PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Comparative analysis of sequence covariation methods to mine evolutionary hubs: Examples from selected GPCR families

Julien Pelé,[1] Matthieu Moreau,[1] Hervé Abdi,[2] Patrice Rodien,[1,3] Hélène Castel,[4] and Marie Chabbert[1]*

[1] UMR CNRS 6214–INSERM 1083, Laboratory of Integrated Neurovascular and Mitochondrial Biology, University of Angers, 49045 Angers, France
[2] The University of Texas at Dallas, School of Behavioral and Brain Sciences, Richardson, TX 75080-3021, USA
[3] Department of Endocrinology, Reference Centre for the pathologies of hormonal receptivity, Centre Hospitalier Universitaire of Angers, 4 rue Larrey, 49933 Angers, France
[4] INSERM U982, Laboratory of Neuronal and Neuroendocrine Communication and Differentiation, DC2N, University of Rouen, 76821 Mont-Saint-Aignan, France

## ABSTRACT

**Covariation between positions in a multiple sequence alignment may reflect structural, functional, and/or phylogenetic constraints and can be analyzed by a wide variety of methods. We explored several of these methods for their ability to identify covarying positions related to the divergence of a protein family at different hierarchical levels. Specifically, we compared seven methods on a model system composed of three nested sets of G-protein-coupled receptors (GPCRs) in which a divergence event occurred. The covariation methods analyzed were based on: $\chi^2$ test, mutual information, substitution matrices, and perturbation methods. We first analyzed the dependence of the covariation scores on residue conservation (measured by sequence entropy), and then we analyzed the networking structure of the top pairs. Two methods out of seven—OMES (Observed minus Expected Squared) and ELSC (Explicit Likelihood of Subset Covariation)—favored pairs with intermediate entropy and a networking structure with a central residue involved in several high-scoring pairs. This networking structure was observed for the three sequence sets. In each case, the central residue corresponded to a residue known to be crucial for the evolution of the GPCR family and the subfamily specificity. These central residues can be viewed as evolutionary hubs, in relation with an epistasis-based mechanism of functional divergence within a protein family.**

Proteins 2014; 00:000–000.
© 2014 Wiley Periodicals, Inc.

**Key words: GPCR; protein evolution; sequence Covariation; correlated mutations; epistasis; protein family; sequence analysis.**

## INTRODUCTION

The sequence analysis of protein families provides a wealth of information on their sequence–structure–function relationship. Typically, the first step of the analysis is the construction of a multiple sequence alignment (MSA). Then, a variety of methods may be used to analyze, for example, sequence conservation,[1] phylogenetic relationships,[2] evolutionary trace,[3] or sequence covariation between residues.[4] The covariation between two positions within an MSA is presumed to reflect structural and/or functional correlations between positions. Constraints on one residue may require a compensatory mutation of another residue to maintain or restore the structure and/or the function of the corresponding protein. Such compensatory constraints may depend on direct physical interactions between two residues or on an indirect interaction through intermediary residues or a ligand.[5]

The analysis of sequence covariation data is not straightforward. The sequences of a protein family are phylogenetically related and the amino acid frequencies computed from the MSA are observed frequencies among sequences that are not independent. In an MSA containing several subfamilies, observed covarying positions may result from different mechanisms that are schematized in Figure 1. When covarying positions are independent of the subfamily (Fig. 1a), it is usually assumed that they correspond to compensatory

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: French Research Agency; Grant number: ANR-11-BSV2-026.

*Correspondence to: Marie Chabbert, UMR CNRS 6214-INSERM U1083, Faculty of Medicine, 3 rue Haute de reculée, 49045 Angers, France. E-mail: marie.chabbert@univ-angers.fr

Received 26 July 2013; Revised 11 March 2014; Accepted 19 March 2014
Published online 27 March 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24570

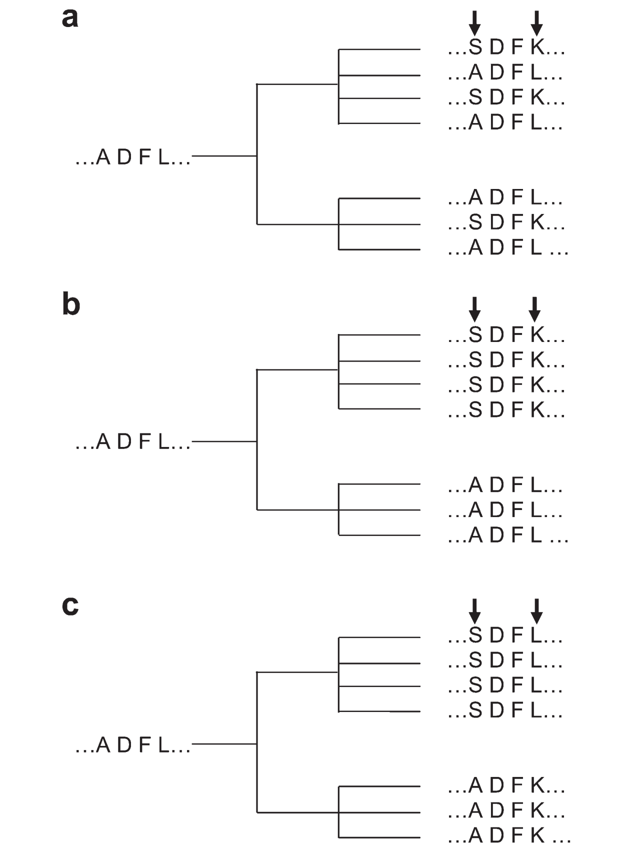

**Figure 1**

Schematic representation of evolutionary mechanisms resulting in observed covarying positions. In (**a**), Ala1 and Leu4 coevolved to Ser1 and Lys4 independently of the subfamily. In (**b**), Ala1 and Leu4 coevolved to Ser1 and Lys4 in subfamily 1 and remained unchanged in subfamily 2. In (**c**), Ala1 was mutated to Ser1 in subfamily 1 and Leu4 was mutated to Lys4 in subfamily 2, in two independent evolutionary events.

mutations important for the protein fold and that their analysis may lead to global structural information on the protein family. However, covarying positions frequently depend on the subfamily. In this case, they may arise either from compensatory mutations within one subfamily (Fig. 1b) or from independent mutations in several subfamilies (Fig. 1c). Both mechanisms lead to a phylogenetic bias, related to the inhomogeneity of the MSA.

Up to now, most covariation studies were aimed at gaining structural information from coevolving positions. Many efforts in the field have been employed to differentiate direct coupling from indirect coupling due to the transitivity effect of pair covariation in terms of structural vicinity[5] and to remove—or at least to reduce—the phylogenetic bias. This bias may be reduced implicitly[6–8] or explicitly by taking into account the relationships between the sequences contained in the MSA.[9,10] The combination of covariation methods with a message-passing approach,[11,12] a Bayesian network model,[5] or a maximum entropy model[13] helps disentangle direct from indirect covariation, and can therefore markedly improve the prediction of interacting pairs. Recently, the maximum entropy method has been successfully used for *de novo* three-dimensional (3D) protein structure prediction with a root-mean-square deviation (RMSD) in the 2.7–4.8 Å range for a variety of soluble or membrane proteins.[13,14] For instance, the structural models of rhodopsin and of the β2-adrenergic receptor—two G-protein-coupled receptors (GPCRs)—predicted *de novo* with this method, differed from the experimental crystal structures with an RMSD of only 3.3 Å.[14]

However, the information provided by the analysis of coevolving residues in an MSA is not limited to the determination of contact pairs and distance constraints. A wide variety of covariation methods have been developed and are used to get information on connectivity pathways,[15–17] protein sectors,[18] or specificity-determining residues.[19–22] Comparison of different methods on artificial and real MSAs[4,6,7] has shown that the performance of these methods to answer specific questions crucially depends on the characteristics of the MSA (e.g., number of sequences, conservation, homogeneity). In this study, we are interested in identifying the correlated mutations that contribute to the subfamily divergence within a protein family. We thus compare several widely used covariation methods for their adequacy with this aim. With this perspective, the phylogenetic bias, related to the intrinsic inhomogeneity of a sequence set containing different subfamilies, is not deleterious but represents a potential source of information about the evolutionary mechanisms of a protein family.

As a model system, we used the human repertoire of non-olfactory class A GPCRs. These receptors have a characteristic topology of seven transmembrane helices (TM) and form a large protein family comprising about 300 members.[23] To test the covariation methods on real MSAs with different characteristics, we considered the full set of receptors and two nested subsets—in which a divergence occurred—of about 100 and 20 sequences. We show that, out of the seven methods analyzed here, two methods favor networks of covarying residues with a hub structure that can be related to the diversification of this receptor family at different hierarchical levels. The implication of the hub structure for protein evolution will be discussed.

## MATERIALS AND METHODS

### Sequence sets

We used three sets of GPCR sequences from *Homo sapiens* to test the different methods. The first set is formed by the non-redundant repertoire of non-olfactory class A receptors that we have developed previously.[24,25] The second set is formed by three GPCR subfamilies that are evolutionary related by divergence: the somatostatin/opioid (SO), chemotaxic (CHEM), and purinergic receptors

(PUR) subfamilies.[24] Finally, the third set is composed of the chemokine receptors that form a subset within the CHEM subfamily, and can be further divided into two phylogenetically related groups.[26] The three sets have 283, 107, and 23 sequences, respectively, and this makes it possible to test the methods on a 10-fold range in the number of sequences around the threshold of 100–125 sequences widely used in covariation studies.[7,27] Two additional sequence sets were used for validation. The PEP-AMIN set comprises the 53 peptide receptors and the 42 amine receptors that are included in Set 1. The PEP-OPN set is formed by the 53 peptide receptors from Set 1 and a nonredundant set of 61 opsins from six different species: *Homo sapiens* (eight sequences), *Drosophila melanogaster* (seven sequences), *Danio rerio* (15 sequences), *Nematostella vectensis* (17 sequences), *Ciona intestinalis* (two sequences), and *Branchistoma floridae* (12 sequences).

All the sequence analyses were carried out on the 236 positions of the MSAs with less than 2% gaps. These positions correspond to the seven transmembrane helices extended to the loop regions wherever possible. The threshold of 2% introduces a small bias in the covariation estimates for the pairs involving these positions but allows monitoring the evolutionary important WXFG motif in the extracellular loop 1.[25] The GPCR positions were numbered by reference to the most conserved position *n*.50 in each helix *n*, according to Ballesteros' nomenclature.[28] The MSAs used for the analysis are available as Supplementary Information.

## Sequence conservation

For each sequence set, the pairwise sequence conservation was analyzed by the cumulative distribution function (cdf) calculated by the program GeneDoc[29] from the 236 positions of the MSAs. For each identity score (ID), the cdf function gives the proportion of the scores (from 0 to 1) as low as or lower than the score being plotted (from 0 to 100% ID).

The sequence conservation at position *i*, $S(i)$, was measured from the MSA by sequence entropy according to the following formula, derived from the Shannon's entropy[30]

$$S(i) = -\sum_{x} p_x(i)\log_{20} p_x(i) \tag{1}$$

where *i* is the position of interest in the sequence alignment, *x* stands for the 20 amino acids, and $p_x(i)$ represents the frequency of residue *x* at the *i*th position. A base 20 logarithm was used to ensure that the entropy takes values between 0 and 1.[7]

## Analysis of sequence covariation

We reviewed the different methods found in the literature to measure the covariation between residues in an MSA and we selected several methods representative of the main four classes: (1) methods based on the $\chi^2$ test,[4,16] (2) methods based on mutual information (MI),[6–8,22,31] (3) methods based on substitution matrices,[32–34] and (4) methods using perturbation of an MSA.[17,35] We detail these methods below.

### Method based on the $\chi^2$ test

A method strictly based on the $\chi^2$ test of homogeneity was initially developed[16] but abandoned because it did not meet Cochran's criterion due to a high number of degrees of freedom and the low expected numbers of elements per class.[36] To circumvent this problem, Fodor and Aldritch[37] developed a variant, called OMES (Observed minus Expected Squared). The OMES method calculates the difference between the observed and expected occurrences of each possible pair of amino acids (*x, y*) at positions *i* and *j* of the alignment as

$$\mathrm{OMES}(i,j) = \frac{1}{N(i,j)} \sum_{x,y} \left(N^{\mathrm{obs}}_{x,y}(i,j) - N^{\mathrm{ex}}_{x,y}(i,j)\right)^2 \tag{2}$$

where $N(i,j)$ is the number of sequences in the alignment with non-gapped residues at positions *i* and *j*, $N^{\mathrm{obs}}_{x,y}(i,j)$ is the number of times that each (*x, y*) pair is observed at positions *i* and *j*, and $N^{\mathrm{ex}}_{x,y}(i,j)$ is the number of times that each (*x, y*) pair would be expected, based on the frequency of residues *x* and *y* in columns *i* and *j*, respectively. The value of $N^{\mathrm{ex}}_{x,y}(i,j)$ for an amino acid pair (*x, y*) at positions *i* and *j* is given by

$$N_{\mathrm{ex}} = p_x(i)p_y(j)N \tag{3}$$

where $p_x(i)$ and $p_y(j)$ are the frequencies of amino acids *x* and *y* at positions *i* and *j*, respectively.

### Methods based on MI

The MI content, MI (*i, j*), between two positions *i* and *j* in an alignment is based on the probability of joint occurrence of events[38] and is given by the following formula[39]

$$\mathrm{MI}(i,j) = \sum_{x,y} p_{x,y}(i,j) \ln \frac{p_{x,y}(i,j)}{p_x(i)p_y(j)} \tag{4}$$

where $p_{x,y}(i,j)$ is the frequency of the amino acid pair (*x, y*) at positions *i* and *j*.

This intuitive formula has been widely applied in the field of sequence covariation but favors pairs with high entropy.[4] To correct this bias, two approaches have been developed. The first one uses statistical models to determine statistically significant correlated pairs.[22,39] The second one corrects the MI scores to reduce the influence of the entropy. We tested the MIr (MI reduced), MIa (MI additive), and MIp (MI product) corrective factors proposed by Wahl and coworkers[6,7] and, as these authors did, we found on our data sets that MIp outperforms both MIa and MIr. Thus, we will use here only the MIp correction in which the average product correction is subtracted from the MI score to obtain

$$\mathrm{MIp}(i,j)=\mathrm{MI}(i,j)-\frac{\mathrm{MI}(i,\bar{j})\mathrm{MI}(\bar{i},j)}{<\mathrm{MI}>} \tag{5}$$

with

$$\mathrm{MI}(i,\bar{j})=\frac{1}{n-1}\sum_{j\neq i}\mathrm{MI}(i,j) \tag{6}$$

$$\mathrm{MI}(\bar{i},j)=\frac{1}{n-1}\sum_{i\neq j}\mathrm{MI}(i,j) \tag{7}$$

and

$$<\mathrm{MI}>\ =\frac{2}{n(n-1)}\sum_{i,j}\mathrm{MI}(i,j) \tag{8}$$

The mutual interdependency (MINT) correction[8] was implemented to differentiate phylogenetically related positions with many interdependencies and functionally related positions with a limited number of interdependencies. The MINT score is given by

$$\mathrm{MINT}(i,j)=\frac{\mathrm{MI}(i,j)}{\mathrm{MS}(i)+\mathrm{MS}(j)}[S(i)S(j)(1-S(i)S(j))] \tag{9}$$

where *S(i)* and *S(j)* are the entropies at positions *i* and *j*, respectively, and *MS(i)* and *MS(j)* denote the Multiple Significant Interdependency at positions *i* and *j*, respectively, that are computed as

$$\mathrm{MS}(i)=\sum_{j\neq i}\mathrm{MI}(i,j) \tag{10a}$$

$$\mathrm{MS}(j)=\sum_{i\neq j}\mathrm{MI}(i,j) \tag{10b}$$

### Methods based on substitution matrices

The McLachlan Based Substitution Correlation method (McBASC) was initially proposed by Valencia and coworkers[32] and relies on a substitution matrix giving a similarity score for each pair of amino acids. The McBASC score is given by the formula

$$\mathrm{McBASC}(i,j)=\frac{1}{N^2\sigma(i)\sigma(j)}\sum_{k,l}(SC_{k,l}(i)-SC(i))(SC_{k,l}(j)-SC(j)) \tag{11}$$

where $SC_{k,l}(i)$ and $SC_{k,l}(j)$ are the scores for the amino acid pair present in sequences *k* and *l* at positions *i* and *j*, respectively, *SC*(*i*) and *SC*(*j*) are the averages of all the scores $SC_{k,l}(i)$ and $SC_{k,l}(j)$, respectively, and σ(*i*) and σ(*j*) are the standard deviations of all the scores $SC_{k,l}(i)$ and $SC_{k,l}(j)$, respectively. We tested this method with both the original McLachlan matrix[40] and with the Miyata matrix.[41] We obtained similar results and, therefore, will only report the results obtained with the McLachlan's formula.

### Perturbations of a MSA

Statistical coupling analysis (SCA) was initially proposed to analyze pathways of energetic connectivity in proteins[17] and subsequently modified to be applied to the analysis of covariations in an MSA.[4] Based on perturbation analysis, the algorithm works by choosing a subset of sequences within an MSA and then compares the characteristics of the subset with those of the full alignment. The subset corresponds to the sequences within the most prevalent amino acid *x* in column *i*. The SCA score is given by

$$\mathrm{SCA}(i,j)=\sqrt{\sum_{y}(\ln P_y(j|\delta_i)-\ln P_y(j))^2} \tag{12}$$

where $P_y(j|\delta_i)$ is the frequency of amino acid *y* at position *j* in the subset defined by the presence of the amino acid *x* at position *i*.

A method similarly based on perturbation of an MSA[35] varies in the measurement of the deviation of the amino acid composition between the subset alignment with $n_y$ sequences having the amino acid *y* at position *j* and the total alignment with *N* sequences. The Explicit Likelihood of Subset Covariation (ELSC) is based on rigorous statistics of covariation in a perturbation-based algorithm. It measures how many possible subsets of size *n* would have the composition found in column *j*

$$\mathrm{ELSC}(i,j)=-\ln\prod_{y}\frac{\binom{N_{y(j)}}{n_{y(j)}}}{\binom{N_{y(j)}}{m_{y(j)}}} \tag{13}$$

where$N_y(j)$, $n_y(j)$ and $m_y(j)$ are, respectively, the numbers of residues *y* at position *j* in the total (unperturbed) sequence alignment, in the subset alignment defined by the perturbation in column *i*, and in the ideal subset (i.e., in a subset with the amino acid distribution equal to the total alignment). The binomial coefficient $\binom{N}{k}$ is computed as

$$\binom{N}{k}=\frac{N!}{k!(N-k)!} \tag{14}$$

where ! denotes the factorial operation (i.e., *N*! = 1 × 2 × . . . × *N*).

All the scripts for the sequence analyzes were written in Perl. The scripts for the MI, SCA, ELSC, and McBASC methods were based on the Java codes provided by Fodor and coworkers at http://www.afodor.net. The code for the MINT method was based on the code provided by Tillier at http://www.uhnresearch.ca/labs/tillier/software.htm#2. Strictly conserved positions were removed from the computation. The total number of covarying pairs was equal to 27,730 for Set 1, 27,497 for Set 2, and 27,659 for Set 3.

### Comparison of the covariation methods

We first analyzed the distribution of the covariation scores to determine the number of top ranking pairs that will be subsequently used for comparison purpose. Boxplots of the *Z*-scores were drawn with R (http://cran.r-project.org). The subsequent analyses were carried out on the top ranking pairs.

#### *Dependence on the entropy*

For each pair ($i$, $j$) of positions in the alignment, we plotted the covariation score between these positions as a function of their entropy, $S(i)$ and $S(j)$. On the $S(i) \times S(j)$ plots, the rank of the covariation score between the positions $i$ and $j$ is given by a color scale with five levels. The top 25 and the next 250 pairs are dark and light blue, respectively. The bottom 25 and the next 250 pairs are red and pink, respectively. The other pairs are grey. These plots visualize the level of conservation of the pairs with top and bottom scores.

#### *Network properties*

We analyzed the network properties of the top 25 pairs. The covariation between the positions was visualized with Cytoscape 2.8, an open-source software for visualizing interaction networks.[42] With this software, the nodes are represented by positions and the edges by covariation signals between them. The connectivity of a residue from these pairs was measured as the number of pairs, within the top 25 pairs, in which this residue is involved.

#### *3D properties of the top pairs*

We visualized the positions of the top 25 pairs on the 3D structure of a typical GPCR, the μ opioid receptor (Protein Data Bank access number: 4DKL), using PyMol, an openGL based molecular visualization system (http://www.pymol.org). The distances between pairs of residues were computed with a Perl script from this structure, using the Cβ atoms as reference atoms, except for glycine residues for which the Cα, atoms were used. A pair of residues was considered in contact when this distance was lower than 8 Å.[33,43]

## RESULTS

### Selection of model sequence sets

To compare covariation methods for their ability to identify correlated mutations that contribute to the divergence of a protein family, we need model sequence sets in which a divergence occurred and led to two sequence subsets consistent in terms of size and conservation. Otherwise, the largest or the most conserved group might bias the signal. Based on our previous phylogenetic analysis of class A GPCRs,[24,25] we searched for GPCR sequence sets for which these conditions were fulfilled.

The full repertoire of human class A GPCRs is composed of about 300 non-olfactory and 400 olfactory receptors.[44] However, the full repertoire is not adapted to test covariation methods, because olfactory receptors have 30 specifically conserved positions[45] that create 450 strongly correlated pairs. Thus, we limited our analysis to non-olfactory receptors. This sequence set includes 283 sequences that are classified into 12 subfamilies ranging from 3 to 53 members.[25,44] We have previously showed that a major pathway of GPCR evolution arose from the deletion of one residue in TM2, very early in metazoan evolution. This pathway led to three subfamilies that are evolutionary related and characterized by a P2.58 pattern.[24,25] The initial deletion in TM2 led first to the SO subfamily that subsequently diverged to give the CHEM subfamily in chordates and the PUR subfamily in vertebrates. The human SO, CHEM, and PUR subfamilies have 14, 47, and 46 members, respectively. The non-olfactory repertoire is thus composed of the SO/CHEM/PUR subfamilies, on the one hand, and of the complementary subfamilies, on the other hand with, respectively, a relative weight of 38 and 62% and an identity median of 26 and 20%.

In the SO/CHEM/PUR set, the PUR receptors are characterized by a FXP pattern in TM6 and a DPXXY pattern in TM7, whereas most GPCRs possess a WXP and a NPXXY pattern at these positions. The SO/CHEM/PUR set is thus composed of the SO and CHEM subfamilies, on the one hand, and of the PUR subfamily, on the other hand with, respectively, a relative weight of 57 and 43% and an identity median of 29 and 27%. The chemokine receptors, which include 23 members, are part of the CHEM subfamily. Phylogenetic analysis of the chemokine receptors showed that they are shared into receptors for homeostatic chemokines and receptors for inflammatory chemokines.[26,46] Sequence analysis reveals that position 2.49 in TM2 is differentially conserved in these two subsets that have, respectively, a relative weight of 56 and 44% and an identity median of 37 and 45%.

We thus selected the three sequence sets formed by the repertoire of non-olfactory receptors (Set 1), the SO/CHEM/PUR subfamilies (Set 2), and the chemokine

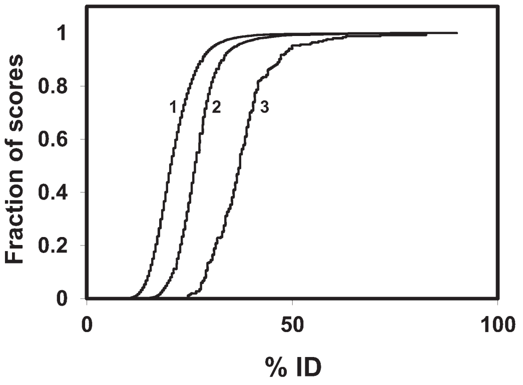

**Figure 2**

Cumulative distribution function of the pairwise IDs. The fraction of the scores as low as or lower than the score being plotted is given for Set 1 (left) to Set 3 (right).

receptors (Set 3), and we compared different covariation methods for their ability to mine sequence covariation in link with the divergence that occurred in these sequence sets. The three sets are nested (i.e., Set 1 includes Set 2 that includes Set 3). The cumulative distribution function of the IDs within each sequence set shows the homogeneity of the sequence sets in terms of sequence identity (Fig. 2). As expected from the hierarchical relationship between the sets, the median of the IDs increases from 20% in Set 1 to 26% in Set 2 and 37% in Set 3, whereas the average entropy decreases from 0.68 in Set 1 to 0.59 in Set 2 and 0.41 in Set 3. These values were determined from the MSAs used for the covariation analysis.

The plot of the sequence entropy for the different positions of the MSA (Fig. 3) shows the periodicity of variable and conserved positions along the helices and extending loop regions and the range of entropy investigated for each set. A highly conserved residue in each helix $n$, indicated by an entropy value close to zero, is used as the positional anchor $n$.50 in the Ballesteros' numbering. The only exception is TM5, for which a proline residue at position 5.50 is conserved in only 77% of the sequences in Set 1. This ratio increases to 91% and 96% in Sets 2 and 3, respectively. In TM2, position 2.58 is variable in Set 1, but corresponds to a highly conserved proline residue in Sets 2 and 3. This is illustrated by the dramatic decrease in the entropy of this position from Set 1 to Set 2 (Fig. 3). In TM7, the anchor residue 7.50 is the proline residue of the NPXXY motif. Most receptors, including the chemokine receptors (Set 3), have an asparagine at position 7.49. However, PUR receptors, which are present in both Set 1 and Set 2 but not in Set 3, have an aspartic acid at this position. This explains the entropy value at position 7.49 that decreases from about 0.28 in Sets 1 and 2 to 0.06 in Set 3.

## Focus on the outliers of the Z-score distributions

To compare the distributions of the scores obtained with the different methods, we normalized the raw scores by computing $Z$-scores (i.e., a score is now represented by its number of standard deviations to the mean[47]). This normalization makes no hypothesis on the shape of the distributions but makes it possible to compare different methods, because it effectively removes the scale of measurement. The boxplots of the distributions are shown in Figure 4. The tails of the distributions obtained with the seven methods are markedly different but the general trends do not depend on the sequence set. The distributions observed with the OMES, McBASC, SCA, and ELSC methods have a marked tail with at least 65 $Z$-scores above 4.0 (as a comparison point, a $Z$-score of 4 corresponds to a $p$-value $< 10^{-4}$ for a Gaussian distribution). Conversely, the MI and MINT methods usually do not lead to well-separated outliers, whereas MIp leads to 20–60 outliers that are larger than 4.0, depending on the sequence set. The $Z$-score of the first ranking pair is always higher with OMES than with any other method. The unusually high values of the top $Z$-scores obtained with OMES for Set 2 (up to 34) will be discussed below.

Several strategies have been used in the literature to determine the number of reliable covarying pairs. They include bootstrap methods,[31,48] a $Z$-score limit of 4.0,[6,7] a number of pairs determined by the MSA length,[49,50] a percentage of covarying pairs,[51] or a limited number of highly scoring pairs.[4,52–54] In any case, the covariation pairs that are investigated are on the tail of the $Z$-score distribution. In most cases studied here, the limit of 275 pairs, corresponding to 1% of all pairs, is among the outliers of the $Z$-score distributions

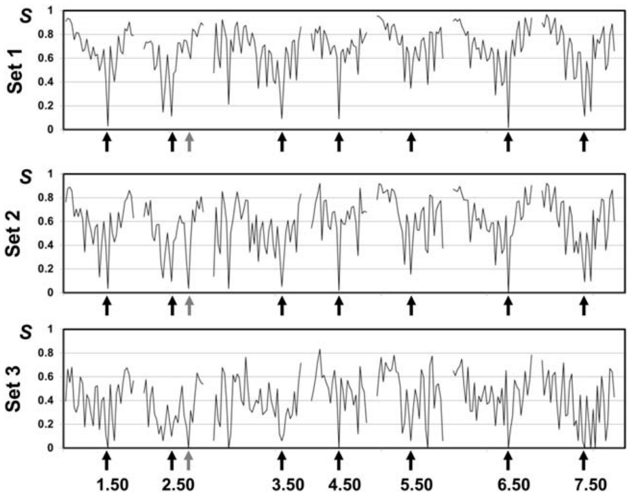

**Figure 3**

Sequence entropy at each position of the MSA for Sets 1–3 (from top to bottom). The black arrows indicate the anchor residues. The grey arrow indicates position 2.58.

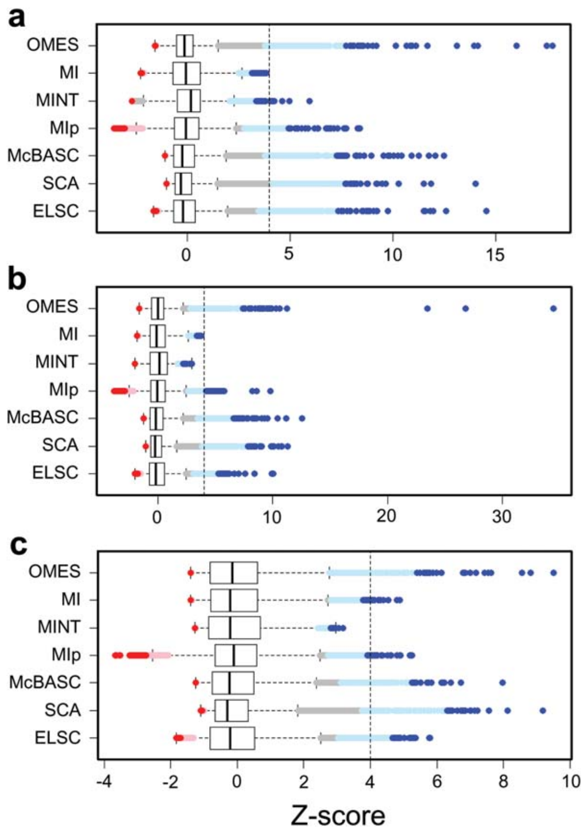

**Figure 4**

Boxplots of the $Z$-scores calculated from the distribution of the covariation scores obtained with the different methods for Set 1 (**a**), Set 2 (**b**), and Set 3 (**c**). The boxes represent the 25th and 75th percentile of the range and include the median as the thick line. The whiskers extend to the lesser of the total range of values or 1.5 times the interquartile range. The top and bottom 275 $Z$-scores and outliers outside the whisker thresholds are shown as circles. The top 25 and the next 250 $Z$-scores are dark blue and light blue, respectively. The bottom 25 and the next 250 Z-scores are red and pink, respectively. Outliers outside the top and bottom 275 $Z$-scores are grey. The dashed line indicates a Z-score of 4.

($Z$-scores larger than 1.5 times the upper quartile) but clearly below the value of 4.0 (Fig. 4). When we strictly limit the number of pairs to 25, these highly scoring pairs are on the very tail of the $Z$-score distribution, usually above the value of 4.0. For the subsequent analysis, we focus on the top 25 pairs and carefully analyze their properties, to determine the parameters underlying the differences between the methods.

### Entropy bias of the covariation scores

The relationship between covariation scores and residue conservation depends on the algorithms that may favor highly conserved or highly variable positions.[4,6] To mine covarying positions in relation with subfamily divergence, a covariation method should favor pairs of positions with an intermediate level of conservation. Highly conserved positions are not informative whereas highly variable positions may lead to spurious results. We used the entropy $S(i)$ to quantify the conservation level of position $i$. Figure 5 displays the covariation scores of each pair $(i, j)$ as a function of $S(i)$ and $S(j)$ for the seven methods and the three sequence sets. The color

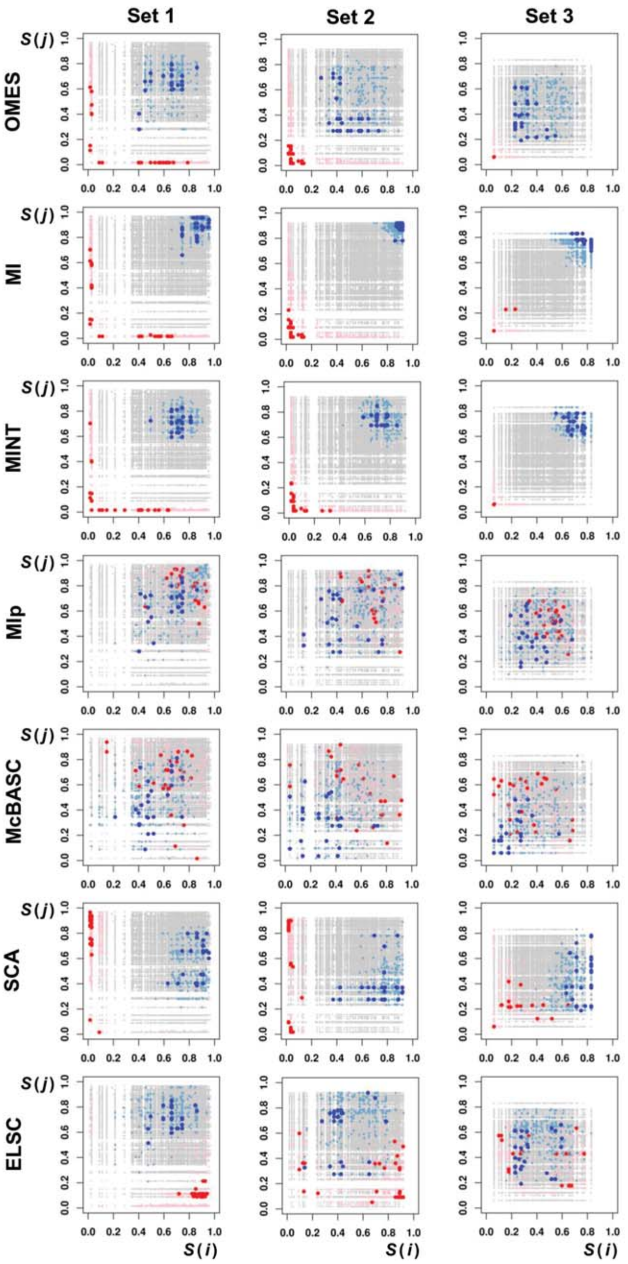

**Figure 5**

Comparison of the covariation methods for their dependence on the entropy of each position, for Sets 1–3. For each pair $(i, j)$ of positions in the alignment, the covariation score between these positions is plotted as a function of their entropy, with a five-level color code indicative of its rank (blue: the top 25 $Z$-scores, light blue: the top 275 $Z$-scores, red: the bottom 25 $Z$-scores, pink: the bottom 275 $Z$-scores, others: grey).

**Table I**
Overlap between the Top Pairs obtained by the Different Covariation Methods[a]

| | MI | MINT | MIp | McBASC | SCA | ELSC |
|---|---|---|---|---|---|---|
| Set 1 | | | | | | |
| OMES | 2 | 11 | 11 | 6 | 2 | 13 |
| MI | | 2 | 3 | 1 | 2 | 2 |
| MINT | | | 14 | 6 | 2 | 11 |
| MIp | | | | 9 | 2 | 11 |
| McBASC | | | | | 1 | 7 |
| SCA | | | | | | 2 |
| Set 2 | | | | | | |
| OMES | 0 | 0 | 11 | 5 | 6 | 11 |
| MI | | 0 | 2 | 0 | 0 | 0 |
| MINT | | | | 0 | 2 | 0 |
| MIp | | | | 8 | 5 | 12 |
| McBASC | | | | | 2 | 5 |
| SCA | | | | | | 5 |
| Set 3 | | | | | | |
| OMES | 0 | 0 | 7 | 4 | 0 | 15 |
| MI | | 4 | 0 | 0 | 2 | 0 |
| MINT | | | 0 | 0 | 0 | 0 |
| MIp | | | | 6 | 0 | 6 |
| McBASC | | | | | 0 | 5 |
| SCA | | | | | | 0 |

[a]Gives the number of pairs within the top 25 pairs that are common to two methods.

code allows visualizing the dependence on the entropy of the pairs with the top and bottom scores.

Three different cases are observed. The first one corresponds to the OMES, MI, and MINT methods. Plots are symmetrical with two characteristics: (1) the bottom pairs possess at least one highly conserved position (entropy close to zero) and (2) the top pairs are markedly separated from the bottom ones. For the OMES method, the top pairs have an intermediate level of entropy. The average values decrease from 0.63 ± 0.13 in Set 1 to 0.34 ± 0.11 in Set 3. The top scoring pairs obtained with the MI method correspond to highly variable positions with a value of the entropy in the 0.8–0.9 range (upper right corner). The MINT method favors positions whose entropy is in the 0.6–0.8 range, in all three sets. These positions correspond to the most variable positions for Set 3, the smallest and most conserved set under consideration, but not for Sets 1 and 2 that are larger and more variable.

The second case is observed for the MIp and McBASC methods. It corresponds to fuzzy data as the top and bottom scoring pairs have overlapping entropies. For MIp, the top pairs have an intermediate entropy, from 0.63 ± 0.14 in Set 1 to 0.39 ± 0.13 for Set 3, similar to those observed for OMES. McBASC is biased toward pairs of positions with low entropy. The average values decrease from 0.49 ± 0.17 in Set 1 to 0.20 ± 0.12 in Set 3.

The third case is observed for methods based on perturbations that are not symmetrical. The bottom pairs for SCA and ELSC are obtained for highly conserved *i* and *j* positions, respectively, at least for Sets 1 and 2. For the top (*i*, *j*) pairs obtained with SCA, the entropy *S*(*i*) is very high (0.86 ± 0.08, 0.74 ± 0.15 and 0.76 ± 0.07 for Sets 1, 2 and 3, respectively), whereas *S*(*j*) is markedly lower (0.59 ± 0.15, 0.39 ± 0.15 and 0.44 ± 0.19). Such an asymmetry is not observed with the ELSC method. In this later case, the average entropy is 0.69 ± 0.09, 0.53 ± 0.20, and 0.40 ± 0.15 for, respectively, Sets 1, 2, and 3.

This entropy bias leads to different top pairs. Table I reports the number of identical pairs in the top 25 pairs for the different methods. In Set 1, the MINT, MIp, OMES, and ELSC methods have about 50% overlap. This overlap is still conserved for the MIp, OMES, and ELSC methods for Set 2, and is only maintained between OMES and ELSC for Set 3. The MIp and McBASC methods have also some overlap ranging from 36% in Set 1 to 24% in Set 3.

### *Variability in the network structures of the top pairs*

It is widely admitted that covarying residues fall into two classes.[27] The first class is composed of residues covarying with only one or two other positions and favors contact pairs. The second class is composed of residues that coevolved with several others and are often located in areas crucial for protein function. We used the Cytoscape software[42] to visualize the connections within the top pairs. The resulting graphs reveal the wide variety of the networking structures obtained by the different methods (Fig. 6).

To quantify these structures, we determined the connectivity of the residues forming the top 25 pairs. In this analysis, the connectivity of residue *k*, *conn*(*k*), is a parameter based on network theory that indicates the number of high-scoring pairs (here 25) in which a residue *k* is included.[54] This parameter might be a robust indicator of evolutionary important residues.[54,55] Table II reports the number of residues with a single partner, with at least two partners, and the maximum number of pairs in which a residue is involved, among the top 25 pairs. The number of residues involved in a single pair is variable among the methods, since it varies from 5 (e.g., MI with Set 3) to 20 and more (e.g., MIp and McBASC with Sets 1 to 3). As previously described,[4,6,56] the MIp and the McBASC methods perform well to find single pairs. This is not the case for the OMES, MI, and MINT methods, whereas the perturbation methods are intermediary.

The number of residues involved in at least two pairs of covarying residues does not show these large variations as it varies from 8 to 13. Nevertheless, the maximum number of pairs in which a residue is involved is variable. For example, for Set 1, it varies from 5 with the McBASC method to 12 with the OMES method. This high connectivity of one or a few residues gives them the

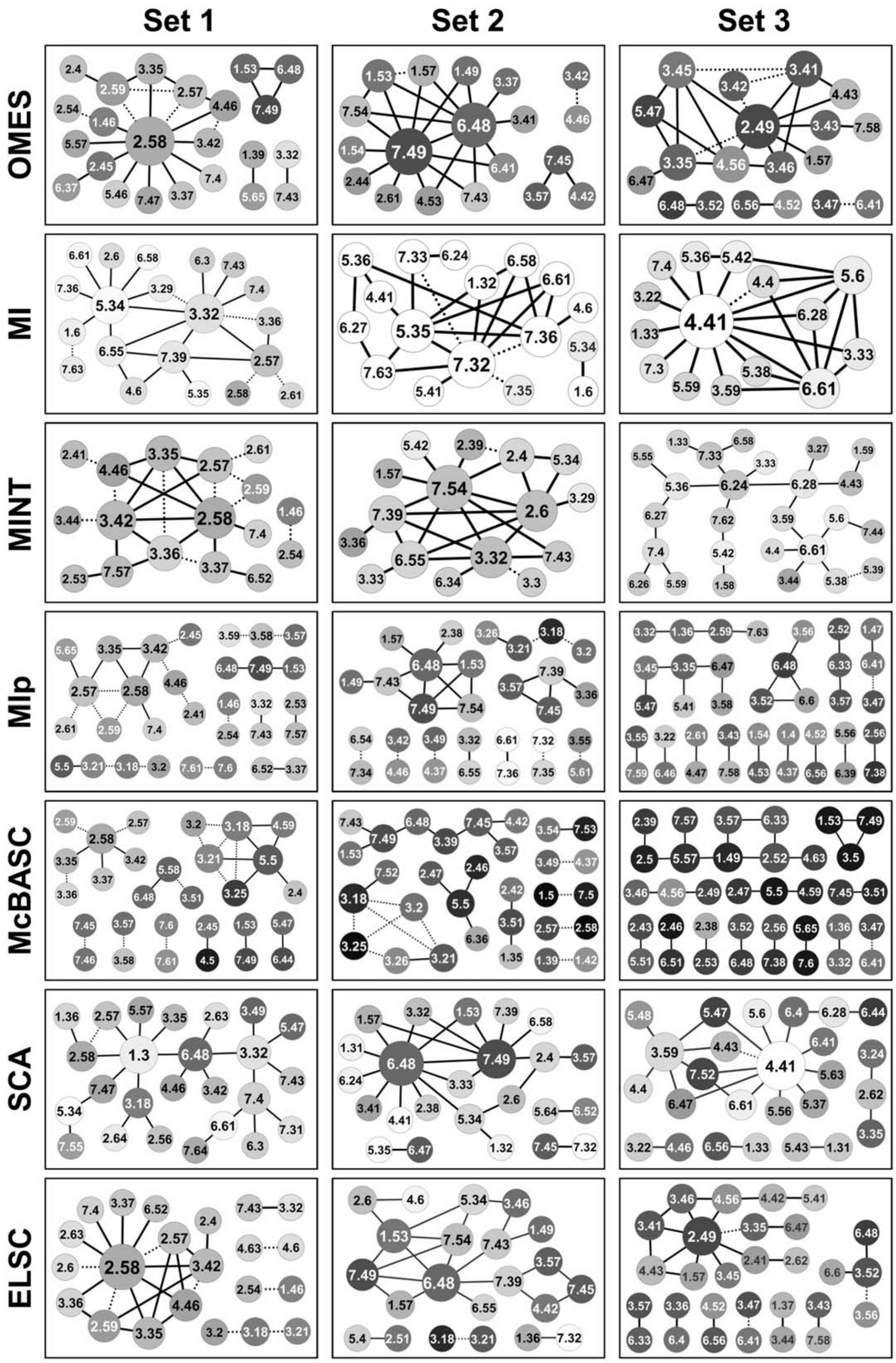

**Figure 6**

Visualization of the network structure of the top 25 pairs of covarying positions, for Sets 1–3. The nodes represent the positions and the edges represent the covariation signal between them. The size of each node is proportional to the connectivity of the corresponding position. The shade of the node indicates the entropy of the position, from black for a fully conserved position to white for the most variable position in the set under investigation. Dotted edges indicate that the distance between the Cβ atoms of the corresponding positions is less than 8 Å.

**Table II**
Connectivity of the Residues in the Top 25 Pairs[a]

| | Set 1 | | | | | Set 2 | | | | | Set 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Connectivity | | | | | Connectivity | | | | | Connectivity | | | | |
| Method | 1 | ≥2 | Max | $i_{\mathrm{Max}}$ | *S*(*i*) | 1 | ≥2 | Max | $i_{\mathrm{Max}}$ | *S*(*i*) | 1 | ≥2 | Max | $i_{\mathrm{Max}}$ | *S*(*i*) |
| OMES | 12 | 11 | 12 | 2.58 | 0.66 | 9 | 10 | 11 | 7.49 | 0.27 | 8 | 9 | 9 | 2.49 | 0.23 |
| MI | 11 | 9 | 9 | 3.32 | 0.86 | 6 | 11 | 9 | 7.32 | 0.92 | 5 | 10 | 14 | 4.41 | 0.83 |
| MINT | 8 | 9 | 8 | 2.58 | 0.66 | 7 | 9 | 9 | 7.54 | 0.69 | 13 | 13 | 5 | 6.61 | 0.78 |
| MIp | 21 | 10 | 5 | 2.57, 2.58 | 0.74, 0.66 | 20 | 10 | 6 | 6.48 | 0.37 | 28 | 10 | 5 | 6.47 | 0.50 |
| McBASC | 20 | 9 | 5 | 2.58, 3.18, 5.50 | 0.66, 0.47, 0.35 | 20 | 11 | 4 | 3.18 | 0.93 | 25 | 11 | 3 | 3.32 | 0.36 |
| SCA | 15 | 10 | 7 | 1.30 | 0.91 | 15 | 9 | 11 | 6.48 | 0.37 | 16 | 10 | 11 | 4.41 | 0.83 |
| ELSC | 14 | 8 | 11 | 2.58 | 0.66 | 9 | 13 | 7 | 6.48 | 0.37 | 19 | 9 | 8 | 2.49 | 0.23 |

[a] The connectivity is the number of pairs (within the top 25 pairs) in which each residue is included. The connectivity of 1 refers to the number of residues included in a single pair. The connectivity larger or equal to 2 refers to the number of residues included in at least 2 pairs. Max refers to the maximum number of pairs in which a residue is included, $i_{\mathrm{Max}}$ is the position of this residue and *S*(*i*) its entropy. When several residues have the maximal connectivity, the positions and the entropies are given in the same order.

role of a hub in the network of covarying residues. However, with the MI, SCA, and, to a lower extend, the MINT method, residues with high connectivity frequently have high entropy (Fig. 6 and Table II). The amino acid variability at these positions suggests that most of these high-scoring pairs might result from spurious correlation. This hypothesis is corroborated by the analysis of the amino acid distribution of the pairs. For example, position 5.34 is sorted out by MI for Set 1. This position accepts any of the 20 amino acids and forms 165 and 135 different amino acid pairs (out of 283 pairs) with positions 3.29 and 3.32. None of these pairs has a weight larger than 4%.

Conversely, the residues with high connectivity observed with the OMES and ELSC methods have intermediate entropy and correspond to positions known to be important in the evolutionary events that led to subfamily divergence (Fig. 6 and Table II). The hub residue observed with both methods corresponds to position 2.58 for Set 1 and position 2.49 for Set 3. For Set 2, the hub residue corresponds either to 7.49 with OMES or to 6.48 with ELSC. Position 7.49 has a connectivity of 4 with ELSC whereas position 6.48 has a connectivity of 10 with OMES, to be compared to the connectivity of 11 for position 7.49. The presence of these two highly correlated residues which share six common partners may explain the unusually high *Z*-score of the top pairs obtained for this sequence set with OMES (Fig. 4).

The hub residue observed with both OMES and ELSC for Set 1 (position 2.58) is present among the top pairs obtained with the other methods and, in most cases, is involved in several pairs (Table III). Conversely, the hub residue observed with OMES and ELSC for Set 3 is not found with the other methods, except McBASC (Rank 21). For Set 2, both positions 6.48 and 7.49 are present among the top pairs with the MIp, McBASC, and SCA methods, whereas neither one is observed with the MI and MINT methods.

A network of covarying positions involving position 5.50 in TM5 and the WXFG motif (positions 3.18 to 3.21) is detected in the top 25 pairs of Set 1 by the McBASC and MIp methods (Fig. 6). Positions 5.50, 3.18, and 3.21 are characteristic of one of the GPCR groups previously observed by multidimensional scaling analysis.[25] The covariation between these positions is detected in the top 275 pairs by OMES and ELSC but not by MI, MINT, and SCA (not shown). These differences may be related to the low entropy of position 5.50 (0.34) and to the entropy bias of each method (Fig. 5).

**Table III**
Connectivity and Rank of the OMES Hub Residue *h*[a]

| | Set 1 | | Set 2 | | Set 3 | |
|---|---|---|---|---|---|---|
| Method | *Conn(h)* | *r(h)* | *Conn(h)* | *r(h)* | *Conn(h)* | *r(h)* |
| OMES | 12 | 1 | 11 | 1 | 9 | 1 |
| MI | 1 | 12 | 0 | 5713 | 0 | 491 |
| MINT | 8 | 2 | 0 | 490 | 0 | 8295 |
| MIp | 5 | 1 | 4 | 1 | 0 | 51 |
| McBASC | 5 | 10 | 3 | 1 | 1 | 21 |
| SCA | 3 | 3 | 8 | 2 | 0 | 65 |
| ELSC | 11 | 1 | 4 | 1 | 8 | 3 |

[a] The hub residue *h* observed with OMES corresponds to residues 2.58, 7.49, and 2.49 for Sets 1, 2, and 3. *Conn(h)* refers to the number of pairs, within the top 25 pairs, in which residue *h* is included; *r(h)* refers to the rank of the first pair, from top, in which residue *h* is included.

**Table IV**
Inter-Residue Distances in the Top Pairs[a]

| | Set 1 | | Set 2 | | Set 3 | |
|---|---|---|---|---|---|---|
| Method | *N* | *d* | *N* | *d* | *N* | *d* |
| OMES | 5 | 17,0 | 2 | 14,3 | 5 | 13,3 |
| MI | 6 | 15,4 | 3 | 24,2 | 1 | 30,5 |
| MINT | **10** | 13,0 | 2 | 18,1 | 1 | 32,0 |
| MIp | **13** | 11,6 | **7** | 13,9 | 1 | 23,8 |
| McBASC | **12** | 13,1 | **10** | 13,1 | 1 | 19,4 |
| SCA | 1 | 21,8 | 0 | 20,5 | 1 | 24,9 |
| ELSC | **8** | 14,1 | 1 | 20,5 | 3 | 17,2 |

[a] *N* refers to the number of contact pairs and *d* to the average inter-residue distance (Å) within the top 25 pairs. Bold numbers indicate that the *Z*-scores for the propensity of contact pairs, estimated from a normal distribution, is above 4.0. The propensity is the ratio of the observed to the expected numbers of contact pairs, estimated from the average in the reference structure (4.1%).

### Localization of the top pairs in receptor structure

For this analysis, we used the structure of the μ opioid receptor as a reference. In this structure, the average distance between any two residues used for the covariation analysis is 24.2 Å and 4.1% of these residue pairs are in contact, based on an 8 Å criterion for the distance between the $C\beta$ atoms.[33,43] This leads to one expected contact pair in the top 25 pairs. The inter-residue distances and the number of contact pairs in the top 25 pairs obtained with each method and each sequence set are reported in Table IV.

With most methods, for Set 1, the average inter-residue distances of the top pairs are significantly below the random distance value of 24 Å, and then increase as the size of the set decreases (Table IV). The exceptions are OMES with average distances $\leq$ 17 Å and SCA with average distances in the 20 Å range for any sequence set. For Sets 1 and 2, the lowest average distances are observed with MIp and McBASC ($d \approx 13 \pm 1$ Å), but these methods lead to a distance of 20 Å or more for Set 3.

Concerning the contact pairs observed among the top 25 pairs, only the SCA method does not lead to an increased propensity for contact pairs in any sequence set (Table V). In Set 1, all the methods, except SCA, lead to a significant increase in contact pairs, with 5–13 contact pairs. In Set 2, however, only MIp and McBASC lead to a significant increase in contact pairs. With these two methods, more than 40% and 30% of the top 25 pairs are within contact distance for Sets 1 and 2, respectively. The performance of these methods to improve residue contact prediction has previously been described in relation with their ability to find isolated pairs[4,6,49,56] and is consistent with the low average inter-residue distances in top pairs. The results obtained for Set 3 are not consistent with those obtained for Sets 1 and 2. None of the methods tested allows an increased detection of contact pairs in this set. This observation can be related to the small size of Set 3 (23 sequences) and corroborates the threshold of about 100 sequences required for the detection of contact pairs.[7]

The top covarying positions were visualized on the 3D structure of the μ opioid receptor (Fig. 7), for the four methods that are not biased toward positions with high entropy (i.e., OMES, ELSC, MIp and McBASC). Figure 7 illustrates the capability of OMES and ELSC to favor a specific position that plays the role of a networking hub for covarying residues, in the three sequence sets under investigation. It is interesting to note that positions 2.58 and 6.48 are located in the ligand binding module,[57,58] whereas most covarying positions detected by OMES are localized in the downstream signaling module. ELSC presents the same trend. The other methods detect covarying residues that are positioned in any domain of the receptor. This observation might result from the specific range of entropy to which each method is biased as the ligand binding domain is more variable that the signaling module (Fig. 3).

## DISCUSSION

The covariation between positions in an MSA can provide different types of information. When the covariation analysis is aimed at detecting contact residues, the phylogenetic bias related to the intrinsic inhomogeneity of a

**Table V**
Detection of Evolutionary Hubs[a]

| Method | Set 1 | Set 2 | Set 3 |
|---|---|---|---|
| OMES | +++ | +++ | +++ |
| ELSC | +++ | ++ | +++ |
| MIp | ++ | ++ | − |
| McBASC | ++ | + | − |
| MI | + | − | − |
| MINT | +++ | − | − |
| SCA | + | +++ | − |

[a] Estimated from the connectivity *C* of the position(s) that correspond to a hallmark in the divergence of the sequence sets under scrutiny (position 2.58 in Set 1, positions 6.48, and 7.49 in Set 2, position 2.49 in Set 3). For set 2, the connectivities of positions 6.48 and 7.49 are summed. The codes for the symbols are: +++ for $\mathrm{Max}(C)/2 < C \leq \mathrm{Max}(C)$; ++ for $\mathrm{Max}(C)/4 < C \leq \mathrm{Max}(C)/2$; + for $1 \leq C \leq \mathrm{Max}(C)/4$; and − for none.

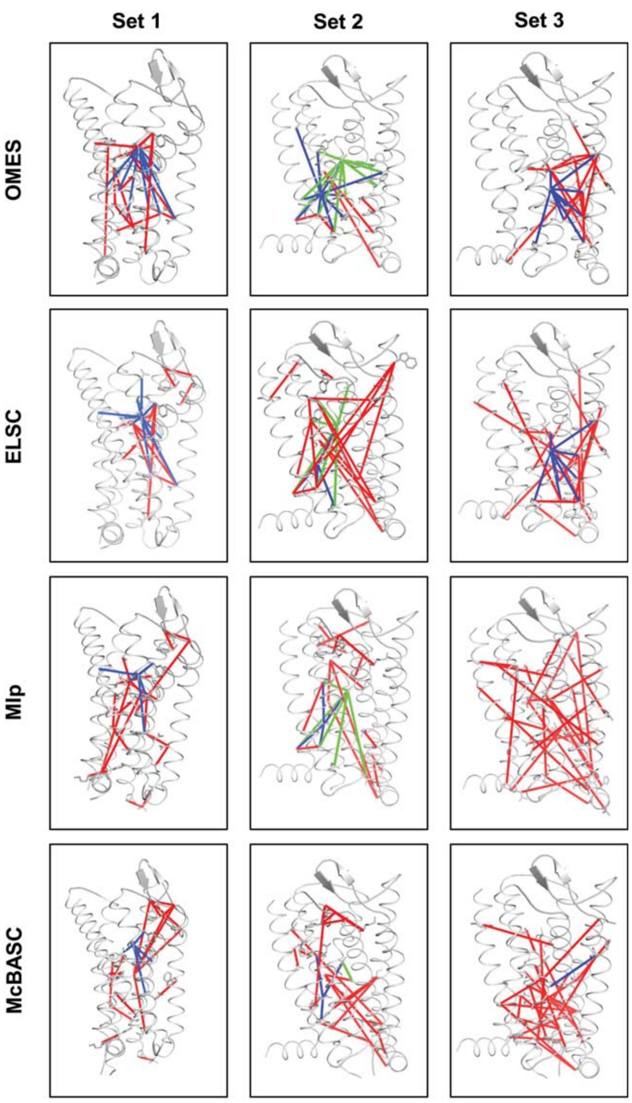

**Figure 7**

Visualization of the top 25 pairs of covarying positions obtained with the OMES, ELSC, MIp, and McBASC methods for Sets 1–3 on the ribbon representation of the μ opioid receptor (4DKL). The blue lines indicate the residues covarying with the hub residue characteristic of OMES (residues 2.58, 7.49, and 2.49 in Sets 1, 2, and 3, respectively). The green lines in Set 2 indicate the residues covarying with residue 6.48 which is also involved in several pairs. The red lines indicate the other pairs of covarying residues.

sequence set containing different subfamilies is deleterious. However, this bias can be exploited to gain information on the molecular mechanisms that drove the divergence of a protein family during evolution. In this article, we searched for a covariation method efficient to identify coevolving residues that can be related to subfamily divergence within a protein family. Thus, we selected seven methods, representative of the main classes of covariation methods and we tested them on three nested sets of about 300, 100, and 20 sequences of human GPCRs. A divergence occurred in each of these sets and is characterized by one or two hallmark positions.

Table V summarizes the covariation analysis and reports the capability of each method to detect covariation pairs involving these hallmark positions. Only two methods, OMES and ELSC, successfully identify such pairs in the three investigated sets. These two methods partially overlap (Table I) and, in both cases, the hallmark residues characterizing the divergence events have a high connectivity and act as covariation hubs (Figs. 6 and 7). MIp and McBASC do not perform as well and fail for Set 3. When hallmark residues are detected with these methods, their connectivity is lower than with OMES and ELSC (Table III). This may be related to the capability of MIp and McBASC to favor pairs that have a low connectivity in the three sets (Table II) and an increased propensity for contact in the largest two sets (Table IV). These data corroborate the adequacy of MIp and McBASC for covariation analysis aimed at enrichment in contact pairs when the number of sequences is above a threshold of about 100 sequences.[4,6] In addition, McBASC may be adapted to specific applications that target low entropy positions. An example is given with the network involving position 5.50 in Set 1 (Fig. 6). Three methods, MI, MINT, and SCA, perform poorly to find pairs involving hallmark residues, except for specific sequence sets (for example SCA with Set 2 or MINT with Set 1). These methods are biased toward pairs with at least one high entropy position (Fig. 5) and, in most cases, the variability of the hub residues (high entropy) indicates the lack of evolutionary constraints and precludes that they can be related to functional divergence.

Several studies indicate that the connectivity of the covarying pairs may be a robust criterion to highlight "true" covarying positions with functional significance.[53–55] Covariation hubs have been observed previously in several families, for example, the Glutamyl-tRNA synthetase catalytic domain, the ferritin-like diiron-carboxylate protein domain, or the hedgehog/intein domain.[59] It must, however, be pointed out that observed covariation hubs do not necessarily imply that these hubs can be interpreted in terms of subfamily divergence. An important criterion to take into account is the entropy of the hub residue and of its partners. Comparison of the different methods reveals that, in our sequence sets, OMES and ELSC are robust methods to detect hub residues in link with subfamily divergence, even when the number of sequences is markedly lower than 100 (Figs. 6 and 7). These methods, which favor pairs with intermediate entropy (Fig. 5), might be powerful to reveal similar evolutionary hubs in other protein families.

What do these hub structures reveal about the evolution of GPCRs? In Set 1, the residues covarying with position 2.58 should be related to the divergence of Set 2 which has a proline at position 2.58 as hallmark. Position

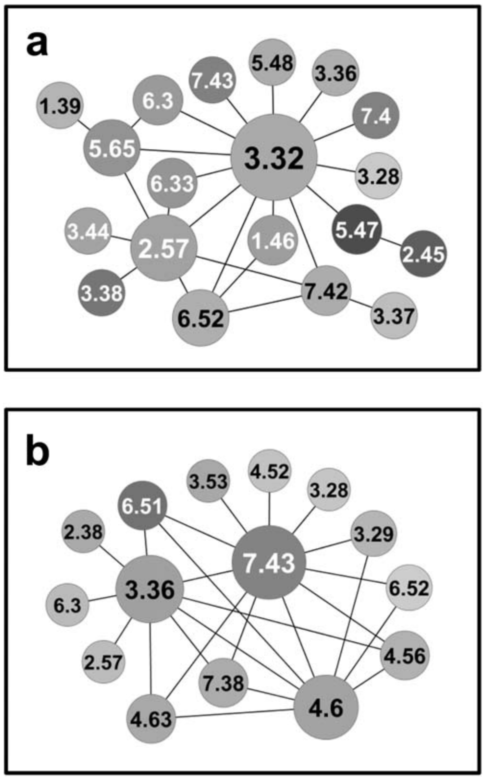

**Figure 8**

OMES analysis of the PEP-AMIN (**a**) and of the PEP-OPN (**b**) sequence sets. The network structure of the top 25 pairs of covarying positions was visualized with Cytoscape. The size of each node is proportional to the connectivity of the corresponding position. The shade of the node indicates the entropy of the position, from black for a fully conserved position to white for the most variable position in the set under investigation.

2.58 is located in the ligand binding module, close to the minor binding pocket,[58] and covaries with positions located in the downstream signaling module[57] (Fig. 7). In particular, position 2.58 covaries with positions 1.46, 5.57, and 7.47 that form consensus contacts between helices to stabilize the receptor fold and with positions 3.35, 3.37, and 3.42 in TM3 that may participate in downstream signaling.[60] Interestingly, numerous receptors from Set 2, including "professional" CHEM receptors,[61–63] receptors from the PUR subfamily[64] and the urotensin II receptor, a member of the SO subfamily,[65] display a chemotaxic activity. The role of the correlated mutations observed in this study in the acquisition of a novel activity remains, however, to be determined.

In Set 2, positions 6.48 and 7.49 differentiate the SO and CHEM subfamilies, on the one hand, and the PUR subfamily, on the other hand. These two residues are specifically mutated from Trp to Phe and from Asn to Asp in the PUR subfamily. Position 6.48 makes contact with diverse ligands across class A GPCRs.[60] Position 7.49 is part of the NPXXY motif in TM7 that participates in the mechanism of receptor activation. Both residues covary with positions 1.49, 1.53, 1.57, 4.53, and 6.41 that participate in consensus contacts between transmembrane helices.[60] The CHEM and PUR subfamilies target different types of ligands. The emergence of a new specificity may have required the correlated mutations of several residues to reorganize the interactions maintaining the receptor scaffold.

In Set 3, the residues present at position 2.49 are only Ala or Ser. The classification of the chemokine receptors based on this pattern match the two groups observed by phylogeny.[26] Receptors from Group 1 (CXCR1–7, CCR6–7, CCR9–10, and CCRL1) appeared in jawless fishes[66] and bind mainly homeostatic chemokines.[46] Receptors from Group 2 (CCR1–5, CCR8, CX3CR1, CCRL2, CCBP2, and XCR) are more recent (jawed vertebrates) and bind mainly inflammatory chemokines.[46] Thus, the network of coevolving positions differentiates the receptors of homeostatic and inflammatory chemokines. Six positions, out of the nine that covary with position 2.49 (Fig. 6), are located in TM3 (3.35, 3.41, 3.42, 3.43, 3.45, and 3.46) and may participate to the downstream signaling from the ligand binding domain to the G protein binding domain.

In the three sets, the residues sorted out as coevolving hubs by OMES and ELSC correspond to hallmarks of the evolutionary processes that drove the GPCR evolution, at different hierarchical levels. However, these networks also indicate that a single key mutation is not sufficient for the divergence to occur. Divergence requires the coevolution of several residues to reorganize the interactions and/or to allow the acquisition of a new specificity or function. The detailed mechanism by which these covariations affect protein function and subfamily specificity remains to be explored.

Can such a model of protein evolution be extended to other GPCR subfamilies? It is worth noting that the analysis of coevolutionary data in terms of subfamily divergence requires not only a suited method, but also adapted sequence sets, with two divergent subsets of consistent size and conservation. Analyzing sequence covariation in link with the divergence of different GPCR subfamilies may require tailored data sets. We give two such examples (Fig. 8) that are based on our model of GPCR evolution by radiation from peptide receptors.[25] The hallmark of the divergence of the amine receptors from the peptide receptors is an Asp residue at position 3.32.[25] The OMES analysis of the PEP-AMIN sequence set including the human peptide and amine receptors from Set 1 reveals the residues covarying with position 3.32 (Fig. 8a). The second example is aimed at detecting covariation in link with the divergence of opsins from peptide receptors. Opsins are characterized by a lysine residue at position 7.43 that is covalently bound to retinal. The small size of the human opsin subfamily (8 members) requires the enrichment of the sequence set with opsin sequences from other species. The analysis of the PEP-OPN sequence set including the human peptide receptors and opsins from several species (see Materials and Methods) allows mining the residues covarying with position 7.43 (Fig. 8b). A similar result was obtained in a recent covariation study of the GPCR family[67] based on a sequence set obtained by Psi-Blast search from a rhodopsin sequence, resulting in the enrichment of the sequences under scrutiny with opsins. The comparison of the networks obtained with different methods for the PEP-AMIN and the PEP-OPN sets corroborate the results obtained previously with Sets 1–3 (Supporting Information, Fig. S1).

Several models of protein evolution, based on the accumulation of mutations within a duplicated gene, have been proposed.[68] They rely on the concept of fitness landscape,[68,69] introduced by Wright in the 1930s.[70] The outcome of one mutation in a gene depends on the presence of other mutations in this gene (epistasis). As a result, new protein functions may arise from the coevolution of several residues. The initial mutation can be neutral or deleterious and protein function will be restored or shifted by one or several additional mutations. The present data on sequence covariation in selected GPCR subfamilies support such an epistasis model of protein evolution.

In a recent article that evaluates the differences between short-term and long-term protein evolution,[71] Kondrashov and coworkers point out that the short-term effects of substitutions may not provide an adequate framework for understanding the effects of accumulated amino acid substitutions on the long-term evolution of proteins. Compensating mutations to maintain structure and/or function, as schematized in Figure 1a, may correspond to short-term evolution, whereas functional divergence following gene duplication is a long-term evolutionary process. In this case, the difference in the evolutionary mechanisms displayed in Figure 1(b,c) does not prevent the mutations to have occurred in the specific context of a subfamily and thus to be related to functional divergence.

In conclusion, the OMES and ELSC methods are well-adapted to find pairs of coevolving residues important for the divergence of a protein family (Table V). Compared to methods aimed at finding specificity-determining positions that are based on the sequence conservation properties of an MSA and that find differentially conserved positions,[72] OMES and ELSC have less stringent requirement of residue conservation and may select differently variable positions. Moreover, compared to the other covariation methods investigated, OMES and ELSC are able to deal with a large range in the size of the sequence set, indicating that these methods can be used to analyze a protein family at different hierarchical levels. Finally, and most importantly, these methods favor residues with high connectivity that can be related with the key evolutionary events which led to the emergence of protein subfamilies. These residues can be viewed as evolutionary hubs, in relation with an epistasis-based mechanism of functional divergence within a protein family.

## ACKNOWLEDGMENT

The authors thank NEC France S.A.S. (Nanterre, France) for the kind provision of a multiprocessor server. They acknowlege the fellowships from CNRS to MM and from Conseil Général de Maine-et-Loire to JP.

## REFERENCES

1. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
2. Whelan S. Inferring trees. Methods Mol Biol 2008;452:287–309.
3. Wilkins A, Erdin S, Lua R, Lichtarge O. Evolutionary trace for prediction and redesign of protein functional sites. Methods Mol Biol 2012;819:29–42.
4. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211–221.
5. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 2010;6:e1000633.
6. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics (Oxford, England) 2008;24:333–340.
7. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics (Oxford, England) 2005;21:4116–4124.
8. Tillier ER, Lui TW. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics (Oxford, England) 2003;19:750–755.
9. Dutheil J, Pupko T, Jean-Marie A, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 2005;22:1919–1928.
10. Tuffery P, Darlu P. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol 2000;17:1753–1759.
11. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 2011;108:E1293–E1301.
12. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 2009;106:67–72.
13. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PloS one 2011;6:e28766.
14. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149:1607–1621.
15. Dima RI, Thirumalai D. Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Sci 2006;15:258–268.
16. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 2002;48:611–617.
17. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
18. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell 2009;138:774–786.
19. Bachega JF, Navarro MV, Bleicher L, Bortoleto-Bugs RK, Dive D, Hoffmann P, Viscogliosi E, Garratt RC. Systematic structural studies of iron superoxide dismutases from human parasites and a

statistical coupling analysis of metal binding specificity. Proteins 2009;77:26–37.

20. Bleicher L, Lemke N, Garratt RC. Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. PloS one 2011;6:e27786.
21. Chakrabarti S, Panchenko AR. Coevolution in defining the functional specificity. Proteins 2009;75:231–240.
22. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol 2002;321:7–20.
23. Chabbert M, Castel H, Pele J, Deville J, Legendre R, Rodien P. Evolution of class A G-protein-coupled receptors: implications for molecular modeling. Curr Med Chem 2012;19:1110–1118.
24. Deville J, Rey J, Chabbert M. An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. J Mol Evol 2009;68:475–489.
25. Pele J, Abdi H, Moreau M, Thybert D, Chabbert M. Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors. PloS one 2011;6:e19094.
26. Lio P, Vannucci M. Investigating the evolution and structure of chemokine receptors. Gene 2003;317:29–37.
27. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44:7156–7165.
28. Sealfon SC, Chi L, Ebersole BJ, Rodic V, Zhang D, Ballesteros JA, Weinstein H. Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor. J Biol Chem 1995;270:16683–16688.
29. Nicholas KB, Nicholas HBJ, Deerfield DWH. GeneDoc: analysis and Visualization of Genetic Variation. EMBNEWNEWS 1997;4:14.
30. Shannon CE. A mathematical theory of communication. Bell Syst Techn J 1948;27:379–423.
31. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17:164–178.
32. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
33. Saraf MC, Moore GL, Maranas CD. Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng 2003;16:397–406.
34. Vicatos S, Reddy BV, Kaznessis Y. Prediction of distant residue contacts with the use of evolutionary information. Proteins 2005;58:935–949.
35. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary covariance in multiple sequence alignments. Bioinformatics (Oxford, England) 2004;20:1565–1572.
36. Cochran WG. Some methods for strengthening the common $\chi 2$ tests. Biometrics 1954;10:417–451.
37. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004;279:19046–19050.
38. Kullback S. Information theory and statistics, New York: Wiley; 1959.
39. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180.
40. McLachlan AD. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. J Mol Biol 1971;61:409–424.
41. Miyata T, Miyazawa S, Yasunaga T. Two types of amino acid substitutions in protein evolution. J Mol Evol 1979;12:219–236.
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504.
43. Horner DS, Pirovano W, Pesole G. Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 2008;9:46–56.
44. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB. 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.
45. Zozulya S, Echeverri F, Nguyen T. The human olfactory receptor repertoire. Genome Biol 2001;2:RESEARCH0018.
46. Nomiyama H, Osada N, Yoshie O. A family tree of vertebrate chemokine receptors for a unified nomenclature. Dev Comp Immunol 2011;35:705–715.
47. Abdi H. Z-scores. In: Salkind NJ, editor. Encyclopedia of measurement and statistics. Thousand Oaks (CA): Sage; 2007. pp 1057–1058.
48. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity determining residues. Genome Biol 2002;3:PREPRINT0002.
49. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D. Co-evolving residues in membrane proteins. Bioinformatics (Oxford, England) 2007;23:3312–3319.
50. Jeon J, Nam HJ, Choi YS, Yang JS, Hwang J, Kim S. Molecular evolution of protein conformational changes revealed by a network of evolutionarily coupled residues. Mol Biol Evol 2011;28:2675–2685.
51. Lee BC, Park K, Kim D. Analysis of the residue-residue coevolution network and the functionally important residues in proteins. Proteins 2008;72:863–872.
52. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics (Oxford, England) 2009;25:1125–1131.
53. Gao H, Dou Y, Yang J, Wang J. New methods to measure residues coevolution in proteins. BMC Bioinformatics 2011;12:206.
54. Merkl R, Zwick M. H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. BMC Bioinformatics 2008;9:151.
55. Dietrich S, Borst N, Schlee S, Schneider D, Janda JO, Sterner R, Merkl R. Experimental assessment of the importance of amino acid positions identified by an entropy-based correlation analysis of multiple-sequence alignments. Biochemistry 2012;51:5633–5641.
56. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;63:832–845.
57. Katritch V, Cherezov V, Stevens RC. Diversity and modularity of G protein-coupled receptor structures. Trends Pharmacol Sci 2012;33:17–27.
58. Rosenkilde MM, Benned-Jensen T, Frimurer TM, Schwartz TW. The minor binding pocket: a major player in 7TM receptor activation. Trends Pharmacol Sci 2010;31:567–574.
59. Chakrabarti S, Panchenko AR. Structural and functional roles of coevolved sites in proteins. PloS one 2010;5:e8591.
60. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature 2013;494:185–194.
61. Mackay CR. Chemokines: immunology's high impact factors. Nat Immunol 2001;2:95–101.
62. Peng Q, Li K, Sacks SH, Zhou W. The role of anaphylatoxins C3a and C5a in regulating innate and adaptive immune responses. Inflamm Allergy Drug Targets 2009;8:236–246.
63. Yokomizo T, Izumi T, Chang K, Takuwa Y, Shimizu T. A G-protein-coupled receptor for leukotriene B4 that mediates chemotaxis. Nature 1997;387:620–624.
64. Nakamura M, Honda Z, Izumi T, Sakanaka C, Mutoh H, Minami M, Bito H, Seyama Y, Matsumoto T, Noma M, Shimuzi T. Molecular cloning and expression of platelet-activating factor receptor from human leukocytes. J Biological Chem 1991;266:20400–20405.

65. Segain JP, Rolli-Derkinderen M, Gervois N, Raingeard de la Bletiere D, Loirand G, Pacaud P. Urotensin II is a new chemotactic factor for UT receptor-expressing monocytes. J Immunol 2007;179:901–909.

66. Bajoghli B. Evolution and function of chemokine receptors in the immune system of lower vertebrates. Eur J Immunol 2013;43:1686–1692.

67. Park K, Kim D. Structure-based rebuilding of coevolutionary information reveals functional modules in rhodopsin structure. Biochim Biophys Acta 2012;1824:1484–1489.

68. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nat Rev Genet 2010;11:572–582.

69. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet 2007;8:675–688.

70. Wright S. The role of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the Sixth International Congress of Genetics, Ithaca (NY) Vol. 1. 1932. pp 356–366.

71. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature 2012; 490:535–538.

72. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14:249–261.