Residue coevolution: modeling and interpretation

Zhengyuan Wang
Louisiana State University and Agricultural and Mechanical College, zwang3@lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/3815

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
RESIDUE COEVOLUTION: MODELING AND INTERPRETATION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
In partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
Zhengyuan Wang
B.S., Wuhan University, 1990
M.S., University of Louisiana at Lafayette, 2002
August 2006
Acknowledgements

I am so grateful for the guidance, encouragement, patience, and financial support from my major advisor, Dr. David D. Pollock, during my pursuit of the degree. I thank the other members of my dissertation committee, Drs. Grover Waldrop, Brian Marx, Guangwen Zhou, and Marlene Janes for their scientific advice and approval of this research. For the assistantships I give my thanks to The Department of Biological Sciences of LSU. I would also like to acknowledge my lab co-workers for their collaboration and friendship. Finally, I appreciate my family, especially my parents and my wife Fenghui. Without their love and understanding I could not have come this far.
# Table of Contents

Acknowledgements ............................................................................................................. ii

List of Tables ..................................................................................................................... iv

List of Figures .................................................................................................................. v

Abstract.............................................................................................................................. vi

Chapter I: Background ........................................................................................................ 1

Chapter II: Context Dependence and Co-evolution among Amino Acid Residues in Proteins ............................................................................................................................. 11

Chapter III: Coevolutionary Patterns in Cytochrome C Oxidase Depend on Domain Structure and Function ...................................................................................................................... 28

Chapter IV: Context Dependent Coevolution in Protein Complex Cytochrome C Oxidase Detected Using Bayes Factor Analysis ............................................................................. 58

Chapter V: Structural and Functional Implications from Evolutionary Analysis of Photolyase/Cryptochrome Proteins ................................................................................... 93

Appendix: Permission ..................................................................................................... 126

Vita.................................................................................................................................. 128
List of Tables

Table 2.1. Expected and observed coevolving percentages and total number of pairs analyzed .................................................................................................................................................................................................................................................. 21

Table 3.1. Coevolving of residues in COI. ............................................................................................................................ 39

Table 3.2. Coevolving percentages of COI to regions ........................................................................................................... 41

Table 4.1. Coevolving percentages within and between different sites categories at BF>4 to the two vectors .................................................................................................................. 72

Table 4.2. Coevolution percentages within and between the two subunits according to polarity and volume vectors ................................................................................................................ 73

Table 5.1. The coevolving pairs detected using the charge vector .................................................................................. 104

Table 5.2. The traced pairwise substitutions .................................................................................................................. 108
List of Figures

Figure 2.1. Structural distance distributions for coevolving residues................................. 22

Figure 3.1. The reduced phylogeny of the 231 vertebrates used in this study constructed from all 13 mitochondrial-encoded proteins................................................................. 35

Figure 3.2. Comparison of cumulative frequency distributions. ................................. 40

Figure 3.3. Frequency distribution of C-alpha distances of coevolving pairs. ............. 42

Figure 3.4. Two views of coevolving sites and proposed proton channels in COI. ...... 44

Figure 4.1. The phylogeny of the 131 vertebrate species. ........................................... 65

Figure 4.2. Frequencies of coevolving pairs within COI or COIII to the vectors of residue polarity (Left) and volume (Right), binned to C-alpha distances. .................................. 75

Figure 4.3. Frequencies of coevolving pairs across COI and COIII to the vectors of residue polarity (Left) and volume (Right), binned to C-alpha distances................................. 76

Figure 4.4. Locations of the sites involved in coevolution within COI............................ 77

Figure 4.5. Locations of the sites involved in coevolution within COIII. ...................... 77

Figure 4.6. Locations of the pairs involved in coevolution between COI and COIII..... 78

Figure 5.1. The phylogeny of the 100 sequences........................................................ 100

Figure 5.2. The eleven coevolving pairs displayed on 1TEZ. ....................................... 104

Figure 5.3. The coevolving pairs located in the region with most of conserved hydrophobic sites. ......................................................................................................................... 106

Figure 5.4. The traced pairwise state substitutions of the pair 224-406. ................. 109

Figure 5.5. The traced pairwise state substitutions of the pair 311-315. ................. 110
Abstract

Coevolution between amino acid residues and its context-dependence are important for exploring protein structure and function, and critical for understanding protein structural and functional evolution. Coevolution has long been ignored because of its complexity and the lack of computing power.

In the research presented here, I developed an efficient coevolution analysis methodology based on likelihood comparisons of statistical models. Likelihood ratios and Bayes factors, calculated using the Markov chain Monte Carlo algorithm, were employed as the statistics. Two types of models, 2-state and 3-state, were developed to allow for the context-dependence of coevolution. Computer programs implementing this methodology were coded in C/C++ and were run on the Beowulf clusters of our laboratory and the super computers of LSU. Using these programs and custom Perl scripts, residue coevolution in cytochrome c oxidase and photolyases/cryptochromes protein superfamily was analyzed.

I found that pairwise coevolution between residues is highly dependent on protein tertiary structures and functions. I detected extensive coevolving pairs in all our analyses, and these pairs were primary localized in regions of known structural and/or functional importance. I also found that coevolution is related to evolutionary rate and concentrated in moderately conserved sites. In supporting the importance of functional constraints, I detected a non-negligible coevolutionary signal between complex subunits and stronger coevolution in proteins of functional importance. I also found that the interaction between subunits can serve as a local coevolutionary constraint on one subunit rather than driving coevolution between two subunits. Based on coevolutionary patterns, I suggested that a
domain without any previously supposed function actually operates as a folding core in the proteins of photolyase/cryptochrome superfamily. The coevolutionary patterns also provided clues regarding the functional evolution of electron transfer in this superfamily. I also found that coevolving sites with double substitutions along a branch tend to occur only at physically contacting sites, and that salt-bridge stabilization and secondary structure stabilization are important forces of residue coevolution.

The methodology and programs developed in this research are powerful tools for coevolutionary analysis, which can provide valuable information for characterization of protein structural/functional domains and exploration of protein structural/functional evolution.
Chapter I: Background
In this post-genomic era, understanding the mechanisms of protein structure and function is a prominent goal. Despite more than 100 years of study, our knowledge about proteins is still limited. For example, the mechanisms of protein folding remain unknown; the relationship between protein structure and its function is largely undefined; and how proteins change their structure and function during evolution is still uncertain. In recent years, many useful techniques have emerged, including molecular simulation and sequence analysis. Though some challenges remain, computational sequence analysis has proven powerful in finding proteins and ascertaining their functions (Pearson and Sierk 2005), predicting the structures of proteins (Moult 2005), and determining the evolutionary relationships of proteins (Brocchieri 2001).

Proteins are generally composed of 20 amino acids. The primary structures of proteins are their amino acid sequences, which are encoded by a stretch of DNA sequence called a gene. DNA is comprised of four nucleotides (A, T, G, and C), and during protein synthesis genes are transcribed into RNA, processed, and then, translated into amino acid sequences. The genetic code governs the correspondence between amino acid sequences and DNA sequences via DNA triplets called codons. For example, AAA codes Lys, and AGA codes Arg. Genes are subject to mutation, recombination, deletion, and insertion during replications and other times. These events can change the primary amino acid sequence of a protein, which then affect protein structure and function. Their effects may be neutral, beneficial, or deleterious. Through natural selection, beneficial changes are propagated and deleterious changes are eliminated because the organisms possessing the altered proteins are more or less able to survive and to reproduce. Meanwhile, through duplication and domain shuffling, new genes may be produced.
These new genes may encode new proteins, which will subsequently be subject to natural selection. In these ways, proteins evolve and acquire their specificity and diversity.

Selective constraints in protein evolution come from both structural and functional aspects. To achieve function, the conformational integrity of active sites is required along with flexibility required in enzymatic reactions. This requirement may involve a couple of residues, a specific domain, or the entire protein. To achieve structure, specific folding has to be quick and accurate. At the molecular level, structural and functional information is determined by amino acid sequences (plus the solution environment). The structural/functional requirements and integrity of proteins are mediated and maintained by proper amino acid interactions including salt bridges, hydrogen bonds, electrostatic interactions, Van der Waal's forces, and hydrophobic interactions. Amino acid replacements that disturb these interactions would induce selective stress from structural or functional constraints. Analyses of amino acid residue replacements have revealed considerable selection on individual residues (Dayhoff, Schwartz, and Orcutt 1978; George, Barker, and Hunt 1990; Jones, Taylor, and Thornton 1994). Surveys on conserved sequence patterns (e.g. zinc finger) demonstrated the selection on clusters of residues (Argos, Rao, and Hargrave 1982; Stormo 2000; Andrade, Perez-Iratxeta, and Ponting 2001; Parry 2005). Indeed, the relationship between residue conservation (resulting from selection) and structural/functional importance has become the foundation of many protein analysis approaches (Hughes 1999; Nei and Kumar 2000).

Although interpretations of the relationship between residue conservation and functional importance have been successful, residue conservation analyses generally ignore the interaction among residues by assuming that the probabilities of replacement at
each site are independent of replacements at other sites. This assumption is not realistic. Since protein structures and functions are mediated by residue interactions, these interactions depend on the physicochemical properties of the interacting residues. Such interdependence of physical interactions seems bound to lead to interdependent evolution, or coevolution. In the process of coevolution, replacements at one site influence the replacements at interacting sites, and these sites then exhibit correlated replacement (evolution). A simple illustrative example is that of a critical salt bridge, where a replacement of a positive residue for a negative residue at one site induces a replacement of a negative residue for a positive residue at interacting site. Most real examples are considerably more complex. Analyzing coevolution can, thus, provide information about residue interactions, leading to clues about what sequence features are most critical for protein structure and function.

Coevolutionary analysis has, however, proven difficult. The main reason is the existence of the genetic drift. Many replacements result from random events that do not involve selective pressure. Coevolutionary events are buried in this sea of “background noise”. Another resource of “noise” is the phylogeny, which means sequences are derived from a common ancestor. The inherent covariation from phylogeny looks like coevolutionary signal, but is not related to protein structures/functions and natural selection. To detect coevolutionary signal above this noise requires a large amount of data and very intensive computing.

Taking the advantage of recent dramatic increases in the computing power, coevolutionary analysis has recently made dramatic improvements. Several methods have been explored for analyzing coevolution (Korber et al. 1993; Neher 1994; Shindyalov,
Kolchanov, and Sander 1994; Taylor and Hatrick 1994; Chelvanayagam et al. 1997; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Fukami-Kobayashi, Schreiber, and Benner 2002; Valencia and Pazos 2002; Govindarajan et al. 2003). The most methods ignore phylogeny by assuming sequences are independent (Korber et al. 1993; Neher 1994; Taylor and Hatrick 1994; Thomas, Casari, and Sander 1996; Giraud 1998; Hoffman, Schiffer, and Swanstrom 2003; Saraf, Moore, and Maranas 2003). This assumption is not valid because sequences are derived from a common ancestor, and hence, are all related to one another. Ignoring phylogeny may weaken the detected signal (Fukami-Kobayashi, Schreiber, and Benner 2002) and make it hard to discriminate the covariation due to common ancestry and the coevolution due to selection (Pollock and Taylor 1997; Wollenberg and Atchley 2000). Another common approach detects coevolutionary information by reconstructing ancestral sequences to determine ancestral replacement patterns (Shindyalov, Kolchanov, and Sander 1994; Fukami-Kobayashi, Schreiber, and Benner 2002). There is a fundamental conflict in such methods, however, since ancestral sequences are reconstructed under the assumption that residue sites are independent, but are used to analyze their dependency. In addition, the reliability and uncertainty of ancestral sequence reconstruction are of significant concern (Tuffery and Darlu 2000; Suzuki and Nei 2001; Krishnan et al. 2004).

Several factors critical to improve coevolutionary analysis have been recognized. Inclusion of phylogeny appears to be critical to reduce noise and increase the sensitivity (Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Fukami-Kobayashi, Schreiber, and Benner 2002). Avoiding overly complex evolutionary models can strengthen the robustness of the method, and the using many sequences and high
sampling density can dramatically influence the power of coevolution detection (Pollock and Taylor 1997). Other factors may include the depth of the evolutionary relationship between the sequences, the structural or functional context of residues in the sequences analyzed, the presence of adaptive bursts or rate accelerations, and the potentially variable and dispersed nature of coevolutionary interactions between residues (Shindyalov, Kolchanov, and Sander 1994; Chelvanayagam et al. 1997; Pollock, Taylor, and Goldman 1999; Wang and Pollock 2005).

Overall, the mechanisms of residue coevolution remain largely unknown and the best methods are focused on robust detection rather than accurate modeling. This dissertation study aims to construct a model-based methodology to detect coevolution and to shed light on the mechanisms of evolution and coevolution. In this study, I chose two proteins as my model proteins, cytochrome c oxidase and photolyase/cryptochrome. These proteins have been sequenced extensively, which allowed me to develop sophisticated models yet to control the problem of over-parameterization. As the terminal enzyme in the respiratory chain, cytochrome c oxidase’s structure and function have been well-studied (Iwata et al. 1995; Tsukihara et al. 1996; Gennis 1998; Yoshikawa et al. 1998; Svensson-Ek et al. 2002; Yoshikawa 2003; Pereira and Teixeira 2004). In addition, cytochrome c oxidase contains multiple subunits. These subunits play different roles in the structure and function of this complex. Photolyase/cryptochrome proteins form a super family that exists in most organisms. The functions of these proteins are diverse (Lin and Shalitin 2003; Sancar 2003; Lin and Todo 2005). Photolyases function in DNA repair while cryptochromes function in circadian clocks and signal transduction. Despite their sequence and functional diversity, their structures, especially the N-terminal domain,
are surprisingly similar (Park et al. 1995; Tamada et al. 1997; Komori et al. 2001; Brautigam et al. 2004; Mees et al. 2004). This feature, similar structure but different function, is critical for the differentiation of the structural and functional driven coevolution. In addition, the characterized structural and functional features of these proteins allow better interpretation and optimization of model-based coevolutionary analysis.

Analyzing these two model proteins allowed me to optimize my algorithms and revealed the power of my methodology. The coevolutionary program that I have developed will greatly benefit protein structural and functional studies, and is currently being used to direct mutagenesis for protein engineering.

References


Chapter II: Context Dependence and Co-evolution among Amino Acid Residues in Proteins*

* Reprinted by permission of Methods of Enzymology
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, in order to knock out function, the best places to mutate the protein are the most conserved sites. This relationship is sometimes viewed almost as a tautology, such that 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 upon a combination of the distance from active sites and the 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 the residues. In a somewhat counter-intuitive twist, accelerated evolution can also be used as a predictor of functional importance, since 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, despite the fact that protein structure and function result from interactions among amino
acids. This assumption obviously 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 (Korber et al. 1993; Neher 1994; Shindyalov, Kolchanov, and Sander 1994; Taylor and Hatrick 1994; Chelvanayagam et al. 1997; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Tuffery and Darlu 2000; Pritchard et al. 2001; Fukami-Kobayashi, Schreiber, and Benner 2002; Valencia and Pazos 2002; Govindarajan et al. 2003). 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 (Lopez, Casane, and Philippe 2002; Gribaldo et al. 2003; 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 consistent as to the conditions and manners in which coevolution apparently occurs. The strongest pairwise signal comes from residues stacked in alpha helices (Pollock, Taylor, and Goldman 1999), but the strength of pairwise coevolution between more distant residues appears to vary (Pollock, Taylor, and Goldman 1999), and interaction between protein subunits has had tantalizing but limited success (Pazos et al. 1997; Fukami-Kobayashi, Schreiber, and Benner 2002; 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 used (Pollock and Taylor 1997). Nevertheless, results from methods that do incorporate phylogeny into the development of their statistics (Shindyalov, Kolchanov, and Sander 1994; Chelvanayagam et al. 1997; Pollock, Taylor, and Goldman 1999) 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, Taylor, and Goldman 1999), we have analyzed the coevolution of cytochrome c oxidase 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. These provide 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. 2006). We also present some results from COII from the same taxa for comparison.

Prior to analysis, we clustered amino acids at each site according to volume, polarity, and hydrophobicity. We analyzed the sites with slow substitution rates in greater detail, in order to focus on sites for which the structural and functional context might not have evolved much during the range of evolutionary time that 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 co-localize 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.

**Materials and Methods**

**Choice of Sequences**

Two critical factors that influence the 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 also 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 be 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 ClustalX (Thompson, Higgins, and Gibson 1994) in our EGenBio database. After removing sites involved in multiple insertions and deletions, a phylogenetic tree was constructed using the neighbor-joining (NJ) heuristic (Saitou and Nei 1987) from distances calculated using PHYLIP’s PRODIST module (Felsenstein 1989). 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 are not central to our discussion of context-dependent change, and we did not consider them here.

Available 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; Tsukihara et al. 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 and colleagues is to evaluate coincident changes along branches
(Shindyalov, Kolchanov, and Sander 1994). 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, however, this method should do well for detecting coevolution that is nearly simultaneous, 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 a 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 (Korber et al. 1993; Atchley et al. 2000; Wollenberg and Atchley 2000), for example, consider whether there are significant associations between states, but for pairs of sites with only five 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, Taylor, and Goldman 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. Since 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 recently extended our models to allow more than two groups, here we consider only the two-group model delineated by Pollock and colleagues (Pollock, Taylor, and Goldman 1999).

A prime reason why 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 gives rise 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,
over-parameterization relative to the amount of data, inaccuracies of the model, and even small inadequacies of the methodology may overwhelm the results. Our approach was to find the sites with coevolution statistics that were greater than pre-specified “significance” levels (i.e., 0.05, 0.01, and 0.002), and consider both whether the number of such sites were 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 could also evaluate the posterior probability that these sites have coevolved or, alternatively, that they have not coevolved (that is, 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, since the chi-square distribution cannot be assumed for coevolution analyses (Pollock, Taylor, and Goldman 1999). Here, we simulated 6,000 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 between 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 2.1). 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 (Figure 2.1). 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 differences between the
coevolved fraction and the total distribution of transmembrane sites were seen for the 0.2% significance level cutoff (Figure 2.1). For higher significance levels, the differences between the distributions are smaller, although still highly significant, and the number of excess close sites is larger than at the 0.2% 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 five-fold between adjacent categories).

Table 2.1. Expected and observed coevolving percentages and total number of pairs analyzed

<table>
<thead>
<tr>
<th>Expected Location1</th>
<th>%</th>
<th>#2</th>
<th>%</th>
<th>#2</th>
<th>%</th>
<th>#2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>S</td>
<td>11.1</td>
<td>1071</td>
<td>12.7</td>
<td>355</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>TM</td>
<td>15.1</td>
<td>5778</td>
<td>16.5</td>
<td>2701</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>11.0</td>
<td>7794</td>
<td>13.3</td>
<td>3031</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>S</td>
<td>3.7</td>
<td>5.4</td>
<td>7.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TM</td>
<td>6.5</td>
<td>6.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>3.7</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2%</td>
<td>S</td>
<td>1.3</td>
<td>2.5</td>
<td>4.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TM</td>
<td>2.8</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1.9</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 COI comparisons were within the surface (S), transmembrane (TM) or between the surface and transmembrane domains (A), whereas comparisons in COII were within the entire protein. 2 The total numbers of sites for each comparison are shown only once, in the top row.

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 (Figure 2.1). This is consistent with earlier results for surface residues of myoglobin (Pollock, Taylor, and Goldman...
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

Figure 2.1. Structural distance distributions for coevolving residues. Distance (C\(^\alpha\))
frequency distributions are shown for all residue pairs (dashed lines) and for hypothetical
coevolving pairs (solid lines) 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,
extcept for the surface domains which had many fewer sites, and for which results at the
1% level are shown.
M domain, rather than the IM domain. The coevolving sites from the across-domain comparison (A) show the smallest effect of physical distance (Figure 2.1), 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 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 (Figure 2.1), in contrast to the COI analyses, were not significantly different from the overall distribution of sites.

**Discussion**

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 and colleagues for vertebrate myoglobin (Pollock, Taylor, and Goldman 1999). 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.

While 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 0.2% 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, since 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 with exposure to the environment. COI is at the functional core of the CO complex, whereas COII, like COIII and the ten 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, Pollock, and Goldstein 2001), 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 (Brasseur et al. 2001). COII may be selected mostly to bind COI and the other
adjacent subunits in the CO complex, such that 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 are capable of discriminating between different
kinds of coevolution, some of which may be of greater interest for a particular goal.

References

Correlation among amino acid sites in bHLH protein domains: An information

suppressor mutation reveals long distance interactions in the bc(1) complex of
Saccharomyces cerevisiae. Biochimica Et Biophysica Acta-Bioenergetics
1506:89-102.

An analysis of simultaneous variation in protein structures. Protein Eng. 10:307-
316.

Symposia on Biocomputing in Protein Evolution and Structural Genomics. World
Scientific Publication Co.


covariation signals in protein evolution using reconstructed ancestral sequences. J.


Chapter III: Coevolutionary Patterns in Cytochrome C Oxidase Depend on Domain Structure and Function
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, and residue conservation is commonly used for predicting the effect of residue mutations. One limitation of such an analysis, however, is that it frequently ignores interactions among residues and assumes that substitutions at different sites are independent. This assumption limits the power of evolutionary approaches. Hydrogen bonds, charge interactions, and van der Waals interactions are all highly dependent on the size and the physicochemical nature of interacting amino acid residues. Such inter-dependency as well as the constraints at the amino acid level caused by the hydrophobic interaction and overall protein stability leads to coevolution (the reciprocal substitutions at interacting residues). Since the nature and strength of residue interactions vary according to the residues involved and their molecular environments in proteins, coevolution exhibits a complex context-dependence (Wang and Pollock 2005). Unraveling this dependence through coevolutionary analysis can provide valuable information related to protein structure and function.

Previous coevolutionary studies, using various types of data and sometimes ad hoc methodologies, have supported the widespread existence of coevolution (Neher 1994; Shindyalov, Kolchanov, and Sander 1994; Taylor and Hatrick 1994; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Tuffery and Darlu 2000; Wollenberg and Atchley 2000; Pritchard et al. 2001; Fukami-Kobayashi, Schreiber, and
Benner 2002; Valencia and Pazos 2002; Govindarajan et al. 2003). The conclusions
drawn from these studies (about the conditions that lead to coevolution and the manners
in which coevolution occur) are, however, inconsistent. Spatial distance between residues
is involved in coevolution, but the strength of coevolution between both distant and
proximal residue pairs appears to vary (Pollock, Taylor, and Goldman 1999). Residues in
alpha helices exhibited the strongest pairwise coevolution in myoglobin (Pollock, Taylor,
and Goldman 1999), while highly mobile loops with ligand-binding functions had the
strongest signal in DHFR (dihydrofolate reductase), cyclophilin, and formyl-transferase
(Saraf, Moore, and Maranas 2003). Charge compensation has been identified as a strong
coevolutionary force (Chelvanayagam et al. 1997; Pollock, Taylor, and Goldman 1999;
Fukami-Kobayashi, Schreiber, and Benner 2002), but Kondrashov et al. (Kondrashov,
Sunyaev, and Kondrashov 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 (Pollock and Taylor 1997; Pollock,
Taylor, and Goldman 1999; Fukami-Kobayashi, Schreiber, and Benner 2002; Fleishman,
Yifrach, and Ben-Tal 2004), and a large amount of data in the form of a large number of
sequences is vital (Pritchard et al. 2001). It helps to obtain appropriate sequence distance
and density, both of which improve the detection of coevolution (Chelvanayagam et al.
1997; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Pritchard et al. 2001;
Fukami-Kobayashi, Schreiber, and Benner 2002). On the other hand, genetic
recombination can add a layer of noise by introducing simultaneous substitutions that are not driven by functional and/or structural constraints.

In consideration of these factors, the present study analyzed a large dataset, 231 cytochrome c oxidase (CO) subunit I (COI) homologous sequences from vertebrates, using the methodology of Pollock et al. (Pollock, Taylor, and Goldman 1999), which incorporates phylogeny and uses the likelihood ratio test. To overcome the problem of local maxima in the likelihood ratio test, we employed Markov chain Monte Carlo (MCMC) (Hastings 1970), in addition to previously described peak-searching methods. To reduce computation and noise, we segregated residues according to their physicochemical characteristics. 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 models would have 20 independent frequencies for each site and 400 dependent frequencies, making analysis computationally infeasible and raising serious questions about over-parameterization.

Cytochrome c oxidase 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; Tsukihara et al. 1996; Svensson-Ek et al. 2002). The mitochondrial CO exists as a dimer, with each monomer comprised of 13 subunits, while the bacterial CO consists of only four subunits. In spite of the differences in the number
of subunits between mitochondrial and bacterial CO, 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. The transmembrane helices jointly form a cylinder-like structure with their ends at the two membrane surfaces. The surface regions consist of mixed alpha helix and beta sheet structures made from the loops that connect the transmembrane helices and the N- and C-termini. These features of COI provide a variety of structural and functional contexts for coevolutionary analysis.

Since COI, COII, and COIII are encoded by the maternally-inherited non-recombining 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 examine the specific coevolutionary pattern of this kind of proteins, compared to that of soluble proteins such as myoglobin analyzed by Pollock et al. (Pollock, Taylor, and Goldman 1999).

The study described here provides further support for context-dependent residue coevolution. Relevant contextual factors analyzed here include protein structure, function, and residue physicochemical characteristics. The likelihood ratio test implemented using
MCMC proved to be powerful for coevolution detection when a large number of sequences are available, and the increased number of sequences improved the predictability of the likelihood ratio distribution under the null model. This study suggests that coevolutionary analysis is a valuable tool for protein structural and functional studies and it may be especially useful in detecting interacting networks of residues, such as those involved in proton or electron channels.

**Materials and Methods**

Mitochondrial sequences from 368 vertebrata were automatically downloaded from our *EGenBio* database (www. egenbio.lsu.edu) and all 13 proteins encoded by the mitochondrial genome were aligned with ClustalX (Thompson, Higgins, and Gibson 1994; Thompson et al. 1997). Sites involved in multiple insertions and deletions were removed. A phylogenetic tree was reconstructed using PROTDIST from the PHYLIP package (Felsenstein 1989). Branch lengths for this topology were re-calculated from COI sequences alone using PROML from the PHYLIP package and the JTT matrix (Jones, Taylor, and Thornton 1992). Sequences that were in particularly egregious conflict with known (or well-supported) phylogenetic relationships, or which had particularly long branches, were removed to reduce potential noise arising from phylogenetic inaccuracies (we note that the remaining phylogeny is almost certainly still incorrect in some details, but it is small enough that it can be safely ignored for the purposes of the current study). To limit the number of short uninformative terminal branches, sequences with short branch lengths (less than 0.5 PAM) were also removed. The final phylogenetic tree was of 231 homologous sequences, including the bovine sequence (Figure 3.1).
To relate the results of 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 a likelihood ratio test with likelihood maxima estimated using MCMC. Residues at each site were segregated into two groups (states) according to their hydrophobicity (Argos, Rao, and Hargrave 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. To avoid sites with limited data for which the model might be over-parameterized, sites that showed less than 2% state variation (i.e., the frequency of the major state at that site was greater than 0.98) were excluded from subsequent analysis. With limited variation, these sites produce artificial likelihood maxima. Including these sites results in a higher false positive rate (data not shown).

Two alternative models, independent and dependent, were constructed and described elsewhere (Pollock, Taylor, and Goldman 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 (to the independent model, two sites are analyzed separately; totally there are 4 parameters), and the independent model is nested within the dependent model, meaning that it is a special case.
Figure 3.1. The reduced phylogeny of the 231 vertebrates used in this study constructed from all 13 mitochondrial-encoded proteins. The topology and branch lengths were computed using FITCH and PROML of PHYLIP respectively. Detailed tree is provided in supplement data.
of the dependent model. Each of these models has an instantaneous rate matrix, $Q$, and the transition probabilities over each branch of length $t$ were calculated through standard matrix manipulation (Pollock, Taylor, and Goldman 1999). The likelihood was computed following the pruning algorithm (Felsenstein 1981), and an MCMC algorithm was implemented to traverse the parameter surfaces and locate the maxima (Hastings 1970). In the MCMC process, stationary frequencies were proposed to a Dirichlet distribution, and rates were proposed to a uniform distribution. The temperature of the chain was tuned to achieve an average acceptance of about 50%. We did 10,000 iterations for each chain, though in most cases the chain can reach equilibrium in less than 2000 iterations. The likelihood maxima of the independent model ($L_i$) and the dependent model ($L_d$) were tracked as natural logarithms ($\ln L_i$ and $\ln L_d$).

A likelihood ratio statistic $LR = -2 \ln(L_i/L_d)$ was used to evaluate the significance of coevolution between sites. It has been previously shown that this statistic cannot be assumed to have a $\chi^2$ distribution (with one degree of freedom) under the null (independent) model (Pollock, Taylor, and Goldman 1999). Therefore, to obtain more accurate distribution estimators, we performed parametric bootstrapping by simulating datasets using the same phylogenetic tree and the maximum likelihood parameter estimates (MLEs) from the independent model (Pollock, Taylor, and Goldman 1999). Since, in this case, distributions estimated from the bootstrap were nearly identical to $\chi_i^2$ regardless of the physicochemical vector used (see Results), the LR significance cutoff values considered (3.8 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 $\chi_i^2$. 

36
Since thousands of comparisons were performed in each analysis, these probability values 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 will 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 more coevolving pairs were observed than were expected, the posterior probability that a particular coevolving pair truly coevolved was calculated as \[ 1 - \frac{\text{expected percentage}}{\text{observed percentage}} \].

The distance between two residues of a pair was defined as the distance between their \( C_\alpha \) 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). Automatic parsing of files was performed using custom Perl scripts.
Results

Pairwise Coevolution in COI

Hypothetically coevolving pairs were identified based on clustering 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 \( \chi^2 \) distribution for each of these clustering criteria (Figure 3.2). Although it was shown previously that the null distribution for the LR statistic under these conditions often does not match \( \chi^2 \), probably due to the limited number of substitutions at each site (Pollock, Taylor, and Goldman 1999), it appears that with this phylogeny of 231 sequences, and with the low variability positions removed, most sites contain sufficient data for the \( \chi^2 \) approximation to hold.

The percentage of coevolving pairs predicted in COI were uniformly much higher than expected based on chance, assuming the null or independent model (Table 3.1). 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 truly coevolving pairs was much higher, however, so only the lower criteria were considered further (Table 3.1). For example, at \( P_{0.5} \) about ten times more coevolving pairs were detected than expected by chance. Since we detected 378, 452, and 451 coevolving pairs for the three respective physicochemical segregations, there are 334, 405, and 408 more coevolving pairs than expected by chance in the multiple comparisons. This result is highly significant (G test: \( P < 0.0001 \)). At \( P_{0.1} \), there are about twenty 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, and is again highly significant (G test: P < 0.0001). These numbers and enrichment factors are greater than previous studies, probably due mainly to the inclusion of a phylogenetic tree with appropriate depth and sequence density, avoidance of over-parameterization and local maxima, and the large numbers of taxa involved.

**Table 3.1. Coevolving of residues in COI.** Three different physiochemical segregations at three probability cutoff values are shown. The total numbers of pairs analyzed in each segregation category are shown in parentheses.

<table>
<thead>
<tr>
<th>Hydrophobicity (8778)</th>
<th>Polarity (9423)</th>
<th>Volume (8515)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage&lt;sup&gt;a&lt;/sup&gt; Posterior&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Percentage&lt;sup&gt;a&lt;/sup&gt; Posterior&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Percentage&lt;sup&gt;a&lt;/sup&gt; Posterior&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>P&lt;0.01 6.7% 0.85</td>
<td>7.3% 0.86</td>
<td>8.4% 0.88</td>
</tr>
<tr>
<td>P&lt;0.005 4.3% 0.88</td>
<td>4.8% 0.90</td>
<td>5.3% 0.91</td>
</tr>
<tr>
<td>P&lt;0.001 2.2% 0.95</td>
<td>2.6% 0.96</td>
<td>2.7% 0.96</td>
</tr>
</tbody>
</table>

<sup>a</sup> the coevolving percentage; <sup>b</sup> the posterior probability of a detected coevolving pair to coevolve

**Coevolution, Residue Property, and Protein Structure**

The number of coevolving pairs varied according to the physicochemical vectors used to segregate the amino acid residues (Table 3.1). Volume segregation consistently yielded the highest percentages, regardless of the probability cutoff, and hydrophobicity segregation resulted in the lowest percentages. The fact that the differences in results between the vectors is much smaller (0.1% - 1% of the total pairs evaluated) than the excess numbers detected for all vectors (2% - 7% of total pairs evaluated) suggests that much of the underlying coevolution may not be directly related to the physicochemical properties measured by these vectors. It is also reasonable that the physicochemical
properties mediated coevolutionary interactions may be different in different structural environment. To assess the effect of structure on coevolution, we divided the coevolving pairs into three categories, TM, S, and AR, according to whether two sites of a coevolving pair were in the same transmembrane, a surface region, or were in different regions. There were about 4000, 600, and 4500 pairs in each of these categories respectively (Table 3.2). Coevolving signals for TM pairs were uniformly stronger (their coevolving percentages are up to 70% higher) than those in S and AR pairs, while percentages for S and AR pairs were similar to one another. Coevolutionary signal in the structural regions also varied according to the segregations. In TM and AR regions, volume segregation resulted in the highest coevolving percentages, but in S region the coevolving percentage in the polarity segregation was the highest (Table 3.2).

![Graph](image)

**Figure 3.2. Comparison of cumulative frequency distributions.** The likelihood ratios from parametric bootstrapping using the same tree and the MLEs of the independent model (solid line) and Chi-square with degrees of freedom as one (dashed line) are compared.
Table 3.2. Coevolving percentages of COI to regions. Pairs classified into regions (transmembrane, TM; within a surface region, S; across regions, AR). The probability cutoff was $P < 0.005$, segregation categories are the same as in Table 3.1, and the numbers of variable site pairs in each segregation category and region combination are shown in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Hydrophobicity</th>
<th>Polarity</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>5.7% (3916)</td>
<td>6.0% (4186)</td>
<td>6.8% (3655)</td>
</tr>
<tr>
<td>S</td>
<td>3.0% (543)</td>
<td>5.3% (606)</td>
<td>4.4% (574)</td>
</tr>
<tr>
<td>AR</td>
<td>3.2% (4319)</td>
<td>3.8% (4661)</td>
<td>4.2% (4286)</td>
</tr>
</tbody>
</table>

Coevolution, Protein Structure, and Residue Physical Distance

Although there are a number of theoretical reasons that pairs of residues may coevolve, it is reasonable to assume that the residues must be physically close in order to be able to interact. Therefore, physically close residues might be expected to exhibit a stronger coevolutionary signal. To test this hypothesis and to determine whether the relationship differs among regions, we grouped the coevolving pairs in COI according to their C-alpha distances and plotted the frequencies of each group as the observed coevolving frequency distribution. We also grouped all pairs in each category and plotted their frequencies as the expected coevolving frequency distribution, that is, the coevolving frequency distribution under random expectation. Overall, the distance distributions of coevolving pairs are biased to the proximal pairs in all three segregations (data not shown), but they are strikingly different in different structural categories (Figure 3.3). 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. In S, coevolving pairs are strongly biased towards short distances compared to expectation. More than 80% of coevolving
Figure 3.3. Frequency distribution of C-alpha distances of coevolving pairs. Results shown are for volume segregation. A, pairs in which both sites are located in the transmembrane region; B, pairs in which both sites are located in one of the two surface regions; C, pairs in which the two sites are in different regions (transmembrane region and surface). The observed frequency distributions are shown in solid line; the expected frequency distributions are shown in dashed line.
sites in S have C-alpha distance less than 20 Å, a finding that is highly significant (G test: \( p < 0.0001 \)). The TM comparisons are also slightly biased towards an excess of proximal pairs (G test: \( p < 0.002 \)), whereas the AR coevolving pairs are no 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.

**Coevolution and Proton Channels**

An important aspect of COI function is proton pumping. Three proton channels have been proposed based on the crystallographic structures of CO (Gennis 1998; Yoshikawa et al. 1998), but evidence for these hypothetical channels is controversial (Yoshikawa 2003; Pereira and Teixeira 2004), 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. Polarity and hydrophobicity of residues are important for the channel function, so we analyzed the distribution of those sites that are significantly coevolving to both the polarity and hydrophobicity vector but not to the volume vector (Figure 3.4). As a comparison, we also checked the distribution of those sites coevolving only to the volume vector but neither to the polarity nor to the hydrophobicity vector. To improve visualization, only the coevolving sites with \( P_{0.1} \) are shown.

The proton channels are generally conserved, and thus should offer little chance for coevolution. So, not surprisingly, very few coevolving residues were detected within the proposed channels. A number of coevolving sites were found in sites adjacent to the channels (Figure 3.4), and there are notable differences in the amount of coevolution (to
both polarity and hydrophobicity) at the adjacent sites to the three hypothetical proton channels. No sites were found directly adjacent to channel D, whereas many sites were detected adjacent to channel H, including two sites that are actually parts of the channel. There are only a few coevolving sites associated with the K channel, but only on the entrance side, not on the exit side of the channel. More of the coevolving sites associated with the H channel are also on the transfer side. On the other hand, the above features are not noticeable for the volume-only coevolving sites.

![Figure 3.4. Two views of coevolving sites and proposed proton channels in COI. The left view shows the close co-localization of the polarity- and hydrophobicity-coevolving sites (in green spheres) with the H channel (in red dots). As comparison, the right view shows the distribution of volume coevolving only sites (in purple spheres), which spread over the whole structure. D and K channels are shown with cyan and yellow dots respectively.](image-url)

**Discussion**

Studies have shown that coevolution occurs in a context-dependent pattern closely related to protein function and structure (Taylor and Hatrick 1994; Wang and Pollock 2005).

This dependency makes detecting coevolution a complex task, which requires a large
amount of sequences and an intensive computing power. At the same time, this
dependency makes coevolutionary analysis an important tool for protein structural and
functional analyses. The present study, which analyzed 231 COI amino acid sequences
using a MCMC-implemented likelihood ratio test, effectively detected a high percentage
of coevolving residues. We found that residue properties, pairwise distance, protein
structure, and protein function play important roles in residue coevolution, and
demonstrated the context-dependency of coevolution in great detail.

**Detecting Coevolution**

One of the main concerns about coevolutionary analysis is how to reduce noise while
maintaining the coevolutionary signal. The advantages of our method are that it
segregates amino acids according to their physicochemical characteristics, incorporates
phylogeny, uses a large number of sequences, and employs MCMC for maximum
likelihood searching.

Segregating residues into two groups, thereby reducing the number of parameters,
not only dramatically reduces the computation, but also helps to avoid over-
parameterization. Comparing coevolution detection with and without residue segregation,
Pupko et al. (Pupko and Galtier 2002) found that segregation of residues decreases noise
and, thus, increases the sensitivity of detection. There was a concern that residue
segregation might lead to the lost of coevolutionary signal. The relatively high
coevolution percentages detected in the present study at conservative $p$-values
demonstrated that the segregation did not lead to signal lost.

The importance of incorporating phylogeny in coevolutionary analysis has been
explored previously by Pollock and Taylor (Pollock and Taylor 1997). Incorporating
phylogeny decreases noise dramatically. Without considering phylogeny, the coevolving percentages reached as high as 50% if taking $p < 0.05$ as the cutoff, and were about 40% even at $p < 0.01$ (Atchley et al. 2000). These percentages appear intuitively too high because the majority of mutations do not have significant structural or functional effects and are selectively neutral (Kimura 1968; Kimura 1983; Jukes 2000), thus, they are not subject to coevolutionary constraints. However, incorporating phylogeny by reconstructing ancestral sequences using standard methods is not a good choice because these methods assume the independent evolution of residue sites thereby reducing coevolution signal by erasing all state-dependency among different sites. These reconstruction methods may also produce strong bias and errors in reconstruction properties (Krishnan et al. 2004). These features of standard ancestral sequence reconstruction methods can weaken the coevolutionary signal. The using of a method based on ancestral sequence reconstruction may explain why Tuffery et al. (Tuffery and Darlu 2000) did not detect significant coevolution signal in more than 50% of their studied proteins. However, some studies did find that ancestral sequence reconstruction improves coevolution detection (Fukami-Kobayashi, Schreiber, and Benner 2002; Fleishman, Yifrach, and Ben-Tal 2004). Whether these improvements are from ancestral sequence reconstruction or from the inclusion of phylogeny could not be distinguished in those studies. We tend to think that they are from phylogeny inclusion. Our method incorporates phylogeny by using the phylogenetic tree as a parameter for likelihood calculation, thus avoiding the disadvantages of ancestral sequence reconstruction.

Many coevolutionary analyses suffer from a limited number of sequences for each dataset. Analyzing a small number of sequences results in a higher degree of false signal
Multiple datasets cannot eliminate the problem from the small number of sequences in each dataset because of the context-dependency of coevolution. In each dataset, the context for coevolution might be different. The benefits of a large dataset become clear when comparing the posterior probabilities of the present study to those of Pollock et al. (Pollock, Taylor, and Goldman 1999), which used the same methodology but analyzed less than 60 myoglobin sequences. The coevolving percentages detected in that study were only about 20-30% higher than the corresponding random expectations, which means the posterior probability of coevolving for a detect coevolving pair be less than 0.4, but in our study they are more than 10 times higher than their corresponding random expectations, thus the posterior probabilities are higher than 0.8 (Table 3.1). This comparison suggests that a large number of sequences might be necessary to improve the sensitivity of coevolution detection.

Meanwhile, the improved sensitivity of the current study may partly come from the implementation of MCMC, which helps to reduce the problem of local maxima. It is interesting to note that the methods used in the current study produced $\chi^2$ distributed likelihood ratios in our parametric bootstrapping, which was not found previously (Pollock, Taylor, and Goldman 1999). When the likelihood ratios to the null hypothesis follow a $\chi^2$ distribution parametric bootstrapping is not necessary, thus, saving computing time.

**Coevolution Related to Pairwise Distance**

The frequency distributions of C-alpha distances in coevolving pairs in our study confirmed that the physical distance between residues affects coevolution. Proximal residues have a stronger tendency to coevolve as illustrated by the increase (compared to
the expectation) of the frequencies of coevolving residues with short distance in both the TM and S regions. This is consistent with most previous studies (Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Wollenberg and Atchley 2000; Pritchard et al. 2001; Fukami-Kobayashi, Schreiber, and Benner 2002; Valencia and Pazos 2002; Govindarajan et al. 2003). There is no question that proximal residues have more opportunity to interact directly and thus coevolve. However, our analyses also showed coevolution of distant residues as well. Obviously, the coevolution of distant residues cannot result from direct interaction. It is well established that bonded residue interactions occur over very short distance and that non-bonded residue interactions, such as hydrophobic interaction and hydrophilic interactions, decay dramatically as distance increases (Bahar and Jernigan 1997; Chelli et al. 2004). One explanation for the coevolution of distant pairs is that they result from global constraints on the free energy required by protein folding. Given the additive nature of free energy, one can expect that energy changes induced by a substitution at one site might be restored by subsequent substitution at other amino acid sites, regardless of the distance between the two sites. More likely, distant residue coevolution is driven by long-range residue interactions mediated by intra-molecule networks of functional importance; for example, the proton channels in COI. Earlier studies have revealed that residue electrostatic interactions can occur over distances as far as 15 Å for catalytic centers through residue interacting networks (Thomas, Russell, and Fersht 1985; Russell and Fersht 1987). Substrate specificity of enzymes affected by distal residues has also been reported through a possible electron tunneling network (Hedstrom, Perona, and Rutter 1994; Perona et al. 1995). Furthermore, due to the extensive existence of residue networks, long-range interactions among proteins might be not rare. Mutations
as far as 25 Å were found frequently affecting each other in a thermolysin-like protease (de Kreij et al. 2002). These evidences are mostly from globular proteins such as proteases. Cytochrome c oxidase is a membrane protein. It has been suggested that this type of long range residue interactions might be more frequent in membrane proteins (Gromiha and Selvaraj 2001). In COI, the proton pumping channel(s) is an example of this type of residue interacting networks (Gennis 1998; Yoshikawa et al. 1998; Pereira and Teixeira 2004). Residues in this network interact to accomplish proton pumping including loading, passing, and release of protons. In contrast, the surface regions are less densely packed and expose to the solvent. The residues in the surface regions are less likely to form interacting networks. So that, these coevolving pairs in these regions exhibited greater proximity (Fig3.3).

Coevolution Related to Proton Pumping Channels and Functional Constraints

Different coevolving percentages (Table 3.2) and coevolving frequency distributions (Figure 3.3) observed in the transmembrane region, surface regions, and across regions within COI, on the other hand, imply the important effect of the functional constraints on residue coevolution. The association of coevolving sites, detected using hydrophobicity and polarity vectors, to the proposed channel H proton channel (Figure 3.4) suggests that analyzing coevolutionary information could be useful in identify functional subdomains or regions.

In the transmembrane region of COI, a large proportion of residues are involved in its enzymatic functions as a proton pump. This region contains the electron/proton channels and redox centers (Gennis 1998; Yoshikawa et al. 1998). Among them, the proton channel(s) spreads over the entire region, and is may be involved in up to 14
residues directly (Gennis 1998; Yoshikawa et al. 1998; Pereira and Teixeira 2004). To fulfill the proton pumping function, precise polarity and conformation of these residues are required. These requirements, in turn, are coevolutionary constraints on the residues proximal to the channel(s), which result in increased coevolutionary percentages in the transmembrane regions (Table 3.2). On the other hand, the extend channels may contribute to the long region coevolution within the transmembrane region and across regions. One the surface regions are the entrance and exit of the channel(s) and they might be just involved in several proximal residues. Consistent with this, the coevolving pairs in the surface regions are strongly biased to the proximal pairs.

The impact of the channel(s) on residue coevolution became clear when comparing the localizations of coevolving residues in both hydrophobicity and polarity segregations to those coevolving only in volume segregation (Figure 3.4). The former residues are heavily located proximal to the proposed H channel while the later residues disperse over the entire protein. As well understood, proton pumping is closely related to residue polarity and hydrophility (the opposite of hydrophobicity; in our analysis segregation to hydrophobicity makes no difference with segregation to hydrophility). Coevolving pairs detected using corresponding segregations would thus more be closely associated with the channel(s) compared to those detected using other segregation as volume.

These associations not only demonstrate the coevolution driven by functional constraints but also suggest that analyzing coevolutionary information would a good method to identify regions (or subdomains) with functional importance, such as proton pumping. Because the complexity of residue interaction in these regions and the poor
understanding about functional related residue interactions, the proton channels in COI has proved difficulty to be identified precisely (Yoshikawa et al. 1998; Namslauer and Brzezinski 2004; Papa, Capitanio, and Capitanio 2004; Pereira and Teixeira 2004). Analysis of residue conservation has been commonly used for these tasks, but frequently does not lead to a clear conclusion. Three channels were proposed nearly ten years ago (Gennis 1998; Yoshikawa et al. 1998), but researches are still debating (Papa, Capitanio, and Capitanio 2004; Pereira and Teixeira 2004). Part of the reasons is the identification using directed mutagenesis information is labor and time consuming, and some times biased depending on experimental conditions. The close association of coevolving sites with the proposed H channel gives clear support that the H channel is functional, at least in vertebrates. Comparing to using direct mutagenesis, identifying the channel(s) with coevolutionary methods is fast and less biased. In a sense, coevolutionary analysis uses “directed mutagenesis information” generated by natural selection. This information is potentially abundant and can be used readily. Of course, this methodology could be used in identification of other functional important regions.

Coevolution Related to Physicochemical Interactions and Secondary Structure

The proton pumping channel(s) plays an important role in residue coevolution of COI. However, it is definitely not the only factor. Protein structures should always be considered.

As shown in Table 3.2, the overall highest coevolutionary percentage is in the transmembrane region, and of using volume segregation (instead of polarity). This may imply the effect of the highly residue packing in protein cores. COI transmembrane region is the center of COI. A large proportion of residues in this region may form the
core of COI, which are then densely packed. It is reasonable to assume that residues interact more when they are closer, thereby, in highly packed regions residues form more interactions than in loosely packed regions (Gromiha and Selvaraj 2001).

The above protein core assumption cannot explain why volume segregation resulted in a higher overall percentage of coevolution than segregation based on hydrophobicity or polarity (Table 3.1). The formation of secondary structure is likely to play a role here. Previous studies have found that coevolution based on charge or polarity is stronger than that based on size (Neher 1994; Pollock, Taylor, and Goldman 1999; Fukami-Kobayashi, Schreiber, and Benner 2002). Although residue interactions maintained by hydrophobicity, charge, or polarity, such as electrostatic interaction, are usually more significant than the interactions maintained by volume (van der Waals force), studies have shown that it is the van der Waals force, rather than hydrogen bonding or electrostatic interactions, that plays the vital role in helical formation during protein folding (Kilosanidze et al. 2004). COI consists primarily of helices. Extensive volume coevolution may thus reflect the evolutionary constraints of COI imposed by helix formation. Consistent with this, we found that volume contributed less than polarity to the coevolutionary signal in the surface regions (Table 3.2), which mainly consist of turns and loops.

It is a surprise to find that hydrophobicity segregation resulted in coevolutionary percentages lower than polarity and volume segregations in all structural categories (Table 3.2). As recognized, hydrophobic interactions are extensive existences and play important role in the beginning of protein folding (Rose and Wolfenden 1993; Tsai, Maizel, and Nussinov 2002). The possible reason for these lower coevolutionary
percentages is that hydrophobic interaction is a broad effect and does not have strong pairwise specificity. Therefore, to any specific pair the constraints from hydrophobic interaction is weaker than other interactions, such as electrostatic interactions. Because the lack of pairwise specificity, the pairwise coevolutionary signal is weak and the coevolutionary percentage is, then, lower.

An additional point worth considering regarding to the relationship of co-evolution to structure and function is that the results may vary depending upon the selective environment. If a protein is undergoing functional divergence, or long-term adaptation, then functional sites may be more prone to both evolution and co-evolution. We have obtained some evidence that this is the case in lineages that were removed prior to this study; in the lineage leading to snakes, for example, there appears to be many physically paired substitutions along the proton channels that are otherwise well conserved among the vertebrates (Jiang, Wang, and Pollock, unpublished data). We have hypothesized here that coevolution surrounding proton channel H is indicative of greater functional importance. Although the regions in and around the proton channels are generally more conserved than the rest of the protein, their functional roles would make them more likely targets of adaptive and coadaptive change. Thus, careful analysis of both conservation and coevolution in large-scale studies of molecular evolution may lead to more useful functional predictions than would either on their own.

References


Chapter IV: Context Dependent Coevolution in Protein Complex Cytochrome C Oxidase Detected Using Bayes Factor Analysis
**Introduction**

The requirements of functional and structural integrity are major sources of natural selection during protein evolution. Deleterious mutations that destabilize structure or harm function are either removed by selection or compensated by subsequent mutation(s). Such compensation results in residue coevolution by definition, since a replacement at one site has affected the tendency to replace the residue at another site. Residue coevolution can also arise from adaptation, when a series of residue replacements are required to effect the structural and functional alterations needed arrive at a new selective equilibrium. No matter how coevolution occurs, the necessity of meeting structural and functional requirements is the ultimate cause. Understanding coevolution, therefore, can augment the understanding of the structural and functional mechanisms of proteins.

Indeed, coevolutionary studies have had strong functional/structural implications (Neher 1994; Shindyalov, Kolchanov, and Sander 1994; Taylor and Hatrick 1994; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Tuffery and Darlu 2000; Wollenberg and Atchley 2000; Pritchard et al. 2001; Fukami-Kobayashi, Schreiber, and Benner 2002; Valencia and Pazos 2002; Govindarajan et al. 2003). Most of these studies have focused on individual peptides, but many (if not most) proteins are integral parts of protein complexes or protein interaction networks. Coevolution appears to arise from the functional constraints of protein networks (Goh et al. 2000; Gladysheva et al. 2003; Fraser et al. 2004; Gildor et al. 2005; Dou et al. 2006), but there are no published studies of coevolution within a protein complex. There may be fundamental differences, since subunits in a complex tend to interact with each other more closely than partner proteins involved in a network. Analyzing and characterizing
the coevolution of subunits in protein complexes will provide greater understanding of
the context dependence of coevolutionary processes, and may shed light on the nature of
subunit interactions in the particular complex studied.

Cytochrome c oxidase (CO) is the end complex of the respiratory chain, located in
the inner mitochondrial membrane. It plays a vital role in oxidative phosphorylation as
the final electron transporter and the proton-pump. In this complex, protons are pumped
from the matrix to the intermembrane space while electrons are transferred and accepted
by oxygen. In vertebrates, CO functions as a dimer with 13 subunits in each monomer,
three of which (COI, COII, and COIII) are encoded by mitochondrial genome (Tsukihara
et al. 1996; Collman et al. 1997). These three subunits are the core of CO because these
subunits are directly involved in catalytic activity while the other ten nuclear subunits are
mainly assembly components (Bratton et al. 2000; Richter and Ludwig 2003). The
sequence, structural, and functional features of these three subunits are conserved from
bacteria to vertebrates, although the bacterial CO has only 4 subunits and there is no
evidence that it functions as a dimer (Iwata et al. 1995; Tsukihara et al. 1996; Gennis
1998; Yoshikawa et al. 1998; Svensson-Ek et al. 2002; Richter and Ludwig 2003). COI
and COII are where the redox centers are located, and they participate in electron and
proton transfer directly. No residues in COIII are directly involved in these activities, but
removal of COIII from the complex destroys the normal functions of CO (Hosler 2004).
COI and COIII consist primarily of parallel and anti-parallel transmembrane helices of
similar sizes, and both subunits are highly compacted. These structural similarities and
functional differences make them ideal for characterizing the coevolutionary mechanism
in the protein complex.
Several methods have been used to analyze coevolution. A common approach is to evaluate the degree of residue correlation or mutual information between sites in a protein alignment (Korber et al. 1993; Neher 1994; Taylor and Hatrick 1994; Thomas, Casari, and Sander 1996; Giraud 1998; Hoffman, Schiffer, and Swanstom 2003; Saraf, Moore, and Maranas 2003). These methods assume that sequences in the alignment are statistically independent. This assumption is technically incorrect because sequences are not independent, but rather are derived from a common ancestor. Ignoring phylogeny can weaken the detected signal (Fukami-Kobayashi, Schreiber, and Benner 2002) because it does not allow discrimination between coevolution and noise from covariation due to common ancestry (Pollock and Taylor 1997; Wollenberg and Atchley 2000). Thus, these methods are most appropriate for use in cases where all of the sequences are highly diverged from one another (although problems with alignment accuracy may arise in such cases).

Another approach is to infer coevolution based on phylogenetic reconstructions of ancestral protein sequences (Shindyalov, Kolchanov, and Sander 1994; Fukami-Kobayashi, Schreiber, and Benner 2002). Excess co-occurrence of substitutions along the same branches of the phylogenetic tree is viewed as evidence for coevolution. Coevolutionary substitutions that occur on adjacent branches are ignored, and these methods may lose some power because the ancestral sequences are reconstructed under the assumption of independent evolution among sites. In addition, the reliability and uncertainty of ancestral sequence-reconstruction is a major concern (Tuffery and Darlu 2000; Suzuki and Nei 2001; Krishnan et al. 2004).
We previously utilized likelihood ratio analysis to detect residue coevolution (Pollock, Taylor, and Goldman 1999; Wang and Pollock 2005; Wang and Pollock 2006). This model-based method utilizes the phylogenetic tree, integrates over all ancestral reconstructions for an explicit dependent model in addition to an independent model, and considers coevolution along adjacent branches as well as along the same branch. It also has good sensitivity and statistical power. Two potential limitations of this approach, however, are that it requires nested dependent and independent models, and that the likelihood values may be inaccurate due to lack of data when sites are evolutionary conserved. Conserved sites lack variation (even though they are not invariant), and the dependent model therefore tends to be over-parameterized. To overcome these limitations and to integrate over model uncertainty, here we employed the Bayes factor (Kass and Raftery 1995) rather than the likelihood ratio.

Bayesian inference has become a powerful tool for evolutionary analysis, including phylogenetic inference, ancestral sequence reconstruction, and detecting positive selection (Huelsenbeck et al. 2001; Huelsenbeck, Larget, and Alfaro 2004; Mau, Newton, and Larget 1999; Nielsen 2001; Lartillot and Philippe 2004). The popularity of Bayesian inference in evolutionary studies comes from its capacity to construct complex models and from its inclusion of model uncertainty. In contrast to traditional phylogenetic methods (parsimony or maximum likelihood based), which select the single best estimator and assess its uncertainty by bootstrapping, Bayesian approaches express uncertainty by providing marginal posterior probability distributions for parameters of interest. In addition, Bayesian inference using the Bayes factor (BF) does not require that models be nested. In analyses of a pair of potentially coevolving sites, an empirical Bayes
approach must be used, since there is insufficient data to estimate the phylogeny from a pair of sites alone. Given the context dependency of residue evolution and coevolution, lack of data is not uncommon for conserved sites and can’t be removed by simply increasing the number of sequences (Wang and Pollock 2006). In addition, Bayesian inference does not require the tested models to be nested. These features of Bayesian inference make Bayes factor a good statistic for coevolutionary analysis once computing power is not a big concern. Non-nested models can be easily tested. Noise from lack of data can be reduced by including model uncertainty. Further, a Bayesian-based likelihood method can avoid local maximum, a major concern when using the maximum likelihood method.

To evaluate the utility of the BF, we used it to analyze coevolution within and between two mitochondrially-encoded subunits (COI and COIII) in the CO of vertebrates. This provided an opportunity to compare the results of the BF analysis with previous results from likelihood ratio analysis of COI (Wang and Pollock 2006). Since the restrictions on model nesting and model complexity were relaxed, we were able to employ a somewhat more general model of coevolution. We were also able to critically evaluate the role of evolutionary rate on coevolutionary processes. Since we used a dense sampling of vertebrate mitochondrial sequences and analyzed a large number of sites, the computational requirements of this approach are extremely high; almost 200,000 pairwise comparisons were evaluated in this study. In all comparisons, we used a fixed phylogenetic tree based on most of the mitochondrial genome to alleviate computational demands. The results provide a window into the role of protein complex interactions and the context of functional constraints on the coevolutionary process.
Materials and Methods

Sequences and Their Prior Processing

421 vertebrata cytochrome c oxidase sequences (COI and COIII) were retrieved from our mitochondrial genome database (www.egenbio.lsu.edu) and aligned with ClustalX (Thompson et al. 1997). In consideration of sequence density and distribution (Pollock and Taylor 1997), the alignment was applied to the decrease redundancy program (Notredame, unpublished) of ExPASy (Gasteiger et al. 2003) by setting the maximum pairwise identity as 95%. 290 sequences were thus filtered out, and an alignment of 131 sequences with pairwise identity less than 95% was left. The topology of these 131 sequences was constructed using all 13 mitochondrial encoded proteins from their corresponding species by PROTDIST of PHYLIP (Felsenstein 1989), and the branch lengths of this topology were modified using PROML of PHYLIP. Figure 4.1 shows the final tree topology. This tree was used as a “known” parameter in our subsequent coevolutionary analysis, although our methodology doesn’t require the phylogeny of the sequences to be known. Fixation of the tree parameter during computing saves a lot of time and the negative effect on the statistical power of the analysis is not be dramatic once the phylogenetic tree is reliable (Pollock and Taylor 1997). Using all the 13 mitochondrial encoded protein, instead of the targeted cytochrome c oxidase subunits, to construct the phylogenetic tree increases the reliability of the tree.

To avoid over-parameterization amino acids were classified into two states according to their physicochemical properties (polarity or volume) in a site specific manner. For example, a state of “small” will be assigned to a residue if its property value is less than the mean value of all residues occurring at that site in the alignment,
otherwise, “big” will be assigned. In this study, two residue quantitative properties, polarity and volume, are used separately.

Figure 4.1. The phylogeny of the 131 vertebrate species. It was reconstructed with all 13 mitochondrial encoded proteins.
In phylogenetic analysis, models characterized by a 20 X 20 empirical substitution matrix (such as JTT) are commonly used. Sometimes another random variable drawn from a gamma distribution is incorporated (Yang 1996). Models with a 20 X 20 matrix are not suitable for coevolutionary analysis because this matrix can result is hundreds of parameters and lead to over-parameterization. Over-parameterization produces over-fitting artifacts in likelihood-based analysis (Felsenstein 2004).

**Evolutionary Model Construction**

To avoid this problem, our study constructed a two-state independent model for each site and a four-state dependent model for each pair of sites. The states of these models are defined by classifying residues according to their polarity (Grantham 1974) or volume (Krigbaum and Komoriya 1979). The independent model, parameterized in terms of a 2 X 2 infinitesimal rate matrix, is the same as the one of Pollock et al. (Pollock, Taylor, and Goldman 1999). We constructed the dependent model by removal of some constraints of the dependent model of Pollock et al. in order to increase its suitability for more scenarios of coevolution. The resulting infinitesimal 4 X 4 rate matrix is:

\[
Q = \begin{pmatrix}
AB & Ab & aB & ab \\
AB & -\sum_{AB} & \lambda_1 \pi_{AB} & \lambda_2 \pi_{aB} & 0 \\
Ab & \lambda_1 \pi_{AB} & -\sum_{Ab} & 0 & \lambda_3 \pi_{ab} \\
aB & \lambda_2 \pi_{Ab} & 0 & -\sum_{aB} & \lambda_4 \pi_{ab} \\
ab & 0 & \lambda_3 \pi_{Ab} & \lambda_4 \pi_{aB} & -\sum_{ab}
\end{pmatrix}
\]

In this modified model are seven free variables, four rate (\(\lambda\)) variables and three free frequency (\(\pi\)) variables (the sum of the four frequencies must equals 1). The dependent model assumes that the substitutions of the two sites are correlated to one another while
the independent model assumes that the evolution of one site is independent of the states of other sites (for more details, please refer to Pollock et al. (1999)). These two models were applied to each pair of sites. If data from any pair of sites support the dependent model more strongly than the independent model, coevolution between this pair can be concluded; otherwise, no coevolution exists between them.

After setting up the models, state transition probabilities $P(t)$ ($P(t) = \exp(Qu)$) were calculated through standard matrix operation. Likelihoods were then computed following the “pruning” algorithm (Felsenstein 1981). For more details please refer to Pollock et al. (Pollock, Taylor, and Goldman 1999).

**Bayes Factor**

Assessing coevolution involves deciding whether the dependent model fits the data better than the independent model. Traditionally, a likelihood ratio statistic is used. The ratio of the two maximum likelihood values for these two models is measured. A significant ratio, supporting a better fit of the dependent model, suggests coevolution of the two sites (Pagel 1994; Pollock, Taylor, and Goldman 1999). Bayes factor is an alternative to likelihood ratio, which has been proposed as a good candidate for model comparison (Kass and Raftery 1995; Larget and Simon 1999; Suchard, Weiss, and Sinsheimer 2001). As defined by Kass and Raftery (Kass and Raftery 1995), the Bayes factor is the ratio of the posterior odds of a model to its prior odds. In cases of two models (model 0, $M_0$, and model 1, $M_1$) and observed data, $D$, the Bayes factor is

$$B_{10} = \frac{pr(D \mid M_1)}{pr(D \mid M_0)}.$$  

$pr(D \mid M_i)$ is the marginal likelihood of model $i$ (either model 0 or model 1). Similar to likelihood ratio, the values of $B_{10}$ greater than 1 indicate model 1 fit data $D$ better than
model 0 while values less than 1 indicate model 1 a worse fit than model 0. In the present study, model 1 (\(M_1\)) is the dependent model and model 0 (\(M_0\)) is the independent model. Because in \(M_0\) we treated the two sites independently the denominator of the Bayes factor was calculated as

\[
pr(D \mid M_0) = pr(D_{s1} \mid M_0) \times pr(D_{s2} \mid M_0)
\]

where \(D_{s1}\) and \(D_{s2}\) correspond to the data observed at the two sites separately. Since our models are not nested, we can’t approximate Bayes factor by Savage-Dickey ratio (Verdinelli and Wasserman 1995). Instead, we used a Markov chain Monte Carlo (MCMC) algorithm (Hastings 1970) to estimate it.

**Bayesian Computation Using MCMC**

Following Newton and Raftery (Newton and Raftery 1994), we used the harmonic mean of the likelihood values of a sample from the posterior distribution to estimate \(pr(D \mid M)\).

\[
pr(D \mid M) = \left\{ \frac{1}{k} \sum_{i=1}^{k} pr(D \mid \theta^{(i)}, M) \right\}^{-1}
\]

This estimation has been verified to converge to the correct value of \(pr(D \mid M)\) when \(k\) is large enough. However, it does not satisfy a Gaussian central limit theorem and is unstable since \(\theta^{(i)}\) with a small likelihood would have a large effect. Newton and Raftery (Newton and Raftery 1994) thus proposed adding some imaginary samples whose likelihood values are the expected likelihood to the \(k\) samples from the posterior. If \(\tau k/(1-\tau)\) samples are added, the estimation of \(pr(D \mid M)\) can be approximated as

\[
pr(D \mid M) = \frac{\tau k/(1-\tau) + \sum_{i=1}^{k} pr(D \mid \theta^{(i)}) \{ pr(D \mid M) + (1-\tau) pr(D \mid \theta^{(i)}) \}^{-1}}{\tau k/(1-\tau) pr(D \mid M) + \sum_{i=1}^{k} \{ pr(D \mid M) + (1-\tau) pr(D \mid \theta^{(i)}) \}^{-1}}
\]
Even if $\tau$ is as small as 0.01, this new estimation overcomes the instability of the previous one.

The computation was carried out with MCMC incorporated with the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). Although Bayesian analysis does not require the phylogenetic tree to be a fixed parameter, we took the phylogeny of the 131 sequences as known to simplify the computation. To ensure proper priors we set rate parameters to be uniformly distributed and the stationary frequency parameters to be Dirichlet random variables (Huelsenbeck et al. 2001; Huelsenbeck, Larget, and Alfaro 2004). In each iteration of the MCMC, a parameter was randomly chosen to update. For rate parameters, we proposed a new value $\lambda^* = \lambda \varepsilon$, where $\varepsilon$ is picked uniformly at random between $1-\delta$ and $1+\delta$. $\delta$ was tuned as 0.5 in this study. To minimize computing time $\lambda^*$ is assured to be in (0, K) (rate should not be less than 0 and not bigger than an appropriate maximum, K). K was 20 in this study. The acceptance of the proposed rate value is

$$\min [1, \frac{\text{Likelihood of the proposed state}}{\text{Likelihood of the current state}}]$$

For stationary state frequencies ($\pi$), new values ($\pi^*$) were derived from a Dirichlet distribution with $\pi$ as expected values. Here $\pi$ and $\pi^*$ are vectors of four or two elements corresponding to the frequencies of the four or two states in the two models. The Dirichlet parameters ($\mu$) of the distribution were set as $B^* \pi$ ($\pi$ is a vector here), $\mu = B^* \pi$. B is a constant number and was tuned 150 in the process. The acceptance of the proposed new values is

$$\min [1, \frac{\text{Likelihood of the proposed state}}{\text{Likelihood of the current state}} \times \frac{D(\pi^* | \mu)}{D(\pi | \mu)}]$$
where $D(\pi^*|\mu)$ and $D(\pi|\mu)$ are the probability densities of the state frequency values from the Dirichlet distribution. Tuning B and $\delta$ was done to maintain good acceptance rates for MCMC.

Besides the Bayes factor values, the parameter values at the maximum likelihood were recorded as their maximum likelihood estimates (MLE). The MLEs of the rate parameters in the independent models were used as the evolutionary rates to categorize sites in the subsequent analysis.

**Coevolving Sites**

Two sites were considered as a putative coevolving pair if the natural logarithm of the Bayes factor of them was bigger than the cutoff we set. Bayes factors bigger than 2.0 were considered as stronger evidence against the independent model by Kass and Raftery, and those bigger than 5.0 were treated as very strong evidence (Kass and Raftery 1995). We took 4.0 and 5.0 as our cutoffs. The percentages of these coevolving pairs in different structural categories (COI, COIII, and between COI and COIII) or evolutionary rate categories were recorded and statistically analyzed, and the physical distance distributions of coevolving pairs were also analyzed. Coevolving pairs with Bayes factors bigger than 5 to the polarity vector were further examined against the crystal structure of bovine CO (Tsukihara et al. 1996) to achieve more insight understanding of the relationship between coevolution and protein structure and/or function. PyMOL (DeLano 2002) was used to display its structure.
Results

coevolution and evolutionary rates

Previous studies have implied that residue coevolution might be related to the evolutionary rate of the residues (Pupko et al. 2002; Fraser et al. 2004; Jordan et al. 2004). To detail this relationship, we classified the residue sites into three categories, highly conserved (HC), moderately conserved (MC), and variable (V), and compared the coevolutionary percentages within each category and between different categories. The HC sites were defined as the sites that have their minor state frequency less than 0.03 (i.e., among the 131 sequences less than 4 sequences have the minor state and all the other sequences have the major state after masked with the designated vector). Because of lack of data, the MLEs of rate parameters (of the independent model) of these highly conserved sites are not reliable (Sites with minor state frequency larger than 0.03 might have the lack of data problem for MLE too). Other sites were divided into two groups, MC and V, based on their MLE rates and the mean rate (the average rate of all the other sites except for those highly conserved sites). Specifically, if one site has its rate less than the mean, it belongs to MC, otherwise, V.

Table 4.1 shows the coevolutionary percentages within each category and between different categories. The coevolutionary percentages within the HC sites are extremely low, reflecting the limited coevolution among these sites. The most extensive coevolution exists within the MC sites. Their coevolutionary percentages are more than 30 times higher than those within highly conserved sites. The MC sites also coevolve with both HC sites and V sites. The coevolution between MC sites and HC or V sites are significantly more extensive than those within HC or V sites correspondently (p<0.0001).
Their percentages are about 2-5 times higher. Only a small amount of coevolution between HC and V sites was detected with percentages less than 0.15% (p<0.0001). This is about a 4-fold decrease compared to the coevolving percentages between MC and HC or V. These results are similar when analyzed by either the polarity vector or the volume vector.

Table 4.1. Coevolving percentages within and between different sites categories at BF>4 to the two vectors (polarity and volume)

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>MC</th>
<th>V</th>
<th>HC-MC</th>
<th>MC-V</th>
<th>HC-V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polarity</td>
<td>0.05%</td>
<td>1.58%</td>
<td>0.14%</td>
<td>0.67%</td>
<td>0.59%</td>
<td>0.19%</td>
</tr>
<tr>
<td></td>
<td>(8783)</td>
<td>(2780)</td>
<td>(2150)</td>
<td>(9975)</td>
<td>(4955)</td>
<td>(8783)</td>
</tr>
<tr>
<td>Volume</td>
<td>0.05%</td>
<td>1.42%</td>
<td>0.25%</td>
<td>0.44%</td>
<td>0.68%</td>
<td>0.06%</td>
</tr>
<tr>
<td></td>
<td>(8516)</td>
<td>(5996)</td>
<td>(6110)</td>
<td>(14410)</td>
<td>(12215)</td>
<td>(14546)</td>
</tr>
</tbody>
</table>

HC: within highly conserved sites; MC: within moderately conserved sites; V: within variable sites; HC-MC: between highly conserved sites and moderately conserved sites; MC-V: between moderately conserved sites and variable sites; HC-V: between highly conserved sites and variable sites.

Coevolution within Subunits and across Subunits

The COI and COIII subunit play different roles in the cytochrome c oxidase complex. COI contains the catalytic site and proton channels, while COIII is not directly involved in the functional activities of the complex. It is interesting to compare the coevolutionary signals within these two subunits. The percentages shown on Table 4.2 are calculated excluding the HC sites because the portions of HC sites in different subunits differ substantially, and the HC sites exhibit lower coevolving percentages (Table 4.1). The coevolving percentages in COI are higher than in COIII and between COI and COIII (Table 4.2). The difference is greater for the polarity vector than for the volume vector. For the polarity vector, coevolutionary percentage within COI is about 80% higher than within COIII while it is about 50% higher for the volume vector. Statistical test confirms
the significance of these differences (P<0.05). It might be expected that coevolving percentages between subunits should be weaker than those within either COI or COIII, In fact, the coevolutionary percentages between COI and COIII are close to those within COIII although both are significantly lower than those within COI (P<0.05).

Table 4.2. Coevolution percentages within and between the two subunits according to polarity and volume vectors (BF>4). Highly conserved sites are excluded

<table>
<thead>
<tr>
<th></th>
<th>COI</th>
<th>COI-COIII</th>
<th>COIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>polarity</td>
<td>1.1%(2556)</td>
<td>0.63%(3403)</td>
<td>0.60%(2346)</td>
</tr>
<tr>
<td>volume</td>
<td>0.98%(7021)</td>
<td>0.67%(6150)</td>
<td>0.64%(5151)</td>
</tr>
</tbody>
</table>

Coevolution and Physical Distances

The degree of coevolution is often related to the physical distance between residues in individual proteins (or interacting ligand and receptor) (Pollock, Taylor, and Goldman 1999; Fukami-Kobayashi, Schreiber, and Benner 2002; Wang and Pollock 2005; Wang and Pollock 2006). To visualize the relationship between residue distance and coevolution, we plotted the C-alpha distance distributions of coevolving pairs and compared these distributions to the distance distributions under the null model (the distributions of all the pairs in the corresponding analysis). Figure 4.2 shows that the frequencies of coevolving pairs of proximal residues within subunits (including those within COI and those within COIII) significantly increased, compared to their expected frequencies under null. These increases are greater at the higher cutoff of Bayes factor, which means proximal residues tend to coevolve more closely (higher Bayes factor values). For example, for the volume vector, nearly 9% of coevolving pairs at BF>4 have pairwise distance less than 8 Å, but at BF>5 this value increase to 16%. These results are very significant (p<0.01) compared to the expected frequency under null, which is about
3%. For either the polarity vector or the volume vector, about 40% of coevolving pairs (with BF >5) have distances less than 16 Å, while the expected frequency is about 15%. At the same time, residues as far as 56 Å apart can coevolve. About 40% of coevolving pairs (BF>5) have distances greater than 24 Å even at BF>5. The physical distance between residues is a significant factor for the coevolution within individual subunits, but long distances do not preclude coevolution. The distance distribution of coevolving pairs between COI and COIII is apparently different. More than 90% of coevolving pairs between COI and COIII are more than 24 Å apart, and the C-alpha distance distributions of the coevolving pairs across subunits are not biased to short distances (Figure 4.3). Tightly coevolving pairs (BF>5) between subunits are also not proximal. Among the about 10 pairs with distances less than 8 Å analyzed, to either the polarity vector or the volume vector, none coevolve (data not shown). The underlying coevolutionary constraints clearly differ within subunits versus between subunits.

**Localization of Coevolving Sites**

To obtain further insight about coevolutionary constraints, we manually checked the location of coevolving pairs in COI and COIII to the polarity vector and at higher cutoff (BF>5). Choosing the polarity vector and a higher cutoff reduced the number of coevolving pairs, especially for the case of coevolving pairs between COI and COIII, thereby facilitating visualization. The coevolving pairs are displayed based on the crystal structure of bovine CO (PDB ID: 2OCC) (Tsukihara et al. 1996). The coevolving pairs within COI are largely distributed around the proposed H proton pumping channel and the proton loading surface (Figure 4.4). This is consistent with our previous study (Wang and Pollock 2006). This result implies that maintaining functions, such as proton
pumping, is an important coevolutionary constraint within this subunit. In COIII, the coevolving pairs are heavily distributed on the three helices contacting COI and the matrix surface (N-side) of the subunit (Figure 4.5). However, the contact between COI and COIII does not result in noticeable coevolution between the residues at the interface of the two subunits (Figure 4.6). Most coevolving pairs across COI and COIII are located near the two cap regions, either on the surfaces (both the intermembrane and the matrix) or one to two helices away from the surfaces (Figure 4.6). Among these coevolving pairs, none are directly in contact with each other and only one is proximal (distance $<$ 16 Å). This leads us to propose that coevolution between COI and COIII is induced by global constraints, which might involve the interactions between the complex and the membrane lipids and the functional constraints on the surfaces.

Figure 4.2. Frequencies of coevolving pairs within COI or COIII to the vectors of residue polarity (Left) and volume (Right), binned to C-alpha distances. The expected frequencies under null hypothesis are in blue line. The observed frequencies at BF $>$ 4 are in yellow bars. The observed frequencies at BF $>$ 5 are in brown bars.
Figure 4.3. Frequencies of coevolving pairs across COI and COIII to the vectors of residue polarity (Left) and volume (Right), binned to C-alpha distances. The expected frequencies under null hypothesis are in blue lines. The observed frequencies at BF>4 are in yellow bars. The observed frequencies at BF>5 are in brown bars.

Discussion

Although the context-dependence of coevolution has been reported previously (Wang and Pollock 2005; Wang and Pollock 2006), the pattern and mechanism of coevolution remains poorly understood, especially in protein complexes. In this study we analyzed the coevolution of two CO subunits using Bayes factor. The differential coevolving percentages among the residues with different evolution rates reveal that the moderately conserved sites are involved in coevolution more than other sites. The distance distributions and localizations of coevolving pairs within subunits and the strong coevolutionary signal within COI suggest that functional integrity plays a crucial role in coevolution. The detection of coevolving pairs between COI and COIII suggests the existence of global constraints. Interestingly, we also found that the interaction between COI and COIII drives coevolution within COIII instead of between COI and COIII. In addition, this work demonstrates that implementing Bayes factor increases model
capacity, a feature which will facilitate multiple model comparisons for detecting coevolution in the future.

**Figure 4.4. Locations of the sites involved in coevolution within COI.** The sites are shown in green spheres; the proposed proton channels (H, K, and D) are shown in dots. The channel H is shown in red; the channel K is shown in yellow; the proposed proton channel D is shown in grey.

**Figure 4.5. Locations of the sites involved in coevolution within COIII.** The sites are shown in green spheres. COIII is shown in blue tubes; COI is shown in grey tubes.
Relative Conserved Sites Are Important for Coevolution

For both evolutionary analyses and structural/functional studies, it is of key interest to know what sites are more likely to coevolve. Our results show that coevolution is significantly concentrated among the moderately conserved sites and there are only very few cases in which the highly conserved sites coevolve. These results are consistent with current understanding about coevolution. It is known that coevolution comes from two scenarios, compensation and adaptation. To coevolve, substitutions must be able to accumulate at both sites in a relatively short period, and must be subject to some selective constraints (or the removal of selective constraints in some cases of adaptation). Fast evolving sites easily accumulate substitutions, but because of their structural/functional insignificance, most of these substitutions are neutral, and chances for them to coevolve
are few. The highly conserved sites, on the other hand, are strictly constrained by structure or function making most mutations at these sites too deleterious to survive. It would be hard enough to get two substitutions accumulated, let alone two compensatory ones. Even if a mutation survived at one site, it may only be compensated by back mutations (Poon, Davis, and Chao 2005) or be compensated by substitutions from other sites with higher mutation rates (these sites produces more substations than the conserved sites). This is supported by that the coevolutionary percentages between highly conserved sites and moderately conserved sites are higher than those within highly conserved sites. In the same way, it is not hard to understand why the moderately conserved sites are most likely to coevolve. The functional or structural constraints may be less stringent. Mutations at these sites therefore are more likely to be tolerated, allowing them to survive long enough for subsequent adaptive or compensatory mutations to occur. Detecting coevolution among moderately conserved sites should be a very valuable tool for protein structural and functional analyses. By including the coevolutionary information among moderately conserved sites, researchers were able to manipulate the structure and function of a peptide by site directed mutagenesis while structure and function of the peptide were most likely destroyed by random mutagenesis without this information (Russ et al. 2005; Socolich et al. 2005).

**Functional Integrities Are Dominant Coevolutionary Constraints**

The differences of the coevolutionary percentages within COI and within COIII illustrate the functional dependency of coevolution. Both subunits consist primarily of alpha helices, but COI contains 3 redox centers and related functional domains such as proton
pumping channels. The stronger coevolutionary percentages within COI indicate the significant role of the functional constraints from these structures.

The distribution of the coevolving sites within COI provides more direct evidence for the important role of functional constraints on coevolution. The coevolving sites within COI are heavily distributed around one of the proposed proton pumping channels and the N-side (the matrix side) surface of the complex (Gennis 1998; Yoshikawa et al. 1998; Pereira and Teixeira 2004), which agrees with our previous analysis using maximum likelihood ratio (Wang and Pollock 2006). But, as shown in Figure 4.3, the residues proposed to be involved in proton pumping don’t participate in coevolution directly except for the one that is located at the proposed entrance. These results suggest the properties of the environment around the channel are significant constraints on amino acid substitutions occurring in those regions. We suggest that appropriate polarity is the constraint in COI. The coevolving pairs located on the N-side surface might be involved in forming a charged surface for proton loading. The coevolving pairs around the channel appear to function in maintaining the potential for proton transport. Detecting this type of coevolution provides information that can be used to better characterize the functional domains.

Coevolution could also be produced by functional constraints which extend over the entire complex or protein rather than on a specific region. This type of coevolution exhibits a global pattern. The coevolution between COI and COIII reflects this type of global coevolution. Manually checking the structural locations of the coevolving pairs across subunits found that these sites are located at the cap regions of COI and COIII. The ends of the helices are located in these regions and are where the head groups of
membrane lipids may interact with protein (Killian and von Heijne 2000). This result suggests that one of the global coevolutionary constraints is the interaction between the protein and the membrane lipids. In vertebrates, functional CO is a dimer. Studies have revealed that the interaction between membrane lipids and CO subunits are involved in cytochrome c oxidase dimerization (Seelig and Seelig 1978; Musatov, Ortega-Lopez, and Robinson 2000; Musatov and Robinson 2002), and that this lipid-facilitated dimerization depends on a full complement of subunits (Musatov and Robinson 2002). Therefore, the coevolution between COI and COIII might be partly driven by the dimerization of the complex mediated by lipids. Membrane lipids also are important for membrane anchoring of the CO complex. It has been found that more than 50 lipids are restricted motionally by a CO molecule in the membrane (Marsh et al. 2002). Because of the lack of details about the location (on CO surface) of these restricted lipids, we can’t specifically define the coevolution driven by them though we know that the anchoring involves the entire complex. There are also some coevolving sites located on the P-side and N-side surfaces. As noted above, these sites might reflect the constraints from proton loading, or from the interaction between CO and its substrate, cytochrome c. Though cytochrome c binding might involve primarily negatively charged residues in COII (Roberts and Pique 1999), the number of lipids motional restricted by CO significantly increases when cytochrome c is bond (Marsh et al. 2002). This indicates cytochrome c interacts with the CO complex globally because cytochrome c binding domain of COII locates on the P-side surface of the CO complex and does not interact with lipids directly. The coevolving pairs on the P-side surface might reflect residues in COI and COIII that participate in cytochrome c binding. Overall, the non-negligible coevolutionary percentages between the two
subunits suggest that functional integrity is a significant coevolutionary constraint at the global level also. Although COIII does not participate in catalytic activity directly, it effects the energy conversion (Wu, MorenoSanchez, and Rottenberg 1995) and product turn over (Bratton, Pressler, and Hosler 1999; Lincoln et al. 2003) of CO. Evidence also shows that COIII influences the conformation of COI and COII (Echabe et al. 1995). These findings underscore that constraints between subunits are quiet different from those on interacting protein partners such as receptors and ligands, where the physicochemical interaction between the contacting residues carry out the function by triggering conformational changes in one protein or both.

Because most coevolution is driven the functional integrity, coevolving pairs are not restricted to proximal residues. Distant pairs within subunits and between subunits coevolve. These coevolving distant residues are good examples of long range interactions in proteins and protein complexes. The effects on the catalytic site by residues 10-20 Å away have frequently been reported (Thomas, Russell, and Fersht 1985; Russell and Fersht 1987; de Kreij et al. 2002; Clarkson and Lee 2004; Dupeyrat et al. 2004; Thorpe and Brooks 2004). Some data have suggested that residues as far as 30 Å are coupled by function (Rajagopalan, Lutz, and Benkovic 2002). Uter and Perona (Uter and Perona 2004) reported that the active site configuration of a glutaminyl-tRNA synthetase depends on enzyme-tRNA contacts about 40 Å away. The present coevolutionary analysis and our previous (Wang and Pollock 2006) study suggest that residues may functionally compensate one another over distances 40 or even 50 Å on functionality.
Within Subunit Coevolution Resulted from Inter-Subunit Interaction

Most coevolving pairs within COIII are located in the three helices close to COI. Because COIII is not involved in catalytic activity and proton pumping directly (Bratton et al. 2000; Yoshikawa 2003; Namslauer and Brzezinski 2004), these coevolving pairs could not have resulted from enzymatic functions of these helices themselves. A reasonable explanation is that the functional integrity of COI constrains these COIII helices. COIII has been found to influence the catalytic function of CO even though that its residues do not physically participate in the enzyme activity (Echabe et al. 1995; Wu, MorenoSanchez, and Rottenberg 1995; Bratton, Pressler, and Hosler 1999; Meunier and Taanman 2002; Hosler 2004). Specifically, COIII might play a role in maintaining the structural integrity of the CuB center (Bratton, Pressler, and Hosler 1999) and in the assembly of CO (Meunier and Taanman 2002). These functions mean that there are interactions between COIII residues and residues of other subunits. It might be expected that these interaction would result in coevolution between COIII and other subunits (such as COI) at their interacting regions. In fact, we did not find notable coevolution between COI and COIII in their interacting regions. However, we found that many coevolving sites within COIII were located in the regions contacting COI. We suppose that the interactions between COI and COIII do not result in coevolution between subunits, instead, they result in coevolution within one of the subunits. It is know that interface mutations sometimes destabilize the subunits instead of the interfaces (Brown, Liao, and Wittung-Stafshede 2005). In the complex any interface destabilizing mutations might more likely be compensated by residues in COIII since that the substitution rate of COIII residues is higher than that of COI residues. This scenario also indicates that the
interactions between subunits of a complex may not be composed of residue-specific interactions, in contrast to the direct residue-interaction seen between protein partners. The other coevolving pairs within COIII, which are away from the interface and near to the N-side surface, may also be related to the interaction between COI and COIII. This region has previously been found to influence the activity of CO (Lincoln et al. 2003).

**Using Bayes Factor Instead of Likelihood Ratio**

Our results for COI using Bayes factor in this study and the results using likelihood ratio in our previous study (Wang and Pollock 2006) are consistent. Both revealed the proposed H channel to be a significant evolutionary constraint, thus, supporting the functionality of this channel. The important advantage of implementing Bayes factor in coevolutionary analysis is the increased model capacity. Bayes factor analysis does not require the models to be nested and allows the comparison of multiple models at the same time. This feature of Bayes factor analysis should facilitate more complete studies. Because of the context dependence of residue coevolution, each pair may coevolve to somewhat a different pattern. To fully characterize these patterns and characterize their coevolution, different possible coevolving models have to be compared. A program utilizing Bayes factor analysis can compare these possible models directly and suggest which one fits the data best. In this way, our program will significantly speed up the study of context dependent evolution and coevolution. It should be noted, however, that a big concern for Bayes factor is the high intensity of computation. It may take up to 10 times longer than a likelihood ratio based test. If computation power is scarce, BIC (Raftery 1986) or AIC (Akaike 1981) might be good alternatives to Bayes factor.
The present study analyzed the coevolution within COI and COIII and that between COI and COIII. We provide more evidence that function is the most significant coevolutionary constraint and that the interaction between subunits might produce coevolution within one subunit instead of between the two interacting subunits. To verify these information we need to develop more complicated models. For this point, our Bayes factor based methodology will be a great help since it does not require the models to be nested.

References


Chapter V: Structural and Functional Implications from Evolutionary Analysis of Photolyase/Cryptochrome Proteins
**Introduction**

With accumulation of data and increasing of computing power, more evolutionary information is becoming available, and scientists are becoming aware of its importance in recent decades. Using evolutionary analysis to facilitate protein structural and functional exploration has become a common practice. Residue conservation is utilized extensively as an indication of structural and functional importance. However, recent studies indicated that residue coevolutionary information is also necessary for the success of protein engineering (Russ et al. 2005; Socolich et al. 2005). Part of the reason why residue coevolution should be considered is that residue conservation alone cannot provide much information about the interactions among residues. Residue coevolution, on the other hand, reveals these interactions because it results from the compensatory and/or adaptive substitutions. It is known that protein structures and functions are usually domain based; that is, residues in a localized region closely interact with each other and form a relatively independent structural and/or functional unit. Substitutions at sites in these localized regions have a strong intendancy to induce corresponding substitutions (compensatory and adaptive) in the sites with which they interact because perturbation of their interactions is associated with changes of protein structure/function and thus under natural selection. These correlated substitutions are the main source of coevolution. Functional or structural changes in the course of evolution alter the selections on residue interactions and subsequently alter their coevolution (Wang and Pollock 2005). Analyzing coevolutionary patterns against the background of knowing structural/functional changes will greatly augment our understanding of the mechanisms that underlie protein structure and function. Obviously, the ability to distinguish the
coevolution of structural constrains from that of functional integrities is important. The ideal design for such distinguish would use either proteins with similar functions but different structures, or, proteins with different functions but similar structures. The former faces severe problems with sequence alignment, because proteins with different structures are usually not homologous and aligning non-homologous proteins leads to nonsense. Analyzing of coevolution over homologous proteins with similar structures but different functions eliminates this problem. Photolyases and cryptochromes form a superfamily of proteins that meet these conditions.

Photolyases are part of DNA repairing machinery in organisms from bacteria to vertebrates, but they are absent in placental mammals (Yasui et al. 1994; Kanai et al. 1997). Based on their substrate specificities, photolyases have been classified into two groups, CPD photolyases and (6-4) photolyases (Sancar 1994; Sancar 1996). The former group binds and repairs DNA with cyclobutane pyrimidine dimers in DNA (Sancar and Sancar 1984; Husain et al. 1988; Sancar 1994; Kanai et al. 1997), the latter group binds and repairs DNA with (6-4) pyrimidine-pyrimidone photoproduct (Mitchell and Nairn 1989; Sancar 1996).

Based on sequence similarity, CPD photolyases are further classified into two classes, class I and class II. Their sequence similarity is usually less than 20%. Cryptochromes share close sequence similarity with class I photolyases. Indeed, the similarity between class I photolyase and cryptochromes is greater than that between class I photolyases and class II photolyases (Kanai et al. 1997; Sancar 2003). Most cryptochromes do not possess DNA repair function. Instead, they are mainly involved in circadian activities, functioning as signal transducers for the regulation of gene
expression (Ahmad and Cashmore 1993; Lin and Shalitin 2003; Lin and Todo 2005). To
date, cryptochromes have been identified in both bacteria and eukaryotes. Within the
eukaryotes, they can be subdivided into two groups, plant and animal cryptochromes,
according to their sequence similarity. Interestingly, the sequence similarity between
these two types of cryptochromes is less than that between animal cryptochromes and (6-
4) photolyases (Lin and Todo 2005).

In recent years, a new group, named cryptochrome_DASH (CRY_DASH), has
been identified in bacteria and plants (Hitomi et al. 2000; Brudler et al. 2003; Daiyasu et
al. 2004). This group is closely related to animal cryptochromes, but has a DNA binding
activity (Brudler et al. 2003). The most distinct feature of the cryptochromes and
CRY_DASH subfamilies is that they have an extended C-terminus. The length of this
terminus varies from 50-500 amino acids (Brudler et al. 2003; Sancar 2003; Lin and
Todo 2005).

Both photolyases and the N-terminal region of cryptochrome have been
crystallized (Park et al. 1995; Tamada et al. 1997; Komori et al. 2001; Brautigam et al.
2004; Mees et al. 2004). The structures are surprisingly similar despite of their sequence
divergence. All structures consist of two domains, α/β domain and α domain. These two
domains are connected by a loop which is relatively conserved in length but variable at
the sequence level. In these structures, the cofactor FAD is buried deeply in the α domain
(thus it is also called the FAD-binding domain), and there is an access cavity connecting
it to the solvent. On the surface around this cavity, there is a positively charged groove in
the photolyases as well as CRY_DASH (Brautigam et al. 2004). In CPD-photolyase, this
cavity is where the substrate DNA binds, The CPD dimer of the substrate DNA has been
proposed to flip into this cavity and become accessible to the FAD catalytic center (Vande Berg and Sancar 1998; Mees et al. 2004). The cavity of cryptochromes is bigger than that found in photolyases or CRY_DASH and the positively charged groove is greatly diminished. In fact, the entire surface of the cryptochrome is predominantly negatively charged. Cryptochrome does not bind to a pyrimidine dimer (Lin and Shalitin 2003; Lin and Todo 2005). Overall, the functions of photolyase/cryptochrome family proteins are primarily associated with the FAD-binding domain because the FAD cofactor is generally considered as the reactive center (Lin and Shalitin 2003; Sancar 2003; Lin and Todo 2005).

The abundant information about the sequence, structure, and function of the photolyase/cryptochrome superfamily and, especially, the functional diversity and structural similarity of its members make them good candidates for exploring the relationship between primary sequence and structure/function using coevolutionary analysis. Aimed at retrieving detailed information about how residue coevolution correlates with structural/functional constraints, the present study investigated the residue coevolution and conservation of the \( \alpha \) domain of the photolyase/cryptochrome protein superfamily, and traced the substitutions of the coevolving sites through their phylogeny. We have constructed a methodology using Bayes factor analysis previously (Wang and Pollock 2006). In that study the Bayes factor of the 2-state independent model and 4-state dependent model was used to detect coevolution. For better detection, we extended our previous 2-state/4-state (independent/dependent) models to 3-state/9-state models utilizing residue charges (positive, neutral, negative) as the vector. Also we focused on the \( \alpha \) domain because its functions, structures, and the differences of its functions over
subfamilies are well understood. Our results suggest that analysis of residue coevolution can be an important tool for explore protein structure and function and that this methodology has the potential to provide further information towards the experimental engineering of photolyase.

**Materials and Methods**

**Sequence and Alignment**

We collected and preprocessed the sequences as following: Photolyase and cryptochrome sequences were originally downloaded from ExPASy (Gasteiger et al. 2003). The α/β domain and α domain sequences were extracted according to the domain limits (PF00875 and PF03441) as defined in Pfam database (Bateman et al. 2004). Sequences from each domain were aligned separately using ClustalX (Thompson et al. 1997). Extracting the sequences for the two domains only and aligning them separately were done to improve alignments, since the sequences of the loop region and the C-terminal extension of photolyase/cryptochrome superfamily proteins are extremely diverse (Kanai et al. 1997; Sancar 2003).

The two separate domain alignments were then combined by connecting the sequences of the same species using a Perl script. To eliminate redundant data, sequences with high identity in the combined alignment were filtered out using the decrease redundancy program (www.expasy.org/tools/redundancy, Notredame, unpublished) of ExPASy (Gasteiger et al. 2003). We set the maximum pairwise identity to 95%. The sequences in the filtered alignment were further manually checked to remove those introduced large insertions and deletions to the alignment.
The final alignment contained 100 sequences, including both the α/β domain and the α domain. A phylogeny was reconstructed based on this alignment using FITCH of PHYLIP (Felsenstein 1989), and its branch lengths were estimated using PROML of PHYLIP (Felsenstein 1989). This phylogeny (Figure 5.1) was used as a “known” parameter for the subsequent analyses in order to simplify the computation although our methodology does not require the fixation of the phylogeny parameter (Pollock, Taylor, and Goldman 1999). Though the present study focused on residue coevolution in the α domain, we constructed the phylogeny using residues from both the α and α/β domains. By using more residues, we expect the phylogeny, which we used as a known parameter in this study, to be more robust. An alignment of the α domain sequences was then extracted from the above-mentioned, reduced 100 species alignment, and was used as the sequence alignment for all subsequent analyses.

**Coevolutionary Analysis**

Residue coevolution in the α domain was analyzed by the method explored by Wang and Pollock (Wang and Pollock 2006). Briefly, two evolutionary models, one assuming evolutionary independency (the independent model) and the other assuming dependency (the dependent model) of two sites, were applied to each pair of sites; the Bayes factor computed from the likelihood of these two models was evaluated; pairs with significant Bayes factor values were screened as coevolving pairs and further analyzed.

Because residue charges may play an important role in the functional adaptation of photolyases and cryptochrome (Sancar 1994; Sancar 2003; Lin and Todo 2005), we used charge as a vector to mask the residues in the analyzed alignment; specifically Lys, Arg, and His are positively charged residues, Glu and Asp are negatively charged
Figure 5.1. The phylogeny of the 100 sequences. The three main clades, proteobacterial photolyases (24 sequences), plant cryptochromes (20 sequences), and animal cryptochromes (22 sequences), are labeled and their most recent common ancestors are marked by circles. These three clades are what we focused on during our substitution tracing.
residues, and all others are neutral (Fauchere et al. 1988). This masking resulted in an alignment of sequences with 3 states, positive (P), neutral (zero, Z), and negative (N).

Subsequently, our coevolutionary models are of three states (P, Z, and N in the independent model) and nine states (PP, PZ, PN, ZP, ZZ, ZN, NP, NZ, and NN in the dependent model). The instantaneous transition probability matrix of the independent model is

$$
P = \begin{bmatrix}
    -\sum p & \pi_Z \lambda_a & \pi_N \lambda_c \\
    \pi_p \lambda_a & -\sum Z & \pi_N \lambda_c \\
    \pi_p \lambda_b & \pi_Z \lambda_c & -\sum N
\end{bmatrix}
$$

which contains two free stationary frequency ($\pi$) parameters (the summation of frequencies equals 1) and three rate ($\lambda$) parameters. The corresponding matrix for the dependent model is

$$
\begin{align*}
PP & \begin{bmatrix}
    -\sum PP & \pi_{pz} \lambda_1 & \pi_{pn} \lambda_2 & \pi_{zp} \lambda_4 & 0 & 0 & \pi_{np} \lambda_2 & 0 & 0 \\
    \pi_{pp} \lambda_2 & -\sum PZ & \pi_{zp} \lambda_4 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0 & \pi_{xn} \lambda_6 \\
    \pi_{pn} \lambda_3 & -\sum PN & 0 & 0 & \pi_{zp} \lambda_5 & 0 & 0 & \pi_{xn} \lambda_6 & 0 \\
    \pi_{pp} \lambda_4 & 0 & 0 & -\sum ZP & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 \\
    \pi_{pz} \lambda_3 & 0 & 0 & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0 \\
    \pi_{zp} \lambda_3 & 0 & 0 & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0 \\
    \pi_{pn} \lambda_6 & 0 & 0 & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0 \\
    \pi_{zp} \lambda_6 & 0 & 0 & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0 \\
    \pi_{np} \lambda_8 & 0 & 0 & \pi_{pp} \lambda_5 & 0 & 0 & \pi_{np} \lambda_5 & 0 & 0
\end{bmatrix}
\end{align*}
$$

, which contains eight free stationary frequency parameters and nine rate parameters. In this model, the rates of simultaneous double substitution are set to 0 (e.g., the rate from state PP to NN is 0 since this requires that two substitutions occur simultaneously at two sites). The dependent model also assumes the rates are the same for transitions as Xa to Xb and aX to bX. This assumption indicates that what mediate the substitutions are the
states at each site instead of the site itself. For the independent model two sites were analyzed separately. Therefore, the dependent model has 7 more parameters than the independent model (17-5*2).

Kass and Raftery (Kass and Raftery 1995) suggested that Bayes factor bigger than 5 can be treated as strong evidence favoring one model. In order to limit the number of coevolving pairs thus to facilitate substitution tracing, we took 8.0 as the cutoff. The pairs with Bayes factor greater than 8.0 were extracted as coevolving pairs. Physical information about these coevolving pairs was obtained using the *A. nidulans* CPD photolyase structure (PDB ID: 1TEZ) (Mees et al. 2004). We numbered the sites according to the *E. coli* CPD photolyase sequences.

**Tracing Substitutions**

Rather than tracing each individual site independently, we traced the substitutions of the coevolving pairs using MacClade (Maddison and Maddison 2000) pairwisely. MacClade traces state substitutions by reconstructing the ancestral states using a parsimony algorithm. Although likelihood based methods are available for ancestral sequence reconstruction, the parsimony-based method implemented in MacClade is the only tool that can be used for pairwise ancestral sequence reconstruction.

To implement pairwise reconstruction, the states of each coevolving pair were acquired by masking the alignment using the same charge vector as in the above coevolutionary analysis. We used N (negative), Z (zero, neutral), and P (positive) to represent the three possible states of each site so that the 9 possible states of a pair of sites were NN, NZ, NP, ZN, ZZ, ZP, PN, PZ, and PP. For example, *E. coli* photolyase has Ala at sites 311 and 315, so both sites are neutral and the state for the pair 311-315 was ZZ.
(neutral and neutral). Since MacClade only recognize single character states, the 9 pairwise states were further simplified into single characters, such as ZZ into T. The final state schema of the nine states for MacClade were T for ZZ, G for ZN, H for NZ, E for ZP, W for NN, C for NP, I for PZ, Y for PN, and Q for PP. One reason for using these 9 characters was to utilize the default amino acid-color scheme of MacClade. In this fashion, each character of the final input stands for the state of a coevolving pair, and it has a total of nine possible states. In this study We used this method for analyzing coevolving pairs, but obviously, this method would be useful for any pairwise ancestral sequence reconstruction based on parsimony algorithm.

**Results**

**Coevolution and Pairwise Distances**

Coevolutionary analysis, comparing the 17-parameter dependent and 10-parameter independent models using Bayes factors, detected 11 coevolving pairs (Bayes factor greater than 8.0) with the charge vector (Table 5.1). Among these 11 pairs, we consider 6 to be very strongly coevolved because their Bayes factors are bigger than 9.0. Consistent with our previous analyses (Pollock, Taylor, and Goldman 1999; Wang and Pollock 2005; Wang and Pollock 2006), the coevolving pairs tend to be physically proximal. 8 of the 11 coevolving pairs are less than 20 Å apart, and 5 of them (300-302, 311-315, 331-421, 354-453, and 358-360) are in structural contact with each other (Figure 5.2). Four of these five coevolving pairs have Bayes factor greater than 10.0 and the fifth value is greater than 9.0. Only three of the coevolving pairs are very close to one another in the primary structure (300-302, 311-315, 358-360).
Table 5.1. The coevolving pairs detected using the charge vector. Each column is of a pair.

<table>
<thead>
<tr>
<th>Sites in pairs</th>
<th>224</th>
<th>300</th>
<th>302</th>
<th>303</th>
<th>311</th>
<th>311</th>
<th>331</th>
<th>354</th>
<th>358</th>
<th>430</th>
<th>431</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>16.7</td>
<td>5.1</td>
<td>15.5</td>
<td>20.2</td>
<td>6.2</td>
<td>20.4</td>
<td>12.7</td>
<td>6.5</td>
<td>5.4</td>
<td>27.9</td>
<td>18.4</td>
</tr>
<tr>
<td>BF</td>
<td>8.2</td>
<td>9.3</td>
<td>8.2</td>
<td>10.0</td>
<td>10.2</td>
<td>8.6</td>
<td>13.4</td>
<td>15.4</td>
<td>11.4</td>
<td>8.2</td>
<td>8.3</td>
</tr>
</tbody>
</table>

a The C-alpha distance of the two residues in 1TEZ of *A. nidulans*
b The value of the Bayes factor for each pair.

Figure 5.2. The eleven coevolving pairs displayed on 1TEZ. They are mainly localized in two regions: one around the proposed electron transfer pathway and one that appears to be isolated from the main body of the protein. The three conserved Trp of the proposed electron transfer pathway are shown as white spheres. The coevolving pairs are shown in dots. Each coevolving pair is displayed with the same color, specifically, pair 224-406 in yellow, pairs 300-302 and 302-363 in chocolate, pair 303-364 in pale cyan, pair 311-315 and 311-367 in red, pair 331-421 in purple-blue, pair 354-453 in violet, pair 358-360 in cyan, pair 430-461 in grey, and pair 431-457 light orange.
Coevolution and Physical Localization

To detail the possible functions of these coevolving pairs, their localizations were visualized using PyMOL (DeLano 2002) on the crystallized structure of *A. nidulans* photolyase, 1TEZ (Mees et al. 2004). Five of the 11 coevolving pairs (pairs 300-302, 302-363, 303-364, 311-367, and 358-360) are proximal to the proposed electron transfer channel (Figure 5.2), which involves three conserved tryptophan residues (W306, W359, and W382) (Cheung et al. 1999). The other 6 coevolving pairs (pairs 311-315, 311-367, 431-457, 430-461, 331-421, 354-453) are located entirely, or have one site, in a region that appears to be separated from the main body of the protein (Figure 5.3). Interestingly, this is also the region where most of the invariant sites and the sites with conserved residue properties (all residues have very similar properties, e.g. Lue and Ilu) are located. In the α domain, a total of 12 invariant and 13 residue property conserved sites were identified in our alignment. Seven of the 12 invariant sites and 8 of the 13 residue property conserved sites are located in this region. These 15 (7+8) are hydrophobic except for one (D327), which is negatively charged. Conservation of the sites in this region indicates either structural or functional importance. We tend to believe this is the hydrophobic core because so far no functional activities have been identified in this region.

Surprisingly, we found only 3 coevolving pairs (pairs 354-453, 431-457, and 430-461) appeared close to the bound DNA, although studies comparing plant cryptochrome to *E. coli* photolyase have suggested that there is a systematic reduction of surface positive charge. Closely checking their coevolution patterns suggests that the coevolution of these three pairs, especially pair 430-461, might not be driven by DNA
binding. Site 461 is close to the bound DNA but is neutral in photolyases (data not shown).

Figure 5.3. The coevolving pairs located in the region with most of conserved hydrophobic sites. The conserved hydrophobic sites are shown as white sticks. The only charge-conserved sites are shown as pink sticks. Coevolving pairs are shown in dots. The color scheme is the same as Figure 5.2 for the coevolving pairs.

Tracing the Substitutions of the Coevolving Pairs

In consideration of sequence density and structural/functional information, we tried to focus on the three well-characterized groups: proteobacterial photolyases, plant cryptochromes, and animal cryptochromes (Figure 5.1). Table 5.2 shows the most recent common ancestor (MRCA) of these 3 clades and the traced substitutions occurring in
each of these groups for the 11 coevolving pairs. Because of the ambiguity in the ancestral state reconstruction of pairs 224-406 and 311-315, some information for these two pairs are not shown in Table 5.2. Instead, they are illustrated by Figures 5.4 and 5.5, respectively.

Of the 90 traced substitutions, 17 are double substitutions at both sites (Table 5.2, in bold). There is no doubt that double substitution is obvious evidence of coevolution. Interestingly, we did not find any charge swapping between the coevolving sites. Sixteen of the 17 double substitutions are from the coevolving pairs with residues contacting physically (pairs 300-302, 331-421, 354-453, and 358-360). In total, 33 substitutions were traced in these contacting coevolving pairs. This means that nearly half of the substitutions traced are double substitutions. Given the conservation of these sites, the frequency of double substitutions is obviously higher than would be expected by random chance. At the same time, the frequency of double substitutions is significantly less (P<0.001, G test) in the coevolving pairs whose residues do not contact each other, where there is only 1 double substitution out of a total of 57 substitutions.

Besides these double substitutions, our substitution tracing revealed the existence of state- and phylogeny-dependent substitutions in the coevolving pairs. This means that the substitution at one site depends on the state of the other site and the phylogeny. Among the 11 coevolving pairs, 6 pairs are state invariant (invariant after charge masking) in animal cryptochromes as compared to 3 state invariant pairs in proteobacterial photolyases and none in plant cryptochromes (Table 5.2). Because the average branch length of plant cryptochromes are close to that of animal cryptochromes, the difference of invariant pairs (6 vs. 0) reveals the dependence of substitutions on phylogeny, which
Table 5.2. The traced pairwise substitutions. Results shown are traced substitutions in the three subfamilies of the photolyase/cryptochrome superfamily and their reconstructed pairwise most recent common ancestral (MRCA) states.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Photolyase(^a) MRCA</th>
<th>Subs</th>
<th>Cryptochrome(^b) MRCA</th>
<th>Subs</th>
<th>Cryptochrome(^c) MRCA</th>
<th>Subs</th>
</tr>
</thead>
<tbody>
<tr>
<td>224-406</td>
<td>ZP</td>
<td>Q</td>
<td>ZP</td>
<td>Q</td>
<td>ZP</td>
<td>0</td>
</tr>
<tr>
<td>300-302</td>
<td>ZP</td>
<td>ZP-&gt;PP(2)</td>
<td>PF-&gt;NZ</td>
<td>NZ-&gt;NN</td>
<td>PP-&gt;ZZ(2)</td>
<td>PP-&gt;NN</td>
</tr>
<tr>
<td>302-363</td>
<td>NN</td>
<td>NN-&gt;PN(3)</td>
<td>NN-&gt;ZP(5)</td>
<td>PN-&gt;ZN(2)</td>
<td>ZN-&gt;PN(2)</td>
<td>PZ</td>
</tr>
<tr>
<td>303-364</td>
<td>ZP</td>
<td>ZP-&gt;ZZ(2)</td>
<td>ZZ-&gt;NZ</td>
<td>ZP-&gt;PP</td>
<td>PZ-&gt;ZP(2)</td>
<td>PZ</td>
</tr>
<tr>
<td>311-315</td>
<td>Q</td>
<td>Q</td>
<td>ZZ</td>
<td>ZZ-&gt;PZ</td>
<td>ZZ-&gt;NZ(2)</td>
<td>NZ</td>
</tr>
<tr>
<td>311-367</td>
<td>NZ</td>
<td>NZ-&gt;ZZ(4)</td>
<td>NZ-&gt;ZP(2)</td>
<td>NZ-&gt;PN</td>
<td>PZ-&gt;ZZ</td>
<td>NN</td>
</tr>
<tr>
<td>331-421</td>
<td>PN</td>
<td>PN-&gt;PZ</td>
<td>PN</td>
<td>PN-&gt;PZ</td>
<td>PN</td>
<td>PN</td>
</tr>
<tr>
<td>354-453</td>
<td>NP</td>
<td>NP-&gt;PP(2)</td>
<td>NP-&gt;ZZ(3)</td>
<td>ZZ</td>
<td>ZZ-&gt;NP</td>
<td>NP</td>
</tr>
<tr>
<td>358-360</td>
<td>NP</td>
<td>0</td>
<td>NP</td>
<td>NP-&gt;ZP</td>
<td>ZP-&gt;PZ(4)</td>
<td>ZN</td>
</tr>
<tr>
<td>430-461</td>
<td>ZZ</td>
<td>0</td>
<td>PZ</td>
<td>PZ-&gt;PN(3)</td>
<td>PN-&gt;PP(2)</td>
<td>ZZ</td>
</tr>
<tr>
<td>431-457</td>
<td>PP</td>
<td>0</td>
<td>PP</td>
<td>PP-&gt;ZP</td>
<td>PP-&gt;PZ(2)</td>
<td>NZ</td>
</tr>
</tbody>
</table>

\(^a\) proteobacterial photolyases; \(^b\) plant cryptochromes; \(^c\) animal cryptochromes; MRCA most recent common ancestor; Subs pairwise substitutions; P positive charged; N negative charge; Z neutral; Q in question because of reconstruction ambiguity; 0 no substitution reconstructed. The numbers in parentheses are the frequency of that specific substitution. In bold are simultaneous substitutions at both sites.
Figure 5.4. The traced pairwise state substitutions of the pair 224-406. Different states are shown in different colors according to the legend bar. NP stands for Negative at the first (site 224) and Positive at the second site (site 406); ZP stands for neutral (Zero) at the first site and Positive at the second site; and so on.
Figure 5.5. The traced pairwise state substitutions of the pair 311-315. Different states are shown in different colors according to the legend bar. Pairwise states are represented by initials the same way as in Figure 4.
implies functional or structural evolution of the proteins. On the other hand, pair 431-457 is invariant in the photolyase group but not in either cryptochrome group, despite the fact that the average distance between sequences in the photolyase group is greater than those in the two cryptochromes. This implies an extraordinary evolutionary constraint on this pair in the photolyases. Overall, pairs 224-406, 300-302, 311-315 and 311-367 exhibited similar substitution patterns: one or no substitutions in animal cryptochromes but frequent substitutions (more than 3) in proteobacterial photolyases and plant cryptochromes (Table 5.2, Figure 5.4, Figure 5.5). Pairs 331-421, 354-453, 358-360, 430-461, and 431-457 exhibit another pattern of substitutions, being generally conserved in all the main groups (Table 5.2). Pairs 302-363 and 303-364 show patterns different from all the above pairs. Both of these are invariant or nearly invariant in plant cryptochromes and animal cryptochromes (Table 5.2). These phylogeny-related substitution patterns imply that the changes in the structural and functional constraints are driving residue coevolution.

Substitution can also be examined from its dependence on the states of each site in the coevolving pair. A typical example is pair 430-461, where site 461 has to be Z when site 430 is Z and, thus, pairwise state ZZ is invariant (in both proteobacterial photolyases and animal cryptochromes). Once site 430 has mutated to P, state Z at site 461 can be substituted by either P or N (Table 5.2). As shown in Figure 5.4, substitution patterns of pair 224-406 depend on whether site 406 is positive or not. If it is positive, site 224 is restricted to be neutral (Z) and no substitutions are allowed. Similarly, substitutions in pair 302-363 exhibit strong state-dependency. When the state of 363 is N, site 302 is relatively less constrained. When it is Z, site 302 is strictly constrained and
substitutions are limited (Table 5.2). For pair 311-315, the state of site 311 is constrained to be N when the state of site 315 is P (Figure 5.5). The other pairs do not clearly exhibit this type of dependency.

Because the substitution patterns of pairs 358-360, 331-421, and 354-453 are relatively constant over the entire phylogeny, and the side chains of their residues directly contact one another (Figure 5.2, Figure 5.3), we examined their amino acid states according to our alignment (supplemental data 5.1). For pair 358-360, pairwise residues S358-E360, D358-R360, P358-K360, P358-R360, and P358-T360 have frequencies of more than 0.95 together (S-E and D-R have frequencies of more than 0.7). P-K, P-R, and P-T combinations are in the plant cryptochrome group. For pair 331-421, R331-E421, T331-V421, T331-K421, and T331-I421 have frequencies of more than 0.95 (R-E has a frequency of more than 0.7). The T-V, T-K, and T-I combinations are in the animal cryptochrome group. For pair 354-453, D354-H453, H354-H453, N354-L453, N354-Y453, T354-L453, V354-L453, V354-I453, F354-I453, F354-L453, and M354-I453 have frequencies of more than 0.95 (D-H has a frequency of more than 0.6). The D-H, N-L, and H-H combinations are in photolyases and plant cryptochromes.

Discussion

Coevolution Depends on Physical Distance, Protein Structure, and Protein Function

In the α domain of photolyase/cryptochrome proteins, coevolving pairs are physically proximal, which verifies the dependence of coevolution on physical distance (Pollock, Taylor, and Goldman 1999; Wang and Pollock 2005; Wang and Pollock 2006). It is known that residue interaction is an import force of coevolution. Proximal residues tend to interact with one another and have more chances to coevolve. Residues interact via
several forces, such as covalent bonding, hydrogen bonding, van der Waals forces, and electrostatic interaction. Charge coevolution, which we detected in the present study using residue charge as vector, may mainly reflect coevolution driven by electrostatic interactions. The short distances (less than 21 Å) between these coevolving residues agree with the results of previous studies of electrostatic interactions. Charged residues have been reported to significantly affect the electronic behavior and catalytic function of catalytic centers that are 10-20 Å away from them (Thomas, Russell, and Fersht 1985; Russell and Fersht 1987; de Kreij et al. 2002; Clarkson and Lee 2004; Dupeyrat et al. 2004; Thorpe and Brooks 2004). Some data have suggested that charged residues as far as 30 Å can be coupled functionally (Rajagopalan, Lutz, and Benkovic 2002). This does not mean that charge residues more than 30 Å apart cannot interact and coevolve. In fact, we found many pairs coevolving at the cutoff of Bayes factor >5.0, which have distances larger than 30 Å (data not shown). Though charge coevolution can occur over a relatively long distance, a specific type of coevolution, double substitution, occurs mostly in the pairs with residues contacting physically (Table 5.2). This finding might reflect the direct interaction of residues, such as the formation of a hydrogen bond or salt bridge as discussed below. If this is true, a coevolutionary analysis that identifies these double substitutions could facilitate the prediction of 3-dimensional protein structures.

The ultimate selective force for evolution and coevolution comes from protein structures and functions. Proteins have to fold efficiently and precisely to function properly. Our previous studies on cytochrome c oxidase have suggested that structurally and/or functionally important regions exhibit more coevolution (Wang and Pollock 2005; Wang and Pollock 2006). In the photolyase/cryptochrome α domain, 5 of the 11 strongly
coevolving pairs are found in the region where three conserved tryptophans (W306, W359, and W382) are located. This implies that the region is of critical importance. Studies suggest that the tryptophans may form an intra-protein electron transfer pathway connecting the buried flavin to the protein surface (Lin and Shalitin 2003; Sancar 2003; Byrdin et al. 2004). Mutants of these tryptophans destroyed electron transfer and repair activity of photolyases in vitro (Li, Heelis, and Sancar 1991; Cheung et al. 1999; Byrdin et al. 2003), while mutating them in plant cryptochromes resulted in abnormal phenotypes related to light responses (Giovani et al. 2003; Zeugner et al. 2005). However, there is evidence that under physiological conditions, these tryptophans are not involved in the normal activity of photolyases at all (MacFarlane and Stanley 2003; Kavakli and Sancar 2004; Kao et al. 2005).

While the locations of the coevolving sites in our study suggest this is a region of importance, tracing the state substitutions of these coevolving pairs suggests that it is unlikely that this electron channel is functional in photolyases. Among the 5 coevolving pairs proximal to the supposed channel, two of them (302-363, 303-364) have multiple substitutions in the photolyases, and show pairwise states with very different properties. If this channel is of functional importance to electron transfer in photolyases, it would be highly unusual that these very significant substitutions did not affect their functionality. Charged residues around electron transfer channels have been shown to strongly affect electron transfer (Gunner, Nicholls, and Honig 1996; Farver and Pecht 1997; Kirmaier, Weems, and Holten 1999; Johnson et al. 2003; Cheng et al. 2005). The distances of each of these 4 residues is less than 10 Å from the channel. A change in their charge would almost certainly change the property of the electron passing, significantly changing the
function. Therefore, we suggest that this channel is not involved in the normal function of CPD-photolyases. The lack of substitutions within these two coevolving pairs in cryptochromes, on the other hand, suggests that the channel might be functional in cryptochromes. The lack of substitutions reflects the different evolutionary constraints under which the cryptochromes operate, and these constraints are most likely related to electron transfer.

The 7 coevolving pairs that are associated with a region that is not proximal to known active and functional centers suggest that there are some as yet unidentified evolutionary constraints in this region. Because this region also exhibits conservation of hydrophobicity in its residues, we suggest that this may be where the folding hydrophobic core is located. Protein folding is the process by which amino acid chains acquire the precise 3-dimensional structures required to execute their functions. The side-chains in the cores of a protein are usually tightly-packed (Tanford 1978; Rose et al. 1985; Ahmad and Cashmore 1993), even subtle substitutions inside the core tend to be destabilizing (Eriksson et al. 1992; Lee 1993; Buckle, Cramer, and Fersht 1996). Thus, it is reasonable that coevolving residues would be in the core. Since protein stabilization can be achieved by mutations in solvent-exposed regions (Eijsink et al. 1995; Spector et al. 2000; Makhatadze et al. 2003), coevolving residues do not need to be in the center of the hydrophobic core. In fact, the center of the core is usually too conserved to allow coevolution. In this study, we found all the coevolving pairs in this region are at its boundary, and sometimes with one site distant from the core. The fact that the coevolving pairs were detected using a charge vector, not a hydrophobicity one, leads us to suggest the following explanation for coevolution in this region of photolyase/cryptochrome
proteins. In the center of this suggested core is a negatively charged residue, which might be stabilizing or destabilizing the structure significantly depending on its residue context (Waldburger, Schildbach, and Sauer 1995; Elcock and McCammon 1998; Giletto and Pace 1999; Loladze, Ermolenko, and Makhatadze 2001). This highly context-dependent effect of a polar residue will result in a driving force for the coevolution of surrounding residues.

Surprisingly, we did not find extensive coevolution related to DNA binding. According to the photolyase crystal structure, two coevolving pairs, 354-453 and 431-457, might be involved in DNA binding. Sites 453 and 457 are both close to a phosphate group in DNA (Figure 5.3) and are both positively charged in photolyases (Table 5.2). In contrast, structural comparisons of photolyase and cryptochrome have suggested a systematic loss of the charges on the DNA binding surface in cryptochromes (Lin and Shalitin 2003; Lin and Todo 2005). The limited coevolution related to DNA binding might be explained by assuming that DNA binding is not a very strong constraint on charge properties of photolyase, and thus, residues involved in DNA binding only exhibit weak coevolution. This kind of weak coevolution would be missed by our stringent Bayes factor cutoff (the coevolution of pairs 354-453 and 431-457 might be driven by other causes as discussed above and below).

**Context Dependency on Residue Interactions**

The coevolution of contacting residues provides detailed information about residue interactions and their relationship with protein structure and function. The coevolution of pair 358-360 reflects the effect of salt bridges on the secondary structure. In photolyase, the positive and negative charges (D358-R360) may form a salt bridge, which terminates
the helix structure. As in animal cryptochromes, this salt bridge can be replaced by the interaction between Ser and Glu (S358-E360), which also have a strong interaction in proteins (Kim, Mao, and Gunner 2005). Once a Pro occurred at site 358, this interaction is no longer a requirement for termination of the helix. Pro can terminate the helix by itself, allowing site 360 to accept different residues such as Thr, Arg, and Lys. This is exemplified by the patterns observed in plant cryptochromes (P358-K360, P358-R360, and P358-T360). The coevolution of pair 354-453 may reflect the need to stabilize the Asp at 354 (D354-H453), whose side chain is buried. As pointed out by previous studies, buried charges are very destabilizing to the structure and must be stabilized by opposite charges or polar residues (Waldburger, Schildbach, and Sauer 1995; Elcock and McCammon 1998; Giletto and Pace 1999; Loladze, Ermolenko, and Makhatadze 2001). When Asp disappears, site 453 becomes free to substitutions (H354-H453, N354-L453, etc.). Because His residue is usually not ionized in proteins (Kim, Mao, and Gunner 2005), the His at site 354 does not require a negative charge in site 453 to stabilize (H354-H453). Both sites 331 and 421 are exposed to the solution. The coevolution of pair 331-421 might not have resulted from the requirement for stabilization as seen in pair 354-453. Instead, it may reflect the structural role of the electrostatic interaction between an Arg and Glu, which appears to bring two separated secondary structures together. If this is true, the structure of animal cryptochrome may be different from those of photolyase or plant cryptochrome because the salt bridge disappears with substitutions of charged residues by non-charged residues at both sites (R331-E421 by T331-V421, T331-K421, or T331-I421, etc.).
It has long been an assumption that a strong pairwise interaction of residues may result in a charge swapping between the interacting ion pairs. However, previous coevolutionary analyses did not find evidence to this assumption (Kondrashov, Sunyaev, and Kondrashov 2002). In our analysis, when considering the residue charges of the 11 strongest coevolving pairs, we did not find any direct charge swapping over the course of their evolution using sequence ancestral reconstruction. Thus, we suggest that charge swapping is not an important aspect of coevolution. It is not hard to rationalize this point given the importance of ion-pair interactions in protein stability. To form a charge swapping, there may have to be two substitutions simultaneously occur at the interacting sites, each substitution changing the residue charge oppositely, because the constraints on ion-pair interactions would eliminate any substitution that destabilize the interaction. A single mutation at one site is definitely destabilizing if it changes the charge property of one site. Therefore, charge swapping occurs only when simultaneous substitutions at interacting sites occur. It is well-known simultaneous substitutions are rare, so that charge swapping is also rare. Our substitution tracing on the tree agree with this (Table 5.2) since we did not find any charge swapping.

Coevolution and Detection

The coevolving pairs detected in this study illustrate two phenomena of coevolution: double substitutions, and phylogeny- and state-dependent pairwise substitutions. The latter is more extensively observed than the former in our coevolving residues. Double substitutions are mainly produced by the coevolving pairs that are contacting one another physically (Table 5.2). Most coevolving residues appear to be subject to substitution patterns that depend on both the states of sites and the phylogeny. Functional evolution
would be reason for the phylogeny-dependent substitutions. During protein evolution, the “birth and death” of functions result in changes in the constraints the relevant regions or domains are under. These constraint changes will then reflected in substitution pattern changes of sites that are either structurally or functionally important to the regions. This understanding of coevolution is useful in helping to determine regions with structural or functional importance. Using this information in the present study, we were able to suggest the regions important for electron transfer and protein folding in our model protein.

Many coevolutionary analyses, such as those based on mutual information theory (Korber et al. 1993; Neher 1994; Taylor and Hatrick 1994; Thomas, Casari, and Sander 1996; Giraud 1998; Hoffman, Schiffer, and Swanstrom 2003; Saraf, Moore, and Maranas 2003), focus on the frequency of pairwise residue states, ignoring the phylogeny. Other studies focus mainly on double substitutions (Fukami-Kobayashi, Schreiber, and Benner 2002; Dimmic et al. 2005). These analyses did not consider the evolutionary context within which the residues coevolved. As revealed by our analysis, coevolution is extensively related to the evolutionary context as well as the states of interacting pairs and the phylogeny. Ignoring these factors might lead to the loss of the coevolutionary signal, especially when the proteins analyzed have a history of structural- or functional-related evolution, such as the photolyase/cryptochrome superfamily. Because analyzing structural and functional evolution is one of the main goals of coevolutionary analysis, our model-based methodology is a better choice for coevolutionary analysis.

References


Appendix: Permission

http://mymail.paws.lsu.edu/scripts/mail/read.mail?print=1&msg_uid=11...

From: "Smith, Stephanie (ELS-OXF)" <ST.Smith@elsevier.com>
To: zwang3@lsu.edu
CC:
Subject: RE: Obtain Permission
Date: Thu, 13 Apr 2006 11:52:58 -0000

We hereby grant you permission to reproduce the material detailed below in your thesis at no charge subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copy.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

   "Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright Year, with permission from Elsevier".

3. Reproduction of this material is confined to the purpose for which permission is hereby given.

4. This permission is granted for non-exclusive world English rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

Steph Smith
Rights Assistant
Elsevier Ltd
The Boulevard
Langford Lane
Kidlington
Oxford OX5 1GB

----Original Message----
From: zwang3@lsu.edu [mailto:zwang3@lsu.edu]
Sent: 12 April 2006 17:25
To: permissions@elsevier.com
Subject: Obtain Permission

This Email was sent from the Elsevier Corporate Web Site and is related to Obtain Permission form:

---------------------------------------------------------------------
Product: Customer Support
Component: Obtain Permission
Web server: http://www.elsevier.com
IP address: 10.16.24.149
Client: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)

5/26/2006 10:11 AM

Vita

Zhengyuan Wang was born in Hunan, People’s Republic of China, on July 12, 1968. He received a Bachelor of Science degree with a major in genetics in July of 1990 from Wuhan University, P. R. China. He worked in the Virology Laboratory at People’s Hospital of Guangxi as a research associate from 1990 to 1999. Mr. Wang entered Graduate School at the University of Louisiana Lafayette in August 1999 and received his Master of Science in biology degree in August 2002. He is currently a candidate for the degree of Doctor of Philosophy in biochemistry with a minor in applied statistics.