Large, structurally-curated alignment and phylogeny of Vertebrate biogenic amine receptors

Stephanie J. Spielman[1,2,3], Ahmad R. Sedaghat[4,5], Keerthana Kumar[1,2,3], and Claus O. Wilke[1,2,3]

[1]**Department of Integrative Biology, The University of Texas at Austin, Austin, U.S.A.**
[2]**Institute of Cellular and Molecular Biology, The University of Texas at Austin, Austin, U.S.A.**
[3]**Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, U.S.A.**
[4]**Department of Otolaryngology–Head and Neck Surgery, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, U.S.A.**
[5]**Department of Otology and Laryngology, Harvard Medical School, Boston, Massachusetts, U.S.A.**

ABSTRACT

Keywords: biogenic amine receptors, phylogenetics, multiple sequence alignment, G-protein coupled receptors

TO DO FOR MS: COME UP WITH ALIGNMENT ACRONYMS!!!!!!

INTRODUCTION

Biogenic amines, such as the molecules serotonin and dopamine, play critical roles in virtually all Metazoa taxa, exerting significant influence on both behavior and physiology particularly within the central nervous system. In Vertebrate species, their actions are mediated primarily through the biogenic amine receptor family, which includes dopamine (DRD), histamine (HRH), trace (TAAR), adrenergic (ADR), muscarinic cholinergic (mAChR), and most serotonin receptors (5HTR). Biogenic amine receptors belong to the broad family of G protein-coupled receptors (GPCR), one of the largest and most diverse receptor families in eukaryota, and indeed Metazoa. Indeed, due to the extensive diversity of biological functions they direct and the ongoing expansion of their ligand repertoire, GPCRs are considered one of the most evolutionarily innovative and successful gene families (Bockaert and Pin, 1999; Lagerstrom and Schioth, 2008).

Biogenic amine receptors fall in to the Rhodopsin-like (family "A") GPCR family, which emerged as a distinct GPCR clade with the origins of the Opisthokont (Fungi and Metazoa) lineage (Krishnan et al., 2012). Subsequently, the Rhodopsin-like family underwent substantial expansion in Metazoa, most notably in Vertebrate lineages (Rompler et al., 2007; Stäubert et al., 2013). Like all GPCRs, these receptors are highly structurally conserved, containing seven transmembrane (TM) domains separated by three extracellular (ECL) and three intracellular (ICL) loops with an N-outside C-inside orientation, and they propagate intracellular signaling through a G protein-mediated pathway. Biogenic amine receptors are frequent targets for a wide array of pharmaceuticals aimed to treat diseases such as schizophrenia, migraines, hypertension, allergies and asthma, and stomach ulcers Schoneberg et al. (2004); **?**); Mason et al. (2012).

In spite of these receptors' biological and clinical importance, studies on their evolution are relatively limited. Evolutionary studies which have focused on a biogenic amine receptors have predominantly been limited to single receptor subtypes, namely TAAR (Gloriam et al., 2005; Lindemann et al., 2005; Hashiguchi and Nishida, 2007), DRD (Callier et al., 2012; Yamamoto et al., 2013), and 5HTR

Anbazhagan et al. (2010). Moreover, many of these studies, and indeed larger-scale studies on the general evolution of the broad Rhodopsin-like GPCR family, have examined very narrow species distributions, for instance specifically teleosts (Gloriam et al., 2005), primates (Anbazhagan et al., 2010), humans and mice (Vassilatis et al., 2003; Kakarala and Jamil, 2014), or even strictly humans Fredriksson et al. (2003). Therefore, studies which account for the full breadth of known bioamine receptor sequences as well as their broad species distributions

This dearth of biogenic amine receptor-specific evolutionary understanding is underscored by the difficulties in establishing a robust multiple sequence alignment (MSA). Commonly used to identify conserved sequence motifs and functionally important residues as well as investigate the sequences' evolutionary history, MSAs provide the foundation for nearly all comparative sequence analyses. As MSA errors readily introduce bias these downstream analyses (Ogden and Rosenberg, 2006; Wong et al., 2008; Jordan and Goldman, 2012), it is crucial to ensure accuracy in the alignment to the extent possible. For GPCR sequences, in particular, any alignment should recapitulate their canonical 7TM structure, which a naive alignment of sequences cannot necessarily accomplish. While there are certain MSA software platforms which explicitly incorporate structural information into the alignment algorithm (e.g. 3DCoffe (O'Sullivan et al., 2004) and PROMALS3D (Pei et al., 2008)), these programs are fairly computationally-intensive and thus unwieldy (ill-suited?) for large-scale (over 1000 sequences) applications. Moreover, such programs require the use of a single crystal structure to guide sequence alignment. While all GPCRs have the same conserved 7TM domains, different GPCRs, and indeed the biogenic amine receptors, feature a wide variety of ICL and ECL sizes. For instance, the ECL3 lengths for human HRH1 and DRD3, respectively, are roughly 27 and 117, and their respective ICL3 lengths are 68 and 14 (as predicted by GPCRHMM (Wistrand et al., 2006)). Thus, aligning with a single crystral structure will not effectively represent the domain variability across biogenic amine receptor subtypes. A desirable alignment strategy would ensure that all sequences are anchored by their conserved 7TM domains without inappropriately constraining the heterogeneous ECL and ICL domains.

We therefore have adopted a novel iterative alignment strategy to create a large (3064 sequences), structurally-curated alignment of vertebrate biogenic amine receptors. In order to ensure proper structural alignment, we employed the software GPCRHMM (Wistrand et al., 2006), which uses a hidden markov model approach to assign each residue in a given GPCR sequence to its respective domain, either extracellular, transmembrane, or intracellular. Previously validated with real GPCR crystal structures (Spielman and Wilke, 2013), GPCRHMM relies on the remarkably-conserved GPCR 7TM structure to predict GPCR domains with exceptional accuracy and without the use of a crystal structure. Our alignment strategy therefore ensured that all predicted 7TM domains aligned appropriately across all sequences. Importantly, this strategy did not require any manual or visual data inspection, thus removing any potentially confounding subjectivity.

We present this biogenic amine receptor MSA as a resource for any group interested in studying the dynamic evolutionary processes, structural, and/or functional constraints operating on this exceptionally important GPCR clade. Moreover, to demonstrate the utility of our alignment, we constructed a phylogeny of our receptors with this alignment as well as a naive alignment which did not undergo this iterative process. We found that our structurally-curated alignment offered dramatic improvements over a standard alignment, and thus we additionally report the most comprehensive and supported vertebrate biogenic amine receptor phylogeny to date. From this phylogeny, we are able to discern the relationships among receptor types with a far increased sensitivity than previous analyses. Even so, we are unable to fully resolve the deeper nodes in the phylogeny, revealing the capabilities and limitations of current-day phylogenetic methods.

Any large-scale study requires a robust sequence alignment to identify protein homology, and then can examine the overlap between receptor-specific trends and broader patterns which fall across all of these things. Furthermore, although biogenic amine receptors are relatively more straightforward to generate pharms for than other GPCRs, their therapeutics still suffer from a lack of complete specificty. Having a thorough understanding of the general patterns and rules governing this clade's sequence evolution, and indeed understanding the permissible amino acids at structurally important positions, may enable substantial progress in homology modeling and pharm development for these important receptors.

RESULTS

Constructing MSAs for biogenic amine receptors

We collected all sequences using PSI-BLAST and filtered data to ensure a high-quality data set (see *Methods* for details). We additionally retained only sequences which the program GPCRHMM unequivocally classified as a true GPCR sequence, leaving a total dataset of 3464 receptor sequences to align.

Using the protein sequences for these 3464 receptors, we built an MSA according to the iterative strategy outlined in Figure 1. The resulting alignment, obtained after four iterations of this algorithm, contained a total of 3039 sequences. However, according to this algorithm, sequences in which up to 5% of positions do not match the consensus domain are retained. Therefore, certain residues which lie at domain boundaries may be improperly aligned with regards to structure. Thus, we created a modified version of this structural alignment in which all residues whose GPCRHMM-assigned domain did not agree with its respective column's consensus domain were masked. In total, 2.69% of all positions in the alignment were masked. We refer to the former MSA as the "structural unmasked" MSA and the latter as the "structural masked" MSA.

Masked structural alignment greatly improves phylogenetic inference

To demonstrate the utility of our bioamine receptor MSA, we constructed five distinct maximum likelihood (ML) bioamine receptor phylogenetic trees, as detailed in Table 1. First, we built a phylogeny using an MSA, again built with MAFFT (Katoh and Standley, 2013), comprised of the full set of 3464 receptors; in other words, this MSA was not subjected to the iterative process shown in Figure 1 and was therefore not explicitly structurally-curated. We refer to this MSA as the "naive" MSA. We additionally constructed two phylogenies each from the structural masked and structural unmasked MSAs, such that one phylogeny contained two distinct partitions, and hence evolutionary models, for the TM and EM domains, whereas the other phylogeny had a single partition representing the whole sequence. As we have the structural info and we know that TM and EM evolve very differently with distinct base frequencies, we can get a way better tree. Typically, alignments are partitioned for phylogenetic analysis when multiple genes have been concatenated, or to allow for different evolutionary regimes at different codon positions. Here, we can partition based on structure!

In general, having partitions was far superior to not having partitions, and the best tree based on AIC was clearly the SM tree, ultimately revealing the dramatic improvement that structurally-aware analyses provide.

Phylogeny allows insight into some, but not all, of the relationships

Here I should have a little paragraph for each clade. I should then have a final paragraph addressing what phylogenetics can and cannot accomplish. Cite that Morrison paper, and cite the

Missing data vs evolutionary event. Current phylo methods treat gaps simply as missing data and assume complete homology in columns. However, these are not missing data. The alignment was XXX long, but the longest GPCR is only YYY long. In reality, esp. given the duplication and neo/subfunctionalization inherent to GPCR evolution, domain sizes changes are key, and hence gaps should be evolutionarily meaningful and in fact understanding indel evolution is critical to understanding this gene family's evol. Other studies have attempted syntenic analyses based on hypothetical ancestral genomes, which has shown some promise. However, a more promising future avenue of research would carefully parse out how to treat gaps evolutionarily. Also can mention PRANK which using this indel info to construct far superior alignments relative to other methods, but that more development is clearly needed since PRANK incurs very ong runtimes and is thus ill-suited for an analysis of this scope. Also it's not a phylogenetic method.

CONCLUSIONS

Please feel free to use our alignment for evolutionary analysis of amine receptors, GPCRs, transmembrane domains. Also have great use for identifying motifs, possible pharmaceutical application.

METHODS

Sequence Collection and Processing

We collected protein sequences using PSI-BLAST (Altschul et al., 1997), specifically from the RefSeq (v2̇2̇9+) database (Pruitt et al., 2013), for 42 distinct human biogenic amine receptor sequences representing the full range of known receptors in the human genome. PSI-BLAST parameters included 5 iterations, an e-value of 10^{-20}, a sequence identity threshold of 25%, and a length difference of 50%. After combining all sequences recovered from the individual PSI-BLAST searches, we discarded duplicate sequences, leaving a total of 4232 PSI-BLAST results. We then filtered this sequence set by removing sequences from non-vertebrate taxa, sequences annotated as low-quality, pseudogene, and/or partial, and sequences which contained more than 1% ambiguous (i.e. B, X, or Z) residues. We additionally removed any sequences that could not be robustly considered GPCRs. We used the program GPCRHMM (Wistrand et al., 2006) to determine whether a given sequence was indeed a GPCR, and we discarded sequences which had either a local or global GPCRHMM score less than 10. Both of these thresholds are extremely conservative. Thus, while it is possible that some true GPCRs were discarded, these high thresholds for both local and global scores provide a very high confidence that all retained sequences were true GPCRs. Taken together, these filters left a total of 3464 receptor sequences.

Sequence Alignment

Before aligning sequences, we used the program GPCRHMM (Wistrand et al., 2006) to assign each residue in all protein sequences to its respective structural domain, extracellular, transmembrane, or intracellular. To ensure that all residues were assigned to a particular domain, we used a GPCRHMM posterior probability threshold of 0.5 to determine each residue's domain.

As all GPCRs are comprised of precise and highly conserved structural domains, it is critical that any sequence alignment of GPCRs maintains this overall homologous structure. In particular, the 7 TM domains should properly align with one another, across all sequences. Unfortunately, current alignment methods which consider structural information, such as 3DCoffee (O'Sullivan et al., 2004) and PROMALS3D (Pei et al., 2008) are ill-suited for such large datasets. Therefore, we adopted a novel alignment strategy to ensure that the resulting alignment accuractely reflected the highly conserved overarching GPCR structure. We began by assigning, again using GPCRHMM (Wistrand et al., 2006),

Our alignment strategy, outlined in Figure 1, consisted of a series of alignment iterations. Within each iteration, we first obtained an alignment using MAFFT (v7.149b) Katoh and Standley (2013). Next, using the GPCRHMM-determined residue domains, we determined the consensus structural domain for each alignment column, and we discarded all sequences in which 5% of columns did not match this consensus structure. We then realigned the remaining sequences, performing this strategy until no sequences were discarded.

It took four iterations to achieve a stable alignment, consisting of a total of 3039 sequences. In addition to this final alignment, we additionally created a "masked" alignment, in which protein residues which did not conform to their respective consensus domains were replaced with a "?", thus effectively replacing these positions with an ambiguous character. Therefore, while the final alignment might contain a few improperly aligned residues, this issue is avoided in the masked alignment as each column contains strictly properly-aligned residues.

ACKNOWLEDGMENTS

Funded by NIH, army, probably.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:6:716–723.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402.

Anbazhagan, P., Purushottam, M., Kiran Kumar, H. B., Mukherjee, O., Jain, S., and Sowdhamini, R. (2010). Phylogenetic analysis and selection pressures of 5-HT receptors in human and non-human primates: Receptor of an ancient neurotransmitter. *Journal of Biomolecular Structure and Dynamics*, 27(5):581–598.

Bockaert, J. and Pin, J. P. (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *The EMBO Journal*, 18(7):1723–1729.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol Method Res*, 33:261–304.

Callier, S., Snapyan, M., Crom, S., Prou, D., Vincent, J.-D., and Vernier, P. (2012). Evolution and cell biology of dopamine receptors in vertebrates. *Biology of the Cell*, 95(7):489–502.

Fredriksson, R., Lagerstrom, M., Lundin, L., and Schioth, H. (2003). The G-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, 63(6):1256–1272.

Gloriam, D. E. I., Bjarnadóttir, T. K., Yan, Y.-L., Postlethwait, J. H., Schioth, H. B., and Fredriksson, R. (2005). The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish. *Molecular Phylogenetics and Evolution*, 35(2):470–482.

Hashiguchi, Y. and Nishida, M. (2007). Evolution of trace amine associated receptor (taar) gene family in vertebrates: Lineage-specific expansions and degradations of a second class of vertebrate chemosensory receptors expressed in the olfactory epithelium. *Mol Biol Evol*, 24(9):2099–2107.

Jordan, G. and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*, 29:1125–1139.

Kakarala, K. K. and Jamil, K. (2014). Sequence-structure based phylogeny of GPCR class A rhodopsin receptors. *Molecular Phylogenetics and Evolution*, 74(C):66–96.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 30:772–780.

Krishnan, A., Almén, M. S., Fredriksson, R., and Schioth, H. B. (2012). The origin of gpcrs: Identification of mammalian like rhodopsin, adhesion, glutamate and frizzled gpcrs in fungi. *PLoS ONE*, 7(1):e29817.

Lagerstrom, M. C. and Schioth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov*, 7:339–357.

Le, S. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol*, 25:1307–1320.

Lindemann, L., Ebeling, M., Kratochwil, N. A., Bunzow, J. R., Grandy, D. K., and Hoener, M. C. (2005). Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics*, 85(3):372–385.

Mason, J. S., Bortolato, A., Congreve, M., and Marshall, F. H. (2012). New insights from structural biology into the druggability of G protein-coupled receptors. *Trends in Pharmacological Sciences*, 33(5):249–260.

Ogden, T. H. and Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, 55(2):314–328.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340:385–395.

Pei, J., Kim, B. H., and Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nuc Acids Res*, 36(7):2295–2300.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., and Ostell, J. M. (2013). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*, 42(D1):D756–D763.

Rompler, H., Staubert, C., Thor, D., Schulz, A., Hofreiter, M., and Schöneberg, T. (2007). G protein-coupled time travel. evolutionary aspects of GPCR research. *Molecular Interventions*, 7(1):17–25.

Schoneberg, T., Schulz, A., Biebermann, H., Hermsdorf, T., H, R., and Sangkuhl, K. (2004). Mutant G protein-coupled receptors as a cause of human diseases. *Pharmacol Ther*, 104:173–206.

Spielman, S. J. and Wilke, C. O. (2013). Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J Mol Evol*, 76(3):172–182.

Stäubert, C., Le Duc, D., and Schöneberg, T. (2013). Examining the dynamic evolution of G protein-coupled receptors. pages 23–43. Humana Press, Totowa, NJ.

Vassilatis, D. K. et al. (2003). The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci USA*, 100(8):4903–4908.

Wistrand, M., Käll, L., and Sonnhammer, E. L. L. (2006). A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Prot Sci*, 15(3):509–521.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476.

Yamamoto, K., Mirabeau, O., Bureau, C., Blin, M., Michon-Coudouel, S., Demarque, M., and Vernier, P. (2013). Evolution of dopamine receptor genes of the D1 class in vertebrates. *Mol Biol Evol*, 30(4):833–843.

FIGURES AND TABLES

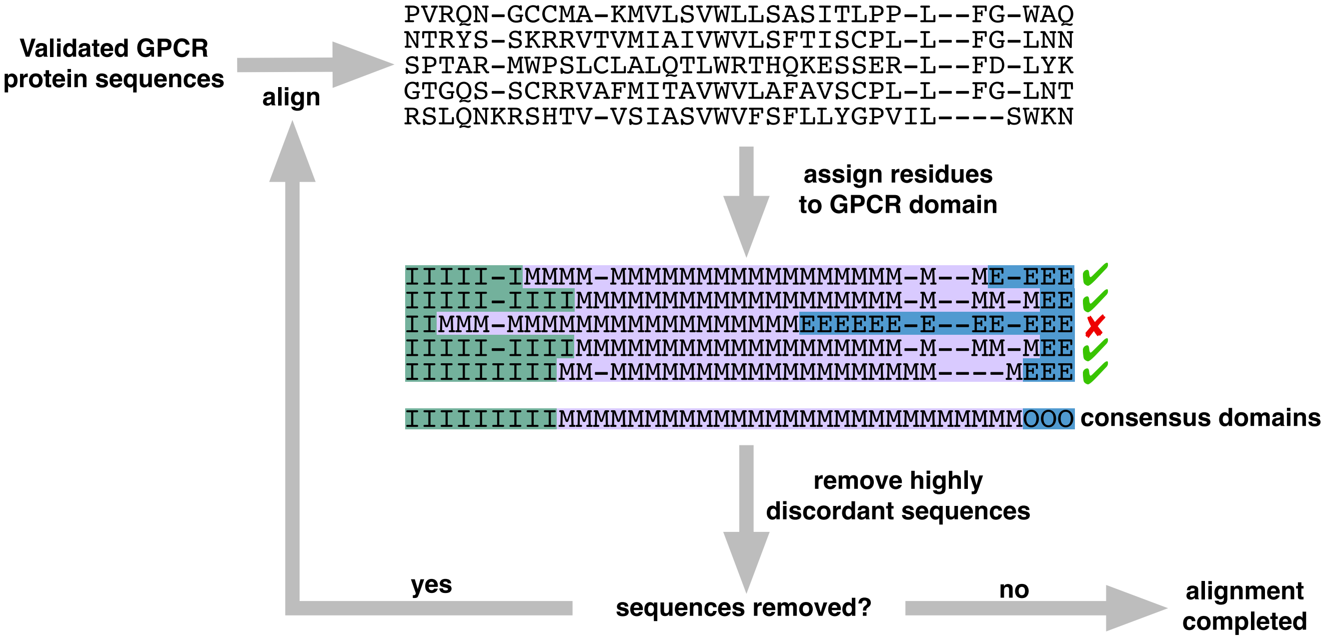
Validated GPCR protein sequences
align
PVRQN-GCCMA-KMVLSVWLLSASITLPP-L--FG-WAQ
NTRYS-SKRRVTVMIAIVWVLSFTISCPL-L--FG-LNN
SPTAR-MWPSLCLALQTLWRTHQKESSER-L--FD-LYK
GTGQS-SCRRVAFMITAVWVLAFAVSCPL-L--FG-LNT
RSLQNKRSHTV-VSIASVWVFSFLLYGPVIL----SWKN
assign residues to GPCR domain
IIIII-IMMMM-MMMMMMMMMMMMMMMMM-M--ME-EEE ✓
IIIII-IIIIMMMMMMMMMMMMMMMMMMM-M--MM-MEE ✓
IIMMM-MMMMMMMMMMMMMMMMMEEEEEE-E--EE-EEE ✗
IIIII-IIIIMMMMMMMMMMMMMMMMMMM-M--MM-MEE ✓
IIIIIIIIIMM-MMMMMMMMMMMMMMMMMMM----MEEE ✓
IIIIIIIIIMMMMMMMMMMMMMMMMMMMMMMMMMMMOOO consensus domains
remove highly discordant sequences
sequences removed?
yes
no
alignment completed

Figure 1. Iterative alignment strategy to create a structurally-curated multiple sequence alignment. Starred sequence is consensus structure. I=inner, M=membrane, O=outer. Sequences were removed if ≥ 5% of columns belonged to a different structural domain than the respective consensus domain.

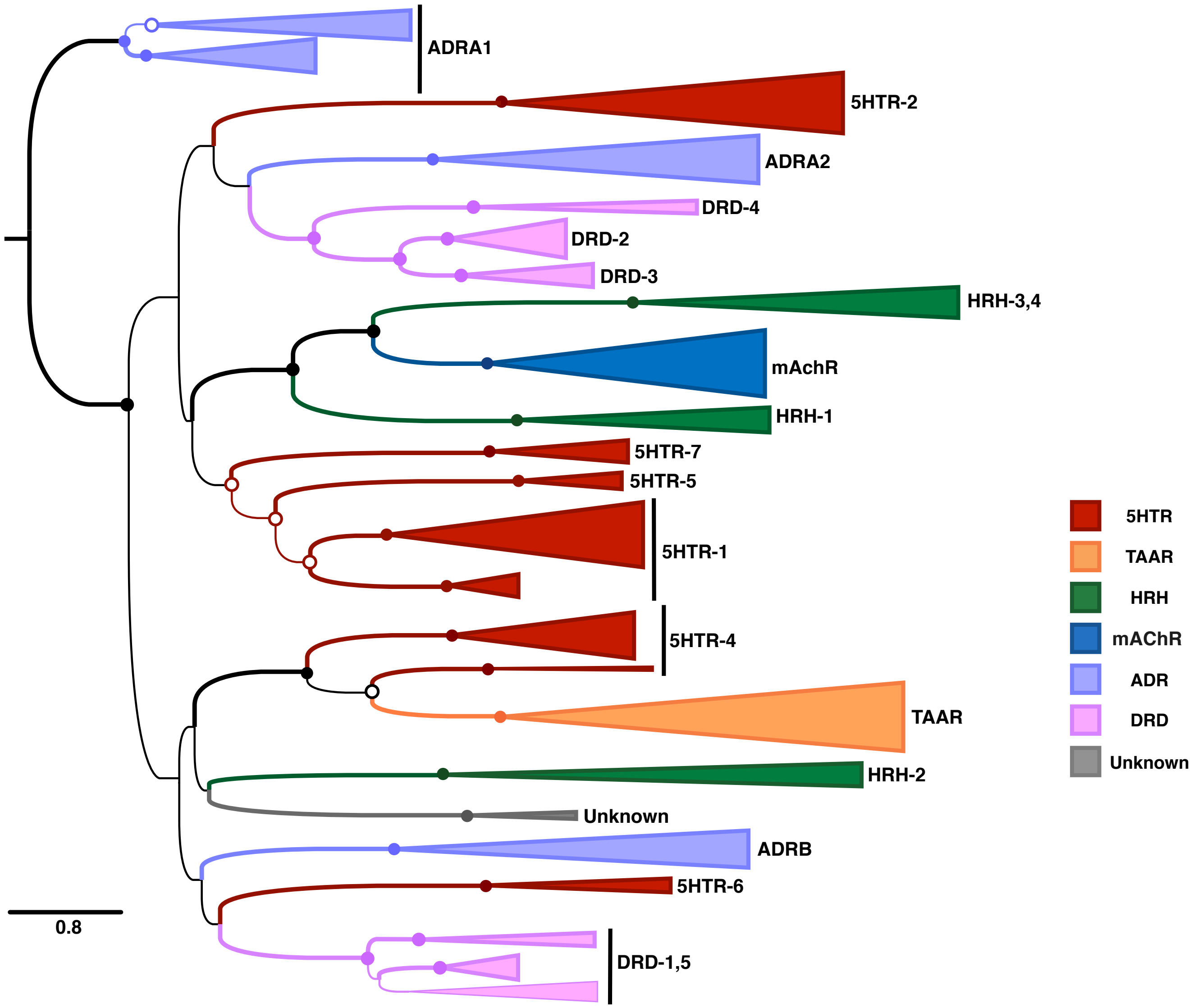

Figure 2. Maximum likelihood phylogeny of Vertebrate bioamine receptors built using the masked structural protein MSA alignment in RAxML. Nodes with open circles indicate ≥ 50% bootstrap support, and nodes with closed circles and thick lines indicate ≥ 90% bootstrap support. Bioaminergic receptors are abbreviated as 5HTR, serotonin receptors; TAAR, trace amine-associated receptors; HRH, histamine receptors; mAChr, muscarinic acetylcholine receptors; ADR, adreneric receptors; and DRD, dopamine receptors. The clade labeled "Unknown" could not be clearly identified as one of the major receptor types. Likely, this is a herp-specific gene...

Alignment	Partitions?	$\ln L$	k	ΔAIC
Structural Masked	Yes	-505500.8	6115	0
Structural Masked	No	-515991.7	6095	1752
Structural Unmasked	Yes	-515343.6	6115	19685
Structural Unmasked	No	-515991.7	6095	20941
Naive	No	-589703.7	6945	170047

Table 1. ΔAIC scores relative to the best performing for phylogenies. AIC is computed as $\mathrm{AIC} = 2(k - \ln L)$, where k is the number of free parameters of the model, and $\ln L$ is the log-likelihood (Akaike, 1974; Burnham and Anderson, 2004).