Using correlation data and network decomposition to obtain sub-class determinants in protein families

Lucas Bleicher, Ney Lemke, Richard Charles Garratt

Introduction

The observation of patterns of correlation at certain positions in a protein family multiple sequence alignment has been described at least since the eighties. It gained considerable attention in the nineties, when a variety of correlation metrics were proposed, mostly to search for contact pairs (Gobel, Sander, Schneider et al., 1994). The search for sets of co-evolving position has also been discussed since the late 90's. Atchley, Terhalle and Dress (1999) used a procedure to find "cliques" of co-evolving positions, while Lockless and Ranganathan (1999) presented two energy-like parameters to quantify positional conservation and inter-positional correlation. In recent work (Bachega, Navarro, Bleicher et al., 2009), we observed that correlated positions could cluster into different groups, related to different properties in a protein family, the Fe/Mn superoxide dismutases (SODs). Our findings correlated well to positions already described as determinants of oligomeric state and metal selectivity in this family (Wintjens, Gilis and Rooman, 2008). This family is used to illustrate a simple method we propose which exploits and quantifies specific correlations in order to detect sub-class determinants.

Methods

Given an alignment of N sequences having nA sequences with a given amino acid x at position i and nB sequences with a given amino y acid at position j, we want to test whether the presence of x in position i has any correlation to the presence of y in j. If they were uncorrelated, the observed frequency of y in j for the subset of sequences having amino acid x in position i would still be nB/N, which is our null hypothesis. If it is not, we measure the corresponding p-value using the cumulative binomial distribution cb(N,n,f) as described:

\[ p = \sum_{n=n_B}^{n_A} \frac{n_A!}{n!(n_A-n)!} \left( \frac{n_B}{N} \right)^n \left( 1 - \frac{n_B}{N} \right)^{n_A-n} \]

If, conversely, the observed number of residues y in position j is less than expected, we use the opposite tail of the cumulative binomial distribution to measure the probability of having no more than this observed value, i.e.:

\[ p = \sum_{n=0}^{n_B} \frac{n_A!}{n!(n_A-n)!} \left( \frac{n_B}{N} \right)^n \left( 1 - \frac{n_B}{N} \right)^{n_A-n} \]

Therefore, we can use -log(p) as a measure of the correlation between two positions. Using log(p) instead of -log(p) in the second case, it is possible to denote anti-correlation.

Networks of residue correlations are built using individual combinations of residue type plus position as vertices, and -logP as the edge weight between two nodes. Pairs presenting anti-correlation have reversed sign. The network can then be submitted to a community detection procedure.

We can also define an adherence measure to quantify the extent to which a given sequence fits into a given community.

\[ Adh(S, A) = \frac{1}{N_A(N_A-1)} \sum_{a_i,a_j \in A} w(a_i, a_j) \delta_S(a_i, a_j) \]

The delta function \( \delta_S(a_i, a_j) \) takes the value 1 if both vertices (i.e., given amino acids in alignment positions) are present in sequence S and 0 otherwise. If the amino acids in community A are related to a given property in the protein family, then high values of Adh(S,A) indicate that sequence S may possess that property, being useful for gene annotation applications.

Results

The resulting decomposed network for Fe/Mn-SODs is shown below:
Community 1 groups six residues which are related to the presence of an active-site manganese, while community 3 groups residues found in SODs which bind iron instead. Since these properties are mostly mutually exclusive, the residues in their communities are linked with negative edges, shown in red. Community 2 groups six residues which are related to dimeric SODs. It should be noted that they do not present either positive or negative links to the other communities, which is compatible with the notion that metal selectivity and oligomeric state are independent properties. Since the Fe/Mn superoxide dismutase family presents a good test set for oligomeric state and metal specificity (Wintjens, Gilis and Rooman, 2008), we can also assess the utility of the adherence parameter $\text{Adh}(S,A)$ as defined above. Histograms are shown below:

For iron binding SODs, virtually all sequences have $\text{Adh}(S,1)$ equal to zero and the maximum value for $\text{Adh}(S,3)$ (Figure 5A). Conversely, all manganese binding SODs have $\text{Adh}(S,3)$ equal to zero, and most have high values for $\text{Adh}(S,1)$. Most cambialistic SODs have low or null values for $\text{Adh}(S,1)$ and $\text{Adh}(S,3)$, suggesting that the lack of residues in communities 1 and 3 are related to non-specificity for manganese or iron. Finally, all tetrameric SODs have $\text{Adh}(S,2)$ equal to zero, while most dimeric SODs show higher values.

**Conclusions**

We present a method based on correlated mutations and network analysis to calculate and analyze groups of amino acids which may be related to functional classes in protein families. Due to nature of the correlation metric and the network decomposition method, the results can be readily interpreted and related to biological features and their inter-relation. We also propose additional parameters and procedures that can be used to further analyze and extract information from the data. We argue that community structure in networks constructed using the described method is an expected feature for protein families presenting functional sub-classes, and therefore could be exploited to identify key residues for specific functional properties. Also, it can be a useful tool for gene annotation, since key residues which are clustered in a community should be more likely to predict function than sequence identity methods, which considers all residues evenly.