Monophyletic clustering and characterization of protein families

Jian Zhang [1,*], **Zhiyuan Zhao** [2], **Jennifer Evershed** [3], **Guoying Li** [2]

[1]Institute of Mathematics, Statistics and Actuarial Science, University of Kent
Canterbury, Kent CT2 7NF, U.K.
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing 100080, China
[3]GlaxoSmithKline, New Frontiers Science Park South, Third Avenue, Harlow, Essex
CM19 5AW, U.K.

Summary

A protein family contains sequences that are evolutionarily related. Generally, this is reflected by sequence similarity. There have been many attempts to organize the set of protein families into evolutionarily homogenous clusters using certain clustering methods. How do we characterize these clusters? How can we cluster protein families using these characterizations? In this work, these questions were addressed by use of a concept called group-wide co-evolution, and was exemplified by some real and simulated protein family data. The results have shown that the trend of a group of monophyletic proteins might be characterized by a normal distribution, while the strength and variability of this trend can be described by the sample mean and variance of the observed correlation coefficients after a suitable transformation. To exploit this property, we have developed a monophyletic clustering method called monophyletic $k-$medoids clustering. A software package written in R has been made available at http://www.kent.ac.uk/ims/personal/jz .

1 Introduction

The rapid advance of genome-wide sequencing of an increasing number of genomes from different species has opened the way for exploiting the phylogenetic information in proteomic studies. It has been previously observed that interacting proteins have more similar phylogenetic trees to what could be expected from a general divergence between the corresponding species under the standard molecular hypothesis [1-4]. An extreme of co-evolution of two interacting proteins would be those cases where both proteins are simultaneously lost in the same species, probably because of the strong functional dependence between these two proteins [5]. Intuitively, if two proteins are functionally coordinative across an evolutionary history, they have probably experienced similar selective pressures during evolution. This provides a rationale for testing functional correlation (interaction) between two protein families via their trends of evolution. Based on this rationale, a correlation-based statistical strategy was designed by Pazos and Valencia [1] and Goh et al.[2] for the prediction of protein-protein interactions. In their method, as a measure of the tree similarity, the correlation coefficient was calculated between the distance matrices used to build the trees. Obviously, the observed correlation coefficient is expected to be close to 1 if two proteins have nearly the same evolutionary trend. On the other hand, if two proteins are unrelated, then the observed correlation coefficient should be near 0. However, two correlated proteins do not necessarily interact because the speciation (i.e. these two proteins come from the same species) and the so-called inter-molecular interaction can also contribute a certain amount of correlation [2, 6, 7]. Various other computational methods have been developed for predicting pairwise functional interactions between protein families. See [6] and the references therein. A common problem of these studies has been that there is a lack of description of co-evolution at a group level, although we believe that there are some interaction-networks where sets of proteins are interacting as certain functional units. In this paper, this problem was addressed by use of a new concept termed *group-wide co-evolution* and was

*Corresponding author, j.zhang@kent.ac.uk

exemplified by some real and simulated datasets including eubacteria ribosomal protein families and the protein structural domains. Here the property that a group of protein families have a similar phylogenetic tree will be referred to as the group-wide co-evolution throughout the paper. The similarity between two trees is gauged by the Pearson correlation coefficients between the observed score matrices. See the next section for more details.

The main methodology proposed in this paper is the monophyletic $k-$medoids clustering, where the proteins are partitioned into several evolutionarily homogenous groups, each with a monophyletic characterization. This characterization is of great importance to the identification of interacting protein families. Taking the ribosomal protein families as examples, we considered the protein sequences sampled from 11 selected eubacteria. These proteins can be biologically divided into 20 small subunit families - namely, S1-S20; and 31 large subunit families-namely L1-L7/12,L9-L11,L13-L24,L27-L29,L31-L36, leading to 1275 potential functionally interacting pairs. Since a large number of coordinating ribosomal proteins are required for a ribosome to carry out the task of protein synthesis, we hypothesize that as a unit these protein families are group-widely co-evolving with many interactions in the group. Unfortunately the current incomplete experimental studies of the bacterial ribosome assembly are not sufficient to confirm this hypothesis, owing to the challenges of monitoring the association of many components simultaneously [8]. In particular, only a few interacting pairs have been confirmed by experiments so far, which are (L10,L7/12), (L2,L16), (S6,S18), (S2,S10), (S2,S11), (S5,S8), and (S5,S10). See [1,9-11] and the references therein. Again these observations are not enough to specify the unknown baseline correlation determined by the speciation. Therefore a few questions arise naturally about whether we can confirm the above hypothesis by use of computational methods and about how high the correlation need to be to justify the rejection of the null hypothesis that the two proteins are not interacting. The proposed cluster analysis allows one to identify a group of protein families, of which the correlations (except a few outliers) are mainly determined by speciation. We selected the cut-off value of significant correlations via the quantiles of the correlation coefficients in this baseline group.

We applied our procedures to both the ribosomal protein families and protein structural domain families. An important finding in this application is that the observed Fisher Z-transformed correlation coefficients could be well fitted by some normal distributions. We hypothesized that this group-wide phenomenon did not hold by coincidence. To confirm this hypothesis, we simulated an ideal scenario with four sets of protein families sampled from four different phylogenetic trees respectively. These trees show varying degrees of evolutionary change. We examined the distribution patterns of the Fisher Z-transformed pairwise correlation coefficients for these families. As we expected, the observed correlation coefficient values for the four datasets did follow normal distributions with different means and variances respectively. Note that this fact is no longer true if we merge any two or more of these datasets together. Therefore, the normality of the Fisher Z-transformed correlation coefficients was adopted as a criterion for identifying the group-wide co-evolution in a set of protein families.

The paper is organised as follows. The datasets and the methodology used in this paper are described in Section 2. Section 3 presents the results for both simulated and real datasets. Some discussions and conclusions are made in Section 4.

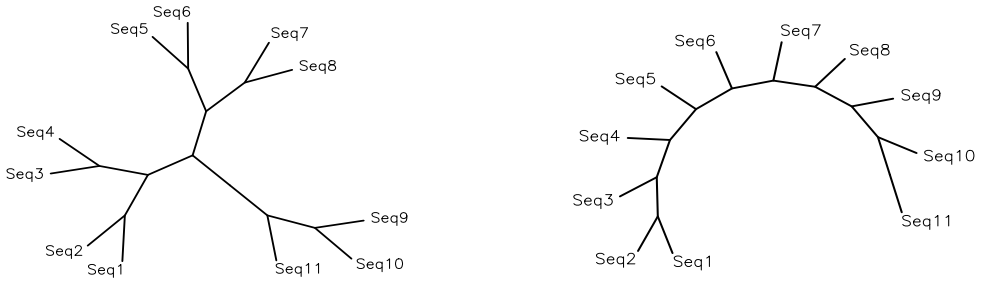
2 Materials and Methods

2.1 Data Sets

Simulated data. Consider the following model trees in the Newick format:

Tree 1:

((((Seq1:0.03,Seq2:0.42):0.03, (Seq3:0.03,Seq4:0.42):0.42):0.015, ((Seq5:0.03,Seq6:0.42):0.03,



(a) Tree 1 (b) Tree 2

Figure 1: The draws of the phylogenetic trees 1 and 2.

(Seq7:0.03,Seq8:0.42):0.42):0.42):0.02, ((Seq9:0.03,Seq10:0.42):0.03, Seq11:0.42):0.02);

Tree 2:

((((((((((Seq1:0.04,Seq2:0.04):0.01, Seq3:0.05):0.01, Seq4:0.06):0.01, Seq5:0.07):0.01, Seq6:0.08):0.01,

Seq7:0.09):0.01, Seq8:0.10):0.01, Seq9:0.11):0.01, Seq10:0.12):0.01, Seq11:0.13);

Tree 3:

((((seq1:0.3,seq2:0.4):0.3,seq3:0.7):0.2,(seq4:0.03,seq5:0.8):0.8):0.4,

(((seq6:0.03, seq7:0.4):0.03, (seq8:0.1,seq9:0.5):0.5):0.2,(seq10:0.07,seq11:0.8):0.2):0.1);

Tree 4:

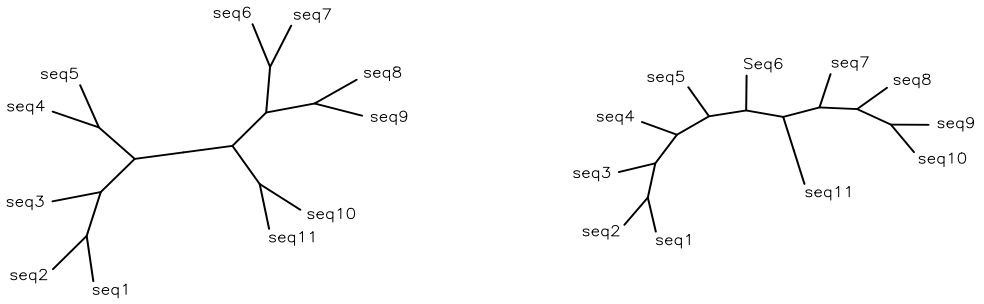
(((((((seq1:0.04,seq2:0.04):0.01,seq3:0.05):0.01,seq4:0.06):0.01,seq5:0.07):0.01,

Seq6:0.08):0.02,(seq7:0.7,(seq8:0.6,(seq9:0.4,seq10:0.4):0.01):0.02):0.01):0.03,seq11:0.8).

These trees were plotted in Figures 1 and 2, respectively using WebPHYLIP [12]. These trees can be ordered according to their pairwise similarities. For example, Tree 2 is evolutionarily more close to Tree 4 because parts of them are very similar. Among these trees, Tree 2 has the highest evolutionary diversity while Tree 4 has the lowest diversity as shown in Section 3.

Consider a sample of 160 protein families, each with 11 protein sequences. Suppose that in truth, there are four groups of 40 families corresponding with the above trees, but this is unknown to the data analyst. To generate such data, for each tree, we sampled 440 protein sequences of length 500 (including indels) each, using the software SIMULATOR [13], these sequences were then randomly divided into 40 families. We will cluster these protein families in Subsection 3.1, pretending we do not know the underlying phylogenetic trees. Using these simulated data we demonstrated a distribution pattern for the transformed correlation coefficients (ρs) in a group of protein families.

Ribosomal protein sequence data. Ribosomal proteins, extremely ancient molecules, are windows into protein evolution [14, 15]. There are two types of ribosomal protein families: small-subunit families and



(a) Tree 3 (b) Tree 4

Figure 2: The draws of the phylogenetic trees 3 and 4.

large-subunit families. They are respectively designated S1, S2,..., and L1, L2,... . The small ribosomal sub-unit is primarily involved in the interaction between the mRNA and tRNA codons during translation whereas the large sub-unit catalyses peptide bond formation and binds translation factors conclusion. These protein sequences were collected from the following eubacteria: Aquifex Aeolicus (abbreviated to aaeoli), Bacillus subtilis (bsub), Mycobacterium tuberculosis (mtub), Borrelia burgdorferi (bburg), Treponema pallidum (tpal), Chlamydia trachomatis (ctra), Escherichia coli (ecoli), Haemophilus influenzae (hinflu), Helicobacter pylori (hyplori), Synechostis species (spcc), and Thermus thermophilus (theth). They are grouped into 51 families: 20 small subunit families-namely, S1-S20; and 31 large subunit families-namely L1-L7/12,L9-L11,L13-L24,L27-L29,L31-L36. The protein sequences in each family bear certain degrees of diversity and homology, showing the traces of their evolution. See [16, 17]. We aimed at giving a characterization of the group-wide evolutionary relationship among these protein families.

Protein structural domain data. This data set collected by Pazos and Valencia [1] is composed of 13 proteins of known structure for which two structural domains in close interaction are clearly visible. These proteins were used to produce a collection of domains. The calculation of the similarity of score matrices was carried out for those pairs of domains with at least 11 sequences from the same species: M.tuberculosis, Rhizobium sp., E.coli, H.pylori, Synechocystis sp., M. thermoautotrophicum, A.aeolicus, B. burgdorferi, P.horikoshii, T.pallidum, B.subtilis, M.jannaschii, H.influenzae, A.fulgidus. The final set contained 133 pairs of domains including 13 pairs of truly interacting domains. See [1] for more details.

2.1.1 Methodology

Calculation of correlation coefficients. Using the software ClustalW with the default setting [18], we first calculated the pairwise alignment scores, which form a score matrix, say (s_{ij}), for each protein family, where n is the number of protein sequences in the protein family under consideration. This score matrix can be linearly changed to a distance matrix so that a distance-based phylogenetic tree can be

produced. Thus for a given pair of protein families with score matrices (s_{ij}) and (t_{ij}) respectively, their correlation coefficient can be defined by the following matrix correlation statistics- namely

$$r = \frac{\sum_{i<j}(s_{ij} - \bar{s})(t_{ij} - \bar{t})}{\sqrt{\sum_{i<j}(s_{ij} - \bar{s})^2}\sqrt{\sum_{i<j}(t_{ij} - \bar{t})^2},}$$

where n is the number of sequences in each family, $\bar{s} = \frac{2}{n(n-1)}\sum_{i<j} s_{ij}$ and $\bar{t} = \frac{2}{n(n-1)}\sum_{i<j}^{n} t_{ij}$. To improve the normality of r, we performed the Fisher Z-transformation as follows:

$$\rho = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right).$$

Obviously, the closer to 0 the ρ is, the less significant the correlation. The sample mean and variance of ρ can be respectively viewed as measures of the overall strength of evolutionary relationships in these families and evolutionary variability between these families. In particular, if ρ follows a normal distribution, then the group-wide co-evolution may be reflected in the mean and variance of ρ.

The monophyletic normality test. In this paper, the normality of ρ was tested for any set of score matrices by use of one-sample Kolmogorov-Smirnov statistics. If the resulting P-value is larger than or equal to t_0, we claim this set to be monophyletically normal. Here t_0 is the prespecified threshold. Our simulation experiences indicate that the test is robust to the choice of t_0 in $(0.15, 0.3)$.

Monophyletic k-medoids clustering. Taking the score matrices as items, we can apply k-medoids clustering to divide these items into k clusters such that the sum of distances over the items to their centres is minimal. The cluster centre is defined as the item which has the smallest sum of distances to the other items in the cluster. The value of k is specified by maximising the so-called average silhouette width [19, 20]. Unfortunately, this algorithm did not work well for monophyletic clustering. Thus we propose a monophyletic k-medoids clustering algorithm as follows.

1. Start with one large cluster, i.e., set $k = 1$, $C_0 = \{$ all the score matrices$\}$. Test the normality of ρs generated from this cluster, resulting the P-value p_{C_0}. If $p_{C_0} \geq t_0$, we believe that C_0 is a monophyletically normal cluster and stop here. Otherwise, set $k = 2$ and go to the next step.

2. Apply the $k-$medoids clustering to C_0. Test the normality of ρs for each cluster. Classify the monophyletic normality according to whether its P-value is larger or equal to the threshold t_0. This leads to n_0 monophyletically normal and m_0 monophyletically non-normal clusters respectively. If $n_0 = k$, stop here. Otherwise, let $D_1, ..., D_{m_0}$ be m_0 monophyletically non-normal clusters. Calculate the cardinal numbers $|D_i|$, $i = 1, ..., m_0$. We merge all D_i with less than t_1 elements in the set POUT. If $|D_i| < t_1$ for all $i = 1, ..., m_0$, stop here. Otherwise, merge all D_i with $|D_i| \geq t_1$ in one set. Let $C_0 = \cup\{D_i : |D_i| \geq t_1, i = 1, ..., m_0\}$ and k_0 be the number of D_i with $|D_i| \geq t_1$. Here we set the threshold $t_1 = 6$ so that we have enough ρs for running the monophyletic normality test on D_i when $|D_i| \geq t_1$. Our simulation experiences indicate that the procedure is robust to the choice of t_1 in $\{5, 6, 7, 8, 9, 10\}$.

3. If $k_0 = 0$, stop here and let n_0 be the total number of monophyletic normal clusters. Back to step 2.

4. If |POUT| ≥ 1 and $n_0 \geq 1$, we test whether we can move some elements in POUT to one of monophyletically normal distributed clusters and generate an outlier set - say, OUT:

   4a). For any $d \in$ POUT, test the monophyletic normality of $D_i \cup \{d\}$ for D_i with $|D_i| \geq t_1$, resulting in a P-vaule p_{di}. Let (d_0, i_0) be the pair of (d, i) in which p_{di} attains the maximum over $\{(d, i) : d \in$ POUT$, |D_i| \geq t_1, 1 \leq i \leq m_0\}$. If $p_{d_0 i_0} \geq t_0$, move d_0 from POUT to D_{i_0}. Otherwise, move d_0 from POUT to OUT.

   4b). Repeat 4a) until POUT is empty.

5. Suppose the previous steps give the following monophyletically normal clusters, $E_1, ...E_{m_1}$.

   5a). Test the normality of $E_i \cup E_j$, $i < j$, leading to the P-values p_{ij}. Let (i_0, j_0) be the pair of (i, j) in which p_{ij} attains the maximum.

   5b). If $p_{i_0 j_0} < t_0$, stop here. Otherwise, merge E_{i_0} and E_{j_0}, reducing the number of monophyletically normal clusters to $m_1 - 1$. For the new set of monophyletically normal clusters, go to the step 5a).

3 Results

3.1 Simulated Data

The Fisher Z-transformed correlation coefficients (ρs) between these 160 simulated protein families were plotted in Figure 3. The top left panel in Figure 3 demonstrated that these protein families can be grouped into four groups. The group 2 has the lowest co-evolution strength as we expected. The correlation coefficients within each group are generally higher than those between groups. Furthermore, the absolute correlations between groups tend to be monotone in the similarities between underlying phylogenetic trees. For example, sharing the half of their phylogenetic trees, groups 2 and 4 have a set of higher correlation coefficients than those between the other pairs of the groups.

However, a naive k-medoids clustering on these score matrices led to only 3 clusters. Two of them are consistent with the underlying group identities but the third one is the merging of the two underlying groups 2 and 4. The Kolmogorov-Smirnov test showed that the ρ values of the first two clusters did have normality (P-values $= 0.979$ and 0.846 respectively) while the third one did not (P-values $= 0$). We thus adjusted the number of clusters, k by setting $k = 4$. The resulting clusters were then fully consistent with the underlying grouping structure except the families 69 and 70 which were wrongly grouped into the cluster 4. In contrast, the monophyletic $k-$medoids clustering method gave the correct groupings of these protein families. The sample means and variances of the observed ρs for the four clusters are listed at columns μ and σ^2 in Table 1.

Table 1: Summary of simulation results

Cluster	μ	σ^2
1	2.1242	0.03271
2	1.3130	0.03977
3	1.9321	0.01967
4	2.5488	0.04892

A very attractive phenomenon was unveiled by the Kolmogrov-Smirnov test as follows. These ρ values in each group were fitted remarkably well by a normal distribution while this is not true if we merge any more than two groups together. See Figures 3 and 4. This implies that when the ρ values of a group of protein families are normally distributed, these families might have very similar underlying phylogenetic trees and thus be group-widely co-evolving.

3.2 Ribosomal Protein Sequence Data

As before, we first calculated the score matrices and the pairwise matrix correlation coefficients for 20 small subunit and 31 large subunit families. The Fisher Z-transformation was then made on these correlation coefficients. The k-medoids clustering suggested a single cluster for these protein families.

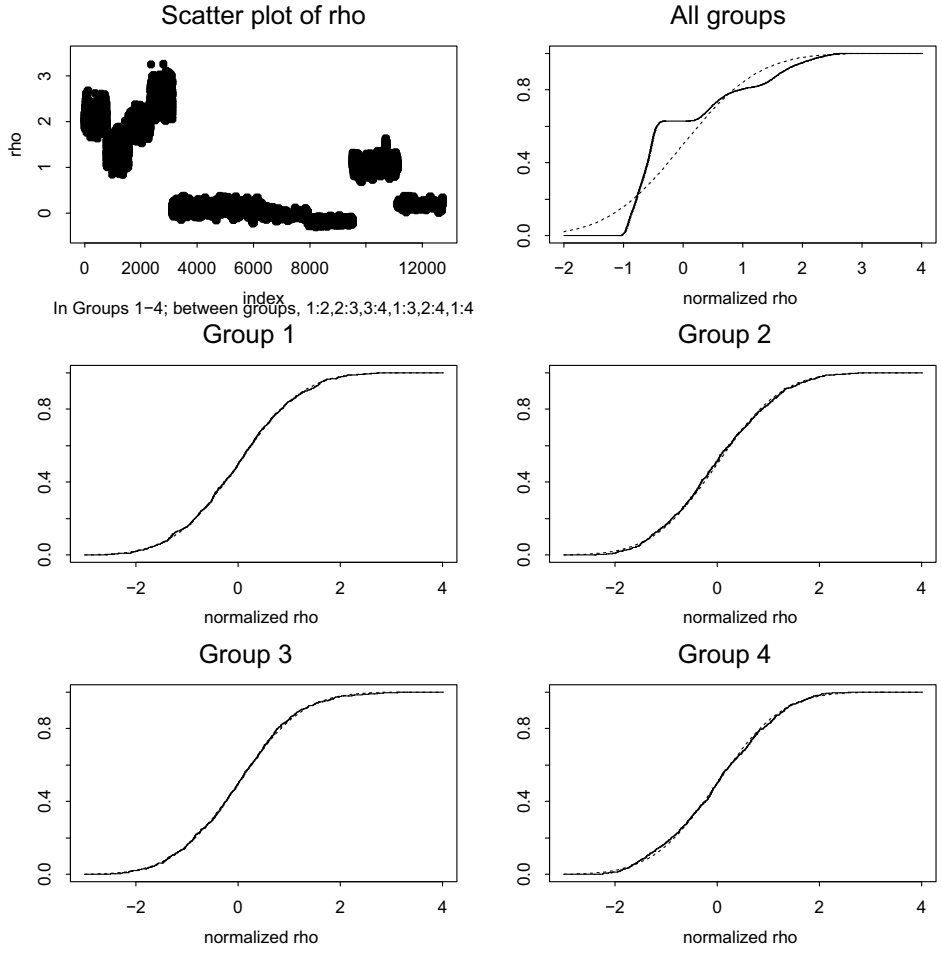


Figure 3: The top row: The left panel is the scatter plot of the Fisher-transformed correlation coefficients (denoted by ρ) for the 160 simulated protein families. These ρs are divided into 10 sets. The first four sets are about the ρs in four groups, say groups 1,2,3,4, of protein families respectively. The remaining 6 sets are the ρs between groups, say between pairs 1:2, 2:3,3:4, 1:3,2:4,and 1:4. These ρs in groups 1,2,3,4 are plotted on the left-hand side of the figure and the remaining are plotted on the right-hand side according to the group pair orders 1:2, 2:3,3:4, 1:3,2:4,and 1:4. The right panel is the comparison between the empirical distribution function of the normalized ρ (i.e., $(\rho - \text{mean}(\rho))/\sqrt{\text{var}(\rho)}$) and the hypothesized normal distribution for all four groups merging together. These four groups were sampled from four different phylogenetic trees plotted in Figures 1 and 2. The middle and bottom rows: The comparisons between the empirical distribution function of the normalized ρ and the hypothesized normal distribution respectively for 4 groups. The solid lines for the empirical distribution functions while the dashed lines for hypothesized normal distribution.

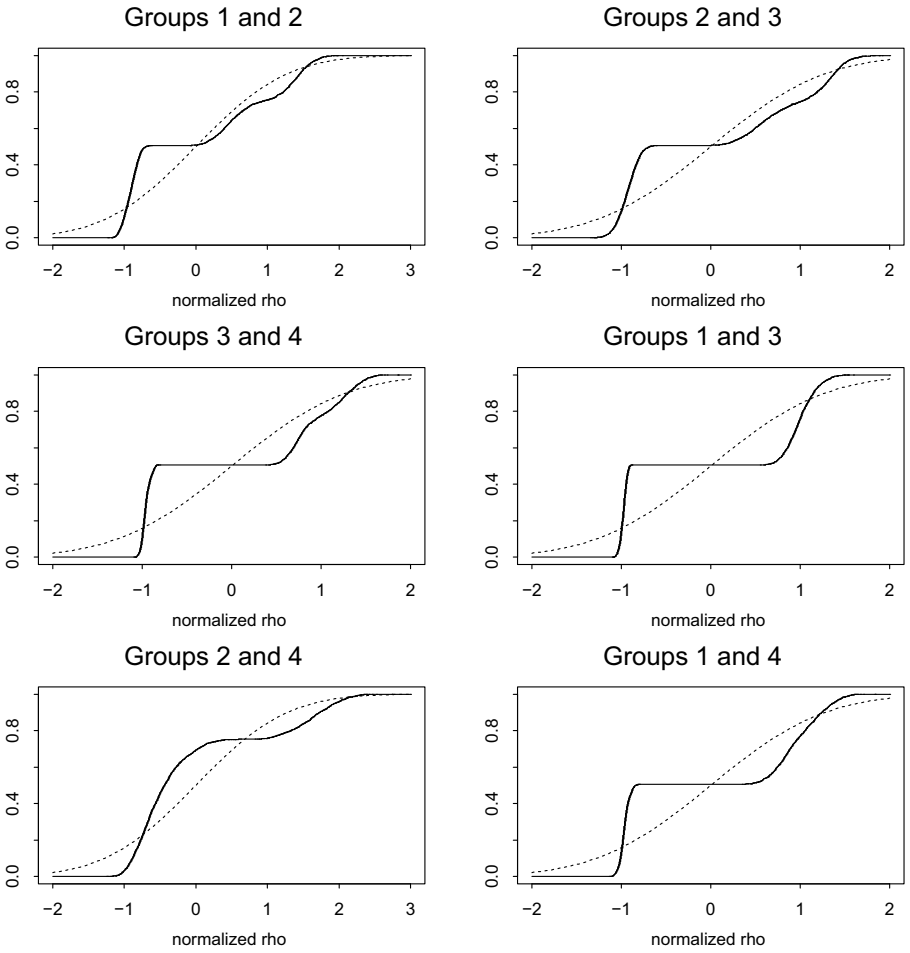


Figure 4: The comparisons between the empirical distribution function of the normalized ρ (i.e. $(\rho - \text{mean}(\rho))/\sqrt{\text{var}(\rho)}$), and the hypothesized normal distribution respectively for each pair of groups merging together. These four groups were sampled from four different phylogenetic trees plotted in Figures 1 and 2.

This fact can also be seen from the scatter plot of these data in Figure 5. In fact, these $\rho-$values are perfectly fitted by a normal distribution with the P-value of 0.9645 and the sample mean and variance $(0.6086177, 0.06670853)$. In light of the simulation results we concluded that these 51 protein families may have very similar underlying phylogenetic trees and are group-widely co-evolving.

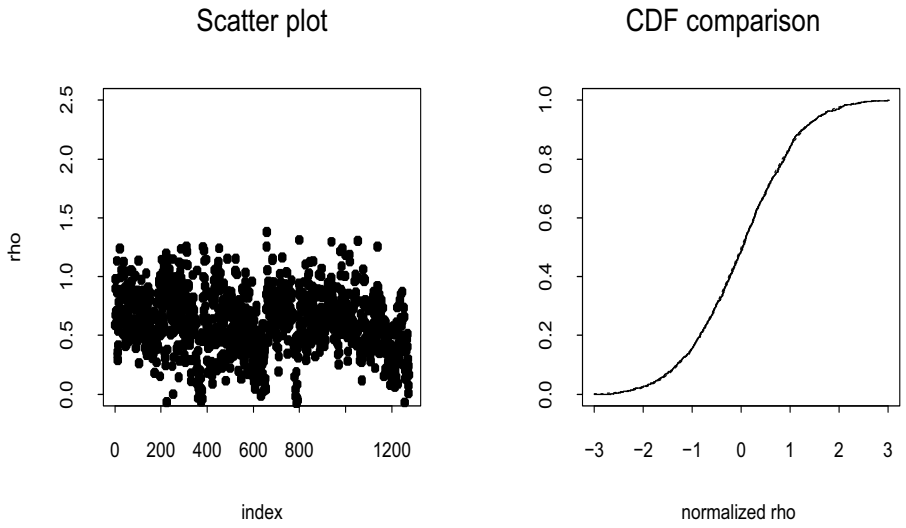


Figure 5: The left panel is the scatter plot of the Fisher-transformed correlation coefficients for the 51 ribosomal protein families. The right panel is the comparison between the empirical distributions of these values after normalization and the hypothesized normal distribution.

Using the cut-off of $\rho > 1.0986$ (i.e.,$r > 0.8$), we obtained 42 significant correlations with the significance level of $P(\rho > 1.0986) = 0.028907$. These are between S1 and S10, S2 and S5, S3 and S4 , S7 and S10, L1 and L17, L2 and L3, L2 and L4, L2 and L5, L2 and L9, L2 and L17, L3 and L5, L4 and L5, L4 and L15, L4 and L19, L5 and L9, L5 and L13, L5 and L15, L9 and L13, L9 and L17, L9 and L20, L9 and L29, L13 and L17, L13 and L20, L15 and L19, L17 and L22, L1 and S3, L1 and S4, L2 and S8, L2 and S9, L4 and S11, L9 and S4, L10 and S11, L11 and S11, L13 and S4, L17 and S4, L19 and S2, L19 and S11, L20 and S4, L21 and S3, L21 and S4, L22 and S19, and L29 and S4. If using $\mu + 2\sigma = 1.125178$ as the cut-off, we found 33 significant pairs with the significant level of 0.02275. They are S1 and S10, S2 and S5, S7 and S10, L1 and L17, L2 and L3, L2 and L5, L2 and L17, L3 and L5, L4 and L15, L5 and L13, L5 and L15, L9 and L13, L9 and L17, L9 and L20, L9 and L29, L13 and L20, L15 and L19, L17 and L22, L1 and S3, L1 and S4, L2 and S8, L2 and S9, L4 and S11, L9 and S4, L13 and S4, L17 and S4, L19 and S2, L19 and S11, L20 and S4, L21 and S3, L21 and S4, L22 and S19; L29 and S4.

If using the cut-off of $\rho > \mu + 2.5\sigma = 1.2493$, we have only 8 significant correlations at the significance level of 0.00656. These are between L5 and L13, L9 and L13, L1 and S3, L1 and S4, L9 and S4, L17 and S4, L22 and S19, L29 and S4.

3.3 Protein Structural Domain Data

Pazos and Valencia [1] calculated the matrix correlation coefficients for 133 pairs of domains mentioned in Section 2. A scatter plot of the Fisher Z-transformed correlation coefficients with their distribution pattern is presented in Figure 6. The result demonstrated that these ρ values were approximately normally distributed. Performing the monophyletic $k-$medoids clustering algorithm on this dataset led to a single group. In light of our simulation results we concluded that that these domains as a group might

be group-widely co-evolving.

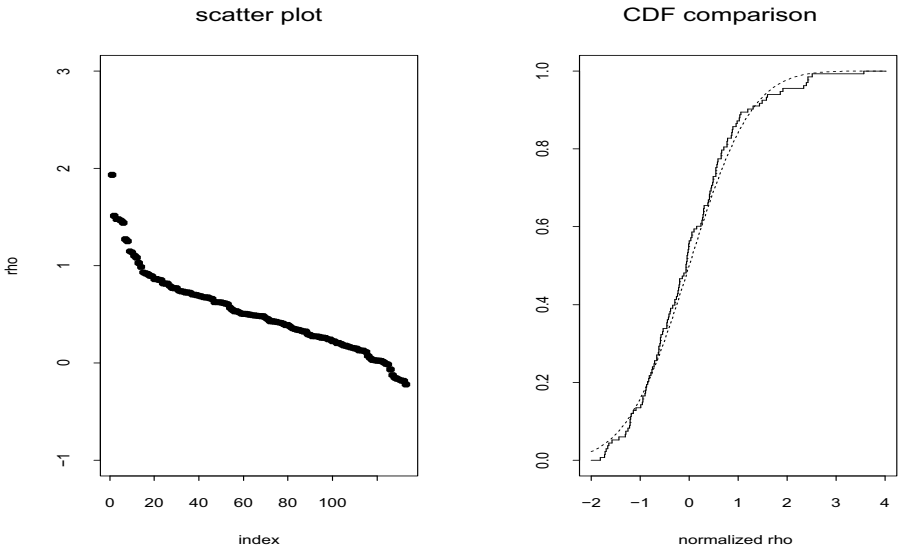


Figure 6: The left panel is the scatter plot of the Fisher-transformed correlation coefficients for the 13 protein structural domain families. The right panel is the comparison between the empirical distributions of these values after normalization and the hypothesized normal distribution.

4 Discussion and Conclusion

It is generally accepted that functionally correlated proteins may have similar evolutionary history. This has made phylogenetic analysis popular in protein function classification and delineation of subfamilies within larger families. See, for example, [6, 21]. A protein family contains sequences that are evolutionarily related. Generally, this is reflected by sequence similarity. There have been many attempts to organize the set of protein families into evolutionarily homogenous clusters using hierarchical clustering with some distance measures. How do we characterize these clusters? How can we cluster protein families using these characterizations? We approached these questions in two ways. Firstly, we have selected four different phylogenetic trees. For each tree, we have simulated 440 protein sequences and randomly divided them into 40 protein families, each with 11 sequences. Thus we have groups of protein sequence families, each derived from a single ancestral line. We have calculated the pairwise alignment score matrix for each family. The Fisher Z-transformed matrix correlation coefficients between these families have been used to measure the evolutionary relationships within and between these groups. We have revealed that the strengths of the evolutionary relationships between protein families within any of these groups can be characterized by a normal distribution although these strengths are not the same. Moreover, this is not true if we merge any two or more of these groups together. This implies that the above distribution characterization may be a unique characterization for a monophyletic group. Exploiting this property, we have proposed a monophyletic $k-$medoids clustering method for dividing a set of polyphyletic protein families into monophyletic groups. Here the monophyletic property called group-wide co-evolution is referred to such that the Fisher Z-transformed matrix correlation coefficients between the trees are approximately normally distributed. Secondly, we have tested our procedure on two real datasets appearing in literature: ribosomal protein families and protein structural domain families. The results showed that both groups are monophyletic. This is biologically meaningful as, for example, we know that the purpose of these ribosomal proteins is relatively simple, that is, stabilising the ribosome and producing proteins.

Studying the monophyletic behaviour of a set of sequence families as opposed to dealing only with the individual sequence family offers some advantages. A well-known problem in proteomics is the identification of interacting proteins based on evolutionary trees. It is hard to set the baseline of evolution within a protein family due to the contribution of the speciation, the inter-molecular interaction of the correlation of two proteins, and the bias in selecting species. This task can be tackled by viewing the baseline as the upper confidence bound for the matrix correlation coefficients. Here we have assumed that the strength of the interaction depends on the evolutionary distance between two proteins in a monotone way. The advantage of such cut-off values is that they have automatically accounted for the baseline differences between the protein families. We have classified the pairs of protein families into three categories with weak, moderate and strong interactions respectively. Assuming that only a few pairs of protein families are interacting significantly in each cluster, the significant interacting pairs have been identified using the bounds at the significance levels of 0.028907, 0.02275 and 0.00659 respectively. Using the first bound we have highlighted 42 ribosomal protein pairs for further investigation. These involve the small subunit proteins S1, S2, S3, S4, S5, S7, S8, S9, S10, S11, S19 and the large subunit proteins L1, L2, L3, L4, L5, L9, L10, L11, L13, L15, L17, L19, L20, L21, L22, L29. This seems consistent with the following biological functions of these proteins: The small ribosomal sub-unit is involved in the interaction between the mRNA codons and tRNA codons. This interaction is imperative to the process of translation. In particular, S1, S4, S7, S8, S15, S17 and S20 are known as the primary binding proteins [15]. The large sub-unit catalyses peptide bond formation between the amino acids and binds factors necessary for the three main stages of translation: initiation, termination, elongation. Most of the large subunit proteins stabilise the ribosomal structure by binding to the 23S rRNA [22]. Many of the proteins in the large sub-unit contain loops in their tertiary structure which penetrate into the sub-unit to stabilise the rRNA. These stabilising proteins are: L1, L2, L3, L6, L9, L11, L14, L15, L16, L20, L21 and L22. Therefore they might be functionally correlated. Another advantage of the monophyletic study lies in providing a rational for clustering protein families based on mixtures of phylogenetic trees. Our proposed clustering procedure can be useful in characterising the existing protein families in literature. More work is needed in this direction.

Web source

WebPHYLIP Version 2.0, http://bioinfo2.ugr.es/WEBPHYLIP/index.htm.

Acknowledgements

The work of Zhiyuan Zhao and Guoying Li was partially supported by National Natural Science Foundation of China under Grants No. 19631040 and No. 90403130.

References

[1] F. Pazos, A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**(9): 609-614, 2001.

[2] C. Goh, A. A. Bogan, M. Joachimiak, D. Walther, F.E. Cohen. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**(2): 283-293, 2000.

[3] A.K. Ramani, E.M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**(1): 273-284, 2003.

[4] U. Stelzl, U. Worm, M. Lalowski, et al. A Human protein-protein interaction network: A resource for annotating the proteome. *Cell*, **122**(6): 957-968, 2004.

[5] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, T.O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**(8): 4285-4288, 1999.

[6] W.K. Kim, D.M. Bolser, J.H. Park. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics,* **20**(7): 1138-1150, 2004.

[7] C. Goh, F.E. Cohen. Co-evolutionary analysis reveals insights into protein-protein interactions. *J. Mol. Biol.,* **324**(1): 177-192, 2002.

[8] M.W. Talkington, G. Siuzdak, J.R. Williamson. An assembly landscape for the 30S ribosomal subunit. *Nature*, **438**(7068): 628-632, 2005.

[9] J.C. Rain, L. Selig, H. De Reuse, et al. The protein-protein interaction map of Helicobacter pylori. *Nature,* **409**(6821):211-215, 2001.

[10] M.I. Recht, J.R. Williamson. Central domain assembly: Thermodynamics and kinetics of S6 and S18 binding to an S15-RNA complex. *J. Mol. Biol.*, **313**(1): 35-48, 2001.

[11] S.H. Tindall, K.C. Aune. Assessment by sedimentation equilibrium analysis of a heterologous macromolecular interaction in the presence of self-association: interaction of S5 with S8. *Biochemistry,* **20**(17): 4861-6, 1981.

[12] Lim, A., Zhang, L., "WebPHYLIP: A Web Interface to PHYLIP," *Bioinformatics*, **15**(12): 1068-1069, 1999.

[13] R. Fiessner. *Sequence alignment and phylogenetic inference*. Logos Verlag, Berlin, 2004.

[14] D.E. Draper, L.P. Reynaldo. RNA binding strategies of ribosomal proteins. *Nucleic Acids Res.*, **27**(2): 381-388, 1999.

[15] S.C. Agalarov. Structure of the S15, S6, S18 rRNA complex: assembly of the 30S ribosome central domain. *Science*, **288**(5463): 107-112, 2000.

[16] J. Zhang. Analysis of information content of biological sequences. *J. Comput. Biol.*, **9**(3): 487-503, 2002.

[17] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**(9): 324-328, 1998.

[18] J.D. Thompson, D.G. Higgins, T.J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**(22): 4673-4680.

[19] L. Kaufman, P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

[20] M.J. van der Laan, K.S. Pollard, J. Bryan. A new partitioning around medoids algorithm. *Working paper series 105*, Division of Biostatistics, University of California at Berkeley, 2002.

[21] A. Krause, J. Stoye, M. Vingron. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics,* **6:** 15, 2005.

[22] N. Ban, P. Nissen, J. Hansen, P.B. Moore, T.A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 *Å* resolution. *Science,* **289**(5481): 905-920, 2000.