Comparing residue-coevolution networks across protein families

Franco L Simonetti*, Cristina Marino-Buslje
Fundación Instituto Leloir, Av Patricias Argentinas 435, Buenos Aires, C1405BWE, Argentina
*To whom correspondence should be addressed: fsimonetti@leloir.org.ar

Background

Coevolution networks carry information about structure, function, interactions, phylogeny and stochastic components. Most works point out the usefulness of estimating coevolution to infer protein residue contacts or protein-protein interactions. Despite some structural constraints in a given protein are critical for performing a specific function, they are not the only constraints that made up coevolutionary information.

We hypothesize that protein families with similar function and a common evolutionary history have similar coevolution networks. We examined a set of 180 domain family hierarchies from the Conserved Domain Database. A domain family hierarchy is a set of related Conserved Domains (CDs) that share a common ancestor, a common set of conserved residues, and a common general function, but differ from each other in their specific functions.

Results

Covariation scores for each CD were calculated using corrected Mutual Information, mfDCA and plmDCA. For a given hierarchy, we compared pairwise the covariation networks of all conserved domains in a hierarchy. Each pair of CDs compared could be categorized as “Highly Related”, “Moderately Related” or “Poorly Related”. To map the alignments between two CDs, PDB structures (contained in the alignments) from each CD were aligned using Tmalign and the best structural alignment was selected.

Graph edit distance was used as a similarity metric between two covariation networks. Given two graphs, the graph edit distance (GED) is defined as the sequence of edit operations that transforms one graph to another and that has minimal editing cost. A GED of 0 indicates identical graph topologies, while a GED of 1 indicates completely disjoint graphs.

Highly related Conserved Domains show the lowest GED scores, with a mean GED score of 0.83±0.08, followed by Moderately related CDs (0.89±0.05). Poorly related CDs display a mean GED of 0.92±0.04. All distributions are significantly different (corrected multiple Kruskal-Wallis comparisons, $p < 0.001$). GED score predictive value can be evaluated by testing its power to separate defined classes. For distinguishing Highly Related and Moderately Related domains from Poorly Related ones, the GED score obtains an area under the ROC curve that ranges from 0.69-0.78 (depending on the covariation method used). Specifically in the case of Moderately related domains, GED score is a better overall score than comparing their contact networks derived from PDB structures.

Conclusions

Using a metric that only describes the topological similarity between covariation networks we were able to show that, within a set of related families with similar functions, those closer in the hierarchy (thus with more specific and similar functions) retain similar covariation networks. This may be the foundation to develop methods adding this information aiming at classifying protein families with a more functional meaning.