# Computers and drug discovery : construction and data mining of chemical and biological databases

Kazius, J.

# Chapter 4

# Spread and Disease Potential of Natural Variants in G Protein-Coupled Receptors

Human G protein-coupled receptors (GPCRs) regulate various bodily functions and many are important drug targets. The substantial inter-individual variation that exists in GPCR genes ranges from rare mutations that cause monogenetic diseases to polymorphisms that have no observable effects on phenotype. Whereas we discussed examples of disease-related genetic variants and other polymorphisms in Chapter 2, in the current Chapter we describe a large-scale statistical analysis of the natural genetic variants from the GPCR NaVa database that was described in Chapter 3. Here, this wealth of data was analysed to relate pathogenic effects of GPCR variants to variant-specific properties, and vice versa. Non-olfactory GPCRs contain more silent polymorphisms and fewer missense polymorphisms than olfactory GPCRs. We also learned that stronger risk factors for disease are either variants that cause more radical amino acid substitutions or variants that occur in transmembrane domains, particularly if they are inaccessible to solvent or membrane according to the three-dimensional structure of these transmembrane domains. Very conserved transmembrane positions also appeared to contain most disease mutations. In conclusion, by profiting from the readily accessible data in the GPCR NaVa database, we were able to reveal structural features of GPCRs that affect disease predisposition.

*J. Kazius, M van Iterson, K Ye, J Kok, T Bäck, A. P. IJzerman, submitted.*

## INTRODUCTION

The 773 currently known *G protein-coupled receptors (GPCRs)* form the largest protein superfamily in man as almost 3% of all human genes encode GPCRs[1]. GPCRs are key interaction points for various bodily functions, and many of today's medicines target a GPCR[2]. In more detail, GPCRs are cell-surface proteins that selectively convert an extracellular signal into an intracellular response. The variety of physiological functions that are regulated through GPCR-based signal transduction is also reflected in the diversity of endogenous ligands that target GPCRs. For instance, endogenous ligands for GPCRs can be photons, protons, small ligands, peptides or large proteins. The extracellular side of the membrane-bound GPCR is connected to the intracellular side via seven α helix-like *transmembrane domains (TMDs)*, which are the most conserved domains within GPCR families. Upon binding, ligands are thought to induce a conformational change in the TMDs and thereby trigger a cascade of intracellular responses, including G protein activation, that alter cell behaviour. While the sizes of TMDs are comparable in GPCRs, the sizes of other regions differ notably between receptors: GPCR protein sequences range from fewer then 300 to over 6000 amino acids in length.

The human GPCR superfamily can be divided into several families, such as the class A, class B, class C, frizzled/smoothened and taste receptor families[3]. Table 4.1 shows that the class A rhodopsin-like GPCR family is the largest subset as it contains 614 GPCRs, which form almost 80% of the human GPCRs and include the 371 olfactory GPCRs. Sequence motifs suggest that class A rhodopsin-like GPCRs are structurally related, and as they form the largest subfamily of comparable GPCRs, they contain the majority of known inter-individual *variants*.

**Table 4.1:** Number of human GPCRs per family.

| Class A (olfactory) | Class B | Class C | Frizzled/ Smoothened | Bitter taste | Vomero-nasal | Putative/ unclassified |
|---|---|---|---|---|---|---|
| 614 (371) | 29 | 20 | 11 | 25 | 5 | 69 |

Rare *mutations*, common *polymorphisms* or other natural variants in GPCR genes can affect their performance. Four studies have analysed variants at the GPCR family level through statistical methods. Small et al.[4] studied 412 SNPs, including 86 missense SNPs in particular, of the 675 variants they detected in 58 non-olfactory and 6 olfactory GPCR genes of healthy humans. Amongst others, they concluded that GPCR genes contain more coding variants than non-GPCR genes. Wahlestedt et al.[5] studied 392 missense variants of a set of 816 candidate polymorphisms in 338 non-olfactory and 89 olfactory GPCR genes. GPCRs that respond to biogenic amines were found to contain fewer variants in the TMDs than GPCRs that respond to peptides. Next to these two studies of (candidate) polymorphisms, two studies analysed characteristics of rare disease mutations in GPCRs. Lee et al.[6] studied 454

missense variants, 153 disease mutations and 301 candidate SNPs, that were located in 62 non-olfactory GPCR genes. One finding was that TMDs contain more disease variants and fewer candidate SNPs. Balasubramanian et al.[7] compared missense variations of species differences with disease-related mutations in non-olfactory GPCRs. They showed that TMDs contain more disease mutations than other variations. A second conclusion was that those amino acid substitutions that cause large physicochemical disruptions are more often disease mutations. Differences between species and disease-related variations were then used to predict the potential for disease association for 464 missense variants from dbSNP.

In Chapter 2, we described the construction of the GPCR NaVa database, which catalogues natural variants that occur at or near human GPCR genes. This database is available from the website http://nava.liacs.nl. Incorporated variants range from *single nucleotide polymorphisms (SNPs)* to *copy number variants*, which are DNA segments of several kb that sometimes occur as common polymorphisms[8]. These variants can be common polymorphisms, rare disease-causing mutations and natural variants of unknown *allele frequency*. The availability and size of the GPCR NaVa database enabled us to perform computational analyses at an unprecedented scale.

The current study aims to derive new relationships between computable and biological properties from the human GPCR variants of our database, and to validate relationships proposed earlier. Such relationships may show characteristics for subclasses of GPCRs or reveal properties that affect the pathogenic potential of variants. A further understanding of inter-individual variation in GPCRs has potential implications for elucidating disease-causing mechanisms and genotype-phenotype relationships.

## MATERIALS AND METHODS

### Data sources

Variants were extracted from the GPCR NaVa database of July 2007[9], which incorporates variants from its underlying databases, such as dbSNP[10] and UniProt/SwissProt[1]. Variants stemming from other sources, such as scientific literature, have also been included in the NaVa database, and were also considered. Protein sequence information in the GPCR NaVa database was derived from UniProt/SwissProt. A large alignment was constructed by concatenating multiple sequence alignments of the subfamilies of class A GPCRs via the numbering schemes for aligned positions from the GPCRDB version 10.0[3].

**Data subsets**

For the study of small natural variants we considered only coding variants that resulted either in no amino acid change (silent variant) or in the substitution of one amino acid into another (missense variant). These criteria excluded uncommon variants, such as in-frame insertions/deletions, nonsense mutations, frame shift mutations, splice-site mutations and other more complex variants. One missense variant was considered per amino acid position of each receptor.

The sequence composition of the therapeutically important non-olfactory class A GPCRs is more similar to olfactory class A GPCRs than to the diverse set of other GPCRs, such as class B and C, see Table 4.1. We therefore chose to compare small variants in non-olfactory class A receptors only with those of olfactory receptors. In addition, to test whether results from class A GPCRs also hold for other GPCRs, similar data were provided where possible for GPCRs that are not from class A. The latter subset of GPCRs is called 'other GPCRs' and contains the diverse remaining GPCR families that, as shown in Table 4.1, are too small in number to allow for statistically significant conclusions on a family level.

For small variants we considered the following data for statistical analysis:

**Minor allele frequency**

The allele frequency was defined as the average fraction of all studied human chromosomes that contain the rarer allele of the variant[10]. A minor allele frequency of 1% was used as a threshold to distinguish between rare mutations (0-1%) and polymorphisms (1%-50%). Moreover, variants with an allele frequency of exactly 0% were deemed unconfirmed upon performing population studies for a candidate polymorphism. Allele frequency data from dbSNP were not available for all variants.

**Disease association**

A variant was considered disease-associated if an association with disease was reported in at least one patient or one association study, even if other studies reported no association. Variants were considered to have little or no effect on phenotype if, to date, no studies reported a disease association. We analysed only missense variants to discern between disease-associated and other variants because most reported disease associations concern missense changes in GPCR sequences. Disease association data were not available for all variants.

**BLOSUM62 score**

BLOSUM62-scores were available for missense variants in all GPCRs. Following Balasubramanian et al.[7], the BLOSUM62-score was used as a measure of how much the substituted amino acid differs from the reference amino acid, where a lower score indicates a larger change in physicochemical properties. Using a threshold of -1 for the BLOSUM62-

score, missense variants were divided into radical (-3 to -1) or mild (0 to 3) amino acid substitutions.

**Transmembrane membership**

Detailed transmembrane membership information was available for small missense and silent variants of class A GPCRs. We first concatenated the multiple sequence alignments of non-olfactory class A GPCRs from the GPCRDB[3]. Then we derived transmembrane membership information from this concatenated alignment to ensure that we used identical TM definitions for each receptor. For other GPCRs, less exact transmembrane membership definitions were taken from UniProt/SwissProt[1].

**Relative solvent accessibility**

Relative solvent accessibility data was only computed for small missense variants that occurred in TMDs of class A GPCRs. We mapped variants in TMDs on the crystal structure of the class A GPCR rhodopsin[11] by using the concatenated multiple sequence alignment from the GPCRDB. The relative solvent accessibility is the fraction of the amino acid surface that can be exposed to a hypothetical solvent[12], as computed from a monomer of the rhodopsin crystal structure that contained retinal. We used a threshold of 10% relative solvent accessibility to distinguish between surface positions (10-100% accessibility to either membrane or solvent) and positions in the receptor core (0-10% accessibility), i.e. those that are enclosed by other amino acids in the crystal structure.

**Ballesteros/Weinstein numbering**

To compare and identify positions of different GPCRs, Ballesteros and Weinstein[13] introduced an indexing method in which a TMD number, is followed by an index. In their method, the most conserved residue in each transmembrane helix is given the index 50 and other residues are numbered relative to this position. For example, the Ballesteros number 2.49 in the second transmembrane helix corresponds with an alanine at UniProt/SwissProt position 82 in rhodopsin, which is located prior to the aspartate (D2.50) that is highly conserved in class A GPCRs.

**Conservation**

Using a threshold of over 75% conservation in the alignment of non-olfactory class A GPCRs, all most conserved positions per TMD (N1.50, D2.50, R3.50, W4.50, P5.50, P6.50 and P7.50) were identified as well as five other highly conserved positions (L2.46, C3.25, Y5.58, F6.44 and Y7.53)[14]. Thus, 12 of the 191 possible TMD positions were considered highly conserved.

**Two-entropies analysis**

A two-entropies analysis can visualize (dis)similarities between the aligned positions in a multiple sequence alignment of a protein family and its subfamilies, as is the case for our concatenated alignment of non-olfactory class A GPCRs. The method was described in detail by Ye, et al.[15] and produces a plot in which each point represents one position from the

alignment. Their paper also used the ligand-based subfamilies of the GPCRDB[3]. For a two-entropies analysis, the conservation of each position was measured in terms of information entropy, where a low entropy indicates high conservation. The averaged entropy of the entire alignment (Y axis) is then plotted with respect to the entropy of the individual alignments of each subfamily (X axis). In short, points in the lower left corner of the plot indicate positions that are globally conserved in the large GPCR alignment and the subfamily-specific alignment(s) (for instance N1.50, D2.50, etc.), while points in the upper right corner indicate positions that are most variable. Points in the upper left corner indicate positions that are conserved within each subfamily-specific alignment, but different between these alignments (where the amino acid is Q in one subfamily, A in another, etc.). As each receptor subfamily recognizes a specific ligand, these positions are likely involved in ligand binding[15].

**Statistical methods**

We used a $\chi^2$ test to assess the partitioning of variants (such as those located in the TM core vs. those at the TM surface) of one dataset (like a disease-associated dataset) with the corresponding partitioning of another dataset (such as a dataset of less harmful variants). The splitting criterion divides each dataset into two subsets, using either the presence or absence of data ('allele frequency data is not available for each variant') or a numerical threshold for available data ('computed accessibility is higher than 10%'). A $\chi^2$ value was computed by comparing the sizes of two subsets of one dataset with those of another dataset. From this $\chi^2$ value, we computed the corresponding $p_\chi$ value, which is the random chance that two distributions are equivalent. A low $p_\chi$ value ($p_\chi < 0.01$) indicates a low chance (<1%) that the two distributions are comparable, which reflects a high chance that they are different. A low $p_\chi$ value also suggests that the splitting criterion (a theoretical measure) correlates with the biological difference between the datasets of variants (disease-related or not).

Following Lee et al.[6] and Balasubramanian et al.[7], we used a Poisson distribution to assess the partitioning of one dataset of variants (such as a disease-associated dataset) among the transmembrane domains (TMDs) and other non-TMD regions of GPCRs. For each coding variant that occurs in a GPCR, the sequence length determines the random chance that the variant occurs in a TMD or not. To estimate the expected number of variants in TMDs, the fractions of all GPCR sequences that consist of TMDs were therefore averaged and then multiplied by the number of missense variants. A Poisson distribution approximates a binomial distribution, and a $p_{poisson}$ value is then computed, which is the random chance the observed number of variants can be observed given this expected number of variants[7]. Here, a low $p_{poisson}$ value ($p_{poisson} < 0.01$) indicates a low chance (<1%) that the observed number of variants (or a more deviating number of variants) could actually be observed in the TMDs as based on the sizes of the TMDs and other regions. A low $p_{poisson}$ value thus implies that TMD membership is an important factor in the distribution of variants over the GPCR structure.

## RESULTS

Using the available data from the GPCR NaVa database, we extracted 3597 coding variants in human GPCRs and analysed the relationships between their computable and biological properties. This section describes the observed correlations.

### Coding variants in GPCRs

Table 4.2 shows the number of receptors as well as the number of missense and silent variants in the three GPCR subsets: olfactory class A GPCRs, non-olfactory class A GPCRs, and other GPCRs (those not part of class A). The following paragraphs detail differences between the distributions of variants in olfactory and other GPCRs.

**Table 4.2:** Number of GPCRs per GPCR subset and their total number of missense and silent variants.

| GPCR subset | GPCRs | Missense variants | Silent variants |
|---|---|---|---|
| Olfactory class A GPCRs | 371 | 783 | 365 |
| Non-olfactory class A GPCRs | 233 | 1043 | 471 |
| Other GPCRs | 159 | 618 | 317 |
| Total | 773 | 2444 | 1153 |

### Non-olfactory vs. olfactory GPCRs: polymorphic vs. rare missense variants

We examined the polymorphic character of variants in the therapeutically important non-olfactory subset of class A GPCRs. Table 4.3 reports for the three GPCR subsets the number of missense variants that occur with an allele frequency that is high (over 1%), low (0-1%), zero (0%) or undetermined. Variants with an allele frequency that was determined as high, low and zero are respectively named polymorphisms, mutations and unconfirmed variants.

Of the 1043 missense mutations in non-olfactory GPCRs, 321 have frequency data from dbSNP. Of the 233 non-olfactory GPCRs, 129 receptors contain missense variants with frequency data, while 49 receptors contain only missense variants that lack frequency data and 55 receptors contain no coding missense variants. 180 of these 321 (56%) are polymorphisms as they have an allele frequency of 1% or higher. This also means that a significant part, namely 44%, of the candidate coding SNPs of these GPCR genes is rare. In fact, our analysis of dbSNP frequency data showed that the minor allele of 79 of the missense candidate SNPs (25%) in non-olfactory class A GPCRs was not confirmed in subsequent population studies.

**Table 4.3:** Number of missense variants with allele frequency classifications per GPCR subset.

| GPCR subset | Polymorphic (>1%) | Rare, confirmed (0-1%) | Rare, unconfirmed (0%) | No frequency measured |
|---|---|---|---|---|
| Olfactory class A GPCRs | 445 | 35 | 57 | 246 |
| Non-olfactory class A GPCRs | 180 | 62 | 79 | 722 |
| Other GPCRs | 202 | 41 | 51 | 324 |

Next, we compared these frequency-related observations with those of variants in the olfactory GPCRs. Olfactory GPCRs contain 537 missense variants with frequency data, of which 445 (83%) have a minor allele frequency of 1% or higher. The fraction of missense mutations that is polymorphic is lower in non-olfactory than in olfactory GPCRs, 56% vs. 83% ($p_\chi << 0.01$, $\chi^2 = 71.7$).

**Non-olfactory vs. olfactory GPCRs: missense vs. silent variants**

We then investigated the allele frequency distribution of silent versus missense variants in the olfactory and non-olfactory subfamilies. Like Table 4.3, Table 4.4 contains the allele frequency classifications, but now for silent variants. In total, allele frequency data was available for 642 coding variants in non-olfactory class A GPCRs, that is, for 321 missense and 321 silent variants. As noted above, 56% of the 321 missense variants with frequency data were common. Of the 233 non-olfactory GPCRs, 143 receptors contain silent variants with frequency data, while 23 receptors contain only silent variants that lack frequency data and 67 receptors contain no silent variants. Because 244 of the 321 silent variants (76%) were common, missense variants are less polymorphic than silent variants in the non-olfactory GPCRs ($p_\chi << 0.01$, $\chi^2 = 27.5$).

**Table 4.4:** Number of silent variants with allele frequency classifications per GPCR subset.

| GPCR subset | Polymorphic (>1%) | Rare, but confirmed (0-1%) | Rare, unconfirmed (0%) | No frequency measured |
|---|---|---|---|---|
| Olfactory class A GPCRs | 174 | 15 | 20 | 156 |
| Non-olfactory class A GPCRs | 244 | 36 | 41 | 150 |
| Other GPCRs | 190 | 12 | 17 | 98 |

As already noted, 83% of missense variants in olfactory GPCRs with frequency data were polymorphisms. Similarly, of the 209 silent variants with frequency data, 174 were common (83%) and 35 were rare. In this subfamily, the fractions of polymorphisms are similar between missense and silent variants. However, where 25% of the missense candidate SNPs in non-olfactory class A GPCRs were not confirmed in subsequent population studies, this was the case for only 41 of the silent candidate SNPs (13%) in this receptor class.

In non-olfactory GPCRs, the coding polymorphisms consist of 180 missense (43%) and 244 silent variants. Coding polymorphisms in the olfactory GPCR repertoire consist of 445 missense (72%) and 174 silent variants. The average fraction of coding polymorphisms that are missense variants is 43% in non-olfactory GPCRs and this is considerably lower than the corresponding 72% in olfactory GPCRs ($p_\chi << 0.01$, $\chi^2 = 89.7$).

Table 4.5 catalogues for each GPCR subset the number of missense variants that are either disease-associated or that lack a clear phenotype. In total, 52 non-olfactory receptors (40 class A GPCRs and 12 other GPCRs) contained one or more disease-associated missense variants. The next paragraphs discuss several descriptive properties that were found to distinguish between missense variants that are disease-associated and those that appear less harmful.

**Table 4.5:** Number of missense variants per GPCR subset that are disease-associated or that lack a clear phenotype.

| GPCR subset | Disease-associated missense variants (radical variants) | Missense variants without phenotype (radical variants) |
|---|---|---|
| Olfactory class A GPCRs | 0 (0) | 783 (348) |
| Non-olfactory class A GPCRs | 366 (221) | 677 (296) |
| Other GPCRs | 107 (65) | 511 (221) |
| total | 473 (286) | 1971 (865) |

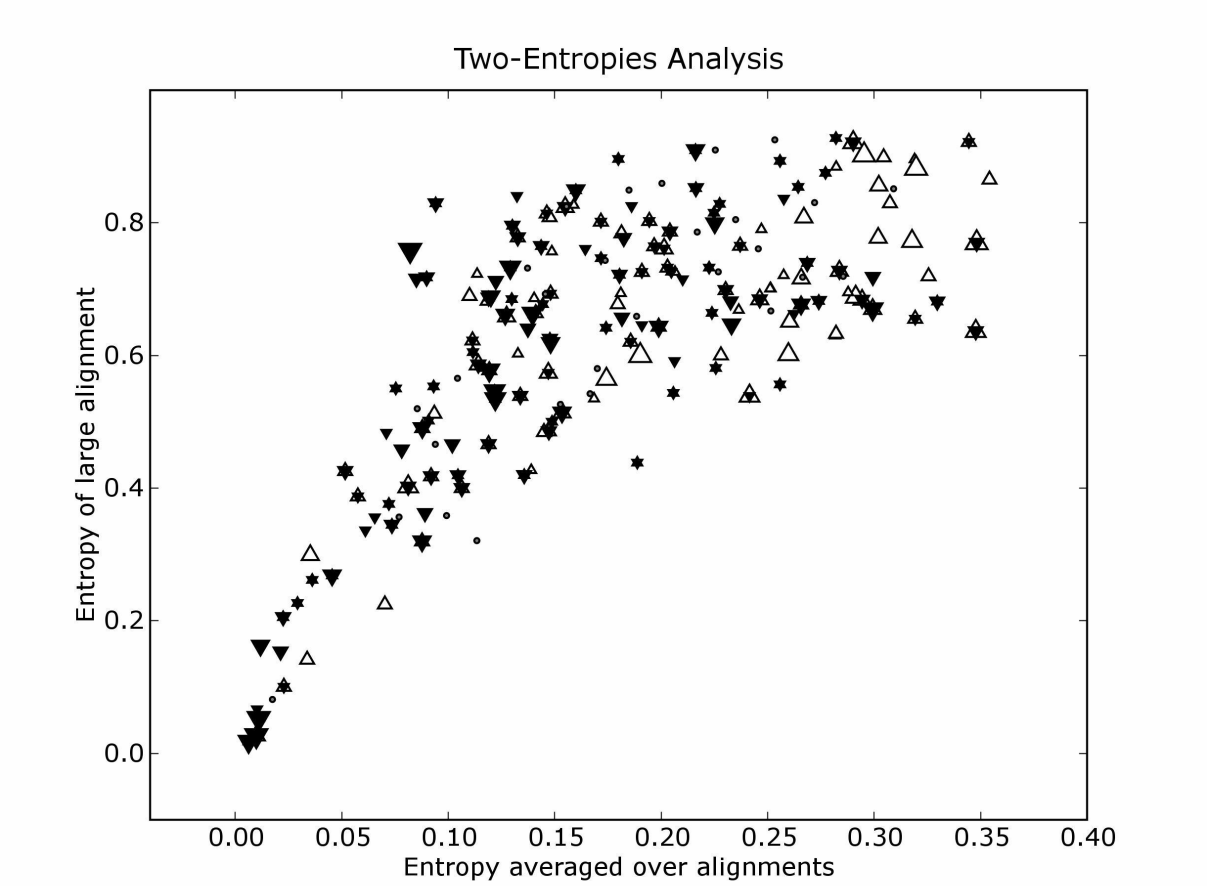


**Figure 4.1:** A two-entropies analysis of positions with disease-related (black triangles) and less harmful variants (white triangles). The size of each triangle reflects the number of disease-related or less harmful missense variants on that position. Positions that contain no variants in non-olfactory class A GPCRs are shown as black dots. For instance, the black dot in the lower left corner represents the conserved Y7.53 residue.

**Two-entropies analysis**

Figure 4.1 plots the two-entropies analysis and shows that 92 positions contain at least one disease-related and one less harmful variant. As a result, the distribution of disease-related variants largely overlaps with the distribution of less harmful variants. One position, Y7.53, in the lower left corner of the plot (which contains the most conserved residues in the GPCR family) stands out as a grey dot as it harbours no reported natural variants in non-olfactory class A GPCRs.

**Amino acid substitution and disease**

According to the physicochemical distortion caused by amino acid substitutions, the 677 less harmful missense variants of the non-olfactory class A GPCRs were categorised into 296 radical (44%) and 381 mild variants. In comparison to these harmless variants, quite many of the 366 disease-associated variants in this subfamily, namely 221 (60%), were radical ($p_\chi << 0.01$, $\chi^2 = 26.3$) while 145 were not.

Similarly, the 511 harmless variants of the other GPCRs that were not part of class A were divided into 221 radical (43%) and 290 mild variants. Again, relatively many of the 107 disease-associated variants in this subfamily, namely 65 (61%), were radical ($p_{\chi}<<0.01$, $\chi^2=8.2$) while 42 were mild.

**Transmembrane membership and disease**

The non-olfactory class A GPCRs contain both many disease variants and many harmless variants and so allow comparison between them. Table 4.6 depicts the observed and the expected number of variants in TMDs or other regions as based on sequence length in the non-olfactory class A GPCRs. Based on the sequence lengths of the transmembrane domains (TMDs) and other regions in these GPCRs, 222 of the 471 silent variants (47%) without obvious phenotype were expected in TM regions and 215 (46%) occurred here, which did not deviate from expectation ($p_{poisson}=0.3$). The distribution of harmless silent and missense variants in olfactory GPCRs also did not deviate from expectation ($p_{poisson}=0.4$, $p_{poisson}=0.4$, data not shown). However, of the 677 missense variants without obvious phenotype in non-olfactory class A GPCRs, 312 (46%) were expected in TM regions, while fewer occurred, namely 265 (39%) ($p_{poisson}=0.01$). Though 156 of the 366 missense disease variants (43%) were expected in TM regions, significantly more disease variants, namely 234 (64%), were observed ($p_{poisson}<<0.01$).

**Table 4.6:** Observed number of variants in TMDs or other regions versus the expected number as based on sequence length in non-olfactory class A GPCRs.

| GPCR subset | Observed variants in TMDs (expected) | Observed variants in regions other than TMDs (expected) |
|---|---|---|
| Silent variants that lack phenotype | 215 (222) | 256 (249) |
| Missense variants that lack phenotype | 265 (312) | 412 (365) |
| Missense variants that are disease-associated | 234 (156) | 132 (210) |

**Table 4.7:** Observed number of variants in TMDs or other regions versus the expected number as based on sequence length in other GPCRs.

| GPCR subset | Observed variants in TMDs (expected) | Observed variants in regions other than TMDs (expected) |
|---|---|---|
| Silent variants that lack phenotype | 63 (53) | 253 (263) |
| Missense variants that lack phenotype | 32 (26) | 75 (81) |
| Missense variants that are disease-associated | 85 (85) | 426 (426) |

We then compared the distributions of disease variants in TMDs and other regions directly with those distributions of missense variants with unknown phenotypes, that is, independently of sequence length. Of the 366 disease variants, 234 were located in TMDs (64%) and 132 were not. Of the 677 variants with unknown phenotypes, 265 occurred in TMDs (39%) and 412 occurred in other regions than TMDs. Disease variants, therefore, occur more often in TMDs than unknown variants ($p_\chi << 0.01$, $\chi^2=58.5$).

Afterwards, we also performed these analyses for the group of other GPCRs and the corresponding numbers are shown in Table 4.7. Though 26 of the 107 missense disease variants (24%) were expected in TM regions, significantly more disease variants, namely 32 (30%), were observed ($p_{poisson}=0.01$). In an analysis that was independent of sequence length, 32 disease variants were located in TMDs (24%) and 75 were not. Of the 511 variants with unknown phenotypes, 85 occurred in TMDs (17%) and 426 occurred in other regions. Disease variants, therefore, also occur more often in TMDs of other GPCRs than unknown variants ($p_\chi << 0.01$, $\chi^2=10.2$).

**Solvent accessibility and disease**

Solvent accessibility data were computed for the 511 variants that occurred in the TMDs of non-olfactory class A GPCRs (see Methods). Of the 234 disease-associated missense variants that occur in the TMDs, 150 were located in the core (64%) of the crystal structure and 84 at exterior positions. Of the 265 less harmful missense variants in TMDs, 112 occurred at such buried positions (43%) and 153 at surface positions. In comparison, disease variants therefore occur more often on the inside of TMDs than on their membrane-facing surface ($p_\chi << 0.01$, $\chi^2=23.8$). Moreover, Figure 4.2 shows a similar observation when the distribution of disease-associated variants (64% are buried in the core) is compared to that of common variants in TMDs, where 21 out of the 59 polymorphisms (36%) are buried ($p_\chi << 0.01$, $\chi^2=15.7$).

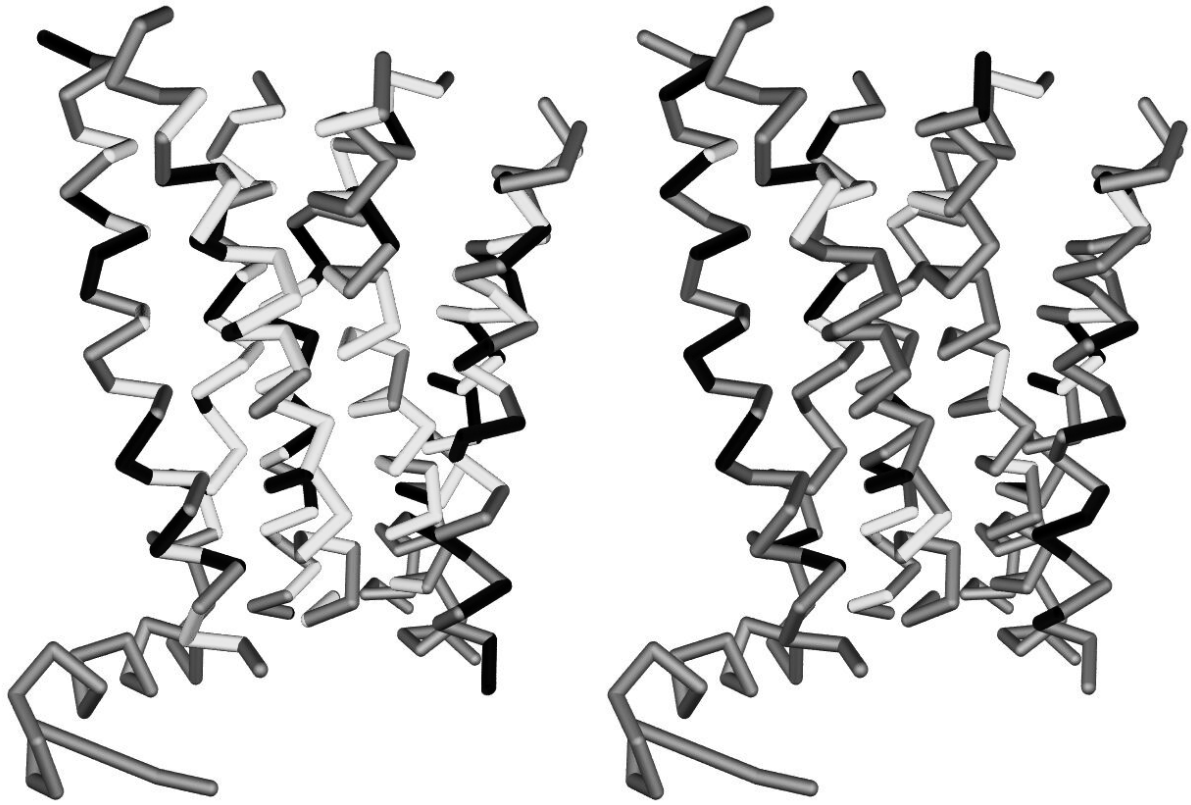

**Figure 4.2:** Disease-related variants (left) and disease-unrelated polymorphisms (right) of the GPCR NaVa database as overlaid on the crystal structure of bovine rhodopsin. From left to right, each crystal structure shows transmembrane domains 1, 2, 3 and 4 in the front while transmembrane domains 5, 6 and 7 are shown in the background. Positions without variants are coloured grey. Variants are coloured white if they occur at transmembrane positions that possess a low accessibility, while variants are coloured in black if they occur on the GPCR-membrane surface and thus possess high accessibility. Transmembrane domains, overall, contain more variants that are linked to disease than variants with known high allele frequency (1% or more). More importantly, the fraction of disease variants that occur at buried positions, which is visible as the portion of white vs. black variants, is visibly higher than the fraction of common variants that occur at buried positions (see Results for more details).

**Disease-prone positions.**

For a more detailed view of critical locations in non-olfactory class A GPCRs, we examined their disease mutations via their conservation and Ballesteros/Weinstein numbering. Of all 191 possible TMD positions from our alignment, 11 positions contained 4 or 5 disease mutations from different non-olfactory class A GPCRs and none contained more. Of these 11 positions, 8 were not conserved (1.41, 2.54, 2.57, 3.40, 5.57, 6.30, 6.34 and 6.41). On the other hand, 3 positions were highly conserved (3.50, 6.50 and 7.50) throughout GPCRs and this amount was considerably more than expected ($p_\chi << 0.01$, $\chi^2 = 8.7$).

## DISCUSSION

We have studied and described the distribution of variants in various subsets of GPCRs. Of the missense candidate SNPs with allele frequency data in non-olfactory class A GPCRs that were considered in population studies, a relatively large fraction of 25% could not be confirmed in any chromosome and another 19% were found to be rare. These findings relate to the earlier observation that only 11 of the 34 investigated missense GPCR variants (32%) reported in dbSNP were confirmed upon retesting in the study by Small, et al.[16]. Moreover, Fredman, et al.[17] analysed samples of 32 individuals and elegantly showed that 40% of the missense variants catalogued in dbSNP could not be confirmed in their validation study. The above observations agree with the explanation from the developers of dbSNP that this database contains many variants of unknown allele frequency rather than polymorphisms alone[10], thus including rare variants that only occur in certain populations. Another possible explanation for the finding that many variants were not confirmed upon frequency determination is that these variants resulted from errors in sequencing or mapping. Interestingly, Fredman, et al.[17] also showed that only 10% of the noncoding variants that were catalogued in dbSNP could not be confirmed, which suggested that missense variants are relatively often rare, or incorrect, with respect to other DNA variants. This agrees with our observation that, with respect to silent candidate SNPs, missense candidate SNPs are more often 'false positive' candidate polymorphisms as they were more often unconfirmed upon follow-up population studies.

The variants with frequency data from dbSNP are distributed over more than half of the non-olfactory class A GPCRs. Table 4.3 suggests that relatively many variants lack frequency data in the subset of non-olfactory GPCRs. Hundreds of these variants cause monogenetic diseases and were therefore probably too rare to be detected during the initial construction of dbSNP and, as a result, they were not considered in subsequent population studies. For consistency purposes, we chose to handle allele frequency values from dbSNP and without assuming low allele frequency values for these variants in our analyses.

Most observed differences between olfactory and non-olfactory GPCRs may be explained by their difference in therapeutic relevance. First, the fraction of the missense variants with allele frequency data that were proven to be polymorphisms was significantly lower in non-olfactory class A GPCRs than in olfactory GPCRs. Second, in the non-olfactory GPCRs, but not in olfactory GPCRs, missense variants are less polymorphic than silent variants. Third, of all coding polymorphisms per subfamily, few cause missense changes in non-olfactory GPCRs. In fact, the estimate that roughly 67% of the random coding variants cause missense variants during neutral evolution[18, 19] appears to hold for olfactory GPCRs, but not for non-olfactory class A GPCRs, where 72% and 43%, respectively, of the observed polymorphisms are missense variants. Taken together, these findings confirm a strong

selection against missense variants, and especially missense polymorphisms, in the non-olfactory class A subfamily. Nature likely limits variability in non-olfactory GPCRs, while allowing, or even promoting diversity in olfactory GPCRs. The latter accompanies variability in the detection of smell, which may be subject to a mild positive selection[20].

In the analysis of disease variants in GPCRs, it is important to note that most disease mutations occur in only a few different GPCRs. The GPCRs with most disease-related mutations are the vasopressin $V_2$ receptor, the calcium-sensing CaS receptor, the probable G protein-coupled receptor 143, rhodopsin, the luteinizing hormone (LH) receptor, the thyroid-stimulating hormone (TSH) receptor and the melanocortin $MC_2$ and $MC_4$ receptors[9]. By excluding the variants of these GPCRs, we were left with less than one third of all disease mutations. Despite such data loss, all the trends that were reported in this analysis were reproduced (data not shown), but without statistical significance. It is therefore important to realise the impact that the variants in these receptors have on our results.

The first observation from the two-entropies analysis of Figure 4.1 was that many positions contain disease-associated as well as less harmful variants. The two-entropies analysis was not able to distinguish between variants that are disease-related and those that are not. However, the plot suggests that disease variants occur more in the bottom half of the plot, at globally conserved positions, while the set of less harmful variants at the upper half of the plot is mildly shifted to the right, to the most variable positions. In the lower left corner of the plot that contains the overall strongly conserved residues, position Y7.53 does not have any natural variants. This residue is part of the NPXXY motif in TMD 7, which is critical for activation of G proteins[21].

The comparison of species differences and disease variants by Balasubramanian et al.[7] suggested that disease mutations in class A GPCRs are relatively radical in terms of amino acid substitution. Not only do our results corroborate their finding, they also suggest a similar observation for GPCRs that are not part of the class A family. Radical GPCR substitutions in any GPCR subclass therefore appear more likely risk factors for disease.

Earlier analyses used the sequence lengths of GPCR domains to determine how many mutations are expected in TMDs with respect to other regions[6, 7]. Assuming a Poisson distribution, both analyses showed that disease variants occur significantly more in TMDs than expected, and we also observed this tendency in our dataset. Both studies observed no such deviation in the distribution of other variants, while we, in our substantially larger dataset, detected a significant trend of harmless missense variants to occur more frequently in regions other than TMDs. In addition, our results suggest that disease variants also occur more often in TMDs of GPCRs that are not part of the class A family. Furthermore, we observed a tendency for disease missense variants to occur more often inside TMDs with respect to less harmful missense variants when their counts were compared directly, rather than using sequence length-based expectations. These findings stress the relative functional

importance of these GPCR domains over other regions and support the common view that the TMDs control the equilibrium between active and inactive receptor states.

Previous studies investigated the three-dimensional distribution of variants in protein collections that lacked GPCRs, such as sets of globular proteins[22-24]. In some[22, 23] but not all studies[24] a tendency was observed of disease mutations to occur more often in the solvent-inaccessible interior of these proteins. As three-dimensional information was only available for TMDs of non-olfactory class A GPCRs, we divided positions in these domains into solvent-inaccessible positions in the interior (core) and membrane-accessible positions in the exterior (surface). The statistical findings from Figure 4.2 suggest that missense variants in the core of the crystal structure are more likely to be predisposition factors for disease than surface variants, polymorphisms or not.

One difference between the earlier studies that analysed candidate polymorphisms and this study is that is we did not only consider candidate polymorphisms as such, but also the variants and their actual allele frequencies. As a result, our hypotheses were tested with actual, confirmed polymorphisms and/or with rare mutations. We moreover analysed differences between the spread of mutations and polymorphisms in olfactory and non-olfactory GPCRs. With our large dataset we indeed confirmed that disease predisposition of missense variants of class A non-olfactory GPCRs relates to a) the distortion of the variant at the amino acid level and/or b) the occurrence of the variant in a TMD. In addition, we showed for GPCRs that are not of class A that the distortion at the amino acid level is a risk factor, whereas TMD membership is not. Another new observation was that missense variants in TMDs of non-olfactory GPCRs are stronger risk factors if they occur in the GPCR core that is formed by the TMDs. If a new GPCR variant is discovered, these conclusions suggest hypotheses on its ability to distort GPCR function.

We also analysed several GPCR positions in more detail. The three highly conserved positions that contain four or more disease mutations, R3.50, P6.50 and P7.50 also contain at least one variant that has not been linked to disease. Two R3.50H variants and one P6.50L variant are described in dbSNP, but not in the literature: while one was not searched for in a population study, two were and both were unconfirmed as they had a minor allele frequency of 0%. So far, only one report[25] describes natural variants at these conserved sites and does not associate them with a severe phenotype. The populations they studied included two apparently healthy individuals, one with a R3.50H variant and one with a P3.50S variant in the melanin-concentrating hormone $MCH_1$ receptor. In general, the R3.50, P6.50 and P7.50 are part of conserved motifs that are deemed critical for GPCR activation[21]. More specifically, side-chain interactions of R3.50 are assumed to be critical for activating G proteins, P6.50 introduces a kink in TMD 6 and so orients its aromatic cluster and P7.50 is important for maintaining the kinked relative orientation between the conserved N7.49 and Y7.53 residues. Our findings support the impact of these positions in various GPCRs. The eight variable

positions that do contain many disease mutations may prove to be important for receptor-specific side-chain interactions that are not conserved in GPCRs.

Perhaps the most striking observation comes from the left-hand receptor structure of Figure 4.2, which is unique to our study as it shows the spread of disease-causing variants in TMDs of non-olfactory class A GPCRs. The first observation from this illustration is that disease-causing variants occur almost everywhere in the TMD structure; they are not restricted to only key sites such as those that are part of conserved GPCR motifs. As the vast majority of disease-causing mutations inactivate GPCR function by hindering cell surface expression[26, 27], this dispersion indicates that distortions anywhere in the TMD structure have the potential to cause such an effect. In contrast, many TMD variants without known effects also exist in non-olfactory class A GPCRs, see for instance Table 4.6. This observation implies that TMD membership alone is not a sufficient criterion for distortion of GPCR function.

## CONCLUSION

In conclusion, we have studied human GPCRs through a very large set of 3597 small natural variants from the GPCR NaVa database. By considering variants with established allele frequency data, we found that non-olfactory class A GPCRs contain fewer missense variants, and much fewer missense polymorphisms, than olfactory GPCRs, which point to an evolutionary preservation of non-olfactory GPCRs. Variants are more often disease-associated if they cause more radical amino acid substitutions, and this also holds for variants that occur in GPCRs that are not of class A. Missense variants are also stronger pathogenic risk factors if they alter TMDs, especially if they occur at positions that are inaccessible to solvent or membrane according to the three-dimensional structure of these TMDs. Moreover, several key positions in TMDs of non-olfactory class A GPCRs correspond to positions that hold many disease mutations, but disease mutations also occur everywhere else in the TMDs. We thus highlighted insightful differences between disease and unknown variants.

## REFERENCES


1. the UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **2007,** 35, D193-7.
2. Hopkins, A. L.; Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **2002,** 1, 727-730.
3. Horn, F.; Bettler, E.; Oliveira, L.; Campagne, F.; Cohen, F. E.; Vriend, G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* **2003,** 31, 294-7.
4. Small, K. M.; Tanguay, D. A.; Nandabalan, K.; Zhan, P.; Stephens, J. C.; Liggett, S. B. Gene and protein domain-specific patterns of genetic variability within the G-protein coupled receptor superfamily. *Am J Pharmacogenomics* **2003,** 3, 65-71.
5. Wahlestedt, C.; Brookes, A. J.; Mottagui-Tabar, S. Lower rate of genomic variation identified in the trans-membrane domain of monoamine sub-class of Human G-Protein Coupled Receptors: the Human GPCR-DB Database. *BMC Genomics* **2004,** 5, 91.
6. Lee, A.; Rana, B. K.; Schiffer, H. H.; Schork, N. J.; Brann, M. R.; Insel, P. A., et al. Distribution analysis of nonsynonymous polymorphisms within the G-protein-coupled receptor gene family. *Genomics* **2003,** 81, 245-8.
7. Balasubramanian, S.; Xia, Y.; Freinkman, E.; Gerstein, M. Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. *Nucleic Acids Res* **2005,** 33, 1710-21.
8. Redon, R.; Ishikawa, S.; Fitch, K. R.; Feuk, L.; Perry, G. H.; Andrews, T. D., et al. Global variation in copy number in the human genome. *Nature* **2006,** 444, 444-54.
9. Kazius, J.; Wurdinger, K.; van Iterson, M.; Kok, J.; Bäck, T.; IJzerman, A. P. GPCR NaVa Database: Natural Variants in Human G Protein-Coupled Receptors. *Human Mutation* **2007**, (in press).
10. Sherry, S. T.; Ward, M.; Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **1999,** 9, 677-9.
11. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A., et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000,** 289, 739-45.
12. Rost, B.; Sander, C. Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* **1996,** 25, 113-36.
13. Ballesteros, J. A., Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neurosciences* **1995,** 25, 366-428.
14. Mirzadegan, T.; Benko, G.; Filipek, S.; Palczewski, K. Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin. *Biochemistry* **2003,** 42, 2759-67.
15. Ye, K.; Lameijer, E. W.; Beukers, M. W.; Ijzerman, A. P. A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins* **2006,** 63, 1018-30.
16. Small, K. M.; Seman, C. A.; Castator, A.; Brown, K. M.; Liggett, S. B. False positive non-synonymous polymorphisms of G-protein coupled receptor genes. *FEBS Lett* **2002,** 516, 253-6.
17. Fredman, D.; Sawyer, S. L.; Stromqvist, L.; Mottagui-Tabar, S.; Kidd, K. K.; Wahlestedt, C., et al. Nonsynonymous SNPs: validation characteristics, derived allele frequency patterns, and suggestive evidence for natural selection. *Hum Mutat* **2006,** 27, 173-86.
18. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G., et al. The sequence of the human genome. *Science* **2001,** 291, 1304-51.

19. Cargill, M.; Altshuler, D.; Ireland, J.; Sklar, P.; Ardlie, K.; Patil, N., et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **1999,** 22, 231-8.
20. Gilad, Y.; Segre, D.; Skorecki, K.; Nachman, M. W.; Lancet, D.; Sharon, D. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat Genet* **2000,** 26, 221-4.
21. Jongejan, A.; Bruysters, M.; Ballesteros, J. A.; Haaksma, E.; Bakker, R. A.; Pardo, L., et al. Linking agonist binding to histamine H1 receptor activation. *Nat Chem Biol* **2005,** 1, 98-103.
22. Ramensky, V.; Bork, P.; Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **2002,** 30, 3894-900.
23. Vitkup, D.; Sander, C.; Church, G. M. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* **2003,** 4, R72.
24. Stitziel, N. O.; Tseng, Y. Y.; Pervouchine, D.; Goddeau, D.; Kasif, S.; Liang, J. Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* **2003,** 327, 1021-30.
25. Wermter, A. K.; Reichwald, K.; Buch, T.; Geller, F.; Platzer, C.; Huse, K., et al. Mutation analysis of the MCHR1 gene in human obesity. *Eur J Endocrinol* **2005,** 152, 851-62.
26. Schöneberg, 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.
27. Tao, Y. X. Inactivating mutations of G protein-coupled receptors and diseases: structure-function insights and therapeutic implications. *Pharmacol Ther* **2006,** 111, 949-73.