doi:10.1016/j.jmb.2008.09.078 J. Mol. Biol. (2009) 385, 91–98

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree

Maricel G. Kann[1]*, Benjamin A. Shoemaker[2], Anna R. Panchenko[2] and Teresa M. Przytycka[2]*

[1]*Department of Biological Sciences, University of Maryland, Baltimore County, MD 21250, USA*

[2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, MD 20894, USA*

Received 31 March 2008; received in revised form 25 September 2008; accepted 27 September 2008
Available online
9 October 2008

Edited by M. Sternberg

It has been observed that the evolutionary distances of interacting proteins often display a higher level of similarity than those of noninteracting proteins. This finding indicates that interacting proteins are subject to common evolutionary constraints and constitutes the basis of a method to predict protein interactions known as mirrortree. It has been difficult, however, to identify the direct cause of the observed similarities between evolutionary trees. One possible explanation is the existence of compensatory mutations between partners' binding sites to maintain proper binding. This explanation, though, has been recently challenged, and it has been suggested that the signal of correlated evolution uncovered by the mirrortree method is unrelated to any correlated evolution between binding sites. We examine the contribution of binding sites to the correlation between evolutionary trees of interacting domains. We show that binding neighborhoods of interacting proteins have, on average, higher coevolutionary signal compared with the regions outside binding sites; however, when the binding neighborhood is removed, the remaining domain sequence still contains some coevolutionary signal. In conclusion, the correlation between evolutionary trees of interacting domains cannot exclusively be attributed to the correlated evolution of the binding sites or to common evolutionary pressure exerted on the whole protein domain sequence, each of which contributes to the signal measured by the mirrortree approach.

Keywords: domain–domain interactions; protein coevolution; protein–protein interactions; phylogenetic tree; mirrortree

Introduction

It has been proposed that interacting proteins should coevolve to maintain their interactions.[1–3] This idea provides the main motivation for the method to predict protein interactions known as mirrortree.[1,3–13] The mirrortree method predicts protein–protein interactions by assessing the extent of agreement between evolutionary distances that could be attributed to correlated evolution. For this purpose, distance matrices are constructed from alignments of orthologous sequences taken from a common set of species. The degree of correlated evolution between families of orthologs is assessed by computing the correlation coefficient between the corresponding distance matrices. The mirrortree method measures the correlation between evolutionary distances and thus, indirectly, the correlation between evolutionary rates along individual branches of evolutionary trees from two families. While correlation between the evolutionary trees of interacting proteins has been well documented,[1,2,4,7,10,14] the principal cause of such correlated changes remains unclear.[15] In particular, it has been proposed that higher correlation values between evolutionary trees of interacting proteins (with respect to noninteracting ones) can be caused by compensatory mutations, in which mutations in one binding partner are being compensated by complementary mutations in another partner to maintain amino acid interactions important for protein function, stability, and foldability.[1,2,4,16–19]

Correlation between evolutionary distances of interacting proteins may also have other sources. For example, Fraser *et al.*[20] used codon adaptation

*Corresponding authors. E-mail addresses: mkann@umbc.edu; przytyck@ncbi.nlm.nih.gov.

Abbreviations used: MSA, multiple sequence alignment; ROC, receiver operating characteristic; CBM, conserved binding mode.

index analysis to infer that the levels of expression of interacting partners are also subject to correlated evolution and that such coexpression could be required for maintaining proper stoichiometry among interacting components. It has been observed that expression levels are correlated with evolutionary rates,[16,21,22] which might contribute to the coevolution signal measured by the mirrortree method. Indeed, Hakes *et al.*[23] demonstrated that mRNA abundance is a good protein interaction predictor. Another important argument against using compensatory mutations to explain the entire coevolutionary signal detected by the mirrortree is that this approach could also identify as interacting the noninteracting proteins within the same protein complex or biological pathway. Indeed, an extension of the mirrortree recently introduced by Juan *et al.* detects proteins within the same metabolic pathways despite the fact that they are not necessarily related by physical interactions.[24]

Challenging previous assumptions about the strong contribution of coevolution of binding interfaces to the correlation signal between evolutionary distances measured by the mirrortree method, Hakes *et al.*[23] suggested that such correlation does not mostly originate from compensating mutations in the interface. In their work, Hakes *et al.*[23] showed that selecting only the surface residues or the interface residues as input for the mirrortree approach yields similar results as using the whole protein sequence. Based on their analysis, they concluded that "correlated sequence evolution is most probably due to interacting proteins being constrained in similar ways and having similar rates of evolution across their entire sequences."

Accounting for the abovementioned considerations, in this work, *correlated evolution* refers to correlated changes in evolutionary rates imposed on a pair of interacting proteins to preserve their interaction properties. As such, this definition of correlated evolution also includes correlated changes to preserve physical binding properties, coexpression, foldability, and all other constraints that are imposed on a pair of interacting proteins to preserve functional properties of interaction. It is important to keep in mind that since the mirrortree technique is based on correlated changes in distances between sequences of interacting proteins rather than on a direct measurement of any of the abovementioned factors, it cannot assess which one of them is a dominating contributor to the signal.

In this work, we analyzed the contribution of the binding sites to the coevolutionary signal of domain–domain interactions measured by the mirror-tree method. For this purpose, we used binding sites together with their spatially surrounding residues, which we refer to as *binding neighborhoods*. We selected a set of protein domains with representatives in one common set of species, thereby avoiding problems related to comparing correlations computed based on different sets of species. Furthermore, to limit the impact of the coevolutionary signal due to common speciation divergence, we applied the speciation subtraction methods of Pazos *et al.*[14] and Sato *et al.*[11] With these controls in place, we developed several tests to compare the relative strength of the coevolutionary signal from binding and nonbinding parts of proteins. In particular, we tested how the coevolutionary signal computed from the binding neighborhood compares with that computed from an equivalent number of nonbinding positions.

In agreement with previous work indicating that coevolutionary signal is not restricted to the binding interface, we found that when the binding neighborhoods are completely removed, the remaining sequences of interacting domains still contain a significant coevolutionary signal. However, we also found that the signal is not distributed uniformly across the sequence. In particular, removing the binding neighborhood significantly reduced the performance of the method. In addition, we found that the binding neighborhood alone provides a stronger coevolutionary signal than the same number of randomly selected residues outside the binding neighborhood. Thus, the correlation between evolutionary distances of interacting protein domains can only be partially explained by the common evolutionary pressure exerted along the whole sequence of interacting protein domains. In particular, our results indicate that the binding neighborhood has a significantly higher contribution to this signal than the rest of the protein domain sequence.

Results and Discussion

To compare the contribution of correlated evolution measured by mirrortree based on binding sites with that of the whole protein domain sequence, we considered a set of columns from the multiple sequence alignment (MSA) corresponding to binding sites and their close neighborhood (binding neighborhood). The rationale for considering this binding neighborhood rather than binding sites alone is as follows: The mirrortree method measures the correlation between evolutionary changes, but the binding sites alone are often (sometimes nearly perfectly) conserved and might not display enough variation to provide detectable coevolutionary signal. Furthermore, it has been found previously that the majority of the coevolving positions are not in direct contact but usually physically close ($\leq$10 Å).[25]

First, we compared the performance of the mirrortree method using MSA columns from the binding neighborhood alone with the performance of the same method when equal numbers of randomly selected nonneighborhood MSA columns were used (Fig. 1). We considered binding neighborhoods at increasing thresholds: 6 Å, 8 Å, 10 Å, and 12 Å (see Materials and Methods). We corrected for the speciation divergence using two methods that we refer to in this article as the "nonorthogonal" and "orthogonal" methods, which were proposed by

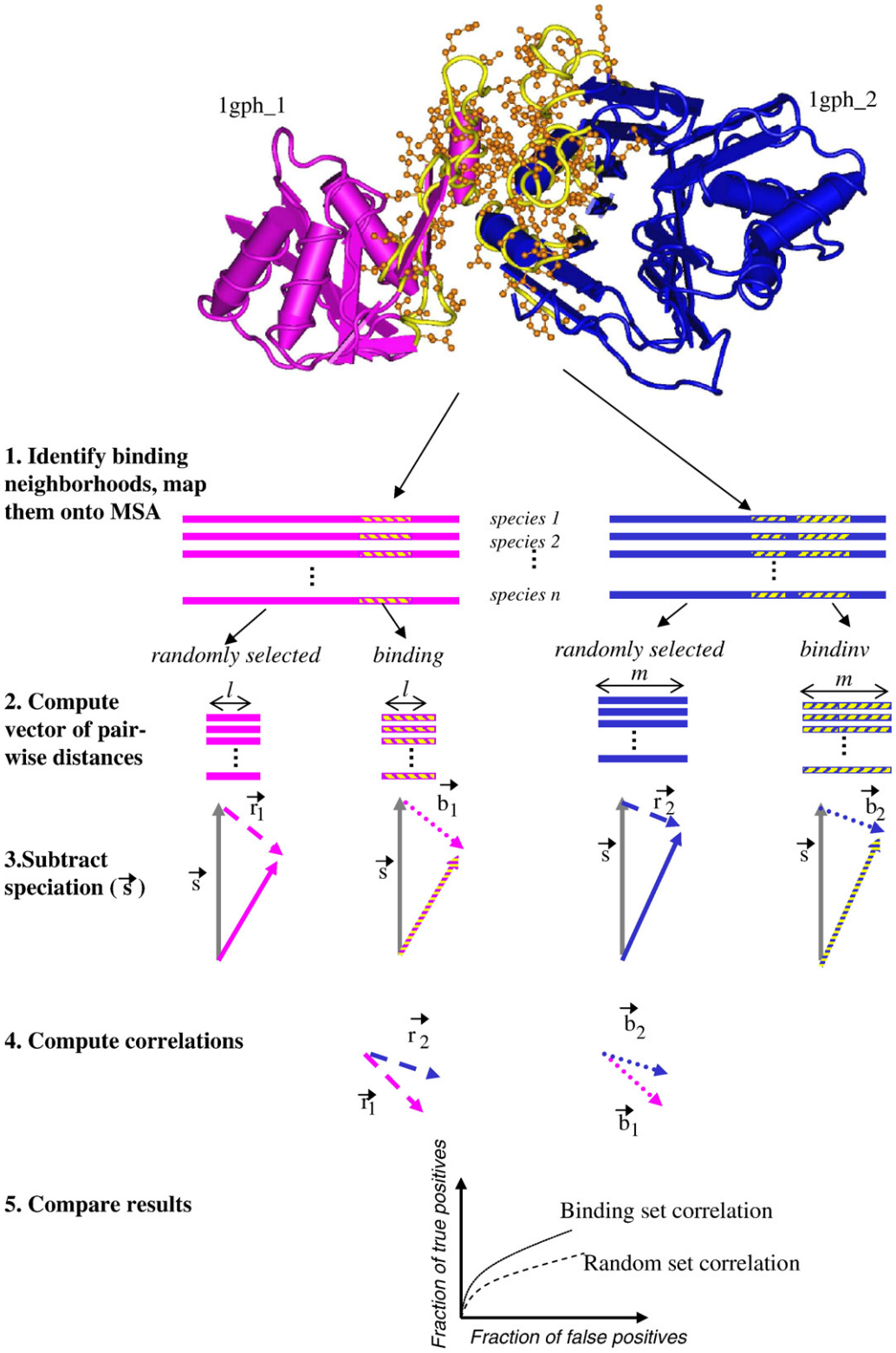

Fig. 1. Comparison of signals from the binding neighborhood with those from randomly selected MSA columns. (1) The binding neighborhoods are extracted from crystal structures of interacting domains and projected onto the MSA of orthologous sequences. (2) The distance matrices are constructed using the MSA columns corresponding to the binding neighborhoods and, separately, for the sequences constructed by randomly selecting the same number of nonbinding MSA columns. The upper triangle of the distance matrix is represented as a vector. (3) Subsequently, each vector is corrected by subtracting the speciation vector ($\vec{s}$, depicted in gray). (4) The correlation coefficient between the resulting vectors is computed (dashed and dotted vectors for randomly selected columns and binding neighborhood, respectively). (5) Finally, it is tested whether the correlation between vectors computed using the binding neighborhood leads to better retrieval results than that between vectors computed using randomly selected MSA columns.

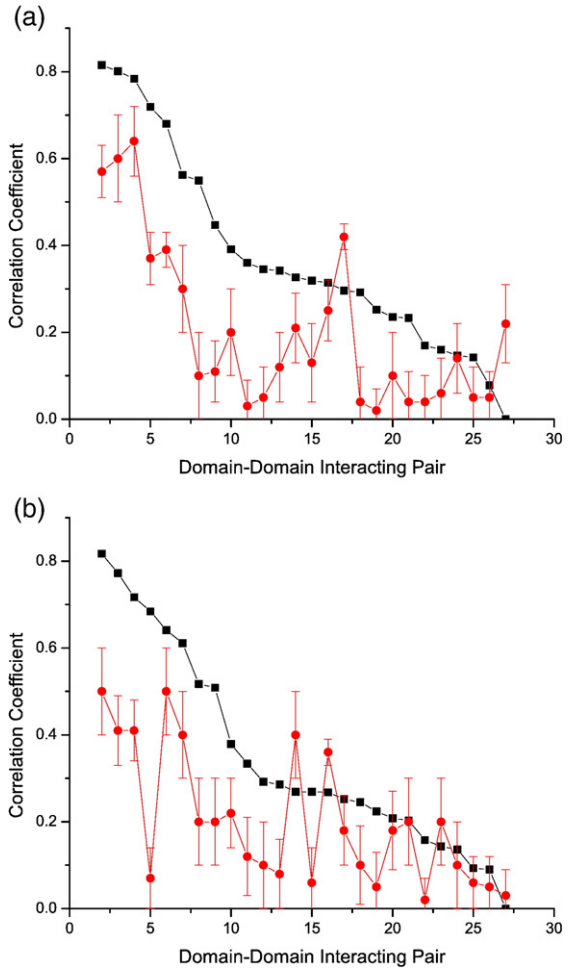

Fig. 2. Comparison of the correlation coefficients for each domain–domain interacting pair using the binding neighborhood (black) and an equivalent number of randomly selected columns (red). The values in the *x*-axis, labeled 1–26, represent each of the domain–domain interacting pairs sorted in descending order by the corresponding correlation coefficient when using the binding neighborhood. Results in panels (a) and (b) were obtained using the orthogonal and nonorthogonal speciation corrections, respectively. For randomized experiments, we plotted the mean value; standard deviations are represented based on 100 trials as the error bar.

Pazos *et al.*[14] and Sato *et al.*,[11] respectively (see Materials and Methods). We refer to our previous study[6] for details about the methodological differences between these methods and for an explanation for this naming convention. We should note that the gold standard used for benchmarking was designed based on a set of domain–domain interactions verified with crystallographic data from Shoemaker *et al.*[26] This data set, with additional constraints (see Materials and Methods), might be biased toward domain pairs that form more stable complexes rather than transient interactions due to the limited sample size. Figure 2 depicts the comparison, for all interacting domain pairs, of the correlation coefficients obtained using the binding neighborhood alone with those obtained using an equal number of randomly selected nonneighborhood MSA columns (see Materials and Methods). Results using the orthogonal and nonorthogonal speciation corrections are depicted in Fig. 2a and b, respectively. For both speciation corrections, the coevolutionary signal strength, represented by the correlation coefficients, derived from the binding neighborhood is predominately higher.

In addition, the accuracy of the different methods was measured using receiver operating characteristic (ROC) curves.[27] We used complete ROC and ROC_n curves (plots truncated after the first n false results[28,29]) that were normalized such that the area under the ROC curve for an ideal retrieval method (one that returns all the true results first) was equal to 1.0. The corresponding ROC curves, for the binding neighborhood of 10 Å, are shown in Fig. 3. Independent of the speciation subtraction method used, exclusive use of the binding neighborhood drastically improves the performance in predicting domain interactions over the set of randomly selected MSA columns outside the binding neighborhood. Figure 4 shows values for ROC_{50} and ROC_{total} for all these experiments, together with the corresponding values from additional experiments discussed below. The values of ROC are given in Table 1. For the experiment using randomly selected columns, we computed the standard deviation based on 100 trials. Note that the results for randomly selected columns and those for the binding neighborhood differ by several standard deviations. We confirmed that the results presented in this article are robust with respect to the definition of binding neighborhood.

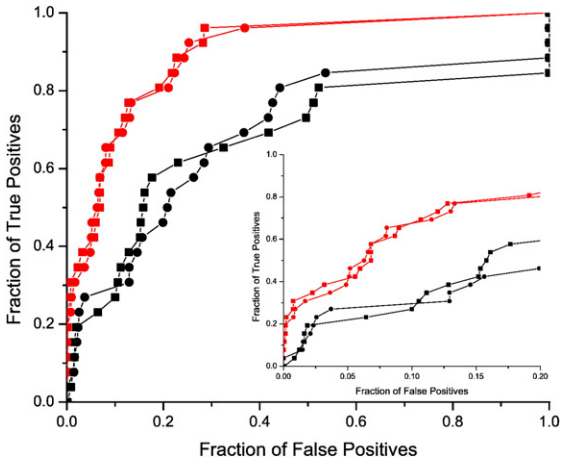

Fig. 3. Comparison of the performance of the mirrortree method on the binding neighborhood with that on the randomly selected MSA column set of the same size. The red lines correspond to the performance using the binding neighborhood with corrections from the orthogonal speciation subtraction (circles) and nonorthogonal speciation subtraction (squares). The corresponding graphs for randomly selected residues are shown in black. Insert shows ROC curves for up to a 20% false-positive rate.

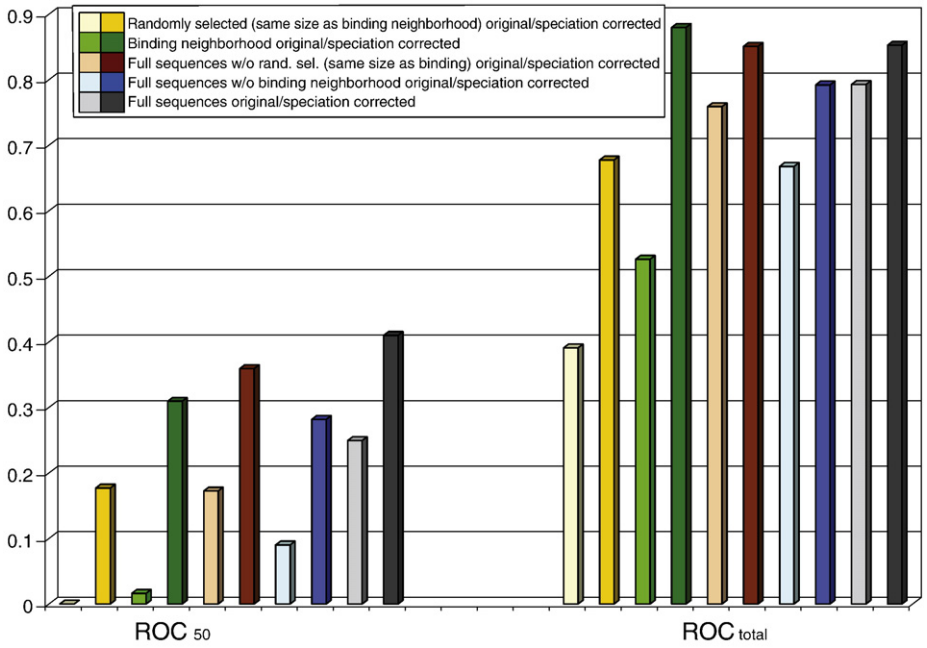

Fig. 4. Dependence of ROC results on the columns used in the alignment and on the speciation correction used in the analysis. Results for ROC_{total} and ROC_{50} show the following trends: Regardless of the region of the sequence used, the performance of the method with the correction for speciation (darker colors) is better than that of the original mirrortree method without the correction (lighter colors). In particular, using the full-length sequence with speciation correction yields, for ROC_{50}, the best results (gray); subtracting a set of randomly selected nonbinding columns (brown) represents only a slight decrease in performance, while subtracting the binding neighborhood (cyan) to the full sequence represents a significant decrease in the overall performance of the method. Finally, performance using the binding neighborhood alone (green) is significantly better than using a randomly selected set of columns of the same size but not belonging to the binding neighborhood (yellow).

Next, we analyzed the effect of removing the binding neighborhood on the performance of the mirrortree method and compared it with the effect of removing randomly chosen columns outside the binding neighborhood. The number of removed random columns was equal to the number of columns in the binding neighborhood, thereby accounting for any effect that the number of columns might have on the method. We applied both orthogonal and nonorthogonal speciation subtraction methods; results are depicted in Fig. 5a and b, respectively. Independent of the applied speciation subtraction, we observed that removal of the binding neighborhood leads to a significant decrease in the performance of the mirrortree method. Yet, our results show that the sequence without the binding neighborhood still provides significant coevolutionary signal. Furthermore, for randomly selected residues, the discriminating power measured by the ROC value increased with the number of selected columns.

One can argue that since Hakes *et al.*[23] found no difference in the discriminating power between the surface region and the whole sequence, the better

Table 1. Summary of ROC_{50} and ROC_{total} values for all experiments

Experiment	Speciation correction	ROC_{50} (std)	ROC_{total} (std)
Randomly selected columns	None	0.000 (0.003)	0.391 (0.006)
Binding neighborhood	None	0.016	0.526
Full without randomly selected residues	None	0.189 (0.02)	0.75 (0.02)
Full without binding neighborhood	None	0.090	0.667
Full sequence	None	0.249	0.793
Randomly selected columns	Nonorthogonal	0.16 (0.03)	0.677 (0.05)
Binding neighborhood	Nonorthogonal	0.309	0.88
Full without randomly selected residues	Nonorthogonal	0.38 (0.02)	0.851 (0.02)
Full without binding neighborhood	Nonorthogonal	0.282	0.792
Full sequence	Nonorthogonal	0.410	0.852
Randomly selected columns	Orthogonal	0.14 (0.03)	0.645 (0.04)
Binding neighborhood	Orthogonal	0.276	0.735
Full without randomly selected residues	Orthogonal	0.31 (0.02)	0.85 (0.02)
Full without binding neighborhood	Orthogonal	0.249	0.735
Full sequence	Orthogonal	0.348	0.838

For randomized experiments, we report the mean value; standard deviations (in parentheses) were computed based on 100 trials.

performance of the binding neighborhood could be due to surface residues that might be contained within the binding neighborhood. To eliminate this possibility, we compared the ROC values obtained from the binding neighborhood with those computed based on surface residues (of the same size as the binding neighborhood and excluding residues from the binding neighborhood). In addition, only interacting pairs that contain sufficiently large numbers of surface residues outside the binding neighborhood were selected (see Materials and Methods). The results for this analysis, depicted in Table 2, show that the ROC values using surface residues outside the binding neighborhood are always smaller than those using the binding neighborhood (independent of the speciation correction used).

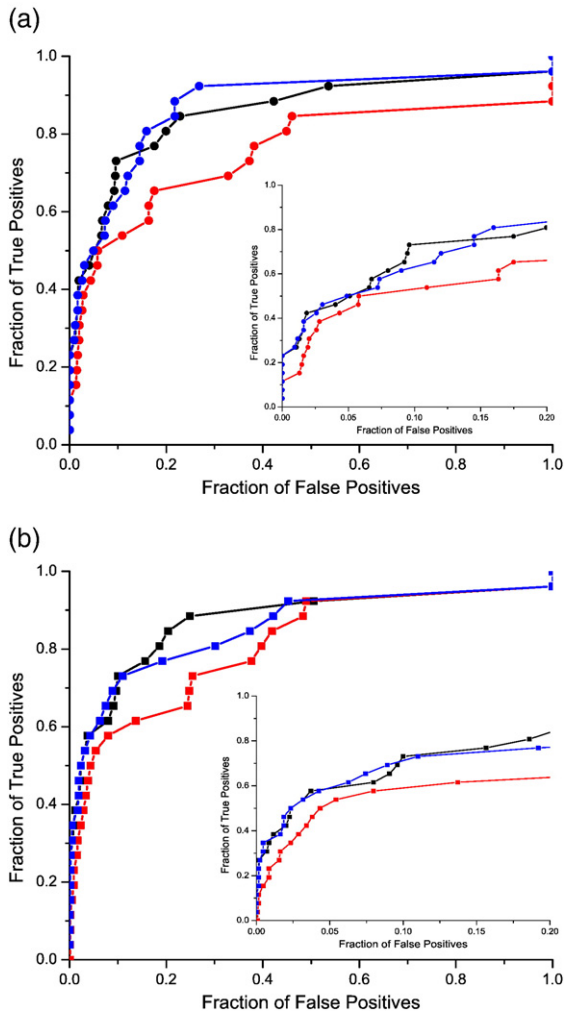

Fig. 5. Comparison of the performance of the mirrortree method using the full sequence (black), full sequence without the binding neighborhood (red), and sequence without the set of randomly selected columns of the same number as the binding neighborhood (blue). Panel (a) corresponds to the orthogonal speciation correction, and panel (b) corresponds to the nonorthogonal speciation correction. Inserts show ROC curves for up to a 20% false-positive rate.

Table 2. Comparison of ROC_{50} and ROC_{100} values for experiments using the full sequence, only the surface, and only the binding neighborhood (radius 10 Å) using set_18

Experiment	Speciation correction	ROC_{50}	ROC_{100}
Full sequence	None	0.196	0.256
Surface	None	0.000	0.000
Binding neighborhood	None	0.000	0.022
Full sequence	Nonorthogonal	0.430	0.548
Surface	Nonorthogonal	0.188	0.261
Binding neighborhood	Nonorthogonal	0.287	0.386
Full sequence	Orthogonal	0.337	0.444
Surface	Orthogonal	0.162	0.211
Binding neighborhood	Orthogonal	0.201	0.312

Finally, Fig. 4 provides a summary of the above-discussed results without speciation subtraction (faded colors) and with the nonorthogonal subtraction (bright colors); results for orthogonal subtraction (not shown) were almost identical. Clearly, the binding neighborhood is a better discriminator of interactions than a randomly selected set of columns of the same size. The relative discriminative powers of the whole sequence, the whole sequence without binding neighborhood, and the binding neighborhood differ in their ROC values, with the whole sequence performing best on the more practical ROC_{50}. In addition to the abovementioned results, our work shows that a larger number of MSA columns provide a stronger signal (the number of columns in the randomly selected set and those in the binding neighborhood sets are smaller than the number of columns in the full-length sequences). Furthermore, from the summary in Fig. 4, one can also appreciate the strongly increased power of the mirrortree method when the correction for speciation is applied.

Conclusions

We have shown that binding neighborhoods of interacting proteins have, on average, higher coevolutionary signal compared with those columns outside binding sites. We also found that the sequences without the binding sites still contain some coevolutionary signal; however, the signal coming from a randomly selected set of columns is weaker than that from the binding neighborhoods. Interestingly, our results also show that the coevolutionary signal of randomly selected MSA columns outside the binding neighborhood increases with the number of columns.

Thus, in agreement with Hakes *et al.* and others, we found that the binding neighborhood alone is not the only contributor to the coevolutionary signal. Hakes *et al.* concluded that "correlated sequence evolution is most probably due to interacting proteins being constrained in similar ways and, consequently, having similar rates of evolution across their entire sequences." Our additional experiments however lead to the conclusion that the signal is not uniform and instead is stronger in the binding neighborhood.

Our results indicate that the binding neighborhood is subject to stronger correlated evolution than other regions of the interacting protein domains. However, since the mirrortree technique is based on common variation of sequence distances between sequences of interacting proteins rather than on a direct measurement of any of the factors that might contribute to this variation, the exact coevolutionary mechanism that leads to the similarity of evolutionary trees of interacting proteins is still not fully uncovered.

Materials and Methods

Data set

The set of interacting and noninteracting domains was selected similar to the work of Kann *et al.*[6] In particular, this test set ensures that all orthologous families (interacting or not) contain sequences from the same set of 70 species. Our set of interacting domain families contains a total of 26 interacting pairs and 1291 noninteracting pairs. The interacting domains have been selected rigorously based on the Conserved Binding Mode (CBM) database of Shoemaker *et al.*,[26] which maps the protein domains from the Conserved Domain Database[30] onto a set of interacting domains from the Protein Data Bank and verifies the interactions via CBMs. The CBM analysis uses conserved geometric interfaces to reduce the possibility of including nonbiological interactions. All interacting domains in this study were extracted from the CBM database. For cases in which multiple CBMs are present for a single interacting pair of domains, we randomly chose one CBM per pair. For each domain pair extracted from the CBM database, an expanded binding neighborhood was determined about the binding interface. Any two residues a and b, from domains A and B, respectively, are included if and only if the distance between the closest atoms from a and b does not exceed a given threshold. Threshold values of 6 , 8 , 10 , and 12 Å were used. The results given in the main text are for 10 Å. In addition, all pairs selected for this study must have contained at least 20 residues satisfying this condition.

Solvent-accessible residues were determined from DSSP[31] calculations available for all structures†. The solvent accessibility of residue "X" is defined as the ratio of its solvent-accessible area in protein structure to that for extended tripeptide Gly–X–Gly. An amino acid is considered as solvent accessible if this ratio is greater than 0.05. To account for the size effect, we chose the same number of columns for both surface and binding neighborhood regions. In addition, we selected only those interacting pairs that contained sufficiently large numbers of surface residues outside the binding neighborhood and did not contain large disordered regions (surface accessibility could not be calculated for disordered regions). After applying all these restrictions, we ended up with 18 interacting pairs (set_18).

Mirrortree method with speciation subtraction

For each domain pair reported in the interacting and noninteracting sets, the correlation between the evolutionary trees was measured by computing the correlation coefficient of the corresponding vectors. The sets of residues and their corresponding MSA columns were selected as follows: All the columns of the protein domain's MSA were used for the "full sequence experiment." For each interacting domain, we randomly selected a set of columns outside the binding neighborhood equal to the number of sites in the binding neighborhood. Only binding neighborhood columns or randomly selected columns were used (or excluded) in the remaining experiments. This selection procedure was repeated 100 times. The correlation coefficient was adjusted using one of the two speciation subtraction approaches cited, orthogonal subtraction and nonorthogonal subtraction, as proposed by Sato *et al*[11] and Pazos *et al.*,[14] respectively. For two domains A and B with n species in common in their MSA, let's denote A_{ij} as the distance between species i and j for protein family A and B_{ij} as that for protein family B.

To implement the speciation signal subtraction, we further modified these distance vectors as follows: First, the background speciation matrix was computed by averaging the evolutionary distance matrices (F) of all protein families (from interacting and noninteracting sets). Thus, for N protein families, the distance between species i and j in the background speciation vector is given by

$$s_{i,j} = \frac{\sum_{k=1}^{N} F_{i,j}^{k}}{N}$$

To reduce the impact of codivergence due to common speciation history, the nonorthogonal speciation subtraction method[14] computes modified distances A'_{ij} and B'_{ij} by $A'_{ij}=A_{ij}-s_{ij}$ and $B'_{ij}=B_{ij}-s_{ij}$, respectively. When using the orthogonal reference, corresponding distances are defined by $A'_{ij}=A_{ij}-P^{\mathrm{A}}_{ij}$ and $B'_{ij}=A_{ij}-P^{\mathrm{B}}_{ij}$, where P^{F}_{ij} is the standard representation of the projection of the distance vector for protein family F = A or B into the speciation vector $\vec{s}_{ij}$ and is given by

$$P^{\mathrm{F}}_{i,j} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{i,j}F_{i,j}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{i,j}^{2}}\, s_{i,j}$$

We point out that the implementation of orthogonal and nonorthogonal subtraction methods is identical with that reported in Ref. 6. Finally, given A_{ij}' and B_{ij}', the correlation between evolutionary histories is estimated by computing a standard Pearson's correlation coefficient of the upper right triangle of the corresponding matrices.

† ftp://ftp.cmbi.kun.nl/pub/molbio/data/dssp

Acknowledgements

Support for this work was provided in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and through funding from the University of Maryland Baltimore County Special Research Initiative Support.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.09.078

References

1. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293.
2. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.* **271**, 511–523.
3. Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614.
4. Goh, C. S. & Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.* **324**, 177–192.
5. Jothi, R., Kann, M. G. & Przytycka, T. M. (2005). Predicting protein–protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, **21**, i241–i250.
6. Kann, M. G., Jothi, R., Cherukuri, P. F. & Przytycka, T. M. (2007). Predicting protein domain interactions from coevolution of conserved regions. *Proteins*, **67**, 811–820.
7. Jothi, R., Cherukuri, P. F., Tasneem, A. & Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J. Mol. Biol.* **362**, 861–875.
8. Ramani, A. K. & Marcotte, E. M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284.
9. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S. & Rothschild, B. (2003). Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, **19**, 2039–2045.
10. Pazos, F. & Valencia, A. (2002). *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.
11. Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. (2005). The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.
12. Tan, S. H., Zhang, Z. & Ng, S. K. (2004). ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. *Nucleic Acids Res.* **32**, W69–W72.
13. Craig, R. A. & Liao, L. (2007). Phylogenetic tree information aids supervised learning for predicting protein–protein interaction based on distance matrices. *BMC Bioinformatics*, **8**, 6.
14. Pazos, F., Ranea, J. A., Juan, D. & Sternberg, M. J. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–1015.
15. Juan, D., Pazos, F. & Valencia, A. (2008). Co-evolution and co-adaptation in protein networks. *FEBS Lett.* **582**, 1225–1230.
16. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
17. Kim, W. K., Bolser, D. M. & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.
18. Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.
19. Chakrabarti, S. & Panchenko, A. (2008). Coevolution in defining the functional specificity. *Proteins*, [Epub ahead of print].
20. Fraser, H. B., Hirsh, A. E., Wall, D. P. & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc. Natl Acad. Sci. USA*, **101**, 9033–9038.
21. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
22. Rocha, E. P. & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**, 108–116.
23. Hakes, L., Lovell, S. C., Oliver, S. G. & Robertson, D. L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl Acad. Sci. USA*, **104**, 7999–8004.
24. Juan, D., Pazos, F. & Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl Acad. Sci. USA*, **105**, 934–939.
25. Yeang, C. H. & Haussler, D. (2007). Detecting coevolution in and among protein domains. *PLoS Comput. Biol.* **3**, e211.
26. Shoemaker, B. A., Panchenko, A. R. & Bryant, S. H. (2006). Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci.* **15**, 352–361.
27. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
28. McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Med. Decis. Mak.* **9**, 190–195.
29. Gribskov, M. & Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **20**, 25–33.
30. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M. *et al.* (2005). CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33**, D192–D196.
31. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.