Bios2cor: an R package integrating dynamic and evolutionary correlations to identify functionally important residues in proteins

Bruck Taddese, Antoine Garnier, Madeline Deniaud, Daniel Henrion, Marie Chabbert

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural Bioinformatics

Bios2cor: an R package integrating dynamic and evolutionary correlations to identify functionally important residues in proteins

Bruck Taddese[1], Antoine Garnier[1], Madeline Deniaud[1], Daniel Henrion[1] and Marie Chabbert[1*]

[1] CNRS UMR 6015 – INSERM 1083, MITOVASC Laboratory, 3 rue Roger Amsler, 49100 ANGERS, FRANCE.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Both dynamic correlations in protein sidechain motions during molecular dynamics (MD) simulations and evolutionary correlations in multiple sequence alignments (MSA) of homologous proteins may reveal functionally important residues. We developed the R package Bios2cor that provides a unique framework to investigate and, possibly, integrate both analyses. Bios2cor starts with an MSA or a MD trajectory and computes correlation/covariation scores between positions in the MSA or between sidechain dihedral angles or rotamers in the MD trajectory. In addition, Bios2cor provides a variety of tools for the analysis, the visualization and the interpretation of the data.
Availability: The R package Bios2cor is available from the Comprehensive R Archive Network, at http://cran.r-project.org/ web/packages/Bios2cor/index.html .
Contact: marie.chabbert@univ-angers.fr

1 Introduction

Covariations in multiple sequence alignments (MSA) of homologous proteins are widely used to gain structural and functional insights on protein families from evolutionary information (de Juan, et al., 2013; Pele, et al., 2014). Dynamic correlations between sidechain motions during molecular dynamics (MD) simulations did not receive the same attention yet. However, both dynamic correlations in MD trajectories and evolutionary covariations in MSAs may reveal functionally important residues and mechanisms. The dynamic structure of proteins is fundamental for function and reflects evolutionary history (Marsh and Teichmann, 2014). Combining both analysis helps elucidate the molecular mechanisms of protein functions (Lakhani, et al., 2017; Mishra, et al., 2019; Stacklies, et al., 2009).

We developed the R package `Bios2cor` to provide a unique framework for the investigation of dynamic and evolutionary correlations. After data organization into matrices (Fig. 1), both correlations are formally similar. Features include (1) the computation of correlation/covariation scores between positions in MSAs or between dihedral angles or rotamers in MD trajectories using different scoring functions and (2) the analysis of the correlation/covariation matrix through network representation and principal components analysis. In addition, several utility functions based on the R graphical environment provide friendly tools for help in data interpretation.

2 Available functionalities

The `Bios2cor` package provides a complete environment for correlation/covariation analysis in the context of either MSA or MD simulations (Fig. 1). Here, we present the main functionalities.

Data import: MSA in fasta and msf formats are read with the `import.fasta` and `import.msf` functions. MD trajectories in dcd formats are read with the `read.dcd` function from the `bio3d` package (Grant, et al., 2006). The `dynamic_structure` function creates the data structure for sidechain dihedral angles with `bio3d`.

Computing scoring matrices: Users can build different scoring matrices. For sequences, scoring methods include `omes` (Observed Minus Expected Squared) (Fodor and Aldrich, 2004), `elsc` (Explicit Likelihood of Subset Covariation) (Dekker, et al., 2004), `mi` (Mutual Information) (Atchley, et al., 2000), `mip` (Mutual Information Product) (Dunn, et al., 2008) and `mcbasc` (McLachlan Based Substitution Corre-

lation) (Casari, et al., 1995). For simulations, correlations between dihedral angles are computed with `dynamic_circular` that uses the `cor.circular` function from the `circular` package (Agostinelli and Lund, 2017) to calculate a circular version of Pearson correlation (Jammalamadaka and SenGupta, 2001). For rotamers, `angle2rotamer` transforms dihedral angles to rotamers using the dynameomics library (Towse, et al., 2016), then covariation scores can be calculated with `dynamic_omes`, `dynamic_mi` or `dynamic_mip`. Boxplots can be visualized with the `scores.boxplot` function.

Analysis of data variability: The sequence variability at each position in the MSA is computed with the `entropy` function based on Shannon entropy (Shannon, 1948). For trajectories, the variability of each dihedral angle (`dynamic_entropy`) is estimated from the number of rotameric transitions during the simulations. The `scores_entropy.plot` function gives 2D plots of scores as a function of variability. A selection of elements based on variability can be obtained with `delta_filter`.

Principal component analysis (PCA): The `centered_pca` function performs PCA of a scoring matrix after double centering and provides coordinates of the elements on the principal components.

Network representation: `network.plot` uses `igraph` (Csardi and Nepusz, 2006) to provide a network representation of the top *n* scoring pairs determined by `top_pairs_analysis`. Files in csv format can be exported for network representation with Cytoscape (Su, et al., 2014).

Dihedral angle visualization: `angles.plot` creates plots of the evolution of the dihedral angles in the top *n* scoring pairs.

Data export: Scores and entropy values are exported with `write.scores` and `write.entropy`. `write.pca` creates a csv file of coordinates in PCA space. The `write.pca.pdb` function creates pdb and pml files for nice visualization of three selected components with the PyMol program (www.pymol.org).

3 Examples

Examples using the `Bios2cor` package for the analysis of MSAs (Taddese, et al., 2018) and MD trajectories (Taddese, et al., 2020) have been published recently. The integration of these two studies reveals that a hallmark residue in the evolution of chemokine receptors plays a pivotal role in the activation mechanism of CXCR4.

4 Conclusions

The R package `Bios2cor` provides a powerful and flexible framework for computation and subsequent analysis of correlation/covariation in MSAs and MD simulations. `Bios2cor` can help decipher complex conformational transition in link with protein evolution (Taddese, et al., 2020). Coupling dynamic and evolutionary correlations will contribute to elucidate protein molecular functionalities.

Acknowledgements

We thank J Pelé and J-M Bécu for their contribution to the `Bios2cor` package.

Funding

This work was supported by grants from The French National Research Agency (ANR-11-BSV2-026 to MC) and from GENCI (100567 to MC) and by institutional grants from CNRS, INSERM and the University of Angers.

Conflict of Interest: none declared.

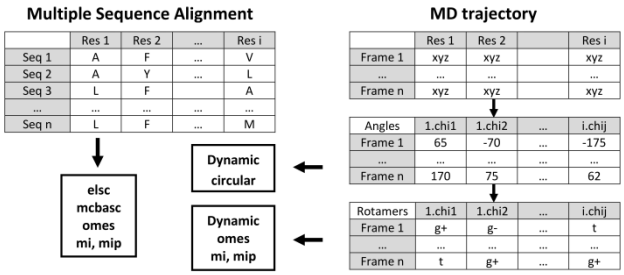

Fig. 1. Schematic representation of the correlation/covariation analyses provided by the Bios2cor package.

References

Agostinelli, C. and Lund, U. (2017) R package 'circular': Circular Statistics (version 0.4-93).

Atchley, W.R., *et al.* (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis, *Molecular biology and evolution*, **17**, 164-178.

Casari, G., Sander, C. and Valencia, A. (1995) A method to predict functional residues in proteins, *Nature structural biology*, **2**, 171-178.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research, *InterJournal Complex System*, **1695**.

de Juan, D., Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution, *Nature reviews. Genetics*, **14**, 249-261.

Dekker, J.P., *et al.* (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments, *Bioinformatics*, **20**, 1565-1572.

Dunn, S.D., Wahl, L.M. and Gloor, G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, *Bioinformatics*, **24**, 333-340.

Fodor, A.A. and Aldrich, R.W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments, *Proteins*, **56**, 211-221.

Grant, B.J., *et al.* (2006) Bio3d: an R package for the comparative analysis of protein structures, *Bioinformatics*, **22**, 2695-2696.

Jammalamadaka, S.R. and SenGupta, A. (2001) *Topics in circular Statistics*. World Scientific Co. Pte. Ltd Singapore.

Lakhani, B., *et al.* (2017) Evolutionary Covariance Combined with Molecular Dynamics Predicts a Framework for Allostery in the MutS DNA Mismatch Repair Protein, *The journal of physical chemistry. B*, **121**, 2049-2061.

Marsh, J.A. and Teichmann, S.A. (2014) Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure, *BioEssays : news and reviews in molecular, cellular and developmental biology*, **36**, 209-218.

Mishra, S.K., Kandoi, G. and Jernigan, R.L. (2019) Coupling dynamics and evolutionary information with structure to identify protein regulatory and functional binding sites, *Proteins*, **87**, 850-868.

Pele, J., *et al.* (2014) Comparative analysis of sequence covariation methods to mine evolutionary hubs: examples from selected GPCR families, *Proteins*, **82**, 2141-2156.

Shannon, C.E. (1948) A mathematical theory of communication, *Bell system technical journal*, **27**, 379-423.

Stacklies, W., *et al.* (2009) Mechanical network in titin immunoglobulin from force distribution analysis, *PLoS computational biology*, **5**, e1000306.

Su, G., *et al.* (2014) Biological network exploration with Cytoscape 3, *Current protocols in bioinformatics*, **47**, 8 13 11-24.

Taddese, B., *et al.* (2018) Evolution of chemokine receptors is driven by mutations in the sodium binding site, *PLoS computational biology*, **14**, e1006209.

Taddese, B., *et al.* (2020) Deciphering collaborative sidechain motions in proteins during molecular dynamics simulations, *Scientific reports*, **10**, 15901.

Towse, C.L., *et al.* (2016) New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities, *Structure*, **24**, 187-199.