STRUCTURAL EVOLUTION OF G-PROTEIN-COUPLED RECEPTORS: A SEQUENCE SPACE APPROACH

Jean-Michel Bécu[1], Julien Pelé[1], Patrice Rodien[1,2], Hervé Abdi[3] and Marie Chabbert[1]*

[1] UMR CNRS 6214 – INSERM U1083, Faculté de Médecine, 3 rue Haute de Reculée, F-49045 Angers, France

[2] Centre de référence des pathologies de la réceptivité hormonale, Service d'endocrinologie, CHU d'Angers, 4 rue Larrey, 49933 Angers, France

[3] The University of Texas at Dallas, School of Behavioral and Brain Sciences.
800 West Campbell Road, Richardson, TX 75080-3021, USA

* To whom correspondence should be addressed. Tel: 33241735873; Email: marie.chabbert@univ-angers.fr

Running title: Sequence space and evolution of GPCRs

Key words: GPCR, evolution, sequence space, multidimensional scaling

ABSTRACT

Class A G-protein-coupled receptors (GPCRs) provide a fascinating example of evolutionary success. In this review, we discuss how metric multidimensional scaling (MDS), a multivariate analysis method, complements traditional tree-based phylogenetic methods and help decipher the mechanisms that drove the evolution of class A GPCRs. MDS provides low dimensional representations of a distance matrix. Applied to a multiple sequence alignment, MDS represents the sequences in a Euclidean space as points whose inter-distances are as close as possible to the distances in the alignment (the so-called sequence space). We detail how to perform the MDS analysis of a multiple sequence alignment and how to analyze and interpret the resulting sequence space. We also show that the projection of supplementary data (a property of the MDS method) can be used to straightforwardly monitor the evolutionary drift of specific sub-families. The sequence space of class A GPCRs reveals the key role of mutations at the level of the TM2 and TM5 proline residues in the evolution of class A GPCRs.

ABBREVIATIONS: GPCR: G-protein coupled receptor; TM: transmembrane helix; MDS: multidimensional scaling; NJ: neighbour-joining; UPGMA: unweighted pair group method with arithmetic mean.

1. INTRODUCTION

G-protein-coupled receptors (GPCRs) are widespread in the animal kingdom and, with around 800 members, form the largest transmembrane receptor family in humans (Bockaert and Pin, 1999). These receptors are involved in most physiological functions and constitute a very important target for the pharmaceutical industry (Overington *et al.*, 2006). By specific binding to endogenous or exogenous ligands, GPCRs undergo a conformational change that activates heterotrimeric G proteins, initiating an intracellular signaling cascade.

GPCRs share a common fold of seven transmembrane helices (TM). This fold seems to have emerged several times independently during evolution and is shared by several families, in eukaryotes or prokaryotes, with no evidence of homology and different coupling and/or functions. In prokaryotes, bacteriorhodopsin, with a chromophore bound to a 7TM fold, is a light sensor which acts as a proton pump (Luecke *et al.*, 1998). In eukaryotes, 7TM proteins may be ligand-gated ion channels, such as the odorant receptors in insects (Sato *et al.*, 2008).

"True" GPCRs, coupled to G proteins, are present in all the animal organisms. In vertebrates, GPCRs are classified—from the name of a typical representative—into five families: *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled,* and *Secretin* (Fredriksson *et al.*, 2003). With the exception of the *Glutamate* family, these receptors share a common ancestor with the cAMP receptors from the social amoeba *D. discoideum* (Nordstrom *et al.*, 2011).

Among the five vertebrate families, the *Rhodopsin* or class A family has undergone the largest expansion and is a landmark example of evolutionary success. In humans, this family includes about 700 GPCRs (400 of which are olfactory receptors). These receptors are characterized by the very high conservation of a few signature residues in each helix, but the overall conservation may be lower than 15%. The about 300 non-olfactory class A receptors found in the human genome respond to a wide variety of ligands as diverse as peptides, proteins, amines, sugars, lipids, nucleotides, and photons (Gether, 2000). This indicates a very robust fold, capable of undergoing extensive mutagenesis, while maintaining its capacity to act as a signal transducer.

The resolution of several crystal structures of G-protein-coupled receptors has marked a breakthrough in the understanding of the structure and function of these receptors. These structures have revealed the conformational details of each receptor that allow recognition and binding of specific ligands, and the mechanism of receptor activation (Deupi and Standfuss, 2011; Katritch *et al.*, 2012).

Understanding the mechanisms that drove the structural evolution of GPCRs could help improve structural prediction and molecular modeling. In this review, we will discuss how

metric (a.k.a classical) multidimensional scaling (MDS)—a multivariate analysis method—complements traditional tree-based phylogenetic methods and provides insights into the main mechanisms that drove the evolution of class A GPCRs.

2. THE PUZZLE OF GPCR EVOLUTION

The classification of the human class A GPCRs into a dozen of sub-families (Table I) is well established, with only minor differences between various studies (Fredriksson *et al.*, 2003; Surgand *et al.*, 2006; Devillé *et al.*, 2009; Pelé *et al.*, 2011a). These differences can usually be explained by the sequence range and/or the method used. The evolutionary relationship between these sub-families, however, is difficult to establish unambiguously because of the large number of sequences and their low identity rates.

Phylogenetic relationships between sequences are usually inferred by tree-based methods that rely on a binary, hierarchical classification of the sequences according to either a distance matrix, such as the Neighbour Joining method (NJ) or the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), or an evolutionary model, such as Maximum Parsimony or Maximum Likelihood. Based on a Maximum Parsimony study (Fredriksson *et al*., 2003), class A GPCRs have been classified into 4 groups. However, several studies based on NJ or UPGMA methods lead to fan-shaped trees for human (Surgand *et al.*, 2006; Devillé *et al.*, 2009), dog (Haitina *et al.*, 2009) and mouse GPCRs (Bjarnadottir *et al.*, 2006). These fan-shaped trees do not display evidence of receptor clustering into four groups. Two NJ trees of human GPCRs, computed with the MEGA4 program (Tamura *et al.*, 2007) from distance matrices based either on the difference score or on the JTT amino acid substitution matrix (Jones *et al.*, 1992) are shown in Fig. 1. These trees clearly lack phylogenetic resolution and give an ambiguous positioning of several sub-families or receptors (e.g. the Mas-related (MRG) and the galanin/kisspeptin receptors).

Several approaches can be used to help decipher the phylogenetic relationships between the GPCR sub-families:

1. Comparison of the GPCR repertoires in different species. This approach has been introduced by Schioth and coworkers (Fredriksson and Schioth, 2005) who showed that “ancient” sub-families are present in different lineages, whereas more “recent” sub-families are lineage specific.
2. Analysis of rare mutational events, such as indels (insertion/deletion) in the TM domain helices. We developed this approach to decipher the origin of the unusual proline pattern of

TM2 and we showed that three sub-families (SO, CHEM and PUR) are evolutionary related and evolved from a deletion in TM2 in PEP receptors (Devillé et al., 2009).

3. Map-based methods. Map-based methods visualize elements in a low dimensional space, according to a distance matrix. Such a method is MDS which transforms a distance matrix between elements into points whose relative distances are as close as possible to the distances in the original matrix (Young and Householder, 1938; Torgerson, 1958; Gower, 1966; Abdi, 2007b). MDS analysis is thus related to distance-based tree methods, but the information given by these methods is different. Tree-based methods perform well for the detailed relationships between closely related sequences, whereas space-based methods may reveal relationships between deep branches corresponding to ancient evolutionary trends (Higgins, 1992).

3. THE EVOLUTION OF CLASS A GPCRs VIEWED BY MULTIDIMENSIONAL SCALING

In this section, we describe the procedure to analyze the GPCR repertoire of a complete proteome by multidimensional scaling.

3.1 Preparation of the sequence set

We first describe the procedure to obtain the non-redundant sequence set of GPCRs from a species whose complete proteome set is available at the UniprotKB database (http://www.uniprot.org).

1. Retrieve the sequences from UniprotKB using a family profile. For class A GPCRs, the PS50262 PROSITE, IPR000276 InterPro and PF00001 Pfam profiles are equivalent. Use the IPR000725 profile to exclude olfactory receptors.
2. Cluster the sequences to avoid redundancy with the perl script *nrdb.pl* (Holm and Sander, 1998). The cut-off value for clustering can be adjusted from 80 to 100%.
3. Choose a representative sequence in each cluster. Do not rely on the automatic choice provided by the perl script but carefully analyze each cluster in order to limit the number of truncated sequences in the final set.
4. Align the sequences with a multiple sequence alignment program such as ClustalW (Thompson *et al.*, 1994) that can handle several hundreds of sequences.
5. Verify and—if needed—correct the alignment with alignment editing programs such as Genedoc (Nicholas *et al.*, 1997) or Jalview (Waterhouse *et al.*, 2009). In class A GPCRs, the high conservation of one residue in each helix *n* facilitates the verification step. This residue is given the number *n*.50 in the Ballesteros' numbering scheme (Sealfon *et al.*,

1995) and serves as a relative reference (N1.50, D2.50, R3.50, W4.50, P5.50, P6.50, P7.50). At this stage, it may be necessary to remove truncated or suspicious sequences. This final manual step is required to obtain a high quality sequence set.

3.2 MDS analysis of the sequence set

The non-redundant set of aligned sequences is now ready to be analyzed by MDS. The next step is the computation of a distance matrix based either on difference scores or on dissimilarity scores obtained with an amino acid substitution matrix, such as JTT. Then, this distance matrix can be analyzed by MDS. MDS analysis corresponds to the principal component analysis (PCA, see Abdi and Williams, 2010) of the cross-product matrix derived from the squared distance matrix (Abdi, 2007b). For a matrix of distances between sequences, MDS provides factor scores for the sequences to evaluate, and these factors scores can be used to create maps that give the best approximation of the original matrix. As in standard PCA, the variance of the factor scores for a given dimension (which gives the variance explained by this dimension) is called the *eigenvalue* associated with this dimension.

Several programs can be used to perform MDS analysis (Table 2). For an exploratory analysis, MDS can be easily done with the *principal components* functions from Jalview (Waterhouse et al., 2009) or MODELLER (Sali and Blundell, 1993) that perform MDS analysis from a multiple sequence alignment. For more detailed analysis, the use of specialized programs written in the R statistical language is recommended. MDS analysis of distance matrices can be performed with several R tools, included either in the R program or in specialized R packages (Table 2). The *bios2mds* package that we have developed (Pelé *et al.*, 2011b) is available at the Comprehensive R Archive Network under the GNU public licence (http://cran.r-project.org/web/packages/bios2mds/ index.html**).** It has been especially designed for the analysis of protein families by MDS from multiple sequence alignments. An example of the MDS analysis of human non-olfactory class A GPCRs (thereafter GPCRs) with *bios2mds* is shown in Fig. 2. In this example, the distance matrix is based on the differences between sequences.

The *bios2mds* package provides a wide choice of options to build distance matrices that are Euclidean or close to Euclidean matrices. Strictly speaking, MDS analysis applies to Euclidean distance matrices (Abdi, 2007b; Abdi, 2007a). Such a matrix can be obtained with distances based of the square root of the difference score (Gower, 1971). However, compared to the sequence space obtained with the difference scores, the sequence space obtained with the square roots is compressed, which results in a lower resolution. On the other hand, amino acid

substitution matrices lead to sequence spaces that are similar to those obtained with difference scores (e. g., the JTT matrix shown in Fig. 3). For these non-Euclidean distance matrices, the MDS analysis will provide negative eigenvalues (i.e. some dimensions are imaginary and therefore have a negative variance) that will correspond to 3-7% of the variance, which does not significantly affect the first components of the MDS analysis.

Whatever the distance matrix used, the relative positioning of the different GPCR sub-families or receptors in the resulting sequence space is similar (compare Fig. 2 and 3). In particular, this is the case for the MRG and the galanin/kisspeptin receptors whose positioning in NJ trees obtained in the same conditions is ambiguous (Fig. 1). This observation points towards the robustness of the MDS method for visualizing distance relationships. The sequence space can thus be confidently computed from the difference scores which are faster to calculate than dissimilarity scores.

3.3 Analysis of the sequence space

The MDS analysis of the aligned sequence set of human GPCRs (Fig. 2) provides a 3D representation of the GPCR sequence space. What does it tell us about GPCR evolution? Using this 3D representation, GPCRs can be clustered into four groups (Table 1), either by visual inspection or by *K*-means clustering (Pelé et al., 2011a). The first component differentiates the groups G1 (SO, CHEM, PUR) and G2 (AMIN, AD), the second component differentiates the group G3 (LGR, PTG, MEC, MRG), whereas the third component differentiates the group G0 (PEP, OPN, MTN). The name of this group, dominated by the PEP receptors, arises from its central location in the plane formed by the first two components of the MDS analysis.

The sequence space strongly supports a model of radiative evolution of class A GPCRs from a node formed by the PEP receptors through three main evolutionary trends, corresponding to the groups G1 to G3 (Pelé *et al.*, 2011a). It is important to note that the clusters are based on a distance matrix and do not depend upon a phylogenetic relationship between their different sub-families.

4. EVOLUTIONARY TRENDS

4.1 Search for hallmark residues

To interpret the groups observed by MDS in terms of evolutionary pathways related to specific sequence determinants, it was necessary to search for specific positions (if any) in each group. We considered two criteria: the first one was related to the correlation between a position and a group, whereas the second one was related to the conservation of this position in

this group or its complement (Pelé *et al.*, 2011a). The first criterion was based on the χ^2 test proposed to measure correlated mutations (Kass and Horovitz, 2002). The second criterion was based on the difference in sequence entropy that measures residue conservation (Mirny and Shakhnovich, 2001), to distinguish highly conserved or highly variable positions.

These tests reveal that the *presence* of a proline residue at position 2.58 is the hallmark of G1 receptors, whereas the *absence* of a proline residue at position P5.50 is the hallmark of G3 receptors (Pelé et al., 2011a). Fig. 4 visualizes the TM2 and TM5 proline pattern in the GPCR sequence space. It also shows the correlation between the absence of proline in TM2 and TM5 (p-values $< 10^{-10}$). It is worth noting that the sequence tests failed to indicate the absence of TM2 proline in G3 receptors because, in this helix, the proline residue can be located at various positions (258, 2.59, or 2.60).

Thus, the GPCR sequence space straightforwardly indicates that group G1 is related to the P2.58 pattern, whereas group G3 is related to the correlated absence of the proline residues in TM2 and TM5. However, the similarities in the proline patterns do not necessarily imply that the sub-families within these groups are phylogenetically related. We discuss this issue below.

4.2 Distinguishing mono vs. polyphyletic groups

The clustering of several sub-families into a group may be due either to divergent evolution from a common ancestor (monophyletic group) or to parallel or convergent evolution (polyphyletic group). Several strategies can be used to differentiate between these two alternatives.

1. Analyze the GPCR sets from a wide variety of species. Several studies have shown that some GPCR sub-families present in the human genome (SO, PEP, AMIN, OPN and LGR) are very ancient and are present in all the animal genomes analyzed to date (Fredriksson and Schioth, 2005; Devillé *et al.*, 2009; Pelé *et al.*, 2011a). Other sub-families are lineage dependent and are present only in bilaterians (AD), chordates (MEC, PTG, CHEM, MTN), vertebrates (PUR) or terrestrial vertebrates (MRG). The phylogenetic relationships between GPCRs must be consistent with the lineage dependence.
2. Compare phylogenetic trees from different species. These trees may lead to evolutionary relationships with significant bootstrap values in some species and not in others. For example, NJ trees indicate evolutionary relationships between the SO, CHEM, and PUR sub-families on one hand and between the AD and MEC sub-families on the other hand

with significant bootstrap values in *D. rerio* (>55%) but not in humans (Devillé *et al*., 2009).

3. Analyze the genomic positioning of the receptors. For example, several PUR and CHEM receptors—albeit initially considered as unrelated sub-families—are located on the same paralogon in the human genome, a pattern that suggests a phylogenetic relationship (Fredriksson *et al*., 2003).
4. Search for infrequent evolutionary events. The absence of phylogenetic relationships between the LGR, MRG, and PTG can be inferred from the TM2 and TM5 proline pattern. As a matter of fact, the very ancestral LGR receptors have no proline in either TM2 or TM5, whereas the PTG and MRG receptors have residual proline residues in TM2 and TM5, respectively.
5. Search for hallmark residues of each sub-family. For example, position 6.44 is usually a Phe/Tyr residue in most GPCR sub-families but, for LGR receptors, it is a hallmark Asp or Asn residue that is highly conserved from *N. vectensis* to *H. sapiens*. This position corresponds to the highly conserved Phe residue in the MRG and MEC sub-families and is variable in the PTG sub-family with residual Phe residues. This sequence property does not support the hypothesis that the G3 sub-families evolved from ancestral LGR receptors.

4.3 Evolutionary drift of sub-families

MDS allows the projection of supplementary elements onto a reference space (Gower, 1968; Abdi, 2007b). The position of the supplementary elements depends only on their distance to the reference elements. This property can be used for a straightforward comparison of orthologous sequences.

1. Prepare aligned sequence sets from the different species to be analyzed. Be careful to align the sequences *within* and *between* sequence sets.
2. Assign the orthologous receptors to the sub-families present in the reference species by sequence homology. Do not rely on the first hit but prefer the rule of four hits out of the first five hits in the same sub-family (Fredriksson and Schioth, 2005).
3. Perform MDS analysis of the reference sequence set to obtain the reference sequence space.
4. Project supplementary elements onto the reference sequence space. The *bios2mds* package provides the *mmds.project* function to perform the projection and different graphical tools to facilitate the analysis.

An example of projection of supplementary elements onto a reference space is shown in Fig. 5. In this figure, the SO and PTG receptors from *C. elegans*, *C. intestinalis* and *D. rerio*

are projected onto the sequence space of human GPCRs and provide a straightforward evidence of the evolutionary drift of these sub-families.

In order to avoid an over-interpretation of these data, however, two points are worth noting. First, supplementary receptors from sub-families with no equivalent in the reference species will be projected towards the centre of the reference space. As a matter of fact, these sub-families evolved in a dimension that is orthogonal to the first three components of the reference space. Second, the positions of sub-families present both in the reference and in the supplementary species correspond to the position for the last common ancestor (because divergence leads to evolution in independent dimensions).

5. STRUCTURAL EVOLUTION OF GPCRs

What can MDS tell us about the structural evolution of GPCRs? To answer this question, we analyze the history of the proline pattern of TM2 and TM5 helices and its consequences for the evolution of GPCRs.

5.1 History of the TM2 proline pattern

The MDS representation of the GPCR sequence space leads to cluster the SO, CHEM, and PUR receptors into a single group. These receptors, characterized by the P2.58 pattern, are phylogenetically related and originate from a deletion in TM2 that led to the split between ancestral PEP and SO receptors (Devillé *et al*., 2009). Such a deletion in TM2 is also observed in arthropod opsins (Devillé *et al*., 2009) and in several peptide receptors (e.g. the motilin receptor), indicating that this deletion can be relatively easily accommodated within the TM2 helix. As a matter of fact, most receptors (e.g. rhodopsin from group G0 and β-adrenergic receptors from group G2) possess a bulge in TM2. The deletion of one residue in this bulge leads to a kinked structure, experimentally observed in the chemokine receptor CXCR4, which is the prototype of P2.58 receptors (Wu *et al.*, 2010).

The MDS analysis provides further insights into the history of the G1 receptors. The projection of orthologous sequences onto the sequence space of human GPCRs supports a three step model: (1) an initial deletion in TM2 from a member of the PEP sub-family, leading to ancestral SO receptors, (2) an evolutionary drift of ancestral SO receptors, leading to vertebrate SO receptors and (3) the differentiation of the vertebrate SO receptors into the CHEM and PUR sub-families (Pelé *et al*., 2011a). It is important to note that the deletion in TM2 alone is not sufficient to initiate a novel sub-family. The SO sub-family arises from the initial deletion *and* from subsequent mutations.

Careful comparison of sequences between PEP and SO receptors points towards the galanin/kisspeptin receptors as the ancestors of the SO receptors (Devillé *et al*., 2009). In phylogenetic trees, these receptors may cluster either with the SO receptors or with the PEP receptors (see Fig. 1). The presence of somatostatin and galanin receptors on the same 7/16p/17q/22 paralogon (Fredriksson *et al.*, 2003) corroborates their evolutionary relationship.

5.2 History of the correlated mutations of the TM2 and TM5 proline residues

The MDS representation of the GPCR sequence space also leads to cluster the LGR, PTG, MEC, and MRG sub-families into one group (G3). These sub-families, characterized by the correlated absence of the proline residues in TM2 and TM5, evolved independently (see above). The LGR receptors have a very ancestral origin because they were present before the bilaterian split. The MEC receptors evolved from AD receptors, with the correlated loss of the proline residues in TM2 and TM5. On the other hand, PTG and MRG receptors have residual proline residues in either TM2 or TM5, pointing toward an independent origin.

The high correlation between proline substitutions in TM2 and TM5 suggests a *covarion* process (Fitch, 1971). The analysis of the proline pattern in the PTG and MRG receptors from different species indicates that the first proline residue to be mutated may be located either in TM2 or in TM5. The P2.59 pattern of PTG receptors from *C. intestinalis* and *D. rerio* is partly lost in mammalians (3 out of 8 human receptors), providing an example of "recently" mutated receptors, in link with the drift of this receptor sub-family (Fig. 5). The mutation of one of the proline residue in TM2 or TM5 seems to facilitate the mutation of the second proline residue. The origin of these correlated mutations for residues that are 25 Å apart in the 3D structure of the receptors remains to be investigated.

Interestingly, substitution of the TM5 proline is also observed in the "recent" PUR sub-family (17 receptors out of 46 in humans). This observation might explain the discrepancies between phylogenetic methods for the relationships between the GPCR sub-families. As a matter of fact, the PUR receptors cluster with the LGR and MRG receptors characterized by the absence of proline in TM5 with maximum parsimony (Fredriksson *et al*., 2003) but these PUR receptors cluster with the receptors characterized by the P2.58 pattern with distance based methods (Surgand *et al.*, 2006; Devillé *et al.*, 2009). This difference might be related to the weight of position 5.50 as an informative site in maximum parsimony, and this, in turn, might explain the differences in the four groups between the MDS and the maximum parsimony approach. It is noteworthy that MDS corroborates the phylogenetic link between P2.58 receptors without information on a particular site.

6. SUMMARY

Evolutionary information is hidden within the sequences. Whatever the method used to recover the history of a protein family, the first step is the careful building of a multiple sequence alignment. Then, MDS analysis gives a three-dimensional representation of the distances between these sequences, the so-called sequence space. Compared to traditional phylogenetic methods that rely on a two dimensional tree, this representation helps remove ambiguity and is robust in regard to distance measures.

MDS provides general trends that may be related to ancient evolutionary events. However, the clustering of sub-families does not necessarily imply an evolutionary relationship resulting from divergent evolution but may also arise from parallel or convergent evolution. Thus, the trends revealed by MDS have to be carefully analyzed, in view of the initial multiple alignments, sequence patterns, and additional information such as genomic location. Projection of supplementary sequences onto a reference sequence space allows a straightforward comparison of orthologous sequence sets and may provide clues to understand the evolution of a protein family.

Applied to GPCRs, MDS emphasizes the role of peptide receptors as a central node of GPCR evolution and reveals evolutionary trends related to the proline patterns of TM2 and TM5. The detailed structural and functional implications of these proline mutations have still to be deciphered.

ACKNOWLEDGMENTS: This work was supported by institutional grants from CNRS, INSERM and university of Angers. We thank NEC Computers Services SARL (Angers, FRANCE) for the kind availability of a multiprocessor server. JP was supported by a fellowship from Conseil Général de Maine-et-Loire. JMB was supported by studentships from the Centre Hospitalier Universitaire (CHU) of Angers and from CNRS.

REFERENCES

Abdi, H. (2007a). Distance. In: Salkind NJ (ed) *Encyclopedia of Measurement and Statistics.* Sage, Thousand Oaks (CA), p 280-284.

Abdi, H. (2007b). Metric multidimensional scaling. In: Salkind NJ (ed) *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks (CA), p 598-605.

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary reviews: Computational Statistics* **2**, 433-459.

Bjarnadottir, T. K., Gloriam, D. E., Hellstrand, S. H., Kristiansson, H., Fredriksson, R., and Schioth, H. B. (2006). Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics* **88**, 263-73.

Bockaert, J., and Pin, J. P. (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *Embo J* **18**, 1723-9.

Charif, D., and Lobry, J. R. (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M (eds) *Structural approaches to sequence evolution: Molecules, networks, populations.* Springer Verlag, p 207-232.

Chessel, D., Dufour, A. B., and Thioulouse, J. (2004). The ade4 package-I: One-table methods *R news* **4**, 5-10.

Deupi, X., and Standfuss, J. (2011). Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr Opin Struct Biol* **21**, 541-51.

Devillé, J., Rey, J., and Chabbert, M. (2009). An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. *J Mol Evol* **68**, 475-89.

Fitch, W. M. (1971). Rate of change of concomitantly variable codons. *J Mol Evol* **1**, 84-96.

Fredriksson, R., Lagerstrom, M. C., Lundin, L. G., and Schioth, H. B. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256-72.

Fredriksson, R., and Schioth, H. B. (2005). The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* **67**, 1414-25.

Gether, U. (2000). Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr Rev* **21**, 90-113.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325-38.

Gower, J. C. (1968). Adding a Point to Vector Diagrams in Multivariate Analysis *Biometrika* **55**, 582-585.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857-871.

Haitina, T., Fredriksson, R., Foord, S. M., Schioth, H. B., and Gloriam, D. E. (2009). The G protein-coupled receptor subset of the dog genome is more similar to that in humans than rodents. *BMC Genomics* **10**, 24.

Higgins, D. G. (1992). Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput Appl Biosci* **8**, 15-22.

Holm, L., and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423-9.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-82.

Kass, I., and Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* **48**, 611-7.

Katritch, V., Cherezov, V., and Stevens, R. C. (2012). Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol Sci* **33**, 17-27.
Luecke, H., Richter, H. T., and Lanyi, J. K. (1998). Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science* **280**, 1934-7.
Mirny, L., and Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J Mol Biol* **308**, 123-9.
Nicholas, K. B., Jr, N. H. B., and Deerfield, D. W. I. (1997). GeneDoc: Analysis and Visualization of Genetic Variation. *EMBNEW NEWS* **4**, 14.
Nordstrom, K. J., Sallman Almen, M., Edstam, M. M., Fredriksson, R., and Schioth, H. B. (2011). Independent HHsearch, Needleman-Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. *Mol Biol Evol* **28**, 2471-80.
Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2011). Vegan: Community Ecology Package. R package version 1.17-11 *http://CRAN.R-project.org/package=vegan*,
Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nat Rev Drug Discov* **5**, 993-6.
Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-90.
Pelé, J., Abdi, H., Moreau, M., Thybert, D., and Chabbert, M. (2011a). Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors. *Plos One* **6**, e19094.
Pelé, J., Bécu, J.-M., Abdi, H., and Chabbert, M. (2011b). bios2mds: From BIOlogical Sequences to MultiDimensional Scaling. *http://cran.r-project.org/web/packages/bios2mds*,
Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.
Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L. B., and Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452**, 1002-6.
Sealfon, S. C., Chi, L., Ebersole, B. J., Rodic, V., Zhang, D., Ballesteros, J. A., and Weinstein, H. (1995). Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor. *J Biol Chem* **270**, 16683-8.
Surgand, J. S., Rodrigo, J., Kellenberger, E., and Rognan, D. (2006). A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* **62**, 509-38.
Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-9.
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80.
Torgerson, W. S. (1958) Theory and methods of scaling. Wiley, New York.
Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-91.
Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F. C., Hamel, D. J., Kuhn, P., Handel, T. M., Cherezov, V., and Stevens, R. C. (2010). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330**, 1066-71.
Young, G., and Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Biometrika* **3**, 19-22.

TABLE 1

Classification of human class A GPCRs[1]

MDS group	Sub-families	Description
G0	PEP	Peptide receptors
	OPN	Opsins
	MTN	Melatonin receptors
G1	SO	Somatostatin/opioid receptors
	CHEM	Chemotactic receptors
	PUR	Purinergic receptors
G2	AMIN	Amine receptors
	AD	Adenosine receptors
G3	MEC	Melanocortin, EDG and cannabinoid receptors
	LGR	Leucine rich repeat receptors
	PTG	Prostaglandin receptors
	MRG	Mas-related receptors

[1] The classification of human non-olfactory class A GPCRs is based on the analysis reported in Pelé *et al.*, 2011a, with the nomenclature adapted from Fredriksson *et al.*, 2003.

TABLE 2

Programs to perform MDS analysis

Program name	Function	Comments
Jalview 2 (Waterhouse et al., 2009)	*principal component analysis* in the *calculate* menu	Performs MDS analysis of a multiple sequence alignment; distance matrix calculated from BLOSUM scores only; fails for large set of sequences.
MODELLER (Sali and Blundell, 1993)	*principal_components*	Performs MDS analysis of a multiple sequence alignment; distance matrix calculated from sequence identities only.
R basic tools (cran.r-project.org)	*cmds*	Performs MDS analysis of a distance matrix; can be used with R packages such as *ape* (Paradis *et al.*, 2004) and *seqinr* (Charif and Lobry, 2007) to read in sequences and calculate distance matrices.
ade4 R package (Chessel *et al.*, 2004)	*dudi.pco*[1]	
vegan R package (Oksanen *et al.*, 2011)	*wcmdscale*	
bios2mds R package (Pelé *et al.*, 2011b)	*mmds*	Provides the tools necessary to perform MDS analysis from a multiple sequence alignment and analyze the data; includes the *project.mmds* function for the projection of supplementary elements onto a reference space.

[1] In a*de4*, MDS is called principal coordinate analysis (PCO)

LEGENDS

Fig. 1: NJ trees of human GPCRs. NJ trees of human non-olfactory class A GPCRs were computed from distance matrices based on either the difference scores (a) or on the JTT amino acid substitution matrix (b). The color code refers to the GPCR sub-families (AD, purple; AMIN: steal; CHEM: dark blue; LGR: brown, MEC: salmon; MTN: grey; MRG: magenta; OPN: orange; PEP: dark green; PTG: cyan; PUR: light green; SO: red; UC: black). The arrows indicate the galanin/kisspeptin receptors. Computed with the MEGA4 program from the sequence alignment reported in Pelé *et al.*, 2011a.

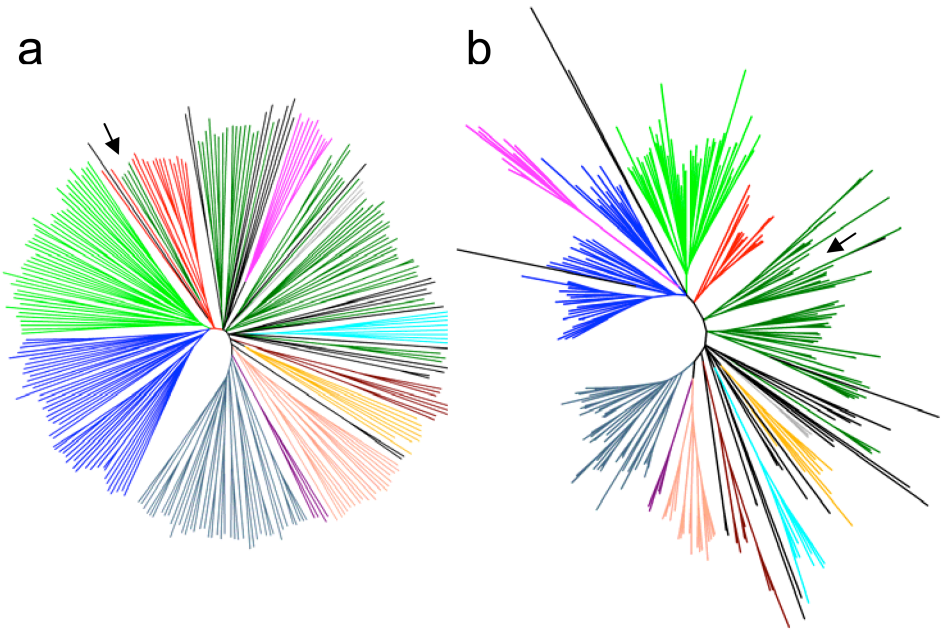
Fig. 2: Sequence space of human GPCRs based on the difference scores. Human non-olfactory class A GPCRs are mapped in the planes formed by the first and second components (a) and by the first and the third components (b) of the sequence space obtained by MDS analysis, with the distances based on the difference score. The color code refers to the GPCR sub-families (AD, purple; AMIN: steal; CHEM: dark blue; LGR: brown, MEC: salmon; MTN: grey; MRG: magenta; OPN: orange; PEP: dark green; PTG: cyan; PUR: light green; SO: red; UC: black). The crosses indicate the galanin/kisspeptin receptors. Spanning ellipses visualize the clusters. Computed with the *bios2mds* package from the sequence alignment reported in Pelé *et al.*, 2011a.

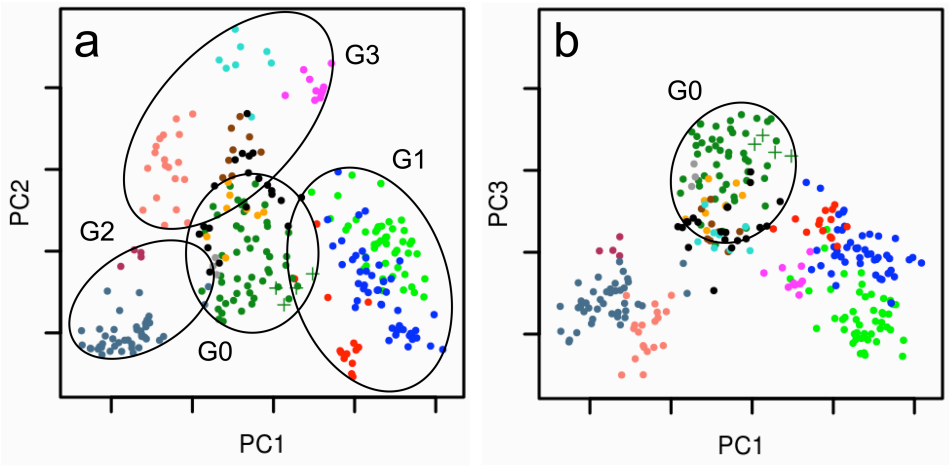
Fig. 3: Sequence space of human GPCRs based on the JTT matrix. Human non-olfactory class A GPCRs are mapped in the planes formed by the first and second components (a) and by the first and the third components (b) of the sequence space obtained by MDS analysis, with the distances based on the JTT amino acid substitution matrix. The color code refers to the GPCR sub-families (AD, purple; AMIN: steal; CHEM: dark blue; LGR: brown, MEC: salmon; MTN: grey; MRG: magenta; OPN: orange; PEP: dark green; PTG: cyan; PUR: light green; SO: red; UC: black). The crosses indicate the galanin/kisspeptin receptors. Spanning ellipses visualize the clusters. Computed with the *bios2mds* package from the sequence alignment reported in Pelé *et al.*, 2011a.

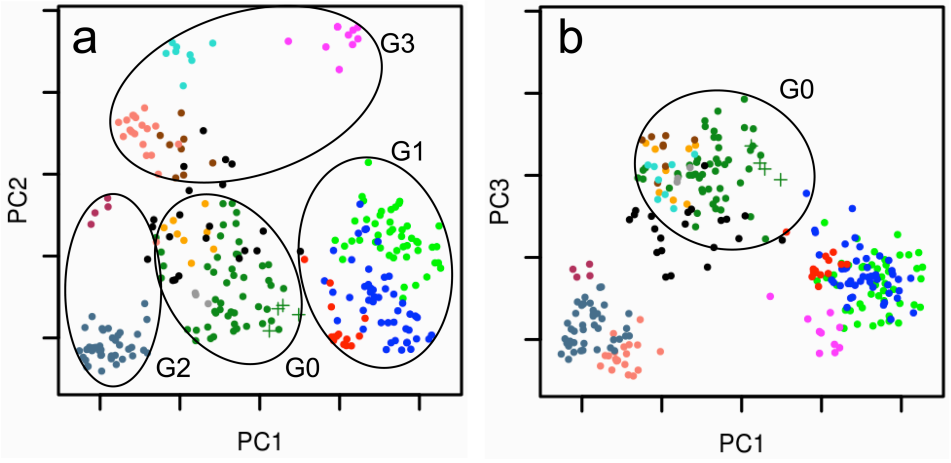
Fig. 4: Proline pattern of human GPCRs. The proline patterns of TM2 (a) and TM5 (b) are shown in the 2D mapping of the GPCR sequence space. The distances are based on the difference score. In (a), the color code refers to the TM2 proline pattern (P2.58: blue; P2.59, red; P2.60, green; no proline, black). For proline doublets, the color code corresponds to the

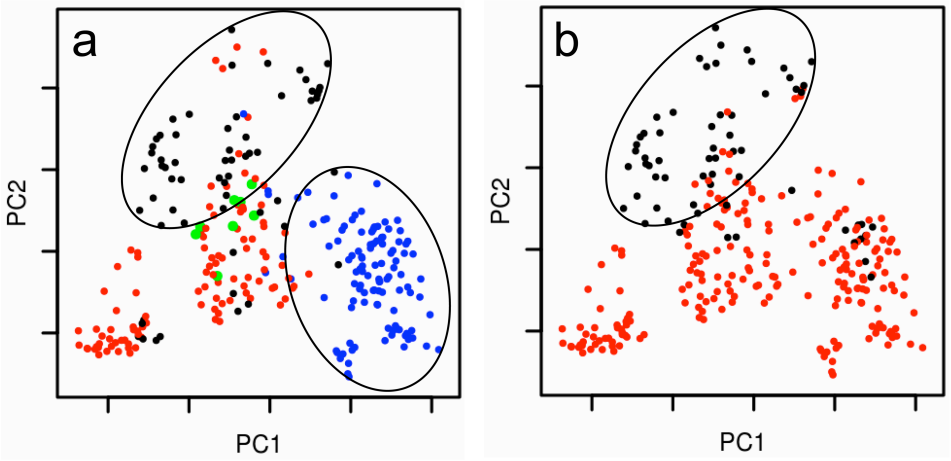
position of the first proline residue. In (b), the color code refers to the TM5 proline pattern (P5.50: red; no proline, black). Spanning ellipses visualize groups G1 and G3. Drawn with the *bios2mds* package from original data in Pelé *et al.*, 2011a.

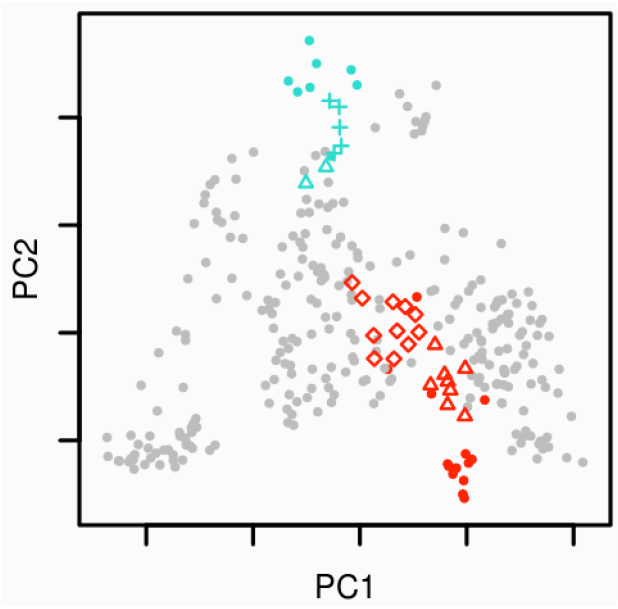
Fig. 5: Evolutionary drift of GPCR sub-families. The SO (red symbols) and PTG receptors (cyan symbols) from different species (*C. elegans*, diamonds; *C. intestinalis*, triangles; *D. rerio*, crosses) are projected onto the 2D sequence space of human GPCRs. The distances are based on the difference scores. The human receptors are indicated by circles. The SO receptors from *D. rerio* are not shown for clarity purpose because they are superimposed on the human SO receptors.Adapted from Pelé *et al.*, 2011a with the *bios2mds* package.


a
b


a
G3
G1
G2
G0
PC2
PC1
b
G0
PC3
PC1


a
G3
G1
G2
G0
PC2
PC1
b
G0
PC3
PC1


a
b
PC2
PC1
PC2
PC1


PC2
PC1