Sullivan, J., Holsinger, K. E., and Simon, C. (1996). The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* **42,** 308–312.

Sullivan, J., Markert, J. A., and Kilpatrick, C. W. (1997). Phylogeography and molecular systematics of the *Peromyscus aztecus* group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst. Biol.* **46,** 426–440.

Swofford, D. L. (1998). "PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)," Version 4.0b10a. Sinauer Associates, Sunderland, MA.

Swofford, D. L., and Sullivan, J. (2003). Phylogenetic inference using parsimony and maximum likelihood using PAUP*. *In* "The Phylogenetic Handbook" (M. Salemi and A. M. Vandamme, eds.). Cambridge University Press, Cambridge, UK.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics," (D. M. Hillis, C. Moritz, and B. K. Mable, eds.), 2nd Ed., pp. 407–514. Sinauer, Sunderland, MA.

Waddell, P. (1995). "Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood." Ph.D. Thesis, Massey University, Palmerston North, New Zealande.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10,** 1396–1401.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39,** 105–111.

Yang, Z., Goldman, N., and Friday, A. (1994). Comparison of models for nucleotide used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11,** 316–324.

[40] Context Dependence and Coevolution Among Amino Acid Residues in Proteins

By ZHENGYUAN O. WANG and DAVID D. POLLOCK

Abstract

As complete genomes accumulate and the generation of genomic biodiversity proceeds at an accelerating pace, the need to understand the interaction between sequence evolution and protein structure and function rises in prominence. The pattern and pace of substitutions in proteins can provide important clues to functional importance, functional divergence, and adaptive response. Coevolution between amino acid residues and the context dependence of the evolutionary process are often ignored, however, because of their complexity, but they are critical for the accurate interpretation of reconstructed evolutionary events. Because residues interact with one another, and because the effect of substitutions can depend on the structural and physiological environment in which they occur, an accurate science of evolutionary functional genomics and a complete understanding of selection in proteins require a better understanding of how context dependence affects protein evolution. Here, we present new evidence from vertebrate cytochrome oxidase sequences that pairwise coevolutionary

0076-6879/05 $35.00

interactions between protein residues are highly dependent on tertiary and secondary structure. We also discuss theoretical predictions that impinge on our expectations of how protein residues may interact over long distances because of their shared need to maintain protein stability.

Introduction

Perhaps the most important and well-known use of evolutionary inference in protein biochemistry is the relationship between functional importance and evolutionary conservation. Beginning graduate students studying a novel protein learn that to knock out function, the best places to mutate the protein are the most conserved sites. This relationship is sometimes viewed as almost a tautology, so conserved sites are believed to be functionally important by definition, but surveys of many proteins have revealed that residue conservation can be well predicted based on a combination of distance from active sites and distance from the hydrophobic core (Dean and Golding, 2000). An important development based on this relationship has been that changes in residue conservation can be viewed (again, sometimes tautologically) as strong predictors of changes in the function of those residues. In a somewhat counterintuitive twist, accelerated evolution can also be used as a predictor of functional importance, because the selective forces underlying accelerated evolution (whether long-term diversifying evolution or short-term adaptive bursts) are unlikely to operate on functionally neutral residues.

Although a simple interpretation of the relationship between divergence rates and functional importance has been highly successful (particularly the relationship between absolute conservation and functional importance), it ignores the potential for interaction among residues and the likelihood that functional importance may change over the normal course of evolution. Most evolutionary analyses rely on the assumption that the probabilities of substitution at each site are independent of substitutions at other sites, although protein structure and function result from interactions among amino acids, and this assumption cannot be true in principle. Although hydrophobic effects may be largely additive, hydrogen bonds, charge interactions, and van der Waals interactions among residues are all highly dependent on the size and physicochemical nature of interacting amino acid residues. Such interdependence of physical interactions seems bound to lead to interdependence, or coevolution, in the evolutionary process, and coevolution has indeed been detected on numerous occasions (Atchley *et al.*, 2000; Chelvanayagam *et al.*, 1997; Fukami-Kobayashi *et al.*, 2002; Gobel *et al.*, 1994; Govindarajan *et al.*, 2003; Korber *et al.*, 1993; Lapedes *et al.*, 1997; Neher 1994; Pazos *et al.*, 1997; Pollock and Taylor, 1997; Pollock *et al.*, 1999; Pritchard *et al.*, 2001; Shindyalov *et al.*, 1994;

Taylor and Hatrick, 1994; Tuff and Darlu, 2000; Valencia and Pazos, 2002; Wollenberg and Atchley, 2000). Interdependence should also lead to changes in rates at individual sites during the normal course of evolution, and such rate changes have been found to occur regularly in the absence of functional change (Gribaldo *et al.*, 2003; Lopez *et al.*, 2002; Philippe *et al.*, 2003), sending a loud warning to those who would define functional divergence as synonymous with rate change.

Despite regular detection of coevolution, results have not been obviously consistent as to the conditions and manner in which coevolution apparently occurs. The strongest pairwise signal comes from residues stacked in alpha helices (Pollock *et al.*, 1999), but the strength of pairwise coevolution between more distant residues appears to vary (Pollock, 2002; Pollock *et al.*, 1999), and interaction between protein subunits has had tantalizing but limited success (Fukami-Kobayashi *et al.*, 2002; Pazos *et al.*, 1997; Pazos and Valencia, 2001; Valencia and Pazos, 2002). One reason for the difficulty in consistently detecting coevolution has been that the majority of methods employed ignore phylogenetic relationships, which adds considerable noise and reduces the power of the methods (Pollock, 2002; Pollock and Taylor, 1997). Nevertheless, results from methods that incorporate phylogeny into development of a statistic (Chelvanayagam *et al.*, 1997; Pollock *et al.*, 1999; Shindyalov *et al.*, 1994) indicate that other factors are also at play. These may include the number of sequences analyzed, the depth of the evolutionary relationship between the sequences, the structural or functional context of residues in the sequences analyzed, adaptive bursts, or rate accelerations, and the potentially variable and dispersed nature of coevolutionary interactions between residues.

Using a phylogeny-based method (Pollock *et al.*, 1999), we have analyzed the coevolution of cytochrome *c* oxidase (Wikström, 2004) subunit I (COI) from a large sample of 231 vertebrates, all of which have had their mitochondrial genomes completely sequenced. The large number of genes available from these species allowed us to obtain phylogenetic trees that were only slightly dependent on substitutions in the gene of interest. As the central functional component of the CO complex, a large portion of COI consists of transmembrane helices, heme-binding regions, electron channels, and proton tunnels, as well as some intermembrane and matrix regions, providing many different structural and functional contexts. We are undertaking a detailed serial investigation of all the mitochondrially encoded members of the oxidative phosphorylation complex, and COI was chosen as the first subject partly because of its functional importance and generally conserved evolutionary rate, which indicates that much of the protein will have been in a similar evolutionary context throughout the vertebrate phylogenetic tree. There has been evidence of adaptive evolution in cytochrome oxidase in primates (Goldberg *et al.*, 2003; Wu *et al.*,

2000). We also present some results from COII from the same taxa for comparison. Before analysis, we clustered amino acids at each site according to volume, polarity, and hydrophobicity, and we analyzed the sites with slow substitution rates in greater detail, again to focus on sites for which the structural and functional context might not have evolved much during the range of evolutionary time we are considering. There was some dependency on the physicochemical vector used for clustering, but our main interests here are the stronger correlation of coevolutionary signal with physical distance in the transmembrane domain than within or between other domains, and the tendency for coevolved sites to colocalize with functionally critical regions. We, thus, present only the results for the polarity vector. The weak physical relation of coevolved sites in some protein regions is discussed in terms of theory on protein stability.

Methods

Choice of Sequences

Two critical factors that influence choice of sequence datasets for context-dependent evolutionary analysis are the number of sequences and their distribution, that is, the relationships among them. For there to be coevolution, there must be evolution, and it is, therefore, pointless to include identical or nearly identical sequences, but beyond that it is useful to include sequences that are closely related so that not too many changes (perhaps only a handful) have occurred along most branches. This allows the pinpointing of most replacement changes along the tree, avoids excess random co-occurrence of change along branches, and allows the presumption that the overall context has not changed too much over the course of evolution being examined. If the context changes dramatically and repeatedly, it is to be expected that coevolutionary relationships between sites will also change, and therefore, the signal will be overwhelmed by noise and difficult to detect. For alignment-only methods that ignore phylogenetic relationships, sequences should be as distant as possible to reduce the influence of phylogenetic relationships and to be consistent with the assumption of the methods that all sequences are independent examples of the protein; distant sequences are incompatible with the goal of a relatively constant contextual environment, however, and issues of alignment accuracy can also become a problem for these methods.

Here, vertebrate protein-coding sequences from complete mitochondrial genomes were obtained from GenBank and underwent automated alignment using ClustalW (Thompson *et al.*, 1994) in our EGenBio database. After removing sites involved in multiple insertions and deletions, a

phylogenetic tree was constructed from distances calculated using Phylip's ProtDist module (Felsenstein, 1989) and using the neighbor-joining (NJ) heuristic (Saitou and Nei, 1987). Branch lengths were modified using Phylip's ProML and PAM matrices. This tree was trimmed to remove as many long branches or obviously incorrect relationships as possible, ultimately resulting in a dataset of 231 species. The accuracy of the tree topology used and whether to consider a distribution of tree topologies (such as derived from a Bayesian posterior probability distribution or from a bootstrap analysis) are important issues, but they are not central to our discussion of context-dependent change, and we do not consider them here.

Availability of Structure

The availability of three-dimensional structural information for proteins under study is essential for interpreting the relationship of coevolutionary interactions and how they are affected by structure and function. Obviously, we sometimes would like to use coevolutionary analyses to predict structural features and interactions, but to study the question of how structural context affects coevolution one or more high-resolution crystal structures are essential, and it is preferable that at least one should be within the phylogenetic tree under consideration. Homology modeling to predict local structure can be performed if only distantly related structures are available, but this reduces the precision of structural inferences. Here, we visualized coevolved residues on the structure of cytochrome oxidase (including all three mitochondrial-encoded subunits) from bovine heart (1OCR) at 2.35 angstroms resolution (Tsukihara *et al.*, 1996;, 2003). The relationship between coevolution and structure was evaluated by calculating the distance between the Cα atoms (Cα distance) and by the location of the pairs, that is whether they were in the transmembrane domain (TM), or one of the surface domains (S), on either the intermembrane (IM) or matrix (M) side, or between the transmembrane and surface domains (Across, A). We also considered whether pairs were part of secondary structure elements (e.g., alpha helices or beta sheets), but the transmembrane regions are almost entirely alpha helix in nature. Cα distances were clustered into bins of 4.0 angstrom width for comparison of total domain distributions with the distributions of proposed "coevolving" sites, and comparisons were carried out with the standard G test.

Analytical Approach and Statistical Considerations

The choice of an analytical approach will undoubtedly affect the outcome of coevolutionary analysis, but because there is so little information about how coevolution occurs in real proteins, the choice is debatable and

not obvious. An approach pioneered by Shindyalov *et al.* (1994) is to evaluate coincident changes along branches. This ignores which amino acids are replaced, although this can be evaluated on an *ad hoc* basis (Chelvanayagam *et al.*, 1997). This method may be strongly affected by inaccuracies in topological inference and by bias in ancestral reconstruction (Krishnan *et al.*, 2004), although such problems can be accounted for, in theory. In principle, this method should do well for detecting coevolution that is nearly simultaneous and if coevolution occurs randomly with respect to physicochemical parameters and amino acids states. Residue-based approaches, in contrast, have potentially greater power to detect coevolution if there is some consistency with regards to amino acids, for example if charge matters, or if there is an energetic need to maintain the volume occupied by hydrophobic side-chain groups in a particular region of the protein structure. The main difficulty with residue-based approaches is that there can be a large number of parameters. Information-theory approaches (Atchley *et al.*, 2000; Korber *et al.*, 1993; Lapedes *et al.*, 1997; Wollenberg and Atchley, 2000), for example, consider whether there are significant associations between states, but for pairs of sites with only 5 of the 20 amino acids each, there are still (at least) 25 parameters to be estimated. Although problems with these methods are often confounded by the absence of phylogenetics in developing the statistic, post-analysis simulations reveal that over-parameterization is a serious hindrance to obtaining reliable results (Wollenberg and Atchley, 2000).

In our "LnLCorr" methodology (Pollock *et al.*, 1999), we avoid problems of over-parameterization by clustering the amino acids into groups or physicochemical "states." The logic behind this is that there may be a primary axis of coevolution with respect to physicochemical properties, and the method will be most powerful if this is so and if the axis is correctly identified. Because the method compares the likelihood ratio of a coevolutionary model of evolution for pairs of sites with that of an independent model, it does not need to estimate ancestral states, and the fewer number of parameters means that it is fairly robust. The power of the method is dependent on the choice of methodology for clustering amino acids, and it is, therefore, generally best to choose at least a few different methods for comparison. Here, for simplicity, we present only the results from clustering according to a vector of polarity. The rate of evolution also matters, both because the rate can affect the ability of a method to detect coevolving sites, and because the same factors that affect the rate of evolution at a site may also affect the likelihood that the site will coevolve with other sites. Here, again for simplicity, we present only the results of coevolutionary analysis among the slowest evolving (most "conserved") half of the sites, which in this analysis had a greater relationship with distance in the

three-dimensional structure. Although we have extended our models to allow more than two groups, here we consider only the two-group model delineated by Pollock *et al.* (1999).

A prime reason coevolutionary analysis is difficult and results are hard to interpret is the large number of comparisons, which increase with the square of the number of sites considered. With thousands of comparisons made, this leads to a large multiple-comparisons problem when evaluating significance. One approach is to consider only sites that are still significant after correcting for the number of comparisons (e.g., a Bonferroni correction), but such approaches sacrifice a great deal of power; it cannot be expected that many coevolving sites will be paired strongly enough to lead to extreme levels of significance, and at such extreme levels of significance the lack of data, overparameterization relative to the amount of data, inaccuracies of the model and even small inadequacies of the methodology may overwhelm the results. The approach we take instead is to find the sites with coevolution statistics that are greater than prespecified "significance" levels (i.e., .05, .01, and .002), and consider both whether the number of such sites is greater than expectation and whether the distribution of such sites in the crystal structure is perturbed relative to the distribution of all sites in the same category. By taking such an approach, we can also evaluate the posterior probability that these sites have coevolved or alternatively that they have not coevolved (i.e., the expected number divided by the observed number). Significance levels for values of the likelihood ratio (or any other statistic) need to be determined by parametric bootstrapping, because the chi-square distribution cannot be assumed for coevolution analyses (Pollock *et al.*, 1999). Here, we simulated 6000 pairs for each data comparison, with values sampled randomly from the maximum likelihood estimators.

Results

The strong dependence of coevolutionary results on structural and functional context was demonstrated by the differences among within-domain analyses, across-domain analyses, and the two different subunits. All comparisons showed significantly greater numbers of residue pairs than expected at all significance levels (Table I). COII had the largest excesses, whereas within the surface domain (S) and between domains (A) in COI had the smallest. The relationship between coevolutionary predictions and structural distance also varied greatly among comparisons (Fig. 1). The transmembrane regions showed the clearest relationship between coevolution and distance, with a large excess of closely paired sites in the coevolutionary fractions. The clearest difference between the coevolved

TABLE I
EXPECTED AND OBSERVED COEVOLVING PERCENTAGES AND TOTAL NUMBER OF PAIRS ANALYZED

		COI all sites		COI conserved		COII conserved	
Expected	Location[a]	%	No.[b]	%	No.[b]	%	No.[b]
5%	S	11.1	1071	12.7	355	17.7	5671
	TM	15.1	5778	16.5	2701		
	A	11.0	7794	13.3	3031		
1%	S	3.7		5.4		7.1	
	TM	6.5		6.7			
	A	3.7		3.9			
0.2%	S	1.3		2.5		4.0	
	TM	2.8		3.9			
	A	1.9		1.9			

[a] COI comparisons were within the surface (S), within the transmembrane (TM), or between the surface and transmembrane domains (A), whereas comparisons in COII were within the entire protein.
[b] The total numbers of sites for each comparison are shown only once, in the top row.

fraction and the total distribution of transmembrane sites was seen for the .2% significance level cutoff (Fig. 1); for higher significance levels, the differences between the distributions are smaller, though still highly significant, and the number of excess close sites is larger than at the .2% level. This indicates that many sites that coevolve due to physical proximity occur within the 5–1% and 1%–0.2% ranges, but that the physically close sites make up a smaller proportion of the sites (this is to be expected, if for no other reason that the expected number of background sites is increasing fivefold between adjacent categories).

Within the surface domain, there are many fewer coevolving sites, but strikingly it appears that coevolution occurs between sites that are close and between those that are distant, but not between those that are moderately close (Fig. 1). This is consistent with earlier results for surface residues of myoglobin (Pollock *et al.*, 1999) and is probably due to maintenance of charge interactions and the charge distribution across the protein. The distance distribution of the closest pairs is different than for the TM analysis, and all four of the more distant coevolving pairs are within the M domain, rather than the IM domain. The coevolving sites from the across-domain comparison (A) show the smallest effect of physical distance (Fig. 1), and most of the excess close sites appear to occur as interactions at the boundary of the transmembrane and surface domain, at the end of the transmembrane helices (unlike many soluble proteins, the domain

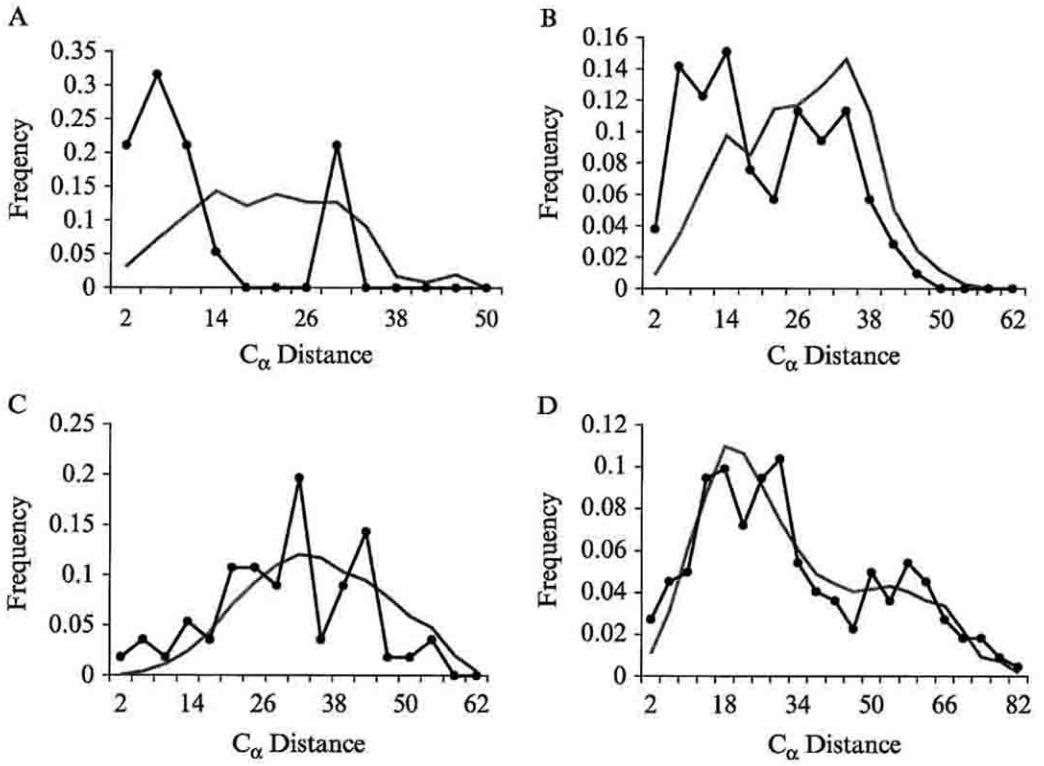

FIG. 1. Structural distance distributions for coevolving residues. Distance (Cα) frequency distributions are shown for all residue pairs (solid lines) and for hypothetical coevolving pairs (circles) within the transmembrane domain of COI (A), within the surface domains of COI (B), across domains of COI (C), and for all pairs in COII (D). The hypothetical coevolving residue pairs shown are for the 0.2% significance level, except for the surface domains that had many fewer sites and for which results at the 1% level are shown.

definitions in COI and many other transmembrane proteins are such that the amino acid chain goes in and out of the different domains repeatedly). As with the TM comparison, for both the S and A comparisons, the larger significance values produce distance distributions more similar to the overall distribution. The distributions of hypothetically coevolved sites in COII (Fig. 1), in contrast to the COI analyses, were not significantly different from the overall distribution of sites.

Concluding Remarks

Pairwise coevolution in vertebrate COI is closely related to distance in the three-dimensional structure, and the correlation with distance is strongest among sites located in the functionally critical transmembrane domain than it is within the two surface domains or across domains. The strongly coevolving pairs were often at the end of helices, echoing the results of Pollock *et al.*

(1999) for vertebrate myoglobin. Interestingly, coevolution appears to be strongest in the functionally critical regions of COI, whereas in COII, which is further from the active site of the CO complex, predicted coevolutionary pairs of sites had no obvious relationship to structural distance.

Although there are clear trends in the relationship of hypothetically coevolving pairs with structural distance, there is also clearly an excess number of coevolving sites that have nothing to do with physical distance. Within COII there was no apparent relationship with distance, even though 7% of the sites were beyond the 1% significance level and 4% were beyond the .2% level (thus, if these predictions are correct, 86% and 95% of the site pairs at these levels have truly coevolved). Possible failures in the topological reconstruction or the model used cannot explain this discrepancy by themselves, because the topology and model are common between the analyses. One possible explanation is that there are adaptive bursts or other forms of variation in the replacement rate along specific branches. Such bursts may tend to be distributed around the protein and would be correlated in evolution only because of a common causal agent. Indeed, in one of the lineages we removed, that of snakes, there were many apparently coevolved pairs that were otherwise conserved throughout vertebrate evolution, and the coevolutionary signal may have been due to an adaptive burst along this lineage. Other explanations may have to do with functionality and exposure to the environment. COI is at the functional core of the CO complex, whereas COII, like COIII and the 10 nuclear-encoded CO complex subunits, is on the periphery, surrounding COI. This may mean that COII has interactions with outside factors that COI is shielded from, and the effect of these outside factors would then be distributed along the elongated COII protein. It may also be that the important functional role of COI, and particularly of the transmembrane helices, leads to tighter pairwise interactions.

Finally, it is worth considering coevolutionary results in an energetic framework. Folding and protein stability may generally be viewed as a global protein variable (Williams *et al.*, 2001; Williams *et al.*, in review; Xu *et al.*, review), and it is easy to conceive that a slightly destabilizing replacement in one part of the protein may be compensated by a replacement leading to greater stability in a distant part of the protein. Certainly mutation studies have long shown that compensatory mutations can occur over long distances in a protein (Brasseura *et al.*, 2001). COII may be selected mostly to bind COI and the other adjacent subunits in the CO complex, so only the overall binding coefficient matters. If this is the case, future developments in coevolutionary analysis should probably be aimed to distinguish which patterns of coevolution are associated with structural distance and which are not, in order to build models that are not only powerful for detecting any kind of coevolution, but also capable of

discriminating between different kinds of coevolution, some of which may be of greater interest for a particular goal.

Acknowledgments

This work was supported by grants from the National Institutes of Health (GM065612-01 and GM065580-01), the National Science Foundation (EPS-0346411), and the State of Louisiana Board of Regents (Support Fund, Research Competitiveness Subprogram LEQSF (2001-04)-RD-A-08 and the Millennium Research Program's Biological Computation and Visualization Center), and Governor's Biotechnology Initiative.

References

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. (2000). Correlation among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol. Biol. Evol.* **17,** 164–178.

Brasseura, G., Ragob, J.-P. D., Slonimskic, P. P., and Lemesle-Meuniera, D. (2001). Analysis of suppressor mutation reveals long distance interactions in the bc1 complex of *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta Bioenergetics* **1506,** 89–102.

Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Connet, G. H., and Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10,** 307–316.

Dean, A. M., and Golding, G. B. (2000). Enzyme evolution explained (sort of). *Pac. Symp. Biocomput.* 6–17.

Felsenstein, J. (1989). Phylogeny inference package. *Cladistics.* **5,** 164–166.

Fukami-Kobayashi, K., Schreiber, D. R., and Benner, S. A. (2002). Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol. Biol.* **319,** 729–743.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins.* **18,** 309–317.

Goldberg, A., Wildman, D. E., Schmidt, T. R., Huttemann, M., Goodman, M., Weiss, M. L., and Grossman, L. I. (2003). Adaptive evolution of cytochrome c oxidase subunit VIII in anthropoid primates. *PNAS.* **100,** 5873–5878.

Govindarajan, S., Ness, J. E., Kim, S., Mundorff, E. C., Minshull, J., and Gustafsson, C. (2003). Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. *J. Mol. Biol.* **328** .

Gribaldo, S., Casane, D., Lopez, P., and Philippe, H. (2003). Functional divergence prediction from evolutionary analysis: A case study of vertebrate hemoglobin. *Mol. Biol. Evol.* **20,** 1754–1759.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc. Natl. Acad. Sci. USA* **90,** 7176–7180.

Krishnan, N. M., Seligmann, H., Stewart, C.-B., de Koning, A. P. J., and Pollock, D. D. (2004). Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference. *Mol. Biol. Evol* **21**(10), 1871–1883.

Lapedes, A. S., Giraud, B. G., Liu, L. C., and Stormo, G. D. (1997). Correlated Mutations In Protein Sequences: Phylogenetic and Structural Effects. Correlated Mutations in Protein. Proceedings of the AMS/SIAM Conference: Statistics in Molecular Biology. Seattle, WA.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19,** 1–7.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. USA* **91,** 98–102.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.* **271,** 511–523.

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14,** 609–614.

Philippe, H., Casane, D., Gribaldo, S., Lopez, P., and Meunier, J. (2003). Heterotachy and functional shift in protein evolution. *IUBMB Life* **55,** 257–265.

Pollock, D. D. (2002). Genomic diversity, phylogenetics and coevolution in proteins. *Applied Bioinformatics* **1,** 25–36.

Pollock, D. D., and Taylor, W. R. (1997). Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.* **10,** 647–657.

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues: Maximum likelihood identification and relationship to structure. *J. Mol. Biol* **287,** 187–198.

Pritchard, L., Bladon, P., Mitchell, J. M. O., and Dufton, M. J. (2001). Evaluation of a novel method for the identification of coevolving protein residues. *Protein Eng.* **14**(8), 549–555.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.

Shindyalov, I., Kolchanov, N., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7,** 349–358.

Taylor, W., and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7,** 341–348.

Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., and Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 A. *Science* **272,** 1136–1144.

Tsukihara, T., Shimokata, K., Katayama, Y., Shimada, H., Muramoto, K., Aoyama, H., Mochizuki, M., Shinzawa-Itoh, K., Yamashita, E., Yao, M., Ishimura, Y., and Yoshikawa, S. (2003). The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process. *Proc. Natl. Acad. Sci. USA* **100,** 15304–15309.

Tuff, P., and Darlu, P. (2000). Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol. Biol. Evol.* **17,** 1753–1759.

Valencia, A., and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12,** 368–373.

Wikström, M. (2004). Cytochrome c oxidase: 25 years of the elusive proton pump. *Biochim. Biophys. Acta Bioenergetics* **1655,** 241.

Williams, P. D., Pollock, D. D., and Goldstein, R. A. (2001). Evolution of functionality in lattice proteins. *J. Mol. Graph Model* **19,** 150–156.

Williams, P. D., Pollock, D. D., and Goldstein, R. A. (in revision). Why are proteins marginally stable? II: Functionality strikes back. Proteins: Structure, function, and genetics.

Wollenberg, K. R., and Atchley, W. R. (2000). Separation of phylogenetic and functional association in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci. USA* **97,** 3288–3291.

Wu, W., Schmidt, T. R., Goodman, M., and Grossman, L. I. (2000). Molecular evolution of cytochrome c oxidase subunit I in primates: Is there coevolution between mitochondrial and nuclear genomes? *Mol. Phylogenet. Evol.* **17,** 294–304.

Xu, Y. O. Hall, R. W., Goldstein, R. A., and Pollock, D. D. (in review). Divergence, recombination, and retention of functionality during protein evolution. Human Genomics.