Advanced Techniques in Biology & Medicine

Pelé et al., Adv Tech Biol Med 2017, 5:4
DOI: 10.4172/2379-1764.1000250

Mini Review Open Access

Co-Variation Approaches to the Evolution of Protein Families

Julien Pelé[1], Bruck Taddese[1], Madeline Deniaud[1], Antoine Garnier[1], Daniel Henrion[1], Hervé Abdi[2] and Marie Chabbert[1]*

[1]UMR CNRS 6015 – INSERM 1083, MitoVasc Laboratory, University of Angers, 49100 Angers, France
[2]The University of Texas at Dallas, School of Behavioral and Brain Sciences. TX 75080-3021, USA

Abstract

In a multiple sequence alignment, sequence co-variations result from structural, functional, and/or phylogenetic constraints. Numerous methods have been developed to calculate co-variation scores, but few studies have compared these methods to identify which methods are best suited for the analysis of protein family divergence. Here, we give an overview of widely used methods and identify simple rules for selection of appropriate methods. Specifically, we found that methods such as OMES and ELSC, which favor pairs with intermediate entropy and co-variation networks with hub structure, are well suited to reveal evolutionary information on family divergence. When applied to G protein-coupled receptors, these methods support an epistasis model of protein evolution in which, after a key mutation, co-evolution of several residues was necessary to restore and/or shift protein function.

Keyword: Sequence co-variation; Protein evolution; Epistasis; GPCR

Introduction

Co-variation in amino acid distribution of any two positions in the multiple sequence alignment (MSA) of a protein family is widely used to obtain structural and/or functional information [1]. It is assumed that, after mutation of one residue, structural/functional constraints can require a compensatory mutation of another residue to maintain the structure or to restore the function of the corresponding protein. Such compensatory mutations can depend on direct physical interactions between two residues or on an indirect interaction through intermediary residues or a ligand [2].

The analysis of sequence co-variation in the MSA of a protein family is not straightforward because the sequences are not independent but phylogenetically related. In the MSA of a protein family containing several sub-families, sub-family independent co-varying positions are assumed to correspond to structurally important compensatory mutations. Sub-family dependent co-varying positions may arise from either compensatory mutations within one sub-family or independent mutations in several sub-families. Both of these mechanisms lead to a phylogenetic bias.

Until now, the goal of most co-variation studies has been to predict contacts in order to gain structural information from co-evolving positions. To achieve this goal, different methods attempt to remove or reduce the phylogenetic bias [3-5] and to differentiate direct coupling from indirect coupling which is created by the transitivity effect of pair co-variation [2]. To this effect, Bayesian network [2], maximum entropy [6] and neural network [7] models markedly improved the prediction of interacting pairs, with successful *de novo* 3D protein structure prediction. Recently, co-evolution based methods coupled to machine learning have significantly improved contact prediction and constitute an area of intense research for *de novo* structure prediction [8]. The analysis of co-varying residues in an MSA has also been used to identify specificity-determining residues [9-12], protein sectors [13] or connectivity pathways [14-16].

Most comparative studies of co-variation methods [3,4,17] aim to find methods that optimize the number of contact pairs to gain structural information. Co-variation approaches that aim to gain information on protein divergence have received less attention. However, these approaches can be very useful to gain information on protein evolution and sub-family specificity. We thus undertook a comparative analysis of widely used co-variation methods on a model system to determine methods best suited to this aim [18].

Characteristics of the model system

To carry out this analysis, we chose the large family of rhodopsin-like G protein coupled receptors as a model system [18]. The human non-olfactory receptor set includes 283 receptors that can be classified into a dozen of sub-families [19] corresponding to three main evolutionary pathways [20]. The median sequence identity of human receptors varies from 20% in the full set to about 30% within sub-families (transmembrane domain only). We compared the different methods on three nested sequence sets. The sets correspond to (1) human non olfactory receptors, (2) receptors characterized by the P2.58 proline pattern (107 members, with 26% sequence identity). The latter receptors diverged after an indel in helix 2 which is one of the main evolutionary pathways of GPCRs [21] and (3) the chemokine receptors (23 members with 37% sequence identity). The characteristics of these three sets are frequently found in a variety of protein families.

Sequence co-variation methods

Most methods that measure co-variation in the amino acid distribution at any two positions in an MSA can be classified into four main classes:

(1) Methods based on the χ^2 test, such as OMES (Observed minus Expected Squared) [17]. The OMES score is computed as:

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

where $N\ (i,\ j)$ is the number of sequences in the alignment with non-gapped residues at positions i and j, $N^{ex}_{x,y}(i,j)$ and $N^{ex}_{x,y}(i,j)$ are the number of times each pair of amino acids (x,y) is, respectively, observed and expected at positions (i,j). Expectation is based on the frequency of amino acids (x,y) at positions (i,j).

***Corresponding author:** Marie Chabbert, MitoVasc laboratory, 3 rue Roger Amsler, 49100 Angers, France, Tel: (33) 2 44 68 82 62; E-mail: marie.chabbert@univ-angers.fr

Received November 18, 2017; **Accepted** November 28, 2017; **Published** November 30, 2017

Citation: Pelé J, Taddese B, Deniaud M, Garnier A, Henrion D, et al. (2017) Co-Variation Approaches to the Evolution of Protein Families. Adv Tech Biol Med 5: 250. doi: 10.4172/2379-1764.1000250

(2) Methods based on mutual information (MI), the probability of joint occurrence of events [22]. The MI score is given by:

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

where, $p_x(i)$, $p_y(i)$, and $p_{x,y}(i,j)$ are the frequencies of amino acids x at position i, y at position j and of the pair (x,y) at positions (i,j). As mutual information formula favors pairs with high entropy, several corrections have been applied to correct this bias. The MIp score (mutual information product) outperforms other MI based methods [3] and is computed as:

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

where $MI(i,\bar{j})$, $MI(i,\bar{j})$ and $<MI>$ are the average of $MI(i,j)$ on, respectively, i, j and both i and j.

(3) Methods based on substitution matrices such as McBASC (McLachlan Based Substitution Correlation method) [23]. The McBASC score is computed as:

$$McBASC(i,j)=\frac{1}{N^2\sigma(i)\sigma(j)}\sum_{k,l}(S_{k,l}(i)-S(i))(S_{k,l}(j)-S(j))$$

where $S_{k,l}(j)$ and $S_{k,l}(j)$ are the similarity scores based on the McLachlan matrix [24] for the amino acid pair present in sequences k and l at positions i and j, respectively, $S(i)$ and $S(j)$ are, respectively, the averages of all the scores $S_{k,l}(j)$ and $S_{k,l}(j)$ and $\sigma(i)$ and $\sigma(j)$ are, respectively, the standard deviations of all the scores $S_{k,l}(i)$ and $S_{k,l}(j)$.

(4) Methods using perturbation of an *MSA*, such as the statistical coupling analysis (SCA) [16] and Explicit Likelihood of Subset Co-variation (ELSC) [25] methods. These latter two methods compare amino acid composition in a subset to the composition in the entire alignment. The SCA score is given by:

$$SCA(i,j)=\sqrt{\sum_y(\ln P_y(j\,|\,\delta_i)-\ln P_y(j))^2}$$

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 most prevalent amino acid x at position i.

The ELSC score measures how many possible subsets of size n would have the composition found in column j. It is computed as:

$$ELSC(i,j)=-\ln\prod_y\frac{\binom{N_y(j)}{n_y(j)}}{\binom{N_y(j)}{m_y(j)}}$$

Where $\binom{}{}$ is the binomial coefficient

and $N_y(j)$, $n_y(j)$ 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).

Entropy biases

To compare the different methods [18], we analyzed co-variation scores as a function of sequence entropy which is a measure of variability based on the Shannon entropy [26] and is given by:

$$S(i)=-\sum_x p_x(i)\log_{20}p_x(i)$$

where, i is the position of interest in the MSA, x represents the 20 amino acids and $p_x(i)$ represents the frequency of amino acid x at the i^{th} position. To do so, we used bi-dimensional plots. In these plots, each dot represents a pair of positions (i,j) in the MSA and is located at the position of the entropies of i and j, with a color code indicating the co-variation score. The bi-dimensional plots obtained with Set 1 (about 300 sequences) are shown in Figure 1. Similar results were obtained with the other data sets. The MI and SCA methods have

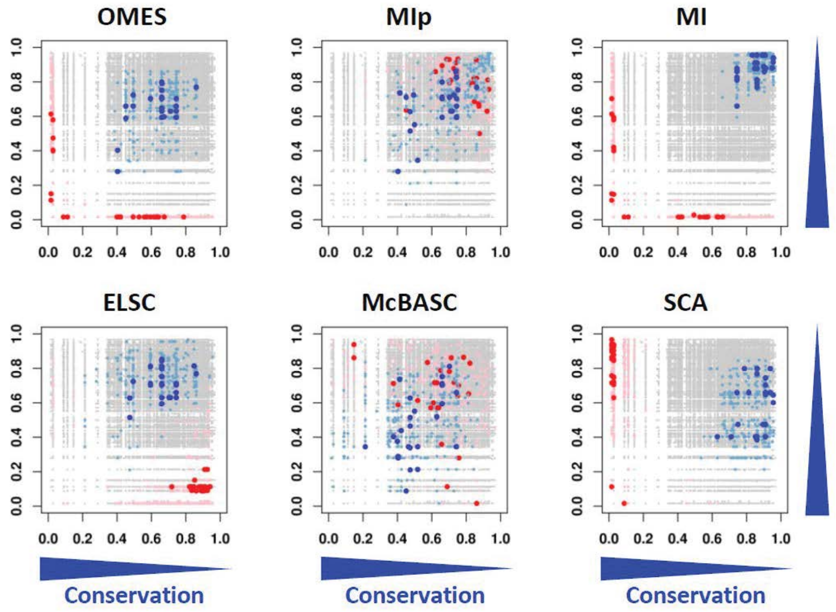

Figure 1: Bi-dimensional plots of the co-variation scores as a function of sequence entropy. Each dot represents a pair of positions (*i*,*j*) in the MSA. The dot position corresponds to the entropy of *i* (*x* axis) and *j* (*y* axis) respectively. The color code indicates the co-variation score. Dark and light blue dots indicate the top 25 and 250 pairs respectively. Red and pink dots indicate the bottom 25 and 250 pairs respectively. The analysis was carried out on the transmembrane domain of human non-olfactory class A GPCRs (283 sequences), adapted from [18].

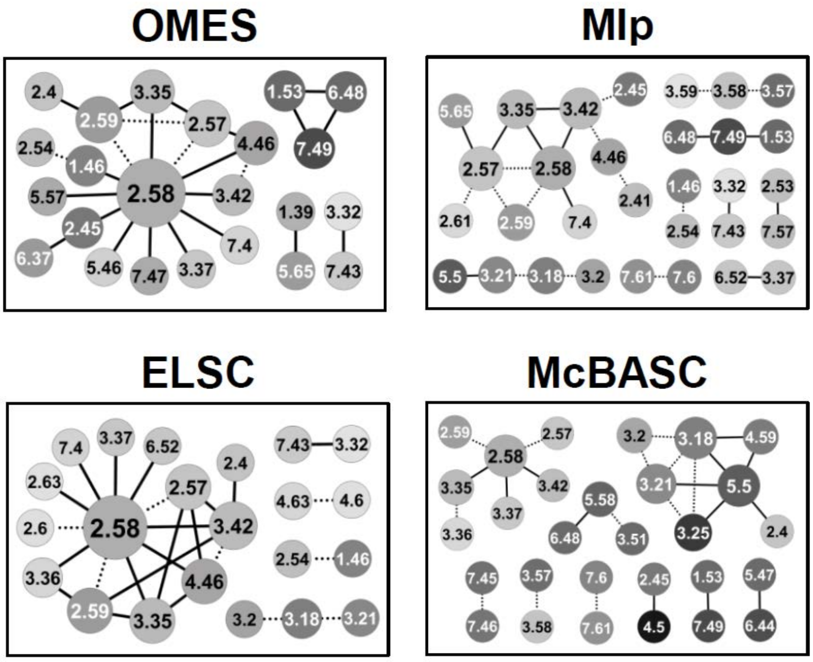

Figure 2: Network representation of the top 25 pairs obtained from co-variation analysis of human non-olfactory class A GPCRs (transmembrane domain only). Positions correspond to nodes and co-variation signals to edges. The size of the nodes is proportional to their connectivity. The color indicates the entropy of the position from black for a fully conserved position to white for the most variable position. Node labels indicate the position with Ballesteros' numbering [29]. Dotted edges indicate distance below 8 Å. Adapted from [18].

strong bias towards pairs with at least one highly variable position (blue dots on right top corner or along right side of the plot) and should thus be considered with caution. The other four methods favor pairs with intermediary conservation (blue dots in the center of the plots). However, with OMES and ELSC, the entropy of the top pairs is clearly separated from the entropy of the bottom pairs (red dots), which is not the case for the MIp and McBASC methods with more fuzzy plots. The plots in Figure 1 clearly indicate that each method favor pairs with different levels of sequence conservation. This entropy bias leads to different results for each method. Consequently, this bias must be taken into account and carefully analyzed upon selection of a method (to do so, we developed the R package *bios2cor* which includes a function for this bi-dimensional analysis [27]).

Networking biases

The network representation of the top pairs obtained with the OMES, ELSC, MIp and McBASC methods for Set 1 (Figure 2) highlights another striking difference between methods. MIp and McBASC methods favor pairs with low connectivity whereas the networks obtained with OMES and ELSC have a hub structure with a highly connected central residue. Most importantly, the central residue (P2.58) obtained with the OMES and ELSC methods is an important determinant of GPCR evolution [20,21]. This networking structure is observed for the three nested sequence sets that we have analyzed [18]. In each case, the central residue corresponds to a residue known to be crucial for the evolution of the GPCR family and the sub-family specificity (divergence of purinergic receptors in the set of P2.58 receptors and of inflammatory receptors in the chemokine receptor set). The hub structure observed for GPCRs strongly supports an epistasis model of protein evolution in which after mutation of a key residue, co-evolution of several residues was necessary to restore/shift protein function.

Selection of a co-variation method

The performance of co-variation analysis to answer specific questions crucially depends on (1) the co-variation method used and (2) the characteristics of the MSA (e.g., number of sequences, homogeneity and conservation) [3,4,17]. The first step in a co-variation analysis is thus to determine the specific questions of interest and then, to select a suited co-variation method and to prepare a sequence set accordingly.

When co-variation analysis aims to gain structural information, the methods favoring pairs with low connectivity are suited because these pairs are enriched in contact pairs [3,4]. This is the case for McBASC and MIp (Figure 2). This latter method was specifically developed with this aim. These methods require large sequence sets, with at least 100 sequences, as homogeneous as possible [3,17]. This last requirement should prompt to build MSA of orthologs whenever possible.

When co-variation analysis aims to gain evolutionary information, OMES and ELSC are well suited to identify the co-evolving residues that contributed to the divergence within a protein family. These methods favor pairs with intermediate conservation and highly connected residues and can work with a very small number of sequences (<25). They require a sequence set with two subsets of similar size and sequence similarity. Size requirement can be fulfilled by tailoring sequence sets with judicious use of paralogs and orthologs [18]. Homogeneity of the subsets is mandatory to avoid bias toward an overwhelming or a highly conserved sub-family.

Conclusion

The phylogenetic bias, related to the intrinsic inhomogeneity of a sequence set containing sub-families, represents a rich source of information on the mechanisms that drove the evolution of a protein family. Two co-variation methods, OMES and ELSC, are adequate to mine evolutionary hubs, in relation with an epistasis-based mechanism of functional divergence within a protein family [28].

Software availability

The R package *bios2cor* is freely accessible at the Comprehensive R Archive Network (http://cran.r-project.org). It includes R-coded OMES, ELSC, MIp and McBASC functions as well as representation tools, such as the 2D entropy-score plots.

Acknowledgment

This study was supported by the grant ANR-11-BSV2-026 from the French Research Agency. JP and BT had a fellowship from Conseil Général de Maine-et-Loire and ANR, respectively.

References

1. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. Nat Rev Genet 14: 249-261.
2. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633.
3. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333-340.
4. Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21: 4116-4124.
5. Tillier ER, Lui TW (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19: 750-755.
6. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766.
7. Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31: 999-1006.
8. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ (2017) Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins.
9. Bachega JF, Navarro MV, Bleicher L, Bortoleto-Bugs RK, Dive D, et al. (2009) Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. Proteins 77: 26-37.
10. Bleicher L, Lemke N, Garratt RC (2011) Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. PLoS One 6: e27786.
11. Chakrabarti S, Panchenko AR (2009) Coevolution in defining the functional specificity. Proteins 75: 231-240.
12. Mirny LA, Gelfand MS (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol 321: 7-20.
13. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774-786.
14. Dima RI, Thirumalai D (2006) Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Sci 15: 258-268.
15. Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 48: 611-617.
16. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295-299.
17. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56: 211-221.
18. Pele J, Moreau M, Abdi H, Rodien P, Castel H, et al (2014) Comparative analysis of sequence co-variation methods to mine evolutionary hubs: Examples from selected GPCR families. Proteins 82: 2141-2156.
19. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups and fingerprints. Mol Pharmacol 63: 1256-1272.
20. Pele J, Abdi H, Moreau M, Thybert D, Chabbert M (2011) Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors. PloS one 6: e19094.
21. Deville J, Rey J, 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: 475-489.
22. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 90: 7176-7180.
23. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309-317.
24. McLachlan AD (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. J Mol Biol 61: 409-424.
25. Dekker JP, Fodor A, Aldrich RW, Yellen G (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 20: 1565-1572.
26. Shannon CE (1948) A mathematical theory of communication. Bell system technical journal 27: 379-423.
27. Taddese B, Garnier A, Deniaud M, Pelé J, Bellenger L, et al. (2017) Bios2cor: From Biological Sequences and Simulations to Correlation Analysis. The Comprehensive R Archive Network.
28. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. Nature 490: 535-538.
29. Sealfon SC, Chi L, Ebersole BJ, Rodic V, Zhang D, et al. (1995) Related contribution of specific helix 2 and 7 residues to conformational activation of the serotonin 5-HT2A receptor. J Biol Chem 270: 16683-16688.