J Mol Evol (2007) 65:485–495
DOI 10.1007/s00239-007-9018-8

Coevolutionary Patterns in Cytochrome c Oxidase Subunit I Depend on Structural and Functional Context

Zhengyuan O. Wang · David D. Pollock

Published online: 23 October 2007

Abstract The strength and pattern of coevolution between amino acid residues vary depending on their structural and functional environment. This context dependence, along with differences in analytical technique, is responsible for the different results among coevolutionary analyses of different proteins. It is thus important to perform detailed study of individual proteins to gain better insight into how context dependence can affect coevolutionary patterns even within individual proteins, and to unravel the details of context dependence with respect to structure and function. Here we extend our previous study by presenting further analysis of residue coevolution in cytochrome *c* oxidase subunit I sequences from 231 vertebrates using a statistically robust phylogeny-based maximum likelihood ratio method. As in previous studies, a strong overall coevolutionary signal was detected, and coevolution within structural regions was significantly related to the C_α distances between residues. While the strong selection for adjacent residues among predicted coevolving pairs in the surface region indicates that the statistical method is highly selective for biologically relevant interactions, the coevolutionary signal was strongest in the transmembrane region, although the distances between coevolving residues were greater. This indicates that coevolution may act to maintain more global structural and functional constraints in the transmembrane region. In the transmembrane region, sites that coevolved according to polarity and hydrophobicity rather than volume had a greater tendency to colocalize with just one of the predicted proton channels (channel H). Thus, the details of coevolution in cytochrome *c* oxidase subunit I depend greatly on domain structure and residue physicochemical characteristics, but proximity to function appears to play a critical role. We hypothesize that coevolution is indicative of a more important functional role for this channel.

Keywords Cytochrome *c* oxidase · Coevolution · Markov chain · Proton pump channels · Residue interaction

Electronic Supplementary Material The online version of this article (doi:10.1007/s00239-007-9018-8) contains supplementary material, which is available to authorized users.

Z. O. Wang · D. D. Pollock
Department of Biological Sciences and Biological Computing
and Visualization Center, Louisiana State University,
Baton Rouge, LA 70803, USA

Z. O. Wang
Genome Sequencing Center, Department of Genetics,
Washington University School of Medicine, St. Louis,
MO 63108, USA

D. D. Pollock (✉)
Biochemistry and Molecular Genetics, University of Colorado
School of Medicine, Aurora, CO 80045, USA
e-mail: David.Pollock@uchsc.edu

Introduction

The structural and functional integrity of proteins serves as a constraint on patterns of amino acid substitution during evolution. Evolutionary analysis can therefore facilitate and extend the study of protein structure and function. The relationship between functional importance and evolutionary conservation, for example, is well established; residue conservation is commonly used for predicting the effect of residue mutations. A common limitation of such analyses, however, is that they frequently ignore interactions among residues and assume that substitutions at different sites are independent. This assumption limits the power of an evolutionary approach. Hydrogen bonds, charge interactions, hydrophobic effect, and van der Waals interactions are all highly dependent on the size and physicochemical nature of interacting amino acid residues. Protein stability is thus dependent on the precise nature of these interactions, and amino acid interdependency can lead to coevolution (dependent substitutions at interacting residues). Since the nature and strength of residue interactions vary according to the residues involved and their local and global environments in proteins, coevolution exhibits a complex context dependence (Wang and Pollock 2005). Unraveling this dependence through detailed coevolutionary analysis of individual proteins can provide valuable information related to protein structure and function.

Previous coevolutionary studies have supported the widespread existence of coevolution in proteins (Atchley et al. 2000; Dutheil et al. 2005; Fukami-Kobayashi et al. 2002; Govindarajan et al. 2003; Neher 1994; Pollock and Taylor 1997; Pollock et al. 1999; Pritchard et al. 2001; Shindyalov et al. 1994; Taylor and Hatrick 1994; Tuffery and Darlu 2000; Valencia and Pazos 2002; Wollenberg and Atchley 2000). Conclusions drawn from coevolution are, however, inconsistent, partly due to the wide variety of data sets and methodologies used. Spatial distance between residues is often involved in coevolution, but the strength of coevolution between both distant and proximal residue pairs appears to vary (Pollock et al. 1999). Residues in α-helices exhibited the strongest pairwise coevolution in myoglobin (Pollock et al. 1999), while highly mobile loops with ligand-binding functions had the strongest signal in DHFR (dihydrofolate reductase), cyclophilin, and formyltransferase (Saraf et al. 2003). Charge compensation has been identified as a strong coevolutionary force (Chelvanayagam et al. 1997; Fukami-Kobayashi et al. 2002; Pollock et al. 1999), but Kondrashov et al. (2002) did not find any swapping of positively and negatively charged residues in their analysis. These conflicting results indicate that current coevolutionary methods and our understanding of the context dependence of coevolution are not mature.

Multiple factors are important for the efficiency of detecting coevolution. Incorporating phylogeny is essential to reduce noise (Fleishman et al. 2004; Fukami-Kobayashi et al. 2002; Pollock and Taylor 1997; Pollock et al. 1999), and a large amount of data in the form of a large number of sequences (adding sequence length is not particularly helpful) is vital (Pritchard et al. 2001). This helps to obtain appropriate sequence distance and sequence density, both of which improve the detection of coevolution (Chelvanayagam et al. 1997; Fukami-Kobayashi et al. 2002; Pollock and Taylor 1997; Pollock et al. 1999; Pritchard et al. 2001). Recombination within nuclear genes may also add noise by allowing different gene regions to have different coalescent histories and, thus, slightly different phylogenies.

In consideration of these factors, in the present study we describe coevolutionary analysis of a large data set of 231 cytochrome c oxidase (CO) subunit I (COI) homologous sequences from vertebrates using the methodology of Pollock et al. (1999), which incorporates phylogeny and uses the likelihood ratio LR test. Some preliminary results from this study were presented previously (Wang and Pollock 2005), but here we present the full results and focus in detail on the relationship of predicted coevolutionary pairs to functional areas in the transmembrane region. To overcome the problem of local maxima in the LR test we employed a stochastic searching algorithm using a Markov chain, similar to Markov chain Monte Carlo (MCMC) processes (Hastings 1970), in addition to previously described peak-searching methods. As in previous studies, we segregated residues according to their physicochemical characteristics to reduce computation and noise. Segregation makes possible a two-state independent model for each site and a four-state dependent model with only one more degree of freedom (see Materials and Methods). If amino acids were analyzed directly, similar independent models would have 20 states for each site and dependent models would have 400 states, making analysis computationally infeasible and raising serious questions about over-parameterization.

CO is the terminal complex of the respiration chain and functions as a redox-driven proton pump utilizing the free energy of oxygen reduction for creation of a proton gradient across a membrane (the inner membrane of mitochondria or the cell membrane of bacteria). The structures of CO from bovine (*Bos taurus*) mitochondria, *Paracoccus denitrificans*, and *Rhodobacter sphaeroides* have been determined (Iwata et al. 1995; Svensson-Ek et al. 2002; Tsukihara et al. 1996). Mitochondrial CO exists as a dimer, with each monomer comprised of 13 subunits, while monomeric bacterial CO consists of only 4 subunits. Despite the differences in the number of subunits between mitochondria and bacteria, the function of CO and the core structure, which is composed of three subunits (COI, COII, and COIII), are conserved. There are four functionally important redox centers (CuA, haem *a*, haem a_3, and CuB); CuA is in COII and the other three centers are in COI. COI is the central functional component of the CO complex and consists of 12 transmembrane helices separated by surface loops, a catalytic site, and electron and proton channels, in addition to two relatively small surface domains (Tsukihara et al. 1996). The transmembrane helices jointly form a cylinder-like structure with their ends at the two membrane surfaces. The surface regions consist of mainly loops and several short α-helices that connect the 12 transmembrane helices and the N- and C-termini. These features of COI provide different structural and functional contexts for coevolutionary analysis.

Since COI, COII, and COIII are encoded by the maternally inherited nonrecombining mitochondrial genome, possible noise introduced by recombination is not a concern. In addition, the conservation of COI means that most of the protein has been in relatively consistent structural and functional contexts over evolutionary time, thereby reducing the complexity of coevolution introduced by changes in structural or functional context. Furthermore, COI is an integral membrane protein. It is of interest to analyze the specific coevolutionary pattern of this kind of protein compared to that of soluble proteins such as myoglobin (Pollock et al. 1999).

The extended study described here provides further details of the structural and functional context dependence of residue coevolution in COI. Contextual factors analyzed here include protein structure and residue physicochemical characteristics but focus on functional context in the transmembrane region. The study suggests that coevolutionary analysis may be especially useful in analyzing interacting networks of residues, such as those involved in the proton and electron channels of COI, and that coevolution can be especially prevalent adjacent to functionally important regions, possibly due to adaptive coevolution. We also show that the LR test implemented using our stochastic searching algorithm worked well for coevolution detection given the large number of sequences that were available. Compared to previous studies, the large number of sequences improved the predictability of the likelihood ratio distribution under the null model.

Materials and Methods

Mitochondrial sequences from 368 vertebrata were automatically downloaded from our EGenBio database (Faith and Pollock 2003; Nahum et al. 2006) and all 13 proteins encoded by the mitochondrial genome were aligned with ClustalX (Thompson et al. 1994, 1997). Sites involved in multiple insertions and deletions were removed. A phylogenetic tree relating these 368 species was reconstructed using PROTDIST (distances based on the gamma distribution) and FITCH from the PHYLIP package (Felsenstein 1989). Branch lengths for this topology were recalculated from COI sequences alone using PROML from the PHYLIP package and JTT matrix (Jones et al. 1992). Part of the reason that we selected JTT was because it gave a higher likelihood for the phylogeny. Sequences that were in particularly egregious conflict with known (or strongly believed) phylogenetic relationships, or which had particularly long branches (e.g., those of snakes, which are more than half again longer than their neighbors), were removed to reduce potential noise arising from phylogenetic inaccurates. We note that the remaining phylogeny is almost certainly still incorrect in some details but that probable errors are small enough that they can be safely ignored for the purposes of the current study. To limit the number of short uninformative terminal branches, sequences separated by short branch lengths (<0.005) were also removed. The final phylogenetic tree had 231 sequences including the bovine sequence (Fig. 1). To relate the results from coevolutionary analysis to structure, the crystal structure of bovine heart mitochondrial CO (PDB ID: 2OCC) (Tsukihara et al. 1996) was obtained from the Protein Data Bank (Berman et al. 2000), and visualized using PyMOL (DeLano 2002).

The 231 aligned sequences, together with the phylogenetic tree, were analyzed using an LR test with likelihood maxima estimated using a stochastic searching process. Residues at each site were segregated into two groups (states) according to their hydrophobicity (Argos et al. 1982), polarity (Grantham 1974), or side-chain volume (hereafter referred to as “volume”) (Krigbaum and Komoriya 1979) by taking the mean value of the amino acids present at that site as the dividing point, as by Pollock et al. (1999). Previous analysis and visual inspection of results from individual pair likelihood surfaces has shown that sites that exhibit little variation tend to appear overparameterized in the dependent coevolutionary model, in that it is difficult to separately identify (and optimize) the rate and frequency parameters. We therefore excluded sites with <2% state variation (i.e., the frequency of the major state at that site was >0.98, and only four or fewer sequences contained the minor state) from subsequent analysis. As shown below, including these sites also contributes to greater deviation from the χ^2 approximation for the LR ratio distribution.

Two alternative models, independent and dependent, were constructed and are described in detail elsewhere (Pollock et al. 1999). In brief, there are two exchange parameters (one for the stationary frequency and one for the rate) for each site in the independent model and five total parameters for the dependent model (three for the stationary frequencies and two for the rates). The dependent model thus has one more parameter than the independent model (5-2*2), and the independent model is nested within the dependent model, meaning that it is a special case of the dependent model. Each of these models is defined by an instantaneous rate matrix, Q, and the transition probabilities over each branch of length t were computed through standard matrix manipulation (Pollock et al. 1999).

The likelihood of the models was computed following the pruning algorithm (Felsenstein 1981), and a stochastic search algorithm was implemented to traverse the

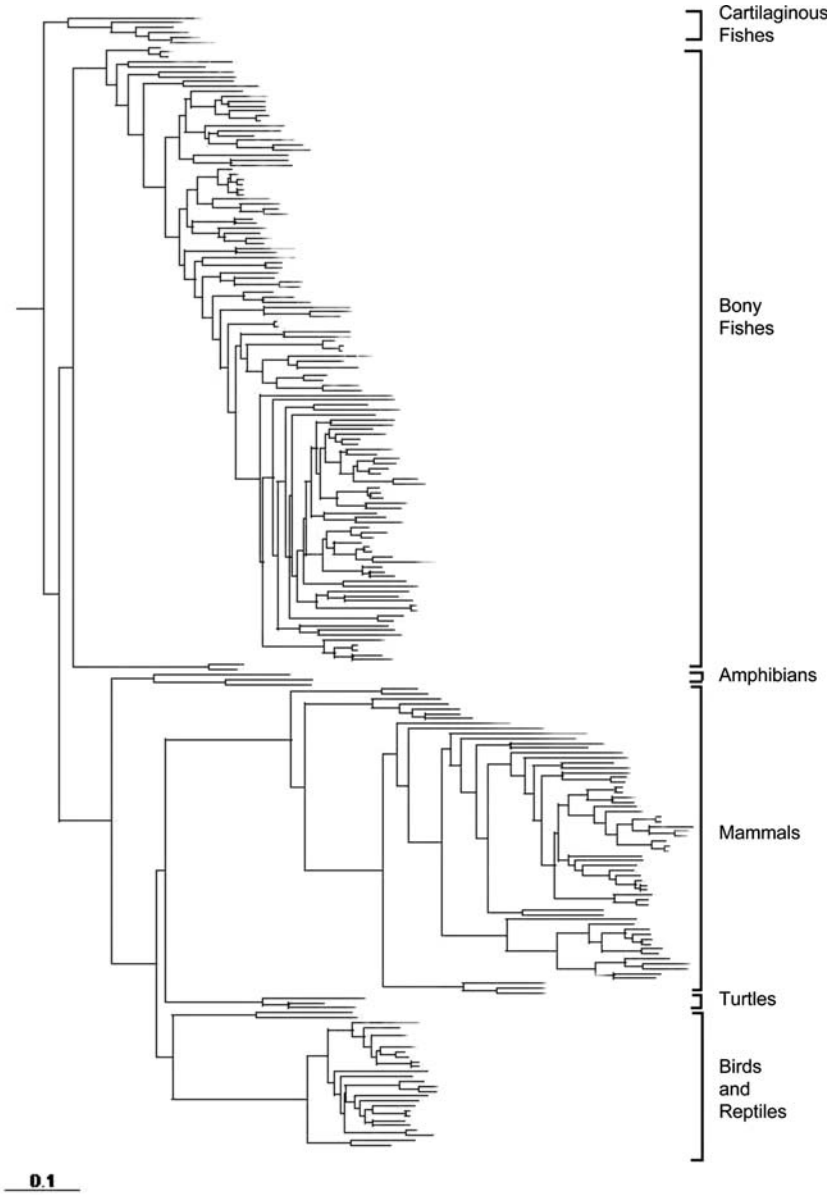


Fig. 1 Phylogeny of 231 vertebrates. The topology and branch lengths were reconstructed based on all 13 mitochondrial-encoded proteins using FITCH and PROML, respectively. A version of this tree with more detailed taxonomic information is provided in the supplementary data

likelihood surfaces and locate the maxima through a Markov chain process similar to MCMC (Hastings 1970). In the Markov process, stationary frequencies were proposed according to a Dirichlet distribution, and rates were proposed according to a bounded uniform distribution. The parameters of these distributions were tuned to achieve an average acceptance of about 50%. For example, the Dirichlet parameter was 120, and the uniform sampling ranged from ±0.25 compared to the current rate when the proposal was made. To be thorough, we performed 10,000 iterations for each chain, though the chain, in most cases, can reach equilibrium in fewer than 2000 iterations. As an aside, it should be made clear that the Dirichlet distributions used in the proposals do not affect calculation of the likelihood, and thus despite similarities to Bayesian search algorithms to define the posterior space or identify the maximum a posteriori (MAP) point, the calculations here are for the likelihood maxima. This search algorithm improved the detected coevolving percentages by more than half (compared to the data in our previous publications (Pollock et al. 1999; Wang and Pollock 2005). Also, we chose not to employ simulated annealing (sequentially lowering the "temperature" to sharpen the likelihood surface) because we felt it would be difficult to find one set of annealing conditions that would be applicable to all of the tens of thousands of likelihood surfaces examined here, and because we wanted to make seamless comparisons with this method and future implementations of Bayesian approaches to coevolutionary study.

A LR statistic, $LR = -2\ln(L_i / L_d)$ was used to evaluate the significance of coevolution between sites (L_i, the likelihood of the independent model; L_d, the likelihood of the dependent model). With coevolutionary analysis, it has been shown that this statistic cannot be assumed to have a χ^2 distribution (with 1 df) under the null (independent) model (Pollock et al. 1999). Therefore, to obtain more accurate distribution estimators, we performed parametric bootstrapping by simulating replicate data sets using the same phylogenetic tree and the maximum likelihood parameter estimates (MLEs) from the independent model (Pollock et al. 1999).

For the complete data set, distributions estimated from the bootstrap were nearly identical to χ^2_1 (see Results), and the LR significance cutoff values considered (6.6 for $p < 0.01$, 7.9 for $p < 0.005$, and 10.8 for $p < 0.001$) were the same as they would be for χ^2_1. We ran further bootstrap simulations to demonstrate what factors were important for the χ^2_1 to approximate the bootstrap well. This included randomly subsampling 116, 58, and 29 taxa to compare to the full 231-taxon data set, sampling 29 taxa from the primates and closely related mammals to demonstrate the topology dependence of the results, and including the conserved sites that had been previously removed.

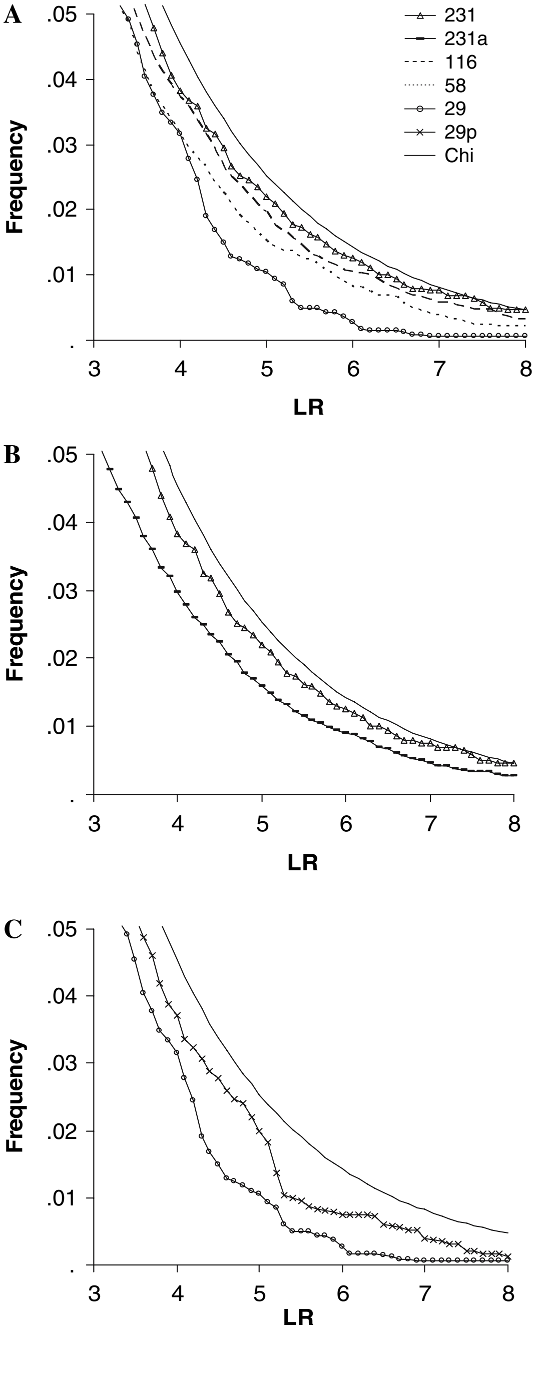
Since thousands of comparisons were performed in each analysis, the probability values at the cutoffs are not accurate estimators of the probability that each pair of sites coevolved. Instead, we considered the pairs with LRs beyond a particular probability cutoff to be a set of hypothetical coevolving pairs (for convenience, we call these simply "coevolving pairs") and compared the observed percentage of coevolving pairs (the "coevolving percentage") to the percentage of false "coevolving pairs" that would have been expected even if no coevolution had actually occurred (1%, 0.5%, and 0.1% for the respective cutoff values). The percentages were calculated as the number of coevolving pairs divided by the total number of pairs analyzed. If significantly more coevolving pairs were observed than were expected (Pollock et al. 1999), the posterior probability that a particular coevolving pair truly coevolved was calculated as 1 – (expected percentage/observed percentage). Here, the number of detected coevolving pairs was always significant. We consider this approach highly preferable to procedural algorithms to target a specific signal-to-noise level, such as "false discovery rate" controlling mechanisms (Benjamini and Yekutieli 2005), which can result in a great loss of power to detect coevolution. Nevertheless, in our case for a "false discovery rate" of 0.05, there were 193 detected coevolving pairs for hydrophobicity and similar numbers of pairs for other physicochemical vectors. Given the extremely large number of comparisons here, the presence of coevolution is highly significant even for LRs with false discovery rates >0.05. For example, if with multiple comparisons 20 pairs are expected beyond the 1% cutoff level but 200 pairs are observed, this is an extremely significant event, although 10% of these pairs probably occurred by chance. We note also that the simple formula used here is only an approximation of the true posterior probability, but under the conditions used (low primary cutoff levels, highly significant levels of excess predicted coevolving pairs), it should be sufficiently accurate.

The distance between two residues of a pair was defined as the distance between their C_α atoms in the monomer unit of the bovine crystal structure. Residues in the bovine structure were classified according to their location in bovine structural regions. The three regions (transmembrane region, intermembrane region, and matrix region) are defined by location, and it should be noted that the amino acid chain traverses back and forth through the membrane. Unlike classic domains, each region is therefore made up of discontinuous stretches of the polypeptide. All coevolved pairs were subdivided into three groups according to the locations of the residues in the pair: both sites located in the transmembrane region (TM), both sites located in one of the two surfaces regions (S), and each site in a different region (across regions, AR).

Results

Pairwise Coevolution in COI

Hypothetically coevolving pairs (or "coevolved pairs") were identified based on classification of alignment positions according to hydrophobicity, polarity, and volume. Estimates for the null distributions of the LR statistic obtained by parametric bootstrapping were approximately the same as the χ^2_1 distribution for these classifications (Fig. 2). Although it was shown previously that the null distribution for the LR statistic under these conditions often does not match χ^2_1, probably due to the limited number of substitutions at each site (Pollock et al. 1999), it appears that with this phylogeny of 231 sequences, and with the low variability positions removed, most sites contain sufficient data for the χ^2_1 approximation to hold. To better understand the reasons for this, we subsampled taxa in the data randomly (116, 58, and 29 taxa) and from primates and close relatives (29 taxa) and sampled the complete data set with the low variability positions put back in. As expected, the χ^2_1 approximation is better and better the more taxa that are included (Fig. 2), but the form of the simulated distribution is also highly dependent on the precise details of the phylogenetic relationship when the number of taxa is held constant. Furthermore, the inclusion of the low variability positions had a large effect on the deviation from the χ^2_1 approximation, equivalent in this case to the deviation seen when dropping to 58 taxa



◀**Fig. 2** Cumulative frequency distributions of the LR statistic. The cumulative frequency distributions shown are based on parametric bootstrapping using a constant phylogenetic tree (Fig. 1) and MLEs based on an independent model of evolution at each site. Simulations are shown with symbols, and the chi-square distribution with one degree of freedom (χ_1^2) is shown as solid gray line (χ). The symbol notations for the three graphs are shown together in the uppper right and are as follows: (**A**) simulated distributions for the full 231-sequence topology (triangles), along with random subsamples of 116 (dashed line), 58 (dotted line), or 29 (open circles) sequences; (**B**) results for the 231-sequence topology with (231a; horizontal bars) and without (triangles; as before) low-variability sites; (**C**) results for the random subsample of 29 sequences (open circles; as before) along with a sample of 29 sequences including all the primates plus a sample of close relatives (29p; line marked with X's)

without the conserved sites. In this case it can be seen that the χ_1^2 approximation is more conservative than the simulations, but this is not guaranteed to be the case with other topologies (Pollock et al. 1999).

The percentage of coevolving pairs predicted in COI was uniformly much higher than expected based on chance, assuming the null or independent model (Table 1). We considered the number of predicted coevolving pairs at three probability cutoffs ($p < 0.01$, p_1; $p < 0.005$, $p_{0.5}$; $p < 0.001$, $p_{0.1}$). At $p < 0.05$, about 17% of pairs were predicted to coevolve (data not shown), meaning that the posterior probability that such pairs were correct was only about 0.7. For lower-significance criteria, the enrichment for predicted coevolving pairs was much higher, however, so only the lower criteria were considered further (Table 1). For example, at $p_{0.5}$ about 10 times more predicted coevolving pairs were detected than expected by chance. Since we detected 378, 452, and 451 predicted coevolving pairs for the three respective physicochemical segregations, there are 341, 407, and 406 more predicted coevolving pairs than expected by chance in the multiple comparisons. This result is highly significant. At $p_{0.1}$, there are about 20 times more pairs than expected, which means that each predicted coevolving pair at this significance level has only about a 5% chance of being in error (posterior probabilities >0.95; Table 1), and is again highly significant. These numbers and enrichment factors are greater than those in previous studies, probably due mainly to the inclusion of a phylogenetic tree with appropriate depth and sequence density, the avoidance of overparameterization and local maxima, and the large number of taxa involved. For simplicity, we henceforth refer to the pairs with LR beyond a particular cutoff as "coevolving pairs," although this is to be understood to mean "predicted coevolving pairs," and there is always a calculated percentage of these "coevolving pairs" that are expected due to chance (i.e., "false discoveries").

Coevolution, Residue Physicochemistry, and Protein Structure

The number of coevolving pairs varied according to the physicochemical vector used to segregate the amino acid residues (Table 1). Volume segregation consistently yielded the highest percentages, regardless of the probability cutoff, and hydrophobicity segregation resulted in the lowest percentages. When different vectors are used, differences between the fractions of predicted coevolving pairs are much smaller than the magnitudes of the fractions detected (2%–9% of the pairs evaluated are predicted to coevolve, whereas differences at the same cutoff are in the

Table 1 Coevolving of residues in COI: three different physiochemical segregations at three probability cutoff values

	Hydrophobicity (n = 8778)		Polarity (n = 9423)		Volume (n = 8515)	
	Percentage[a]	Posterior[b]	Percentage[a]	Posterior[b]	Percentage[a]	Posterior[b]
$p < 0.01$	6.7%	0.85	7.3%	0.86	8.4%	0.88
$p < 0.005$	4.3%	0.88	4.8%	0.90	5.3%	0.91
$p < 0.001$	2.2%	0.95	2.6%	0.96	2.7%	0.96

Note. The total number of pairs analyzed in each segregation category is given in parentheses

[a] The coevolving percentage

[b] The posterior probability of a detected coevolving pair to coevolve

range of 0.1%–1%). This suggests that much of the underlying coevolution may not be directly related to the physicochemical properties measured by these vectors. Different physicochemical properties may also have mediated different coevolutionary interactions in different structural environments, and the average effect may not be a good reflection of the diversity of effects at different pairs of sites. Further analyses using a larger number of combined and transformed metric scores (Atchley et al. 2005) may be a fruitful way to discriminate more different kinds of coevolution, but here we are focused on determining the relationship to structure and function for the distributions of coevolving pairs robustly detected by a few simple well-known metrics.

To assess the effect of structure on coevolution, we divided the coevolving pairs into three categories, TM, S, and AR, according to whether the two sites of each pair were in the 12 transmembrane helices, in the same surface region, or in different regions. There were about 4000, 600, and 4500 pairs in each of these categories, respectively (Table 2). Coevolving signals for TM pairs were uniformly stronger (up to 70% more coevolving pairs) than those in S and AR pairs, while percentages for S and AR pairs were similar to one another. We note that the results for the polarity vector are qualitatively similar to the preliminary results presented by Wang and Pollock (2005), although conserved sites were included in that study and the stochastic search of likelihood space was not employed, the cutoff is slightly different than those presented here, and the sensitivity (predicted signal to noise ratio) is increased in the present study. Coevolutionary signal in the structural regions also varied according to the physicochemical segregation used. In TM and AR regions, volume segregation resulted in the highest coevolving percentages (volume is significantly different from hydrophobicity, p = 0.04 and p = 0.01 for the TM and AR regions, respectively, based on G test), but in the S region the coevolving percentage due to polarity was the highest (Table 2; polarity is significantly higher than hydrophobicity, p = 0.046).

Table 2 Coevolving percentages of COI to regions: coevolving pairs are classified according to region (transmembrane, TM; within a surface region, S; across regions, AR)

	Hydrophobicity	Polarity	Volume
TM	5.7% (3916)	6.0% (4186)	6.8% (3655)
S	3.0% (543)	5.3% (606)	4.4% (574)
AR	3.2% (4319)	3.8% (4661)	4.2% (4286)

Note. The probability cutoff was $p < 0.005$, segregation categories are the same as in Table 1, and the numbers of variable site pairs in each segregation category and region combination are given in parentheses

Although there are a number of theoretical reasons that pairs of residues may coevolve, it is plausible that under some conditions physically close residues are more likely to interact. Therefore, physically close residues are expected to exhibit a stronger coevolutionary signal, and a correlation between coevolving sites and physical proximity in at least some cases provides a good independent test of the functioning of the statistical prediction procedure. As previously (Pollock et al. 1999; Wang and Pollock 2005), we grouped the coevolving pairs in COI according to their C_α distances and plotted their frequency distribution and compared this to the frequency distribution of random residue pairs.

As shown before for the polarity segregations (Wang and Pollock 2005), the distance distributions of coevolving pairs are biased to the proximal pairs, but they are strikingly different in different structural categories (Fig. 3). The results for segregation according to volume are shown because they have the strongest relationship to distance in the three comparisons, but the results are qualitatively similar for hydrophobicity and polarity (data not shown). Results were also not qualitatively different if adjacent positions in the linear sequence were excluded (data not shown). In S, coevolving pairs are strongly biased toward short distances compared to expectation. The observed degree of proximity for adjacent sites in this category is strong independent evidence that the predicted coevolving sites are reflecting underlying biological processes, rather than unidentified statistical noise, such as inaccurate phylogenetic relationships, or rate fluctuations along branches. More than 80% of coevolving sites in S have a C_α distance

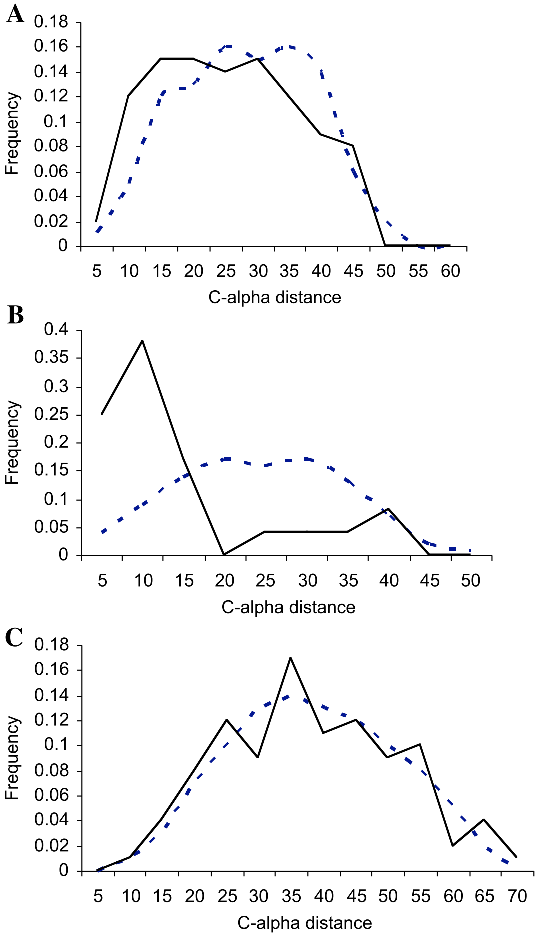


Fig. 3 Frequency distribution for C_α distances between coevolving sites. Sites in coevolving pairs are grouped according to structural region, and results are shown for volume segregation only. Pairs shown have both sites located in the transmembrane region (**A**); both sites located in the same surface region (**B**); or both sites in different regions (**C**). The observed frequency distributions (solid line) are compared to the frequency distribution expected based on all sites pairs (dashed line)

<20 Å, a finding that is highly significant (*G* test: $p <$ 0.0001). The TM comparisons also show that there are more proximal pairs coevolving than expected ($p = 0.02$), whereas the AR coevolving pairs are not differently distributed than random expectations. In summary, S contains a smaller percentage of coevolving pairs, but the S coevolving pairs are considerably closer in the three-dimensional structure.

Relationship to Function in the Transmembrane Region

A vital aspect of COI function is pumping protons across the membrane. Three proton channels have been proposed based on the crystallographic structures of CO (Gennis 1998; Yoshikawa et al. 1998), but evidence for these hypothetic channels is controversial (Pereira and Teixeira 2004; Yoshikawa 2003), and different channels may function differently in different lineages. It therefore is of interest to study the tendency of amino acid residues to coevolve depending on their proximity to the three channels.

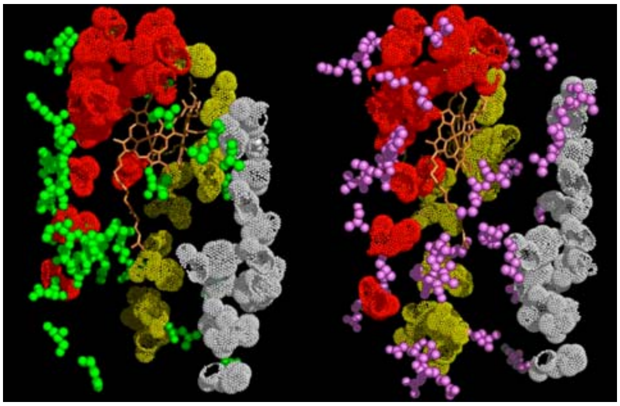


Fig. 4 Two views of coevolving sites and proposed proton channels in the COI crystal structure. The close colocalization of site coevolving according to polarity and hydrophobicity (green spheres) with the H channel (red dots) is shown at the left. The right view shows the distribution of sites coevolving according to volume but not according to polarity or hydrophobicity (purple spheres). The D and K channels are shown with grey and yellow dots, respectively. Short movies of these two views are available as Supplementary Figure S2.

Polarity and hydrophobicity of residues are important for channel function, so we analyzed the distribution of sites involved in coevolved residue pairs using the polarity and hydrophobicity vectors (Fig. 4). These were compared to the distribution of sites involved in coevolving pairs according to the volume vector but not according to the polarity vector and not according to the hydrophobicity vector. To improve visualization (by limiting the number of sites to only those in the best-supported coevolving pairs), only sites in polarity- and hydrophobicity-derived coevolving pairs with $p_{0.1}$ are shown.

Amino acid residues in the proposed proton channels are generally conserved and, therefore, have few opportunities to coevolve; it is thus not surprising that few coevolving sites were detected within these channels. Nevertheless, many coevolving sites were found adjacent to the channels, and there were notable differences among the three hypothetical proton channels (Fig. 4). Only three residues were found directly adjacent to channel D, whereas many coevolving residues were detected adjacent to channel H, along with two residues that are thought to be part of the channel. There are only a few coevolving residues directly adjacent to the K channel, and these are only on the entrance side and not on the exit side of the channel. More of the coevolving residues associated with the H channel are also on the entrance side. In contrast, the volume-derived sites appear to be distributed evenly around the transmembrane region. Comparing the distributions of residues within 10 Å of each of the channels, the bias of polarity- and hydrophobicity-derived coevolving sites toward the H channel is significant, whereas that of volume-derived coevolving sites is not notably biased compared to the variable (candidate coevolving sites) near these channels (see Supplemental movies Figure S2A and S2B).

Discussion

The present study, which analyzed 231 COI amino acid sequences using stochastic searching, a phylogenetic tree based on complete mitochondrial genomes, and a LR test, detected much higher than expected percentages of predicted coevolving residues. This study adds important details to previous studies indicating that coevolution tends to occur in a context-dependent pattern closely related to protein function and structure (Taylor and Hatrick 1994; Wang and Pollock 2005). As in previous studies, the bias of predicted coevolving residues toward pairs of residues that are adjacent in the crystal structure is strong evidence that the statistical coevolutionary predictions are selecting for biologically relevant and meaningful interactions. We found that certain coevolutionary effects in the transmembrane region were closely associated with just one of the three hypothetical proton channels in COI, indicating that the function of this channel was important in distributing coevolutionary interactions along its length.

In an earlier study of fewer than 60 myoglobin sequences (Pollock et al. 1999), coevolving percentages were only about 20%–30% higher than random expectation, meaning that the posterior probability of coevolution in a detected pair was < 0.4. In contrast, the coevolving percentages reported here are more than 10 times greater than their corresponding random expectation, with posterior probabilities up to 0.96 (Table 1). This suggests that increasing the number of sequences improves the sensitivity of coevolution detection. A further benefit is that the large number of sequences utilized (as well as the elimination of low-variability sites) causes the LR statistic to be reliably χ_1^2 distributed (Fig. 2), avoiding the computationally expensive parametric bootstrapping procedure required for smaller data sets (Pollock et al. 1999).

The apparently improved sensitivity of the current study may also partly arise from the implementation of Markov chain and stochastic process in the search for likelihood maxima, which helps avoid errors due to the search becoming stuck on local maxima. The stochastic search is, however, computationally expensive. As noted previously (Pollock et al. 1999), the methodology used here might be improved by integrating over uncertainty in the phylogenetic tree. This could be done, for example, by repeating the analysis for a sample (perhaps hundreds or thousands) from a posterior prediction of phylogenetic tree space, as by Dimmic et al. (2005). For such an approach to be computationally feasible, simplifying assumptions need to be made in other areas, or the data set analyzed needs to be reduced. It is uncertain how much benefit in terms of noise reduction phylogenetic tree integration may provide, or which simplifying assumptions would be most acceptable and lead to introduction of the fewest errors.

Coevolution Related to Pairwise Distance

As with most previous studies (Atchley et al. 2000; Fukami-Kobayashi et al. 2002; Pollock et al. 1999; Pritchard et al. 2001; Valencia and Pazos 2002; Wollenberg and Atchley 2000), proximal residue positions in COI have a stronger tendency to coevolve, whether or not they are adjacent in the linear sequence (see supplemental materials). Nevertheless, distant residues also coevolve. Since bonded residue interactions occur over very short distances, and nonbonded residue interactions such as hydrophobic and hydrophilic interactions decay dramatically as distance increases (Bahar and Jernigan 1997; Chelli et al. 2004), an explanation for coevolution of distant residue positions is desirable. Some distant coevolutionary interactions may be effected via intermediate residues, while other distant pairs may be linked by their effect on the global free energy of the protein or domain in which they reside. Coevolution between distant residues may also be mediated by networks of functional importance within a molecule. In catalytic centers, electrostatic interactions among residues can occur over distances as great as 15 Å through residue interaction networks (Russell and Fersht 1987; Thomas et al. 1985). The substrate specificity of enzymes can also be affected by residues distal to the active site, via hypothetical electron-tunneling networks (Hedstrom et al. 1994; Perona et al. 1995). Such long-distance interactions may be common, since, for example, mutations as far as 25 Å apart frequently affected each other in a thermolysin-like protease (de Kreij et al. 2002).

Coevolution Related to Proton Pumping Channels and Functional Constraints

Long-range residue interactions may be even more frequent in membrane proteins (Gromiha and Selvaraj 2001). The proton pumping residue interaction network in COI (as a membrane protein) requires residues to have precise polarity and conformation (Gennis 1998; Pereira and Teixeira 2004; Yoshikawa et al. 1998). This provides numerous potential opportunities for fine adjustments through coevolution in proximal residues (Table 2 and Fig. 4). The close association of coevolving sites with just one of the three putative proton channels, the H channel, suggests that the H channel may have greater functional importance, at least in vertebrates. Confirmation of the three COI proton channels (Gennis 1998; Yoshikawa et al. 1998) has proved difficult (Namslauer and Brzezinski 2004; Papa et al. 2004; Pereira and Teixeira 2004; Yoshikawa et al. 1998). Thus, coevolutionary analysis appears to augment other methods for predicting regions of functional importance.

Coevolution, Amino Acid Physicochemistry, and Structure

Although polarity and hydrophobicity coevolution appear to be related to functional aspects of COI, volume coevolution is more prevalent overall (Table 1). Considering only the surface regions, however, volume contributed less to the coevolutionary signal than polarity (Table 2). This is consistent with previous studies on soluble proteins (Fukami-Kobayashi et al. 2002; Neher 1994; Pollock et al. 1999). The greater prevalence of volume coevolution in the transmembrane regions, plus the greater number of coevolving pairs (Table 2), indicates that the character of coevolution in the COI transmembrane region is different from that in the soluble region. The transmembrane region differs in that it is not exposed to aqueous solvent, but also in that it is the core of the protein (Gennis 1998; Pereira and Teixeira 2004; Yoshikawa et al. 1998). Residues in protein cores tend to be highly packed, and highly packed residues form more interactions than in loosely packed regions (Gromiha and Selvaraj 2001). It is plausible that these packing interactions should create coevolutionary pressure based primarily on volume.

Secondary structure may also be important in transmembrane region coevolution. COI consists primarily of helices, and van der Waals forces play a vital role in helical formation during protein folding (Kilosanidze et al. 2004). Extensive volume coevolution may thus reflect the evolutionary constraints of COI imposed by helix formation. In soluble proteins, charge coevolution in helices is a critical factor (Pollock et al. 1999), but there is little opportunity for charge coevolution in transmembrane regions since charged residues are rare outside of the proton channels. It is worth remembering, of course, that the physicochemical cause of any true coevolution detected using these vectors may have nothing or little to do with the actual vector used.

Acknowledgments We thank Judith Beekman for useful comments on the manuscript and Matt Reynolds for contributions to the E-GenBio database. This work was supported by grants from the National Institutes of Health (GM065612 and GM065580), the National Science Foundation (through Louisiana EPSCOR and the Center for Biomodular Multi-scale Systems), and the State of Louisiana Board of Regents (Research Competitiveness Subprogram LEQSF 2001-04-RD-A-08 and the Millennium Research Program's Biological Computation and Visualization Center) and Governor's Biotechn ology Initiative.

References

Argos P, Rao JK, Hargrave PA (1982) Structural prediction of membrane-bound proteins. Eur J Biochem 128:565–575

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlation among amino acid sites in bHLH protein domains: An information theoretic analysis. Mol Biol Evol 17:164–178

Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102:6395–6400

Bahar I, Jernigan RL (1997) Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 266:195–214

Benjamini Y, Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. Genetics 171:783–790

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Chelli R, Gervasio FL, Procacci P, Schettino V (2004) Inter-residue and solvent-residue interactions in proteins: a statistical study on experimental structures. Proteins: Structure, Function, and Bioinformatics 55:139–151

Chelvanayagam G, Eggenschwiler A, Knecht L, Connet GH, Benner SA (1997) An analysis of simultaneous variation in protein structures. Protein Eng 10:307–316

de Kreij A, van den Burg B, Venema G, Vriend G, Eijsink VGH, Nielsen JE (2002) The effects of modifying the surface charge on the catalytic activity of a thermolysinlike protease. J Biol Chem 277:15432–15438

DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos, CA

Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21:I126–I135

Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22:1919–1928

Faith JJ, Pollock DD (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics 165:735–745

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

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

Fleishman SJ, Yifrach O, Ben-Tal N (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J Mol Biol 340:307–318

Fukami-Kobayashi K, Schreiber DR, Benner SA (2002) Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 319:729–743

Gennis RB (1998) Protein structure: cytochrome c oxidase: one enzyme, two mechanisms? Science 280:1712–1713

Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C (2003) Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 328:1061–1069

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Gromiha MM, Selvaraj S (2001) Role of medium and long-range interactions in discriminating globular and membrane proteins. Int J Biol Macromol 29:25–34

Hastings WK (1970) Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 57:97–109

Hedstrom L, Perona JJ, Rutter WJ (1994) Converting trypsin to chymotrypsin-residue-172 is a substrate-specificity determinant. Biochemistry (Mosc) 33:8757–8763

Iwata S, Ostermeier C, Ludwig B, Michel H (1995) Structure at 2.8?resolution of cytochrome c oxidase from Paracoccus denitrificans. Nature 376:660–669

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Kilosanidze GT, Kutsenko AS, Esipova NG, Tumanyan VG (2004) Analysis of forces that determine helix formation in {alpha}-proteins. Protein Sci 13:351–357

Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883

Krigbaum WR, Komoriya A (1979) Local interactions as a structure determinant fro protein molecules: II. Biochim Biophys Acta 576:204–248

Namslauer A, Brzezinski P (2004) Structural elements involved in electron-coupled proton transfer in cytochrome c oxidase. FEBS Lett 567:103–110

Nahum LA, Reynolds TR, Wang ZO, Faith JJ, Jonna R, Jiang ZJ, Meyer TJ, Pollock DD (2006) EGenBio: A data management system for evolutionary genomics and biodiversity. BMC Bioinformatics 7(Suppl 2):S7

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

Papa S, Capitanio N, Capitanio G (2004) A cooperative model for proton pumping in cytochrome c oxidase. Biochim Biophys Acta Bioenerg 1655:353–364

Pereira MM, Teixeira M (2004) Proton pathways, ligand binding and dynamics of the catalytic site in haem-copper oxygen reductases: a comparison between the three families. Biochim Biophys Acta Bioenerg 1655:340–346

Perona JJ, Hedstrom L, Rutter WJ, Fletterick RJ (1995) Structural orgins of substrate discrimination in trypsin and chymotrypsin. Biochemistry (Mosc) 34:1489–1499

Pollock DD, Taylor WR (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng 10:647–657

Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287:187–198

Pritchard L, Bladon P, Mitchell JMO, Dufton MJ (2001) Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng 14:459–555

Russell AJ, Fersht AR (1987) Rational modification of enzyme catalysis by engineering surface-charge. Nature 328:496–500

Saraf MC, Moore GL, Maranas CD (2003) Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng 16:397–406

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

Svensson-Ek M, Abramson J, Larsson G, Tornroth S, Brzezinski P, Iwata S (2002) The X-ray crystal structures of wild-type and EQ(I-286) mutant cytochrome c oxidases from Rhodobacter sphaeroides. J Mol Biol 321:329–339

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

Thomas PG, Russell AJ, Fersht AR (1985) Tailoring the Ph-dependence of enzyme catalysis using protein engineering. Nature 318:375–376

Thompson J, Gibson T, Plewniak F, Jeanmougin F, Higgins D (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Thompson J, Higgins D, 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, Yoshikawa S (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272:1136–1144

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

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

Wang ZO, Pollock DD (2005) Context dependence and coevolution among amino acid residues in proteins. Methods Enzymol 395:779–790

Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional association in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A 97:3288–3291

Yoshikawa S (2003) A cytochrome c oxidase proton pumping mechanism that excludes the O2 reduction site. FEBS Lett 555:8–12

Yoshikawa S, Shinzawa-Itoh K, Nakashima R, Yaono R, Yamashita E, Inoue N, Yao M, Fei MJ, Libeu CP, Mizushima T, Yamaguchi H, Tomizaki T, Tsukihara T (1998) Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science 280:1723–1729