Membrane Environment Imposes Unique Selection Pressures on Transmembrane Domains of G-protein Coupled Receptors

Stephanie J. Spielman[1] and **Claus O. Wilke**[1]

[1]Section of Integrative Biology and Center for Computational Biology and Bioinformatics. The University of Texas at Austin, Austin, TX 78712, USA.

Abstract

We have investigated the influence of the plasma membrane environment on the molecular evolution of G protein-coupled receptors (GPCRs), the largest receptor family in Metazoa. In particular, we have analyzed the site-specific rate variation across the two primary structural partitions, transmembrane (TM) and extramembrane (EM), of these membrane proteins. We find that transmembrane domains evolve more slowly than do extramembrane domains, though TM domains display increased rate heterogeneity relative to their EM counterparts. Although the majority of residues across GPCRs experience strong to weak purifying selection, many GPCRs experience positive selection at both TM and EM residues, albeit with a slight bias towards the EM. Further, a subset of GPCRs, chemosensory receptors (including olfactory and taste receptors), exhibit increased rates of evolution relative to other GPCRs, an effect which is more pronounced in their TM spans. Although it has been previously suggested that the TM's low evolutionary rate is caused by their high percentage of buried residues, we show that their attenuated rate seems to stem from the strong biophysical constraints of the membrane itself, or by functional requirements. In spite of the strong evolutionary constraints acting on the transmembrane spans of GPCRs, positive selection and high levels of evolutionary rate variability are common. Thus, biophysical constraints should not be presumed to preclude a protein's ability to evolve.

Protein evolution G protein-coupled receptors Membrane proteins Positive selection

Introduction

A protein's evolution may be constrained by various functional or biophysical requirements. Membrane proteins, in particular, should be heavily constrained by the hydrophobic environment inside the membranes where they reside, specifically with regards to their transmembrane (TM) domains. This structural constraint biases amino acids found in TM domains towards non-polar, or hydrophobic, residues; polar amino acids comprise roughly 60% of TM segments, compared to a 30% frequency in extramembrane (EM) regions, whereas polar amino acids make up a mere 5% of the TM [Tourasse and Li, 2000]. Although a protein's amino-acid composition is not a robust determinant of evolutionary rate, the underlying biophysical constraints yielding this bias presumably enforce a lower rate of evolution in TM regions relative to globular proteins or to EM regions of the same protein [Tourasse and Li, 2000, Julenius and Pedersen, 2006]. The high concentration of buried residues in TM domains has additionally been proposed to be a dominant contributor to their low evolutionary rate [Stevens and Arkin, 2001, Oberai et al., 2009], as highly buried protein residues are known to correlate with low evolutionary rates [Franzosa and Xia, 2009, Ramsey et al., 2011].

---

*Corresponding author: stephanie.spielman@utexas.edu.

Although the general patterns associated with membrane evolution have been loosely characterized, the evolutionary variability within the TM and EM spans, particularly across individual residues, is largely unknown. Previous studies investigating the evolution of membrane proteins have focused primarily on average evolutionary rates, neither addressing rate heterogeneity nor site-based evolutionary parameters [Tourasse and Li, 2000, Gilad et al., 2000, Clark et al., 2003, Julenius and Pedersen, 2006]. Additionally, those studies used either orthologous pairs or trios of sequences, which hindered statistical robustness and precluded any analysis of site-rate variation due to a dearth of data [Tourasse and Li, 2000, Gilad et al., 2000, Clark et al., 2003, Julenius and Pedersen, 2006].

To obtain a more complete picture of membrane protein evolution, we have analyzed the evolutionary constraints acting on one of the most diverse membrane protein gene families in Metazoa, the G protein-coupled receptor (GPCR) family. GPCRs are the frequent targets of structural and biochemical studies; over 40% of pharmaceuticals target GPCRs, and a multitude of diseases are caused by mutant GPCRs [Dorsam and Gutkind, 2007, Schoneberg et al., 2004, Kristiansen, 2004, Fredriksson et al., 2003]. Phylogenetic analyses have shown that GPCRs form five main families, with the vast majority of human receptors belonging to the Rhodopsin-like (Family A) clade [Fredriksson et al., 2003, Fredriksson and Schioth, 2005]. Owing to their enormous diversity of biological functions and the ongoing expansion of their ligand repertoire, GPCRs have been described as one of the most evolutionarily successful gene families [Bockaert and Pin, 1999, Lagerstrom and Schioth, 2008]. Although protein sequences among, and indeed within, GPCR families are widely divergent, all GPCRs share a common structure characterized by a N-outside C-inside orientation with seven TM alpha helices spanning the plasma membrane, separated by three intracellular and three extracellular loops.

GPCRs accept a wide variety of ligands, both endogenous (e.g. hormones, amines, or ions) and exogenous (e.g. odorants), and facilitate signal transduction through a G protein mediated pathway [Kristiansen, 2004, Lagerstrom and Schioth, 2008, Rosenbaum et al., 2009]. Although some larger ligands do bind the extracellular portion of GPCRs, nearly all family A GPCRs, as well as many members of other GPCR families, bind ligands within their TM [Vaidehi et al., 2002, Kristiansen, 2004, Bywater, 2005, Surgand et al., 2006, May et al., 2007, Park et al., 2008]. The notable exceptions to this trend are family C GPCRs, whose ligand-binding domains lie primarily in their extensive and diverse N-termini [May et al., 2007, Park et al., 2008, Lagerstrom and Schioth, 2008]. However, allosteric modulators acting on all GPCR families bind within the TM. This commonality highlights the key role that the TM plays in the regulation of protein activity [May et al., 2007, Lagerstrom and Schioth, 2008].

The TM domain is also a critical determinant of a GPCR's conformational state. Mutational studies have shown that altering specific residues in GPCR TM spans results in structural modifications that induce constitutive activity, regardless of ligand presence [Spalding et al., 1998, Lu and Hulme, 2000]. Maintaining the integrity of TM structure and sequence, then, is necessary for GPCRs to function properly.

As suggested by the strong biophysical, structural, and functional constraints imposed on GPCRs, one would expect that strong purifying selection dominates TM domain evolution. Alternatively, given the continued expansion of the GPCR gene family, notably of Rhodopsin family members such as olfactory receptors [Lagerstrom and Schioth, 2008, Niimura and Nei, 2003, Nei and Niimura, 2007], and of the array of ligands they receive, some positive selection should be detectable throughout GPCRs. As ligands tend to bind the TM, it is possible that positive selection there could drive the evolution of the GPCRs' expanding ligand repertoire. Here, we define positive selection as the ratio of the rate of

nonsynonymous substitutions to synonymous substitutions, *dN/dS*, also known as ω. When ω > 1, positive selection may be inferred; alternatively when ω < 1, there is evidence for purifying selection. Neutral evolution is indicated by ω = 1.

Through a large-scale analysis of 359 mammalian GPCRs, we show that, on average, the TM evolves more slowly than does the EM, a result which should apply to all membrane proteins. Analysis of site rate variation across all GPCRs reveals that, unexpectedly, the average evolutionary rate heterogeneity of the TM is greater than that of the EM, in spite of the stronger biophysical and functional constraints the TM experiences. We additionally find evidence of positive selection in roughly half of the proteins studied here, in both their EM and TM domains. Chemosensory receptors, which includes all GPCRs (olfactory, taste, and vomeronasal receptors) that interact with exogenous chemical stimuli [Mombaerts, 2004], exhibit accelerated evolution relative to non-chemosensory GPCRs. This effect is highly pronounced in chemosensory GPCR TM spans. Finally, contradictory to previous reports [Oberai et al., 2009], we show that the lowered evolutionary rate of TM domains cannot solely be attributed to increased residue burial by other protein residues, but instead seems to stem from the membrane environment itself.

Results

Extracellular and intracellular domains evolve under similar selective pressures

We implemented the Goldman Yang codon evolutionary model (GY94) to estimate an average evolutionary rate $\bar{\omega}$ for each protein using the HyPhy batch language [Goldman and Yang, 1994, Kosakovsky Pond et al., 2005]. We compared fits between three models—one with a single partition forcing both the TM and EM to evolve at an equal rate, one with two partitions (EM and TM), and one with three partitions (extracellular, TM, and intracellular). The latter two models allowed for each partition to have unique parameter values for $\bar{\omega}$, κ, *t*, and equilibrium codon frequency, where κ is the ratio of transition to transversion rates and *t* is time, or branch length (see Methods for details). For each gene, the three models were compared using Akaike Information Criterion (AIC) scores such that models with lower AIC scores were preferred [Akaike, 1974]. AIC scores are reported here as the difference of AIC scores (ΔAIC) between two competing models, averaged across all genes. A larger ΔAIC indicates more support for the preferred model.

The two-partition model, on average across all genes, performed significantly better than the model which considered all domains as a single evolutionary unit (ΔAIC ~ 100), and the three-partition model performed slightly better than the two-partition model (ΔAIC ~ 5). However, there was no evidence that intracellular and extracellular regions had different average evolutionary rates in the three-partition model (paired t-test between extracellular and intracellular $\bar{\omega}$ values, $p = 0.589$). Therefore, the three-partition model was likely preferred due the marked difference in κ between intracellular and extracellular regions (paired t-test between extracellular and intracellular κ values, $p = 4.628 \times 10^{-07}$). Because no difference was detected between intracellular and extracellular $\bar{\omega}$, the two-partition model was used for all subsequent evolutionary rate analyses for all proteins. In terms of selection pressures, therefore, EM domains should be viewed as a single evolutionary unit. Our finding contradicts previous studies which claimed that intracellular regions of membrane proteins evolved more slowly than extracellular regions [Julenius and Pedersen, 2006]. Our analysis shows no support for that hypothesis, likely due to our increased data sampling and more precise methodology; previous results may have been false positives.

TM domains evolve more slowly than EM domains

We first broadly assessed rate differences between the evolution of TM and EM domains for each protein by estimating a single global $\bar{\omega}$ for each partition. Results from this analysis supported the hypothesis that, on average, EM regions evolve faster than their respective TM regions (Figure 1A). 94% of the genes studied here (338 of 359) showed TM $\bar{\omega}$ values less than their gene's EM $\bar{\omega}$ (exact binomial test, $p < 10^{-15}$). A paired t-test comparing log-transformed EM and TM $\bar{\omega}$ values across each gene showed that EM rates are on average 0.094 greater than TM rates ($p < 10^{-15}$). We additionally found that the correlation between log-transformed EM rates and TM rates was highly significant ($r = 0.75$, $p < 10^{-15}$), indicating that each protein likely has its own characteristic rate of evolution.

Elevated evolutionary rate in chemosensory receptors

Roughly one-third of receptors we analyzed were chemosensory receptors (127 of 359), of which 4 were taste receptors and the remainder olfactory receptors. We found that, relative to non-chemosensory GPCRs, chemosensory receptors exhibit significantly elevated evolutionary rates in both TM regions (t-test between log-transformed chemosensory TM and non-chemosensory TM $\bar{\omega}$ values, $p < 10^{-15}$) and EM regions (t-test between log-transformed chemosensory EM and non-chemosensory EM $\bar{\omega}$ values, $p < 10^{-11}$), as shown in Figure 1B. The $\bar{\omega}$ values for chemosensory receptor TM domains are, on average, ~ 0.092 greater than those of non-chemosensory receptors, and the $\bar{\omega}$ values for chemosensory receptor EM domains are, on average, ~ 0.077 greater than in non-chemosensory EM domains.

To determine whether the TM or EM domains experience a greater evolutionary rate increase from chemosensory to non-chemosensory receptors, we compared the mean ratios of TM rate to EM rate between the two receptor types. We recovered a chemosensory ratio of 0.68 and a non-chemosensory ratio of 0.52 (independent samples t-test $p = 2.7 \times 10^{-8}$). That the chemosensory TM:EM rate ratio is, on average, significantly greater than the non-chemosensory TM:EM rate ratio demonstrates that the TM$\bar{\omega}$ increase from non-chemosensory to chemosensory receptors exceeds the EM $\bar{\omega}$ increase. Additionally, we performed a regression analysis with a TM $\bar{\omega}$ response and two predictors: EM $\bar{\omega}$ and receptor type (chemosensory or non-chemosensory). Both EM rates and receptor types have highly significant effects ($p < 10^{-15}$ and $p < 10^{-8}$, respectively) on TM rates. This result further supports our conclusion that TM rates increase more dramatically than do EM rates between non-chemosensory to chemosensory GPCRs.

We then examined whether it was more likely for TM or EM domains to exhibit a higher evolutionary rate in chemosensory receptors compared to non-chemosensory receptors. From an exact Fisher test, we recovered an odds ratio of 2.11 ($p = 0.02$) in favor of the TM. This result demonstrates that chemosensory receptors are twice as likely to have elevated $\bar{\omega}$ in TM spans than in EM regions, compared to non-chemosensory receptors.

We further sought to examine whether the elevated evolutionary rate of chemosensory receptors could be attributed to differential tissue expression. Indeed, evolutionary rates tend to be higher for proteins with a lower expression breadth, as may be the case for chemosensory receptors [Duret and Mouchiroud, 2000, Liao et al., 2007, Pal et al., 2006]. Though it was once presumed that olfactory receptor expression was restricted to olfactory epithelium [Buck and Axel, 1991], recent studies have revealed that olfactory receptors are expressed in a multitude of diverse tissues in mammals [Vanderhaeghen et al., 1997, Feldmesser et al., 2006, Zhang et al., 2007]. However, whether these receptors function in non-olfactory capacities is unknown. Thus, their activity may be limited to sensory tissue, which could cause their elevated evolutionary rates.

To assess the influence of expression breadth on evolutionary rate in GPCRs, we first obtained microarray expression data for 169 of our GPCRs from the Human Protein Atlas (http://www.proteinatlas.org) and regressed each gene's evolutionary rate on expression breadth and receptor type. We did not recover a significant relationship between evolutionary rate and expression breadth for G protein-coupled receptors (EM $p = 0.684$ and TM $p = 0.722$). However, the microarray data which we were able to collect was highly biased towards non-chemosensory receptors—only 12 of the genes for which we had expression data were chemosensory (1 taste and 11 olfactory). Therefore, that limited amount of chemosensory expression data relative to non-chemosensory expression data may have biased our conclusions regarding the influence of expression breadth on $\bar{\omega}$. Possibly, then, chemosensory receptor expression breadth may contribute to their higher $\bar{\omega}$ values, but we lacked the statistical power to detect such an effect here.

TM domains display increased rate heterogeneity

To assess evolutionary rate variation among sites, we calculated an ω for each residue of our 359 proteins using a random effects likelihood model (REL). From these rates, we determined the coefficient of variation for ω [CV(ω)] across partitions. We used CV(ω) as a proxy for rate heterogeneity. We found that the mean CV(ω) for TM domains was 0.402 greater than for EM domains (paired t-test between each protein's TM and EM CV(ω) values, $p < 10^{-15}$). This increased spread of rates in the TM regions revealed their more extensive rate heterogeneity relative to their EM counterparts (Figure 1C). This effect holds for both chemosensory and non-chemosensory receptors.

While the majority of sites in GPCRs are under strong purifying selection, we identified 157 proteins (over two-fifths of our data set) which show evidence of positive selection at some sites. Positively selected sites were identified as those residues with an $\omega > 1$. Of all proteins analyzed, 31.5% had EM residues with $\omega > 1$, and 20.9% had TM residues with $\omega > 1$. Figure 2 depicts the selective regimes for several genes.

To assess bias in the location of positively selected residues, we conducted a Cochran-Mantel-Haenszel Test, a stratified contingency table analysis of association, across all genes. Our overall contingency table was comprised of an array of 2×2 contingency tables for each gene, wherein each 2×2 table compared the number of positively and negatively selected sites in each partition. We recovered an overall odds ratio of 2.25 ($p < 10^{-15}$) in favor of EM. This result strongly suggested that positively selected residues were more than twice as likely to occur in the EM than in the TM. This trend held for both chemosensory and non-chemosensory receptors. Thus, even though there are more positively selected sites in EM domains relative to TM domains, we emphasize that positively selected residues are not uncommon in the TM. A list of all genes with positively selected residues can be found in accompanying Supplementary Information.

Slowed TM evolution is not caused by structure

Finally, we assessed the extent to which structure influences the evolutionary rate in GPCR TM domains. For this analysis, we calculated each residue's relative solvent accessibility (RSA) from ten empirical crystal and one theoretical GPCR structure (see Methods for details). These structures represent all the currently known GPCR structures from the PDB. This effort was motivated by previous studies which have suggested that TM domains evolve slowly due to their relatively high percentage of buried residues [Stevens and Arkin, 2001, Oberai et al., 2009]. In this context, being buried refers to burial by other protein residues in the polypeptide, not by the plasma membrane itself. Buried residues are known to correlate strongly with a lower evolutionary rate [Franzosa and Xia, 2009, Ramsey et al.,

2011]. RSA directly measures how buried or exposed residues are within a protein structure, making it an ideal metric for this analysis.

After RSA was calculated for residues of the aforementioned eleven proteins, we regressed each residue's ω on RSA and partition (TM or EM). Results from this regression are shown in Table 1. We systematically checked for interaction effects in each regression, and found that only two of the eleven proteins showed a significant RSA $\times$ partition interaction. Partition had a highly significant effect in eight of the remaining nine structures. These results demonstrate that the lowered rate of TM domains is not caused entirely by the higher percentage of buried residues they contain (Figure 3), as had previously been hypothesized [Stevens and Arkin, 2001, Oberai et al., 2009]. Rather, it seems that the membrane environment, rather than protein structure itself, contributes to the lowered ω values characteristic of TM residues.

Discussion

We have demonstrated that the average evolutionary rate $\bar{\omega}$ of GPCR TM domains is significantly less than that of EM domains, mirroring results of previous studies which have suggested this trend across several types of membrane proteins [Tourasse and Li, 2000, Julenius and Pedersen, 2006]. Additionally, we have found that rate heterogeneity in TM spans exceeds that in EM regions and that many GPCRs experience positive selection across both structural domains. The average evolutionary rate of chemosensory receptors is also significantly greater than that of non-chemosensory receptors, specifically in the TM domains. Finally, we find no evidence, contrary to previous hypotheses, that increased residue burial influences the attenuated evolutionary rate of TM residues. Many of these results are summarized with a representative protein, the nociceptin receptor OPRL1, in Figure 4.

Although we found that the TM does evolve more slowly than does the EM, we emphasize that residues under positive selection were not uncommon across TM regions. Indeed, we identified 157 proteins, 55 of which are olfactory receptors, out of the 359 proteins we studied whose TMs contained residues with $\omega > 1$. Thus, while biophysical constraints may have limited amino acid diversity in the TM, they did not preclude high rates of evolution at certain sites. Knowledge of positively selected sites within GPCRs may be useful for future biomedical research endeavors, as positive selection may be an indicator of a residue's functionality and potential use in drug development. A list of all GPCRs in this study with positive selected residues can be found in the Supplementary Information.

That TM rate heterogeneity exceeded EM rate heterogeneity was an unexpected result. Given the aforementioned structural and functional constraints, one might instead expect less variation across ω values of individual TM residues. Alternatively, while some key TM residues may experience strong selective constraints, other residues will be much less important to protein structure and/or function. The former residues should be under exceedingly strong purifying selection, while the latter residues should be under weak purifying selection. In this dichotomy, there will be a strong difference in ω values between the highly constrained residues and the weakly constrained residues. In the EM, however, even the most constrained residues are, on average, under weaker negative selection than are the most constrained TM residues. Thus, the difference between strongly and weakly negatively selected EM residues should be less than the difference between TM strongly and weakly negatively selected residues. Therefore, although somewhat unintuitive, the spread of evolutionary rates in the EM is smaller than in the TM.

Although other studies have previously investigated the evolutionary regimes in membrane proteins and olfactory receptors, our approach represents a dramatic methodological

improvement. First, while previous studies of membrane proteins, including GPCR olfactory receptors, have focused either on ortholog duos or trios [Tourasse and Li, 2000, Clark et al., 2003, Julenius and Pedersen, 2006, Gimelbrant et al., 2004, Nielsen et al., 2005], we have included up to 27 mammalian species per phylogenetic analysis (one phylogeny was created per gene). This increased breadth of species sampling should yield more robust conclusions. Specifically, we were able to infer the selective pressures at each residue rather than a single average ω for the whole protein. Had we not included that many species in our analyses, it would not have been possible to infer site-based evolutionary rates, the extent of rate heterogeneity, or positive selection at the residue level. Furthermore, previous studies of membrane proteins did not conduct paired analyses, but rather compared average rates among all TM domains to average rates among all EM domains [Tourasse and Li, 2000, Julenius and Pedersen, 2006]. As we have demonstrated, there is a strong and highly significant correlation ($r = 0.75$, $p < 10^{-15}$) between the TM and EM evolutionary rates within a single protein. Therefore, EM and TM $\bar{\omega}$ values within a single protein are not statistically independent, and a paired analysis as we have conducted is necessary to obtain statistically valid results.

Previous work has shown that transmembrane domains generally contain an increased proportion of buried residues relative to globular proteins or EM domains. This phenomenon is likely due to the highly packed arrangement of the TM span's constituent α-helices [Stevens and Arkin, 2001, Oberai et al., 2009]. Typically, residue burial has been determined using the metric relative solvent accessibility (RSA), which measures the extent to which a residue in a protein structure is buried or exposed by other residues in the protein (not by the plasma membrane). Thus, RSA characterizes the local environment of a residue based on the extent of inter-residue contact, such that lower RSA values indicate increased burial by nearby protein residues. RSA is also a robust constraint on protein evolution, with buried residues evolving more slowly than exposed residues [Franzosa and Xia, 2009, Ramsey et al., 2011]. It has thus been hypothesized that the lowered evolutionary rate of TM domains could be attributed to their high percentage of buried residues [Oberai et al., 2009]. Our evolutionary analysis of ten empirical and one theoretical GPCR structures, however, largely refutes this claim. We instead demonstrate that, while TM residues do display lower RSAs than do EM residues, this factor alone cannot explain the TM's lower evolutionary rate. Instead, we presume that the extreme biophysical constraints of the membrane environment as well as functional constraints are the leading factors which impose a lowered evolutionary rate on TM domains. As more empirical GPCR structures become available, this effect should be confirmed with larger data sets.

We have further demonstrated that chemosensory receptors exhibit increased rates of molecular evolution relative to other GPCRs. Although there are three main groups of chemosensory receptors (olfactory, taste, and vomeronasal receptors), we were only able to obtain mammalian orthologs for olfactory and taste receptors. As vomeronasal receptors specialize in detecting pheromones [Mombaerts, 2004], they should have highly species-specific sequences, thus making ortholog inference difficult.

Previous studies on chemosensory receptor evolution have specifically investigated olfactory receptor evolution, the most common and diverse chemosensory receptors. In general, olfactory receptors are one of most rapidly evolving gene families in human and other mammalian lineages [Gilad et al., 2000, Clark et al., 2003, Nielsen et al., 2005]. Indeed, mammals contain at least 1000 olfactory receptors, and lineage-specific evolution of olfactory receptor families has been documented in primate splits [Mombaerts, 2004, Gimelbrant et al., 2004, Gilad et al., 2005]. Although the olfactory receptor families are rapidly evolving, it has been suggested the the receptors themselves evolve primarily under weak purifying selection, and that there is no robust evidence for positive selection stronger

than would be expected for any gene family [Gimelbrant et al., 2004]. Our results indicate that, while weak purifying selection does dominate mammalian chemosensory receptor evolution, as noted by Gimelbrant et al. [Gimelbrant et al., 2004] with regards to olfactory receptors, their average evolutionary rate is still significantly greater than the mean rate for their GPCR parent gene family. However, we also found that chemosensory receptors are not enriched for positively selected sites relative to other GPCRs, despite their increased $\bar{\omega}$.

Given the rampant evolution of the number of olfactory receptors across species [Niimura and Nei, 2003, Nei and Niimura, 2007], their elevated $\bar{\omega}$ was not unexpected. From an ecological standpoint, a mammal's ability to sense a diverse array of odor- ant and taste compounds is key for survival and species recognition. Such selection pressures are widely presumed to cause the high rate of olfactory gene turnover in animals, and we further this argument to include these genes' elevated rate of molecular evolution. The environmental selective pressures which cause frequent changes in the number of olfactory receptors likely also lead to the increased evolutionary rates of chemosensory receptors. Although both the TM and the EM domains evolve more quickly than do other GPCRs, we emphasize that the TM domains exhibit a more dramatic rate increase. This difference in protein domains could be explained by the ligand-binding pockets in chemosensory receptors. As both odorants and taste molecules bind chemosensory receptors within the TM region [Mombaerts, 2004, May et al., 2007, Park et al., 2008, Lagerstrom and Schioth, 2008], positively selected residues in the TM span should broaden the diversity of odorants and tastes which mammals can sense. This widened diversity could contribute to key evolutionary processes, such as species recognition and speciation.

Based on our analysis of receptor protein evolution, we conclude that structural constraints do not always translate to constraints in evolutionary rate. Although biophysical considerations are important when assessing evolutionary parameters of different proteins, it should not be assumed that strong biophysical requirements limit a protein's ability to evolve, as reflected by the presence of positively selected residues in both the EM and TM. Our findings also shed light on the significant role that membranes play in constraining protein evolution, such that the hydrophobic environment imposes strong purifying selection on membrane proteins.

Materials

Data collection and processing

Human genes associated with the Gene Ontology annotation "G protein-coupled receptor activity" (accession GO:0004930) were collected from Ensembl Biomart. Using Ensembl's gene orthology prediction method [Vilella et al., 2008], we obtained orthologs from 27 other mammalian species with available genomes in the Ensembl database, and retained those sequences which contained no ambiguous residues. Subsequent analyses included all genes with at least 10 orthologs. Protein alignments were performed using Mafft within the Guidance package, to ensure high alignment quality [Katoh et al., 2002, Penn et al., 2010]. As recommended by Privman et al. 2012 [Privman et al., 2012], we masked any residues in the resulting alignment with a guidance confidence score < 0.9 by changing their codons to "NNN". Phylogenies for each alignment were built using RAxML [Stamatakis, 2006] with 100 tree inferences, and the resulting best tree was kept.

Each human protein sequence was partitioned into three structural partitions— intracellular, transmembrane, and extracellular domains—using the software package GPCRHMM, which gave individual posterior probabilities for each site belonging to one of those three partitions [Wistrand et al., 2006]. Each site was categorized as either extracellular, intracellular, or transmembrane if its associated posterior probability was 0.95. All sites with posterior

probabilities below 0.95 were discarded. Each protein's partitions, as derived from the human sequence, were applied to all of its respective orthologs. Only genes with at least 50 amino acids per partition and whose TM comprised at least 15% of their total length were kept. Additionally, any sequences with less than 40% sequence identity to their orthologous human sequence were removed from alignments to ensure that all orthologs shared a common structure with the human protein. Positions corresponding to gaps in the human aligned sequence were removed. Sites belonging to each partition were concatenated such that each protein had a separate alignment for each region. Ultimately, 359 GPCR genes, averaging 18 sequences per alignment, were included in our analysis. Of these, 127 were chemosensory receptors (4 taste and 123 olfactory).

Evolutionary modeling to determine ω values

We calculated the site-based evolutionary rate $\bar{\omega}$ for each protein with the HyPhy batch language, using the Goldman Yang codon evolutionary model (GY94) [Goldman and Yang, 1994, Yang et al., 2000, Kosakovsky Pond et al., 2005]. This Markov process model for codon substitution of i to j (for $i \neq j$) is given by the instantaneous rate matrix

$$Q_{ij} = \begin{cases} 0 & \text{more than one nucleotide changes} \\ \pi_j & \text{synonymous transversion} \\ \kappa\pi_j & \text{synonymous transition} \\ \omega\pi_j & \text{nonsynonymous transversion} \\ \kappa\omega\pi_j & \text{nonsynonymous transition} \end{cases} , \quad (1)$$

where π_j is the frequency of codon j, κ is the ratio of transition to tranversion substitutions, and ω is the ratio of nonsynonymous to synonymous substitution rates. The indices i and j include all 61 sense codons. The transition probability matrix additionally considered time, or branch length t, as measured by the expected number of substitutions for each codon across all residues [Goldman and Yang, 1994, Yang et al., 2000].

To begin, we calculated an average evolutionary rate $\bar{\omega}$ for each protein to infer the optimal partitioning strategy for analyzing TM v.s. EM evolution. In this case, the ω in our GY94 matrix corresponded to an average ω ($\bar{\omega}$) over all sites. Three models of protein evolution were examined; the first considered the entire protein a single evolutionary unit (single partition model), the second partitioned the protein into two distinct regions of transmembrane and extramembrane residues (two-partition model), and the third model partitioned the protein into three regions of transmembrane, intracellular, and extracellular regions (three-partition model). Models allowed each partition its own $\bar{\omega}$, κ, and t parameters. To identify the optimal number of partitions for GPCRs, we compared model fits with the Akaike Information Criterion [Akaike, 1974]. AIC scores were calculated for each model of each gene and compared. The preferred model was the three-partition model. However, as there was no statistical difference between intracellular and extracellular $\bar{\omega}$ values in this model, the two-partition framework was used for all subsequent analyses.

We then implemented a random-effects likelihood (REL) model [Yang et al., 2000, Kosakovsky Pond and Frost, 2005], again using the GY94 rate matrix, to discern an ω value for each residue across all proteins. In particular, we followed the RSA-independent model described in Meyer et al. [Meyer and Wilke, 2012]. To determine the optimal number of rate categories for each protein's partition, we ran the model 25 times, allowing the number of rate categories in each of the two partitions to vary from one to five in all possible combinations. AIC scores were calculated for each model, and the model with the lowest resulting AIC score was selected as the best-fitting model for that protein.

To assign each site to a rate class, we employed an empirical Bayes approach [Nielsen and Yang, 1998] to calculate the posterior probability for each site belong to each rate class. Each site's rate was a weighted average over all rate classes by the associated posterior probability. To calculate an average evolutionary rate $\bar{\omega}$ for each protein's partition, we took the weighted average, by the model's prior probabilities, of the ω values from each rate class. The standard deviation of ω values per partition was calculated using all residue ω values and the average ω value in a partition. Subsequently, we calculated the coefficient of variation for each protein's partition by dividing each partition's standard deviation of ω by its respective mean rate, $\bar{\omega}$.

Structural analysis

Relative solvent accessibility (RSA) was calculated for residues of 10 empirical and 1 theoretical GPCR structures obtained from the protein data bank (PBD). These PDB IDs, along with their respective gene names in parentheses, are 2rh1 (ADBR2); 3uon (CHRM2); 4daj (CHRM3); 3oe6 (CXCR4); 3pbl (DRD3); 3rze (HRH1); 4ej4 (OPRD1); 4ea3 (OPRL1); 1f88 (RHO); 3v2w (S1PR1); and theoretical structure 1kpn (OPN1SW). For each structure, we calculated the surface area for each residue using DSSP [Kabsch and Sander, 1983] and normalized each value by its respective amino acid's maximum surface area value, as determined by Tien et. al. [Tien et al., 2012].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974; 19:716–723.

Bockaert J, Pin Jean Phillipe. Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J. 1999; 18:1723–1729. [PubMed: 10202136]

Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell. 1991; 5:175–187. [PubMed: 1840504]

Bywater, Robert P. Location and nature of the resiudes important for ligand recognition in G protein-coupled receptors. J Mol Recognit. 2005; 18:60–72. [PubMed: 15386622]

Clark AG, Glanowski S, Nielsen R, Thomas Paul D, Kejariwal A, Todd Melissa A, Tanenbaum DM, Civello D, Lu Fu, Murphy B, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 2003; 302:1960–1963. [PubMed: 14671302]

Dorsam RT, Gutkind JS. G coupled-protein receptors and cancer. Nat Rev Genet. 2007; 7:79–94.

Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 2000; 17:6874.

Feldmesser E, Oldener T, Khen M, Yanai I, Ophir R, Lancet D. Widespread ectopic expression of olfactory receptor genes. BMC Genomics. 2006; 7:121138.

Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 2009; 26:2387–2395. [PubMed: 19597162]

Fredriksson, Robert; Schioth, Helgi B. The repertoire of G protein-coupled receptors in fully sequenced genomes. Mol Pharmacol. 2005; 67:1414–1425. [PubMed: 15687224]

Fredriksson, Robert; Lagerstrom, Malin C.; Lundin, Lars-Gustav; Schioth, Helgi B. The G protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol. 2003; 63:1256–1272. [PubMed: 12761335]

Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat Genet. 2000; 26:221–224. [PubMed: 11017082]

Gilad Y, Man Orna, Glusman Gustavo. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res. 2005; 15:224–230. [PubMed: 15687286]

Gimelbrant, Alexander A.; Skaletsky, H.; Chess, Andrew. Selective pressures on the olfactory receptor repertoire since the human–chimpanzee divergence. Proc Natl Acad Sci USA. 2004; 101:9019–9022. [PubMed: 15178761]

Goldman, Nick; Yang, Ziheng. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994; 11:725–736. [PubMed: 7968486]

Julenius, Karin; Pedersen, Anders Gorm. Protein evolution is faster outside the cell. Mol Biol Evol. 2006; 23:2039–2048. [PubMed: 16891379]

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]

Katoh, Kazutaka; Misawa, Kazuharu; Kuma, Kei-Ichi; Miyata, Takashi. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30:3059–3066. [PubMed: 12136088]

Kosakovsky Pond, Sergei; Frost, Simon DW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005; 22:1208–1222. [PubMed: 15703242]

Kosakovsky Pond, Sergei L.; Frost, Simon DW.; Muse, Spencer V. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; 12

Kristiansen, Kurt. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. Pharmacol Ther. 2004; 103:21–80. [PubMed: 15251227]

Lagerstrom, Malin C.; Schioth, Helgi B. Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat Rev Drug Discov. 2008; 7:339–357. [PubMed: 18382464]

Liao BY, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 2007; 24 20722080.

Lu, Zhi-Liang; Hulme, Edward C. A network of conserved intramolecular contacts defines the off-state of the transmembrane switch mechanism in a seven-transmembrane receptor. J Biol Chem. 2000; 275:5682–5686. [PubMed: 10681552]

May, Lauren T.; Leach, Katie; Sexton, Patrik M.; Chistopoulus, Arthur. Allosteric modulation of G protein-coupled receptors. Annu. Rev. Pharmacol. Toxicol. 2007; 47:1–51. [PubMed: 17009927]

Meyer AG, Wilke CO. Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol. 2012 In press.

Mombaerts, Peter. Genes and ligands for odorant, vomeronasal and taste receptors. Nat Rev Neurosci. 2004; 5:263–278. [PubMed: 15034552]

Nei, Masatoshi; Niimura, Yoshihito. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS ONE. 2007; 2:e708. [PubMed: 17684554]

Nielsen, Rasmus; Yang, Ziheng. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998; 148:929–936. [PubMed: 9539414]

Nielsen, Rasmus; Bustamante, Carlos; Clark, Andrew G.; Glanowski, Stephen; Sackton, Timothy B.; Hubisz, Melissa J.; Fiedel-Alon, Adi; Tanenbaum, David M.; Civello, Daniel; White, Thomas J., et al. A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biol. 2005; 3:e170. [PubMed: 15869325]

Niimura, Yoshihito; Nei, Masatoshi. Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci USA. 2003; 100:12235–12240. [PubMed: 14507991]

Oberai, Amit; Joh, Nathan H.; Pettit, Frank K.; Bowie, James U. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. Proc Natl Acad Sci USA. 2009; 106:17747–17750. [PubMed: 19815527]

Pal C, Papp B, Lercher MJ. An integrated view of protein evolution. Nat Rev Genet. 2006; 7 337348.

Park, Paul S-H.; Lodowski, David T.; Palczewski, Kryzysztof. Activation of G protein-coupled receptors: beyond two-state models and tertiary conformational changes. Annu. Rev. Pharmacol. Toxicol. 2008; 48:107–141. [PubMed: 17848137]

Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol. 2010; 27:1759–1767. [PubMed: 20207713]

Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012; 29:1–5. [PubMed: 21772063]

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics. 2011; 188:479–488. [PubMed: 21467571]

Rosenbaum, Daniel M.; Rasmussen, Søren GF.; Kobilka, Brian K. The structure and function of G protein-coupled receptors. Nature. 2009; 459:356–363. [PubMed: 19458711]

Schoneberg T, Schulz A, Biebermann H, Hermsdorf T, Rompler H, Sangkuhl K. Mutant G protein-coupled receptors as a cause of human diseases. Pharmacol Ther. 2004; 104:173–206. [PubMed: 15556674]

Spalding, Tracy A.; Burstein, Ethan S.; Henderson, Shaw C.; Ducote, Karin R.; Brann, Mark R. Identification of a ligand-dependent switch within a muscarinic receptor. J Biol Chem. 1998; 273:21563–21568. [PubMed: 9705286]

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–2690. [PubMed: 16928733]

Stevens, Timothy J.; Arkin, Isaiah T. Substitution rates in alpha-helical transmembrane proteins. Prot Sci. 2001; 10:2507–2517.

Surgand JS, Rodrigo J, Kellenberger E, Rognan D. A chemogenomic analysis of the transmembrane binding cavity of human G protein-coupled receptors. Proteins. 2006; 62:509–538. [PubMed: 16294340]

Tien, Matthew; Meyer, Austin G.; Spielman, Stephanie J.; Wilke, Claus O. Maximum allowed solvent accessibilites of residues in proteins. 2012. arXiv:1211.4251 [qbio.BM].

Tourasse, Nicolas J.; Li, Wen-Hsiung. Selective constraints, amino acid composition, and the rate of protein evolution. Mol Biol Evol. 2000; 17:656–664. [PubMed: 10742056]

Vaidehi N, Floriano Wely B, Trabanino Rene, Hall Spencer E, Freddolino Peter, Choi Eun Jung, Zamanakos Georgios, Goddard William A III. Prediction of structure and function of G protein-coupled receptors. Proc Natl Acad Sci USA. 2002; 99:12622–12627. [PubMed: 12351677]

Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species. Genomics. 1997; 39:239–246. [PubMed: 9119360]

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2008; 19:327–335. [PubMed: 19029536]

Wistrand, Markus; Käll, Lukas; Sonnhammer, Erik LL. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. Prot Sci. 2006; 15:509–521.

Yang, Ziheng; Nielsen, Rasmus; Goldman, Nick; Pedersen, Anne-Mette Krabbe. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000; 155:431–449. [PubMed: 10790415]

Zhang, Xiaohong; De la Cruz, Omar; Pinto, Jayant M.; Nicolae, Dan; Firestein, Stuart; Gilad, Yoav. Characterizing the expression of the human olfactory receptor gene family using a novel DNA microarray. Genome Biol. 2007; 8
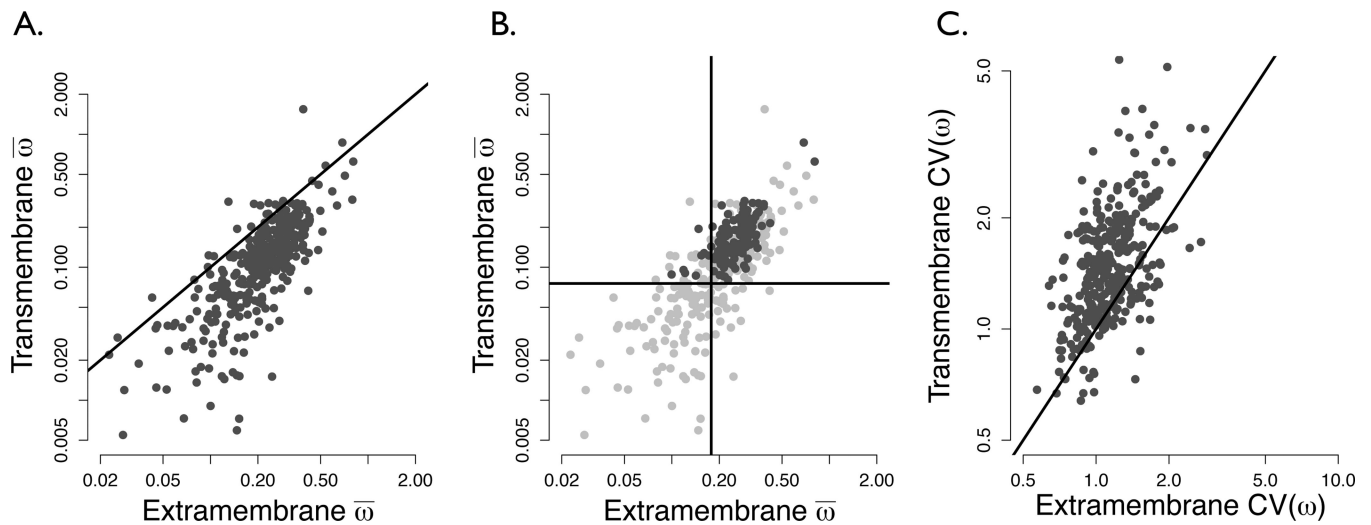
A.

B.

C.

Figure 1.
(**a**) TM $\bar{\omega}$ plotted against EM $\bar{\omega}$, both calculated by the REL model, shown on a log-log scale for all 359 proteins. The solid line indicates points where the EM rate equals that of the TM. The vast majority of proteins lie below this line (exact binomial test $p < 10^{-15}$), showing that TM domains evolve slower than EM domains. (**b**) Average TM ω plotted against average EM ω, showing different types of GPCRs, on a log-log scale. Dark gray points represent chemosensory receptors, and light gray points non-chemosensory receptors. Vertical and horizontal lines lie at the mean of nonchemosensory TM and EM average rates, respectively. The majority of chemosensory points lie fall in the top-right quadrant of this plot (exact binomial test $p = 6.85 \times 10^{-4}$), indicating their elevated evolutionary rate relative to non-chemosensory receptors. (**c**) Regression of TM against EM coefficients of variation of ω values [CV(ω)] on a log-log scale. If the spread of rates between partitions were equal, all points would lie roughly on the $x = y$ line shown. However, the majority of points lie on the TM side of the line, demonstrating the increased rate heterogeneity in TM domains of receptors proteins. This shift is highly significant at $p < 10^{-15}$ by the exact binomial test.
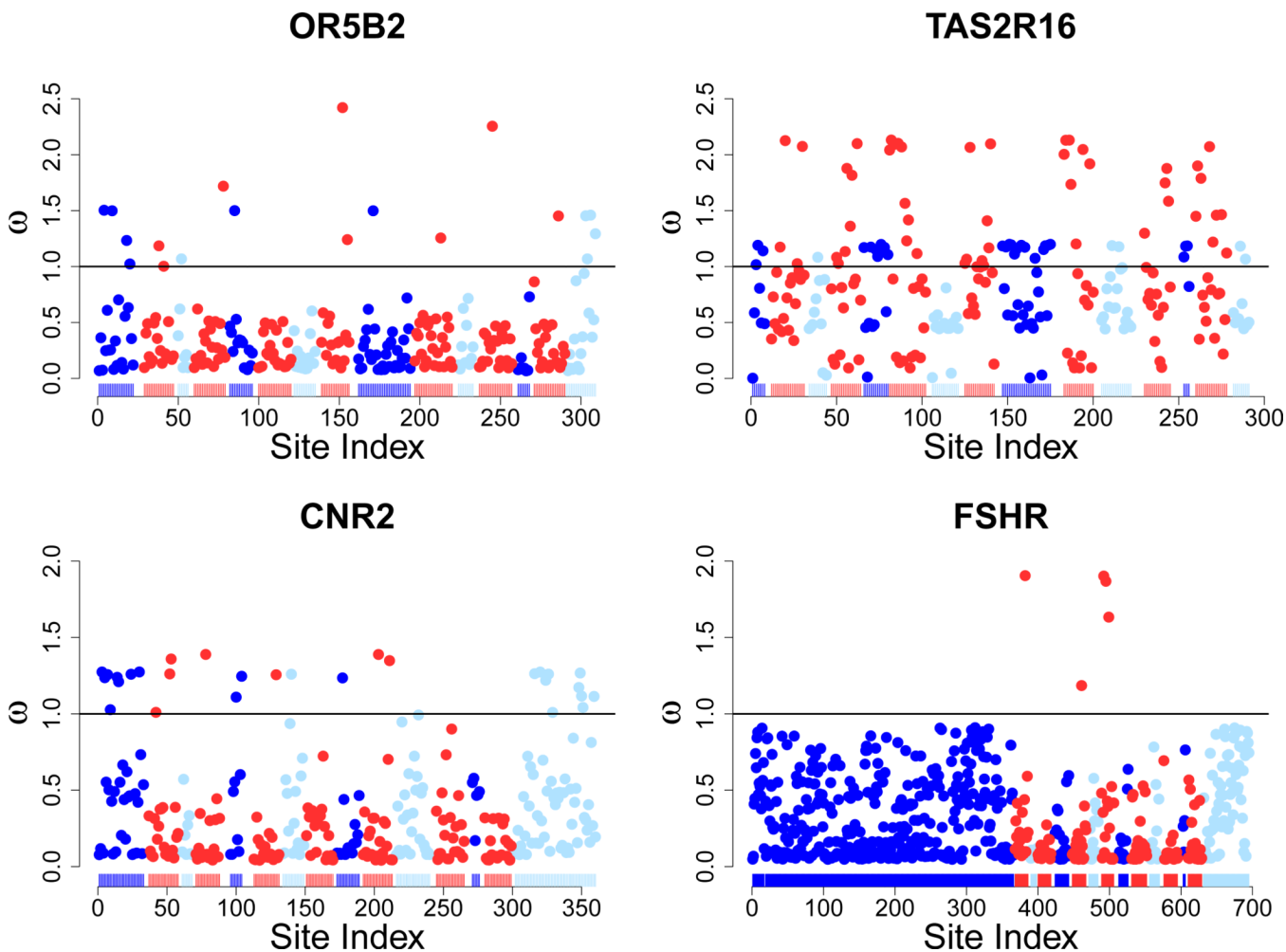
Figure 2.
Distribution of residue ω values across four example proteins that contain positively selected residues. OR5B2 and TAS2R16 are both chemosensory receptors (olfactory and taste, respectively), CNR2 is cannabinoid receptor 2, and FSHR is the follicle-stimulating hormone receptor. Each of these proteins contain positively selected residues, in particular the chemosensory proteins. Red points represent TM residues, light blue points represent intracellular residues, and dark blue points represent extracellular residues, all as predicted by GPCRHMM [Wistrand et al., 2006]. The bar at the bottom of the plots signifies the overall structure of the protein, with the same color-coding as the points. White spaces in between regions along the bottom bar indicate that the residue was not included in analysis, either due to lack of confidence in structure or alignment. Note that in each figure the entire structure of a GPCR is clearly visible - an extracellular N-terminus, intracellular C-terminus, three intracellular loops, three extracellular loops, and seven TM domains.
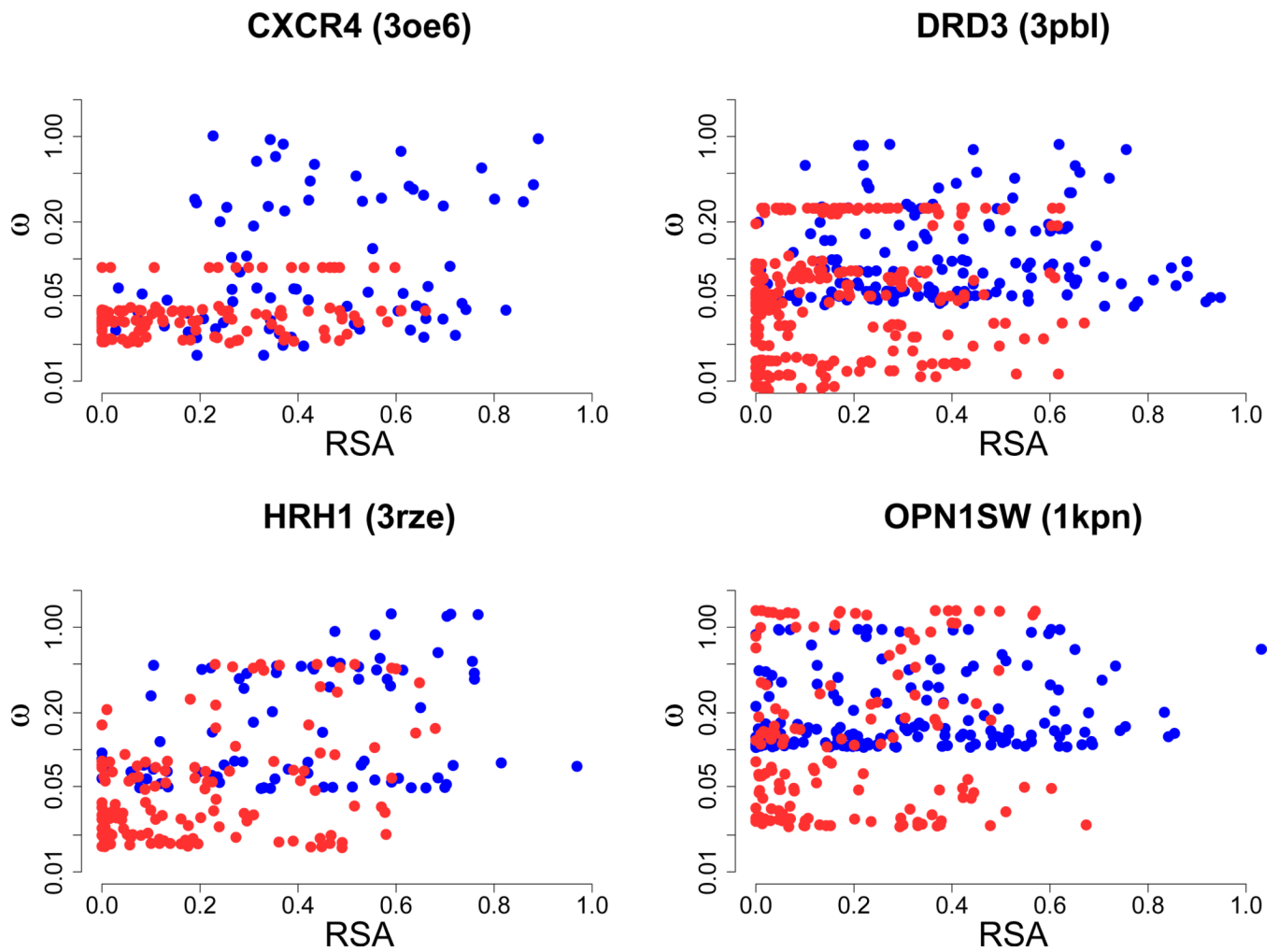
Figure 3.
Regression of the *log*(ω) against RSA and residue partition. Red points represent TM residues and blue points represent EM residues. The gene name with its PDB ID in parentheses is shown above each graph. Linear regressions show that partition has a highly significant effect for each protein shown here (see Table 1 for more details). Further, the TM points display a noticeable shift towards lower RSA values, reflecting the increased burial by neighboring residues experienced in that domain.
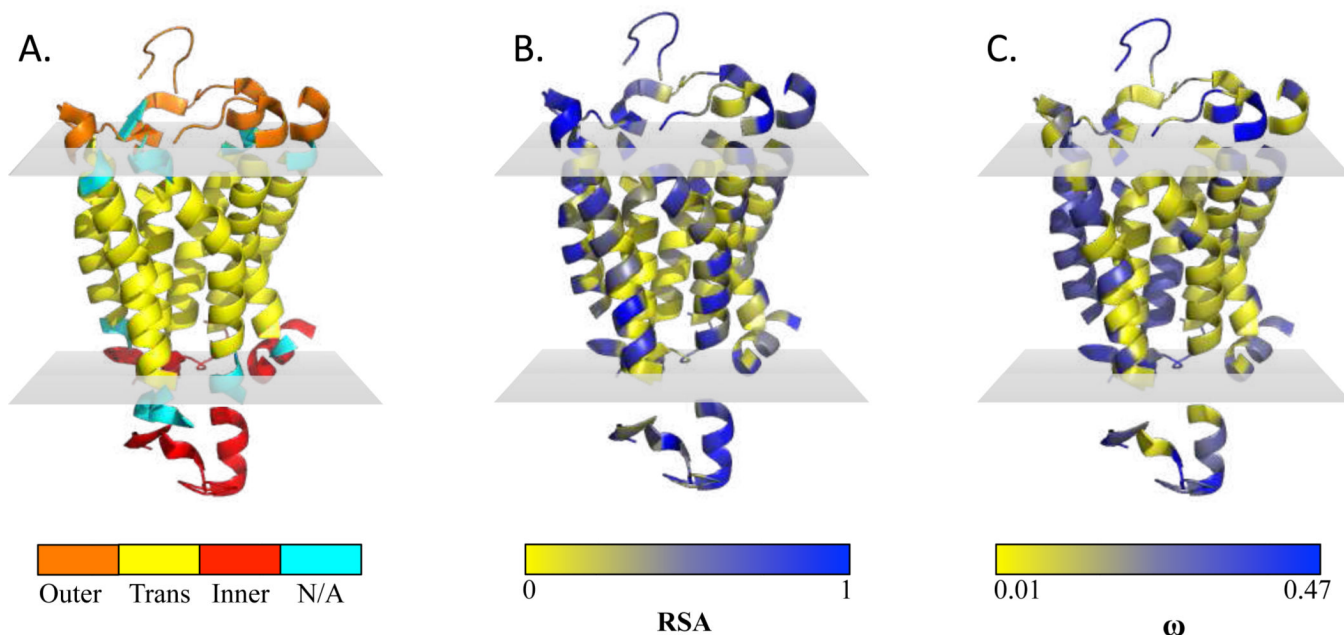
Figure 4.
Structure of nociceptin receptor OPRL1 (PDB ID 4ea3), where gray planes represent borders of the plasma membrane. (**a**) Predicted extracellular, TM, and intracellular regions by GPCRHMM, which match the true structure nearly perfectly. Average TM ω for this protein is 0.0686, and average EM ω is 0.0837. Cyan residues marked 'N/A' were excluded from analysis as GPCRHMM could not assign these residues with high confidence to a structural partition. These residues are not shown in parts B and C. (**b**) RSA (relative solvent accessibility) for each residue of OPRL1 analyzed. RSA values range from completed buried (0) to completed exposed (1). The vast majority of residues in the TM domain are buried whereas nearly all residues of the EM portions are highly exposed. (**c**) ω value at each residue of OPRL1 analyzed. Although each residue of this protein experiences purifying selection (ω < 1), rate heterogeneity is still pervasive throughout the protein.

Table 1

Results from the regression of log(ω) on RSA and partition (TM and EM) for each residue in 11 GPCR structures from the PDB. Significant values for partition and RSA are shown in bold. Empty values in the RSA × Partition area of the table indicate that no significant interaction effect was detected.

Gene Name	PDB ID	r^2	Partition	p-value	RSA	p-value	RSA × Partition	p-value
ADBR2	2rh1	0.16	**-0.28**	0.0295	**1.47**	1.62×10^{-6}		
CHRM2	3uon	0.82	**-1.58**	$<2 \times 10^{-16}$	0.12	0.34		
CHRM3	4daj	0.093	**-0.19**	3.61×10^{-5}	**0.68**	1.02×10^{-9}		
CXCR4	3oe6	0.26	**-0.77**	4.78×10^{-8}	**0.76**	0.0069		
DRD3	3pbl	0.14	**-0.58**	5.11×10^{-8}	**0.70**	0.0024		
H1R1	3rze	0.33	**-0.81**	1.99×10^{-7}	**1.77**	8.51×10^{-8}		
OPN1SW	1kpn	0.055	**0.46**	4.98×10^{-4}	0.49	0.11		
OPRD1	4eje	0.009	**0.22**	0.048	0.16	0.53		
OPRL1	4ea3	0.046	-0.044	0.647	**0.91**	5.91×10^{-5}		
RHO	1f88	0.017	-0.084	0.51	0.30	0.75	**1.33**	0.00467
S1PR1	3v2w	0.818	2.93	$<2 \times 10^{-16}$	1.02	1.66×10^{-5}	**-1.45**	9.59×10^{-5}