################## (2369)
3508.78000 |################################################# (2547)
3514.84000 |################################################### (2619)
3520.90000 |######################################################### (2952)
3526.96000 |######################################################## (2904)
3533.02000 |####################################################################### (3682)
3539.08000 |################################################################ (3299)
3545.14000 |################################################################# (3396)
3551.20000 |################################################################## (3414)
3557.26000 |##################################################################### (3599)
3563.32000 |#################################################################### (3546)
3569.38000 |################################################################### (3496)
3575.44000 |###################################################################### (3609)
3581.50000 |################################################################### (3492)
3587.56000 |################################################################# (3389)
3593.62000 |################################################################# (3383)
3599.68000 |############################################################ (3105)
3605.74000 |######################################################## (2906)
3611.80000 |##################################################### (2740)
3617.86000 |################################################# (2523)
3623.92000 |############################################# (2332)
3629.98000 |######################################## (2065)
3636.04000 |########################################## (2158)
3642.10000 |############################## (1533)
3648.16000 |######################## (1258)
3654.22000 |#################### (1045)
3660.28000 |################ (828)
3666.34000 |############# (659)
3672.40000 |########## (511)
3678.46000 |####### (347)
3684.52000 |#### (230)
3690.58000 |### (169)
3696.64000 |## (89)
3702.70000 |# (64)
3708.76000 |# (35)
3714.82000 | (16)
3720.88000 | (5)
3726.94000 | (1)
3733.00000 | (2)
           \----------------------------------------------------------------------
```

# Acknowledgements

The University of Skövde, PhD Björn Olsson, and supervisor PhD  Dan Lundh kindly made this project possible and the Friedrich Schiller University of Jena and secondary supervisor Dr. Franco Weth provided much support to carry it out. Sincere thanks. The work was conducted in the research group of neurogenetics, Universitätsklinikum Jena (BMBF FKZ 01 ZZ 0405).

# References

1. Buck, L. B. (2004). **The search for odorant receptors**. *Cell* 116, Suppl. S117–S119
2. Lane, R.P., Cutforth, T., Young, J., Athanasiou, M., Friedman, C., Rowen, L., Evans, G., Axel, R., Hood, L., and Trask, B.J. (2001). **Genomic analysis of orthologous mouse and human olfactory receptor loci.** *Proc. Natl. Acad. Sci.* 98: 7390–7395.
3. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 420: 520–562.
4. Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L. & Trask, B. J. (2002). **Different evolutionary processes shaped the mouse and human olfactory receptor gene families.** *Hum. Mol. Genet.* 11, 535–546.
5. Glusman, G., Bahar, A., Sharon, D., Pilpel, Y., White, J. & Lancet, D. (2000). **The olfactory receptor gene superfamily: data mining, classification and nomenclature.** *Mamm. Genome* 11, 1016–1023.
6. Dulac, C. & Torello, A. T. (2003). **Molecular detection of pheromone signals in mammals: from genes to behavior.** *Nat. Rev. Neurosci.* 4, 551–562.
7. Halpern, M. (1987). **The organization and function of the vomeronasal system.** *Annu. Rev. Neurosci.* 10, 325–362
8. Halpern, M. & Martínez-Marcos, A. (2003). **Structure and funtion of the vomeronasal system: an update.** *Prog. Neurobiol.* 70, 245–318
9. Malnic, B, Hirono J, Sato T, Buck LB (1999). **Combinatorial receptor codes for odors**. *Cell* 96, 713
10. Buck, L. & Axel, R. (1991). Buck, L. and Axel, R. (1991). **A novel multigene family may encode odorant receptors: a molecular basis for odor recognition.** Cell 65, 175-187 *Cell* 65**,** 175–187.
11. Dulac, C. & Axel, R. (1995). **A Novel Family of Genes encoding Putative Pheromone Receptors in Mammals.** *Cell* 83, 195–206.
12. Ryba, N. J. & Tirindelli, R. (1997). **A new multigene family of putative pheromone receptors.** *Neuron* 19, 371–379.
13. Matsunami, H. & Buck, L. B. (1997). **A multigene family encoding a diverse array of putative pheromone receptors in mammals.** *Cell* 90, 775–784.
14. Herrada, G. & Dulac, C. (1997) **A Novel Family of Putative Pheromone Receptors in Mammals with a Topographical Organized and Sexually Dimorphic Distribution.** *Cell* 90, 763–773.

15. Boschat, C., Pelofi, C., Randin, O., Roppolo, D., Luscher, C., Broillet, M. C. & Rodriguez, I. (2002**). Pheromone detection mediated by a V1r vomeronasal receptor.** *Nat. Neurosci.* 5, 1261–1262.
16. Del Punta, K., Leinders-Zufall, T., Rodriguez, I., Jukam, D., Wysocki, C. J., Ogawa, S., Zufall, F. & Mombaerts, P. (2002). **Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes.** *Nature* 419, 70–74.
17. Karlson, P., and Lüscher, M. (1959). **"Pheromones": A new term for a class of biologically active substances.** Nature *183*, 55–56.
18. Schaal, B. *et al.* (2003). **Chemical and behavioural characterisation of the rabbit mammary pheromone.** *Nature* 424, 68–72
19. Novotny, M. V. (2003). **Pheromones, binding proteins and receptor responses in rodents.** *Biochem. Soc. Trans.* 31, 117–122
20. Leinders-Zufall, T., Lane, A.P., Puche, A.C., Ma, W., Novotny, M.V., Shipley, M.T., and Zufall, F. (2000). **Ultrasensitive pheromone detection by mammalian vomeronasal neurons.** *Nature* 405: 792–796.
21. Hudson, R. and Distel, H. (1986). **Pheromonal release of suckling in rabbits does not depend on the vomeronasal organ.** *Physiol. Behav.* 37: 123–128.
22. Dorries, K.M., Adkins-Regan, E., and Halpern, B.P. (1997). **Sensitivity and behavioral responses to the pheromone antrostenone are not mediated by the vomeronasal organ in domestic pigs.** *Brain Behav. Evol.* 49: 53–62.
23. Fernandez-Fewell, G.D. and Meredith, M. (1998). **Olfactory contribution to fos expression during mating in inexperienced male hamsters.** *Chem. Sens.* 23: 257–267.
24. Fewell, G.D. and Meredith, M. (2002). **Experience facilitates vomeronasal and olfactory influence on Fos expression in medial preoptic area during pheromone exposure or mating in male hamsters.** *Brain Res.* 941: 91–106.
25. Hudson, R. & Distel, H. (1986). **Pheromonal release of suckling in rabbits does not depend on the vomeronasal organ.** *Physiol. Behav.* 37, 123–128
26. Okada T, Fujiyoshi Y, Silow M, Navarro J, Landau EM, Shichida Y. (2002) **Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography.** Proc Natl Acad Sci U S A. 99(9): 5982-5987.
27. Rodriguez, I., Greer, C. A., Mok, M. Y. & Mombaerts, P. (2000) **A putative pheromone receptor gene expressed in human olfactory mucosa.** *Nat. Genet* 26, 18–19.
28. Wakabayashi, Y., Mori, Y., Ichikawa, M., Yazaki, K. & Hagino-Yamagishi, K. (2002). **A Putative Pheromone Receptor Gene Is Expressed in Two Distinct Olfactory Organs in Goats**. *Chem. Senses* 27, 207–213.
29. Sam, M., Vora, S., Malnic, B., Ma, W., Novotny, M.V., and Buck, L.B. (2001). **Neuropharmacology: Odorants may arouse instinctive behaviours.** *Nature* 412: 142.
30. Trinh, K. & Storm, D. R. (2003). **Vomeronasal organ detects odorants in absence of signaling through main olfactory epithelium**. *Nature Neurosci.* 6, 519–525
31. Brennan, P. A. & Keverne, E. B. (2004). **Something in the air? New insights into mammalian pheromones.** *Curr. Biol.* 14, R81–R89
32. Filipek S. et al. (2003). **The crystallographic model of rhodopsin and its use in studies of other G protein-coupled receptors.** Annu. Rev. Biophys. Biomol. Struct.; 32: 375–397.
33. Takami, S. (2002). **Recent progress in the neurobiology of the vomeronasal organ.** *Microsc. Res. Tech.* 58, 228–250.

34. Dennis, J. C., Allgier, J. G., Desouza, L. S., Eward, W. C. & Morrison, E. E. (2003). **Immunohistochemistry of the canine vomeronasal organ.** *J. Anat.* 203, 329–338.
35. Zhang, X., Rodriguez, I., Mombaerts, P. & Firestein, S. (2003) **Odorant and vomeronasal receptor genes in two mouse genome assemblies.** *Genomics* DOI 10.1016/j.ygeno.2003.10.009.
36. Xie, S. Y., Feinstein, P. & Mombaerts, P. (2000). **Characterization of a cluster comprising ~100 odorant receptor genes in mouse.** *Mamm. Genome* 11, 1070–1078
37. Chess, A., Simon, I., Cedar, H. & Axel, R. (1994). **Allelic inactivation regulates olfactory receptor gene expression.** *Cell* 78, 823–834
38. Ishii, T. *et al.* (2001). **Monoallelic expression of the odourant receptor gene and axonal projection of olfactory sensory neurons.** *Genes Cells* 6, 71–78
39. Strotmann, J. *et al.* (2000). **Local permutations in the glomerular array of the mouse olfactory bulb.** *J. Neurosci.* 20, 6927–6938
40. Mombaerts, P. (2004). **Odorant receptor gene choice in olfactory sensory neurons: the one receptor–one neuron hypothesis revisited.** *Curr. Op. Neurobiol.* 14, 31–36
41. Mombaerts, P. *et al***.** (1996). **Visualizing an olfactory sensory map.** *Cell* 87, 675–686
42. Wang, F., Nemes, A., Mendelsohn, M. & Axel, R**.** (1998). **Odorant receptors govern the formation of a precise topographic map.** *Cell* 93, 47–60
43. Bozza, T., Feinstein, P., Zheng, C. & Mombaerts, P. (2002). **Odorant receptor expression defines functional units in the mouse olfactory system.** *J. Neurosci.* 22, 3033–3043
44. Del Punta, K., Puche, A., Adams, N. C., Rodriguez, I. & Mombaerts, P. (2002). **A divergent pattern of sensory axonal projections is rendered convergent by second-order neurons in the accessory olfactory bulb.** *Neuron* 35, 1057–1066
45. Mombaerts, P. (2004**) Genes and ligands for odorant, vomeronasal and taste receptors.** *Nat. Rev. Neurosci.* 5, 263–278.
46. Loconto, J., Papes, F., Chang, E., Stowers, L., Jones, E.P., Takada, T., Kumanovics, A., Fischer-Lindahl, K., and Dulac, C. (2003). **Functional expression of murine V2R pheromone receptors involves selective association with the M10 and M1 families of MHC class 1b molecules.** *Cell* 112, 607–618.
47. Ishii, T., Hirota, J., and Mombaerts, P. (2003). **Combinational coexpression of neural and immune multigene families in mousevomeronasal sensory systems.** *Curr. Biol.* 13, 394–400.
48. Wess J. (1997). **G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of G-protein recognition.** FASEB J (11): 346-54,
49. Wistrand, M., Sonnhammer, E (2005). **Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER.** *BMC Bioinformatics* 6:99 doi:10.1186/1471-2105-6-99
50. Speca, D. J., Lin, D. M., Sorensen, P. W., Isacoff, E. Y., Ngai, J. & Dittman, A. H. (1999). **Functional identification of a goldfish odorant receptor.** *Neuron* 23, 487–498.
51. Kunishima, N., Shimada, Y., Tsuji, Y., Sato, T., Yamamoto,M., Kumasaka, T., Nakanishi, S., Jingami, H. and Morikawa, K. (2000). **Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor.** *Nature*, 407, 971–977

52. Hermans, E., CHALLISS, J. (2001). **Structural, signalling and regulatory properties of the group I metabotropic glutamate receptors: prototypic family C G-protein-coupled receptors**. *Biochem. J.* 359 (465–484)
53. Eisthen, H.L. (2004) **The goldfish knows: olfactory receptor cell morphology predicts receptor gene expression.** *J. Comp. eurol.* 477, 341–346
54. Broman, I. 1920 **Die Organon vomero-nasale Jacobsoni - ein Wassergeruchsorgan!** *Anat. Hefte* **58**, 143-191.
55. Bertmar G. (1981). **Variations in size and structure of vomeronasalorgan in reindeer (*Ranfiger tarandus tarandus*L.).** *Archives of Biology* (Bruxelles), 92, 343–366.
56. Schmidt, A., and Roth, G.. (1990), **Central olfactory and vomeronasal pathways in salamanders,** *J. Hirnforsch.*31:543-553.
57. Wallace, I., Blackshields, G., Higgins, D. (2005). **Multiple sequence alignments.** *Current Opinion in Structural Biology* 2005, 15:261–266
58. Eisthen, H.L., D.R. Sengelaub, D.M. Schroeder, and J.R. Alberts (1994). **Anatomy and forebrain projections of the olfactory and vomeronasal organs in axolotls (*Ambystoma mexicanum*).** *Brain Behav. Evol. 44:*108–124.
59. Panchen, A. L. (1991) **The early tetrapods: classi¢cation and the shapes of cladograms. In *Origins of the higher groups of tetrapods: controversy and consensus*** (ed. H.-P. Schultze & L. Trueb), pp. 100-144. Ithaca, NY: *Cornell University Press.*
60. Lebedev, O.A. and Coates, M.I. (1995). **The postcranial skeleton of the Devonian tetrapod Tulerpeton curtum Lebedev.** *Zoological Journal of the Linnean Society,* 114: 307- 348.
61. Golding, G.B., and A.M. Dean, 1998. **The structural basis of molecular adaptation.** *Mol. Biol. Evol.* 15: 355-369.
62. Kimura, M. (1983). **The Neutral Theory of Molecular Evolution**. *Cambridge University Press*, Cambridge, MA.
63. Yang Z.( 1997): **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci.* 13(5):555-6.
64. Emes RD, Beatson SA, Ponting CP, Goodstadt L (2004) **Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents.** *Genome Res* 14: 591-602.
65. Katada, S., T. Hirokawa, Y. Oka, M. Suwa, and K. Touhara (2005). **Structural Basis for a Broad But Selective Ligand Spectrum of a Mouse Olfactory Receptor: Mapping the Odorant-Binding Site.** *J. Neurosci.* 25: 1806-1815
66. Hall, S., Wely B. Floriano, Nagarajan Vaidehi and William A. Goddard III **Predicted 3-D Structures for Mouse I7 and Rat I7 Olfactory Receptors and Comparison of Predicted Odor Recognition Profiles with Experiment** *Chem. Senses* 29: 595–616, 2004
67. Man et al. (2004) **Prediction of the odorant binding site of olfactory receptor proteins by human–mouse comparisons** *Protein Sci* 13 (1): 240.
68. Vaidehi, N et. al**. Prediction of structure and function of G proteincoupled receptors** *PNAS (99)* 12622–12627
69. Floriano, W. B., Vaidehi, N., Singer, M. S., Shepherd, G. M., Goddard III, W. A. (2000).**Molecular Mechanisms underlying differential odor responses of a mouse olfactory receptor**. *P. Natl. Acad. Sci. USA* 97, 10712-10716.
70. Singer MS. (2000) **Analysis of the molecular basis for octanal interactions in the expressed rat 17 olfactory receptor**. *Chem Senses*. (2):155-65.

71. Afshar, M., Hubbard, R.E. and Demaille, J (1998**), Towards structural models of molecular recognition in olfactory receptors**, *Biochimie,* 80, 129-135
72. http://www.cbs.dtu.dk/services/TMHMM/
73. Lecompte,O., Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2001) **Multiple alignment of complete sequences (MACS) in the postgenomic era.** *Gene*, 270, 17–30.
74. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) **ProbCons: probabilistic consistency-based multiple sequence alignment.** *Genome Res.*, 15, 330–340.
75. 3. Katoh,K., Kuma,K., -i., Toh,H. and Miyata,T. (2005) **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res*., 33, 511–518.
76. 5. Edgar,R.C. (2004) **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res*., 32, 1792–1797.
77. R. Hughey and A. Krogh. SAM (1995) : **Sequence alignment and modeling software system.** *Technical Report UCSC-CRL-95-7,* University of California, Santa Cruz,
78. http://hmmer.wustl.edu/
79. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Research,* 24:4876-4882.
80. K. Katoh, K. Kuma, H. Toh and T. Miyata (2005) **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res.* 33: 511-518
81. Edgar,R.C. (2004) **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res*., 32, 1792–1797.
82. http://ibivu.cs.vu.nl/programs/pralinewww/
83. http://probcons.stanford.edu/
84. Lassmann, T, Sonnhammer, E.(2005). **Automatic assessment of alignment quality.** *Nucleic acids research*, vol 33, No. 22 web ref
85. Strimmer, K. and von Haeseler, A. (1996) **Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies**. *Mol. Biol. Evol.*, 13, 964–969.
86. Felsenstein, (1978a) **The number of evolutionary trees**. *System. Zool*., 27:27-33.
87. Crooks GE, Hon G, Chandonia JM, Brenner SE, (2004) **WebLogo: A sequence logo generator,** *Genome Research*, 14:1188-1190 http://weblogo.berkeley.edu/
88. http://paup.csit.fsu.edu/
89. Wang,L. and Jiang,T. (1994) **On the complexity of multiple sequence alignment.** *J. Comput. Biol*., 1, 337±348.
90. Waterman,M.S., Smith,T.F. and Beyer,W.A. (1976) **Some biological sequence metrics.** *Adv. Math*., 20, 367±387.
91. Hogeweg,P. and Hesper,B. (1984) **The alignment of sets of sequences and the construction of phyletic trees: an integrated method.** *J. Mol. Evol*., 20, 175±186.
92. Feng,D.F. and Doolittle,R.F. (1987) **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.** *J. Mol. Evol*., 25, 351±360.
93. Thompson JD; Plewniak F, Poch O, J (1999) **A comprehensive comparison of multiple sequence alignment programs.** *Mol Biol* 1999, 27:2682-2690

94. Sokal, R. R., and P. H. A. Sneath. (1963). **Principles of Numerical Taxonomy**. *W. H. Freeman*, San Francisco.

95. J. Felsenstein. (1988) **Phylogenies from molecular sequences: inference and reliability**. *Annual Review of Genetics*, 22:521--65,

96. Huelsenbeck, J. P. and D. M. Hillis. (1993). **Success of phylogenetic methods in the four-taxon case**. *Systematic Biology* 42(3):247-265.

97. Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. (1996.) **Phylogenetic inference.** *Pages 407-514 in D. M. Hillis, C. Moritz, and B. Mable (eds.) Molecular Systematics* (2nd ed.), Sinauer Associates, Sunderland, Massachusetts.

98. Saitou, N., and M. Nei. (1987). **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular Biol. Evol*.4:406–425.

99. http://www.ebi.ac.uk/interpro/

100. Eisthen HL, Wyatt TD. (2006) **The vomeronasal system and pheromones.** *Curr Biol.* Feb 7;16(3):R73-4.

101. Schmidt, H.A., K. Strimmer, M. Vingron, and A. von Haeseler (2002) **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics*. 18:502-504.

102. J. Adachi and M. Hasegawa. (1996) **MOLPHY version 2.3 programs for molecular phylogenetics based on maximum likelihood.** *Computer Science Monographs*, 28:1--150.Univ. Press, Cambridge.

103. Niimura Y, Nei M. (2005) **Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods.** *Proc Natl Acad Sci U S A.* 102(17):6039-44.

104. Hughes, A. L. and M. Nei. (1988). **Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.** *Nature* 335:167-170.

105. Yang, Z., and R. Nielsen. (1998). **Synonymous and nonsynonymous rate variation in nuclear genes of mammals.** *Journal of Molecular Evolution*, 46: 409-418.

106. Yang, Z. (1997). **PAML: a program for package for phylogenetic analysis by maximum likelihood.** *CABIOS* 15: 555-556. (http://abacus.gene.ucl.ac.uk/software/paml.html).

107. Yang, Z., W.J. Swanson, and V.D. Vacquier. (2000**). Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites.** *Molecular Biology and Evolution* 17: 1446-1455.

108. Yang, Z. (1998). **On the best evolutionary rate for phylogenetic analysis.** *Systematic Biology* 47:125-133.

109. Kishino, H. and M. Hasegawa. (1989). **Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea.** *Journal of Molecular Evolution* 29: 170-179.

110. Shimodaira, H., and M. Hasegawa. (1999). **Multiple comparisons of log-likelihoods with applications to phylogenetic inference**. *Mol. Biol. Evol.* 16:1114-1116

111. Yang, Z. (2002) **Inference of selection from multiple species alignments**. *Current Opinion in Genetics and Development* 12: 688-694.

112. Attwood, TK, Findlay JB (1994). **Fingerprinting G-protein coupled receptors.** *Protein Eng.* (2) 195 – 203

113. R. Hughey and A. Krogh. SAM (1995): **Sequence alignment and modeling software system.** *Technical Report* UCSC-CRL-95-7, University of California, Santa Cruz

114. Schneider TD, Stephens RM. 1990. **Sequence Logos: A New Way to Display Consensus Sequences.** *Nucleic Acids Res*. 18:6097-6100

a. Pantages, E, Dulac, C (2000): **A Novel Family of Candidate Pheromone Receptors in Mammals** *Neuron*, Vol. 28, 835–845

b. Grus W., Peng S., Ya-ping Z., Jinzhi Z (2005): **Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals**, *PNAS* vol. 102 , 5767-5772

c. Niimura, Y and Nei, M. (2005) : **Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods,** *PNAS*, vol. 102, no. 17, 6039-6044

d. Weth F (2001) **Molekulare Aspekte der funktionellen Architektur des Geruchssystems beim Zebrabärbling, danio rerio**. Dissertation, Fakultät für Biologie der Eberhard-Karls-Universität Tübingen

e. Olender T, Fuchs T, Linhart C, Shamir R, Adams M, Kalush F, Khen M, Lancet D., (2004) **The canine olfactory subgenome**, *Genomics*. 2004 Mar;83(3):361-72.

f. Mundy NI, Cook S. (2003): **Positive selection during the diversification of class I vomeronasal receptor-like (V1RL) genes, putative pheromone receptor genes, in human and primate evolution.** *mol Biol Evol*.(11):1805-10.

g. Pfister, P. & Rodriguez, I. (2005) **Olfactory expression of a single and highly variable V1r pheromone receptor-like gene in fish species** *Proc. Natl. Acad. Sci*. USA 102, 5489–5494

h. Rodriguez, I., Punta, KD., Rothman, A., Ishii, T., and Mombaerts, P. (2002). **Multiple new and isolated families within the mouse superfamily of V1r vomeronasal receptors** . *Nature Neurosci* 5, 134.

i. Zhang X, Rodriguez I, Mombaerts P, Firestein S (2004). **Odorant and vomeronasal receptor genes in two mouse genome assemblies**. *Genomics*. 83(5):802-11.

j. Berghard A, Dryer L (1998) **A novel family of ancient vertebrate odorantreceptors.** *J Neurobiol* 37, 383–392

Freitag J, Beck A, Ludwig G, von Buchholtz L, Breer H (1999) **On the origin of the olfactory receptor family: receptor genes of the jawless fish (*Lampetra fluviatilis*).** *Gene 226*, 165–74

## Additional files

**Additional file 1 – filter.pdf**
Self-written program with comments. (Original perl file available from author upon request.)(PDF file)

**Additional file 2 – clustal_logo.pdf**
Sequence logo of MSA by CLUSTAL, excluding alignment regions that feature more than 20% gaps.(PDF file)

**Additional file 3 – mafft_logo.pdf**
Sequence logo of MSA by MAFFT, excluding alignment regions that feature more than 20% gaps.(PDF file)

**Additional file 4 – muscle_logo.pdf**
Sequence logo of MSA by MUSCLE, excluding alignment regions that feature more than 20% gaps.(PDF file)

**Additional file 5 – praline_logo.pdf**
Sequence logo of MSA by PRALINE, excluding alignment regions that feature more than 20% gaps.(PDF file)

**Additional file 6 – probcons_logo.pdf**
Sequence logo of MSA by PROBCONS, excluding alignment regions that feature more than 20% gaps.(PDF file)

**Additional file 7 – muscle_alignment.pdf**
Picture of multiple alignment by MUSCLE as seen in the editor ALNEDIT.
(PDF file)

**Additional file 8 – trees.pdf**
Topologies of the top ten trees according to PAML estimations.
(PDF file)