STRUCTURAL DETERMINANTS OF REPLACEMENT RATE HETEROGENEITY

STRUCTURAL DETERMINANTS OF AMINO ACID REPLACEMENT RATE HETEROGENEITY

by
FRANCES RAFTIS, B.SC. (HONS.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2006) McMaster University
(Biology) Hamilton, Ontario

TITLE: STRUCTURAL DETERMINANTS OF REPLACEMENT RATE HETEROGENEITY

AUTHOR: Frances Raftis, B.Sc. Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: [x], 114

ABSTRACT

Protein sequences display replacement rate heterogeneity across sites. In an earlier work, half of the causal site-wise variation in replacement rates was explained by a simple linear regression model consisting of terms for the solvent exposure of each residue, distance from the active site, and glycines in unusual main-chain conformations. Replacement rates vary not only across sites, they may also vary over time. In this study, we apply the linear regression model to phylogenies divided into subtrees to see if lineage-specific rate shifts have a structural basis that can be detected by the model. This approach is applied to two different data sets. The first set consists of phylogenies containing two representative structures, divided into subtrees such that one structure is present in each subtree. These structures have little or no obvious functional divergence between them. The model is tested with permutations of subtrees and structures from each subtree. While there is a slight effect of the specific structure on the fit of the model, the specific subtree has a greater effect. The second data set involves homologous structure pairs where the quaternary structure has changed at some point in the phylogeny. These pairs are examined to see how the change in constraint on the new interface sites affect the replacement rate, and its relationship with other structural factors. We find that the unique interfaces are as conserved as the shared ones, and they exhibit a different relationship between replacement rates and indicators of constraint than the shared interfaces or other protein sites. We also find that the unique interfaces display characteristic amino acid preferences that may identify interfaces which are still in the process of stabilizing.

ACKNOWLEGEMENTS

I would like to thank my advisor, Dr. G. Brian Golding, who provided the inspiration to work on a fascinating topic and exposed me to whole new areas of the evolutionary process, and who also kept faith in me which helped me persevere through the difficult times. Thanks to Melanie Huntley, whose support was invaluable and without whose help I would have not been able to complete this work. I cannot express deep enough thanks to my husband, Graydon Hoare, whose continual support, patience, love, and assistance were indispensible. Thanks to Dr. Dara Torgerson, Brett Whitty, Weilong Hao, and Heidi Musters for moral support and entertainment. Special thanks to Dara and Brett in particular for the shared caffeine dependence. I would like to thank Ying Fong for company during the long evenings in the lab and the wonderful tea, and to Pat Hayward for making the beauracratic processes much less painful.

Special thanks to Morgan Hay for continual reality checks and goading as required, and to Charles Mingus and Ken Vandermark for providing the soundtrack.

Contents

List of Figures

List of Tables

Part I

INTRODUCTION

Kimura's neutral theory of molecular evolution (Kimura 1989) proposed that the majority of evolutionary changes at the molecular level are selectively neutral. This theory predicts that substitutions will occur by a molecular clock for sequences that are not under selective pressure. Where substitution rates vary, the variation is thought to be largely due to constraint and purifying selection rather than positively selected adaptation. To some degree, constraint can be described in terms of general structural features of the protein for coding sequences, and evolutionary rates at each protein site often change to reflect the structural constraints that they are subject to.

There have been many works seeking to exploit this link between protein structure and evolutionary rates. A fuller understanding of this relationship has uses both in refining structural prediction, and in distinguishing those residues under actual positive selection from those that are merely under low constraint.

One approach is the creation of amino acid transition matrices built with a consideration of structural factors (Wako and Blundell 1994; Goldman, Thorne and Jones 1998; Mizuguchi and Blundell 2000; Shi, Blundell and Mizuguchi 2001; Robinson *et al.* 2003). While this approach is likely to be useful for creating more accurate alignments for coding sequences, it is often computationally demanding (particularly when multiple interactions of factors are considered), and of limited explanatory power when proteins are introduced that are outside the scope of the set which were used to create the matrix. Another approach is to model some metric of evolutionary rates as a product of various structural factors (Dean and Golding 2000; Dean *et al.* 2002; Bustamante, Townsend and Hartl 2000). With this approach, the focus is on explanation of rates or polymorphisms rather than on prediction of structural factors. It is generally less computationally demanding, and allows for easy tests of new hypotheses about which factors introduce constraint. Additionally, some factors exhibit a continuous relationship with evolutionary rates, and linear models readily allow expression of this relationship.

Various works have tested a variety of structural factors, such as secondary structure (Saunders and Baker 2002; Shi, Blundell and Mizuguchi 2001; Mizuguchi and Blundell 2000; Bustamante, Townsend and Hartl 2000; Goldman, Thorne and Jones 1998; Thompson and Goldstein 1996b; Thompson and Goldstein 1996a; Dean and Golding 2000), length of secondary structural elements (Mizuguchi and Blundell 2000; Goldman, Thorne and Jones 1998), interactions between residues (Robinson *et al.* 2003; Shi, Blundell and Mizuguchi 2001; Dean and Golding 2000; Dean *et al.* 2002), specific physiochemical properties (Saunders and Baker 2002;

Bustamante, Townsend and Hartl 2000; Thompson and Goldstein 1996b; Thompson and Goldstein 1996a; Dean and Golding 2000; Dean *et al.* 2002), and solvent exposure of residues (Robinson *et al.* 2003; Saunders and Baker 2002; Shi, Blundell and Mizuguchi 2001; Mizuguchi and Blundell 2000; Bustamante, Townsend and Hartl 2000; Goldman, Thorne and Jones 1998; Thompson and Goldstein 1996b; Thompson and Goldstein 1996a; Dean and Golding 2000; Dean *et al.* 2002), and various interactions between these factors. The solvent exposure is almost universally a significant factor in these studies, with buried residues tending towards greater conservation. Secondary structural elements are significant in many studies, but not in some (Dean and Golding 2000; Dean *et al.* 2002). Dean and Golding (2000) tested the explanatory power of various structural factors with a linear regression model. They found three factors that were significant sources of variation in the replacement rates at a protein site: solvent accessibility, distance of the residue from the active site of the protein, and whether or not the residue was a glycine in a conformation that could not be adopted by other residues. Dean *et al.* (2002) expanded on this work, and demonstrated that this minimal model could account for half of the causal rate variation between sites in a variety of proteins. In the following chapters, we develop this relationship between site-specific replacement rates and structural factors further. In the first chapter, we test the linear model on a sub-divided phylogeny to see if it is capable of detecting lineage-specific changes in constraints. In the second chapter, we investigate changes in constraint that occur as a result of changes in quaternary structure.

Chapter 1

Evolutionary rates do not just change over protein sites, they also change throughout time. One of the earliest works recognizing this pattern was the covarion model of Fitch (1976), but the topic has recieved renewed interest. The most obvious causes of these temporal rate changes is functional divergence, due to gene duplication, or speciation resulting in strongly varied functional requirements on the protein. However, Lopez, Casane and Philippe (2002) found evidence of these rate shifts in vertebrate mitochondrial cytochrome b, which is unlikely to be undergoing such divergence. Rate shifts at a site may therefore also occur in a drift-based fashion, and it is possible that these shifts are often due to minor changes in structural constraints across lineages. We test this by splitting large phylogenies into subtrees, and comparing changes in the fit of the model across these subtrees, deriving structural parameters from a representative structure from each subtree.

In this chapter, we test whether or not these changes in constraint are detectable

with the basic linear model. We also attempt to improve upon this basic model. The hydropathy of the current residue is found to be another significant factor in rate variation. Additionally, Dean and Golding (2000) found that the distribution of replacement rates was atypical for the large subunit of RUBISCO, which is chloroplast encoded. We find that the nuclear-encoded small subunit does not share this atypical pattern. Overall we do find that using the structure native to the specific subtree has a small positive effect on the fit of the model, but there is a much greater effect due to subtree-specific effects. The distribution of site-specific rates becomes more stochastic at smaller evolutionary scales, making the current form of the model unsuitable for detecting smaller-scale constraint changes.

Chapter 2

Protein-protein interfaces are more conserved than other surface residues (Elcock and McCammon 2001; Landgraf, Xenarios and Eisenberg 2001; Glaser *et al.* 2003; Halperin, Wolfson and Nussinov 2004; Ma *et al.* 2003; Hu *et al.* 2000; Teichmann 2002). In the first chapter, there was generally little prior evidence for functional divergence between the subtrees and associated structures in a phylogeny. In this chapter, we select a set of homologous protein pairs where the quaternary structure has changed between the two forms, which offers clear and strong functional divergence between the sites involved in the new interface. The sites in each protein are divided into those which participate in an interface for only one structure, the aligned sites that do not participate in an interface in their native structure, sites which participate in interfaces in both structures, and sites which do not participate in any interface. The replacement rates and the relationship between the rates and the solvent accessibility and hydropathy are compared for these various categories. The amino acid compositions for each alignment site in the subtrees are investigated in a similar manner. We find differences in the overall replacement rates, hydrophobicity, and amino acid preferences of each category. These differences offer an insight into how novel interfaces evolve and change over time.

Chapter 1

Patterns in Amino Acid Replacement at Smaller Evolutionary Scales

1.1 Abstract

Protein sequences display replacement rate heterogeneity across sites. In an earlier work, half of the causal site-wise variation in replacement rates was explained by a simple linear regression model consisting of terms for the solvent exposure of each residue, distance from the active site, and glycines in unusual main-chain conformations. Replacement rates vary not only across sites, they may also vary over time. In this study, we apply the linear regression model to phylogenies divided into subtrees to see if lineage-specific rate shifts have a structural basis that can be detected by the model. The model is tested with permutations of subtrees and structures from each subtree. While there is a slight effect of the specific structure on the fit of the model, the specific subtree has a greater effect. We conclude that the model is more appropriate for larger phylogenetic scales, as differences in constraints become more apparent as the number of taxa increase. A new hydropathy term is added to the linear model, and the atypical distribution of replacement rates for RUBISCO is anaylzed in further detail.

1.2 Introduction

The neutral theory of evolution (Kimura 1989) states that most substitutions have no or nearly no effects on the fitness of the organism. It proposes that new alleles become fixed in a population at the same rate as the appearance of new neutral mutations. However, most protein sites are not truly selectively neutral, as general hydrophilic and hydrophobic interactions are necessary to ensure that the protein does not precipitate out of solution, and the general form must be maintained to bring the catalytic residues into the active site. These structural constraints still fit within the neutral model, as the protein sites which are less important for maintaining the structure and function of the enzyme can be expected to evolve more freely, and so the actual neutral mutation rate will vary between sites. Working under a neutral framework allows one to test which structural factors are useful determinants of constraint by measuring the effects these have on the replacement rate at each site in a protein. As each site in a protein experiences different degrees of constraint, the constraints on a site can also change over time. The most obvious causes of such change are gene duplication and major speciation events.

After a gene duplication event, it is generally expected that one copy retains the original function and associated constraints, and the other copy (if it does not become a pseudogene) experiences a temporary relaxation of constraint until it evolves a new function and gains a new set of constraints. There are several methods currently being developed to detect such changes in constraint from the evolutionary rates at a site in different lineages (Gu 1999; Gu 2001; Gu 2003; Yang, Swanson and Vacquier 2000; Gaucher *et al.* 2002; Knudsen and Miyamoto 2001; Knudsen *et al.* 2003; Susko *et al.* 2002). In general, these follow the approach described in Golding and Dean (1998), which focuses on only a few sites which are likely to be responsible for the functional divergence. This approach has the advantage of creating testable hypotheses about which replacements are adaptive. This can be carried out with biochemical assays that reveal how historical adaptations have taken place at the molecular level. Where these studies consider the protein structure, it is typically only at a small number sites which are identified as playing a pivotal role in functional divergence.

Speciation events can result in changes in constraint in a manner similar to that of gene duplication, but that is more likely when the protein experiences new functional demands due to the new environment or lifestyle of the organism (or if it is a protein which is typically under positive selection, such as those related

to sex (Yang, Swanson and Vacquier 2000) or pathogen evasion). In these cases, one might similarly expect that only a few sites are experiencing acute changes in constraint. However, there are also causes of more subtle changes in constraint, such as co-operative drift between interacting proteins (Lopez, Casane and Philippe 2002), and changes in population size that change selective pressures overall (Fay and Wu 2003).

Dean and Golding (2000) used a simple linear model to explain rate variation between protein sites in terms of the structural factors that are likely to introduce constraint. The model consists of terms for solvent accessibility, distance of the residue from the active site, and glycine residues in unusual main-chain conformations, and it can explain half of the causal variation in a number of proteins (Dean *et al.* 2002). Here we wish to determine if the model can detect lineage-specific constraints as well. This approach is not site-specific, so it will not point to individual sites which have experienced great changes in constraint. It does consider structural factors over the entire protein, and gives a numerical estimate of which structural factors are responsible for the changes in constraints across lineages over the whole protein.

We have chosen five α/β barrel proteins from a previous study (Dean *et al.* 2002), based on their large phylogenies and room for improvement in the linear model. We have also chosen three new non-α/β barrel proteins to see if their rates are similarly amenable to explanation by the linear model.

1.2.1 The Linear Model

The linear regression model used in previous works (Dean and Golding 2000; Dean *et al.* 2002) models the replacement rate at a protein site as a function of the solvent exposure of the residue, its distance from the active site, and whether or not it is a glycine with an unusual main-chain conformation. These three factors were able to explain about 50% of the causal rate variation for 25 functionally unrelated α/β barrel proteins. Dean and Golding (2000) tested a number of other structural factors as well (main-chain torsion angles, involvement in hydrogen bonding, secondary structure, flexibility, individual amino acid identity), but these were found to have insufficient explanatory power for the degrees of freedom required. The hydropathy of the residue was not tested, and we add it to the model here.

Since mutations are a largely stochastic, Poisson process, one cannot expect all

of the variation in replacement rates to be deterministic. Dean *et al.* (2002) introduced the concept of partitioning the site-wise variation in rates into its Poisson-distributed component and a causal component. The fit of the model is compared to the expected amount of causal variation (the Poisson estimated coefficient of determination, PECD) which is calculated from the mean and variance of the replacement rates across the protein:

$$\hat{\rho}^2 = 1 - \frac{\bar{y}}{s_y^2}$$

Dividing the $\hat{r}^2$ from the linear regression by the PECD gives the normalized coefficient of determination (NCD).

1.2.2 The Model at a More Local Evolutionary Scale

To detect structure-based changes in constraint for different lineages, we divide a large phylogeny into subtrees and determine if the fit of the linear model (as measured by the NCD) is improved at this smaller evolutionary scale. Where available, a second structure from the other subtree will also be used to model the replacement rate heterogeneity both for the subtree that it is a member of, and for the other subtree. If there are changes in constraint that are due to structural variation, we expect the NCD to be higher for a subtree when a native structure is used than if a strucutre from another subtree is used.

1.2.3 Enzymes Studied

The enzymes used in this study are summarized in Table 1.1. The Root Mean Square (RMS) distances between α-carbons of structures provided in the table were obtained from with the Swiss-PdbViewer (Guex and Peitsch 1997).

Enolase

Enolase is a glycolytic α/β barrel enzyme which catalyzes the reversible dehydration of 2′-phosphoglycerate to phosphoenolpyruvate. It is cytoplasmic, and

it is a homodimer in all of the species used in this study. Two divalent cations (Mg^{2+} *in vivo*) are absolutely required for function. One induces a conformational change in the enzyme, and the second is catalytic and binds in the presence of substrate (Duquerroy, Camus and Janin 1995). Our tree for enolase (Figure 1.2) spans the Fungi/ Metazoan group, and we use structures from *Saccharomyces cerevisiae* (2ONE:Zhang *et al.* 1997) and *Homarus gammarus* (lobster, 1PDZ:Duquerroy, Camus and Janin 1995). The two structures are fairly similar, with a RMS distance of 0.83 Å between α-carbons, and consistent secondary structure. Yeast enolase may be more tolerant of a monomeric state, which has been observed in the absence of divalent cations and at low enzyme concentration. A monomeric lobster enolase has not been observed (Duquerroy, Camus and Janin 1995). As the enzyme catalyzes an essential and conserved reaction, we do not expect functional divergence between the two structures.

Fructose 1,6 Bisphosphatase (Class I)

Fructose-1,6-bisphosphate aldolase (ALDO) is another α/β barrel, glycolytic enzyme catalyzing reversible aldol cleavage of fructose-1,6-bisphosphate to dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. Our phylogeny (Figure 1.3) has representatives from plants, alveolates, and metazoans. All are class I, which do not use a divalent cation cofactor. The representative structures for ALDO are from an Alveolate, *Plasmodium falciparum* (human malarial parasite, 1A5C:Kim *et al.* 1998), and an invertebrate, *Drosophila melanogaster* (1FBA: Hester *et al.* 1991).

We expect some degree of functional divergence, particularly in the metazoan subtree. There are two plant isoforms, a cytosolic one and one which is plastid-targeted. These two isoforms are monophyletic. In the metazoan subtree, vertebrates have three forms (A, B, and C) with tissue-specific expression patterns and different substrate specificities. Type A is expressed in muscle and erythrocytes. The *Drosophila* ALDO has the greatest sequence similarity with vertebrate type A, so this is most likely the more ancestral form. The *Drosophila* and *Plasmodium* forms have an RMS distance of 1.5 Å between them. Drosophila and vertebrate type A have an RMS of 3.89 Å, but Plasmodium and vertebrate type A are more distant, but have a much lower RMS distance of 1.4 Å (Kim *et al.* 1998). As *Plasmodium falciparum* lives in human erythrocytes, it's possible that there are selective pressures towards PfALDO maintaing a more similiar structure to the human erythrocyte form.

5-Aminolevulinate Dehydratase

5-Aminolevulinate dehydratase (ALAD) is an octameric α/β barrel enzyme that catalyzes the dimerization of two 5-aminolevulinic acid molecules to form porphobilinogen, which leads to the biosynthesis of tetrapyrroles such as chlorophyll and heme. It is found across archaea, bacteria, and eukaryotes, with high sequence similarity. Our phylogeny (Figure 1.4) is limited to bacteria and a small archaeal cluster, and our structure is from *Escherichia coli* (1B4E:Erskine *et al.* 1999). The enzyme requires two metal ions for function, which can be Zn^{2+} or Mg^{2+} depending on the enzyme. Our tree is divided into Zn^{2+} and Mg^{2+} dependent subtrees. There are a number of functional differences between the Zn^{2+} and Mg^{2+} dependent forms, so some functional divergence is likely. The Zn^{2+}-dependent form can be inactivated by lead, and the Mg^{2+}-dependent form is less susceptible to oxidation. There are also differences in kinetics and pH dependence between forms. This is the only enzyme for which only one subtree had a structure available, so we can only test for an elevated NCD in the Zn^{2+}-dependent subtree.

3-α-Hydroxysteroid Dehydrogenase

3-α-hydroxysteroid dehydrogenase (3αHSD) is a monomeric, α/β barrel liver enzyme that reversibly inactivates circulating steroid hormones. Our first structure is from *Rattus norvegicus* (1LWI:Bennett *et al.* 1996). However, our phylogeny (Figure 1.5) for this enzyme spans the vertebrate Aldo-keto reductase superfamily, with enzymes of many different functions. Our alternate structure is for human Aldose reductase (ALR), which converts glucose to sorbitol as the first step of the polyol pathway (1PWM:El-Kabbani *et al.* 2004). It has no known physiological role. Neither enzymes use a metal cofactor, though 3αHSD binds NADP+ and ALR binds NAP^{+}. We expect functional divergence for these enzymes, as the two structures carry out distinct biological roles, and most of the enzymes in the phylogeny have different functions. Dean *et al.* (2002) found a very low NCD (0.113) for 3αHSD, which they attributed largely to a cluster of hydroxysteroid dehydrogenases. Removal of this cluster raised the NCD to 0.42. The HSD cluster showed the fastest rates clustered around the substrate-binding cleft, whereas the rest of the enzymes in the tree had a more typical conserved pattern around the active site.

Enzyme	PDB	Species	Quaternary Structure	RMS Distance (Å)	Ligands	Ions	PROSITE
Enolase	2ONE	*Saccharomyces cerevisiae*	homodimer	0.75	2′-Phosphoglycerate	Mg^{2+}, Li^{+}	D320-P327 L342-S355
	1PDZ	*Homarus gammarus*	homodimer		Phosphoenolpyruvate 2-Phosphoglycolic Acid	Mn^{2+}	D319-P326
Fructose-1,6-Bisphosphatase	1A5C	*Plasmodium falciparum*	tetramer	0.91			V228-N238
	1FBA	*Drosophila melanogaster*	tetramer				V221-N231
5-Aminolevulinate Dehydratase	1B4E	*Escherichia coli*	octamer	n/a	Glycerol Levulinic Acid	SO^{4-} Zn^{2+}	G240-Y252
3-α-Hydroxysteroid Dehydrogenase	1LWI	*Rattus norvegicus*	monomer	0.96	NADP+		M151-F168 L268-V283, G45-G62
	1PWM	*Homo sapiens*	monomer		NAP+ Fidarestat	Cl^{-}	
RUBISCO	8RUC	*Spinacia oleracea*	8L8S	0.34 LSU 0.64 SSU	2-Carboxyarabinitol-1,5-diphosphate	Mg^{2+}	KCX201, D203 E204
	1IR2	*Chlamydomonas reinhardtii*	8L8S		2-Carboxyarabinitol-1,5-Diphosphate	Mg^{2+}	
Superoxide Dismutase	1YAZ	*Saccharomyces cerevisiae*	homodimer	0.73	Azide	Cu^{2+} Zn^{2+}	G44-T54 G138-I149
	1HL5	*Homo Sapiens*	homodimer			Cu^{2+} Zn^{2+}	G44-T54 G138-I149
Calmodulin	1CLM	*Paramecium tetraurelia*	monomer	0.41		Ca^{2+}	D20-L32, D56-F68 D93-L105, D129-F141
	1CLL	*Homo sapiens*	monomer		Ethanol	Ca^{2+}	D20-L32, D56-F68 D93-L105, D129-F141
SRC Tyrosine Kinase	2SRC	*Homo sapiens*	monomer	1.31	Phosphoaminophosphonic acid-Adenylate Ester Phosphotyrosine		L273-K295 Y382-V394, Y527
	1AD5	*Homo sapiens*	monomer		Phosphoaminophosphonic acid-Adenylate Ester Phosphotyrosine	Ca^{2+}	L273-K295 Y382-V394, Y416

Table 1.1: Enzymes used in this study.

Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO) is an α/β barrel protein that fixes atmospheric carbon dioxide (or oxygen) to ribulose-1,5-bisphosphate as the first step of the dark reactions of photosynthesis. The protein is a heterohexadecamer (8 large and 8 small subunits). The small subunit (RSSU) is encoded in the plant nuclear genome and only plays a modulating role, whereas the large subunit (RLSU) is encoded in the chloroplast genome and is catalytic. Each large subunit binds a Mg^{2+} cofactor, and an activator CO^2 molecule in addition to the reactant CO^2. Our trees include only eukaryotic species, ranging from Chlorophytes to Angiosperms (Figures 1.6 and 1.7). We use structures from *Spinacia oleracea* (8RUC:Andersson 1996) and *Chlamydomonas reinhardtii* (1IR2:Mizohata *et al.* 2002). The sequence for the large subunit is very highly conserved and slowly-evolving, and has been frequently used in plant taxonomics. The structure is also very similar for the large subunit. The structure for the large subunit is also very conserved. The spinach and Chlamydomonas structures have a RMS distance of only 0.33 Å (Mizohata *et al.* 2002). By comparison, the small subunit varies much more in both sequence and structure. We do not expect functional divergence for RUBISCO, as the enzymatic function is mostly unchanged between cyanobacteria and angiosperms, though there are minor kinetic differences between the spinach and Chlamydomonas forms (Mizohata *et al.* 2002).

The previous work found an atypical pattern of replacements in the large subunit of RUBISCO. While most sites of the large subunit of RUBISCO are very strongly conserved, some sites have very high replacement rates, leading to an atypically high variance and an a very low mean-to-variance ratio. Previous investigation focused on the large subunit, which is chloroplast-encoded. The small subunit is nuclear-encoded. Our refinement was motivated by a desire to separate out any rate variation which might have occurred due to selective pressures at a particular encoding location.

Superoxide Dismutase

Superoxide dismutase (SOD) converts superoxide radicals to hydrogen peroxide and molecular oxygen, and is necessary for life in oxygenic environments. Superoxide radicals are created as a by-product of photosynthesis and oxidative respiration, they can cause damage to the cell. SOD has an alternating parallel / anti-parallel

folded hairpin Greek key β-barrel fold. Each monomer binds one Cu^{2+} and one Zn^{2+} ion. These ions are both catalytic, though they are also necessary for proper folding (Strange *et al.* 2003). Our tree (Figure 1.8) includes species from plants and fungi/metazoans. SOD is a homodimer in all eukaryotes. Our representative structures are from *Saccharomyces cerevisiae* (1YAZ:Hart *et al.* 1999) and *Homo sapiens* (1HL5:Strange *et al.* 2003). Though we would not expect functional divergence in SOD, (Miyamoto and Fitch 1995) used a covarion approach and found that SOD had different variant and invariant sites for plant and vertebrate SODs. The channel leading to the active site has a conserved set of charged amino acids to lead superoxide ions to the active site, and these were invariant in both groups. The differently-variable sites were mostly in the beta-strand hairpins and random coils on the surface of the protein.

Calmodulin

Calmodulin (CaM) is a small protein that modulates the activity of a variety of proteins. It has four EF-hands, each of which binds a calcium ion. It is found in all eukaryotes, and our tree (Figure 1.9) includes representatives from plants, fungi, metazoa, and protists. Our structures come from *Homo sapiens* (1CLL:Chattopadhyaya *et al.* 1992) and *Paramecium tetraurelia* (1CLM:Rao *et al.* 1993). The structure is very similar across our two structures, with an RMS distance of 0.52 Å. This protein is unlike the others in this study, in that it is fairly short (148 residues) and has a long fully-exposed α helix connecting two smaller globular domains. Our structures are also from the same subtree, so it will show if structures that are closer can also lead to differences in NCDs. We do not expect functional divergence for this protein. It is generally very conserved (Rao *et al.* 1993). Its role in modulating the activity of a number of proteins will also limit the number of sites with low constraint. Even if one of these proteins is no longer regulated by CaM in one lineage, CaM will still be constrained by the requirement for interaction with the remaining proteins.

SRC Tyrosine Kinase

SRC tyrosine kinase (*c-src*) is a non-receptor protein tyrosine kinase involved in a number of signalling pathways and is implicated in carcinogenesis. The *c-src* gene family has nine members (*blk*/p56, *c-fgr*/p58, *fyn*/p59, *hck*/p59, *lck*/p56, *lyn*/p53,

p56, *c-src*/p60, *c-yes*/p62 and *yrk*/p60) with various functions and tissue expression patterns. These are monomeric, and they all share three domains which are similar among the family members, and a variable N-terminal domain. Some of the family members also have splice variants. Our tree (Figure 1.10) includes members of all these families, as well as some invertebrate *src* genes. Our structures are both from *Homo sapiens*, one is *c-src*/p60 (2SRC:Xu *et al.* 1999), and the other is a haematopoietic cell kinase (*hck*, 1AD5:Sicheri, Moarefi and Kuriyan 1997). The *c-src* family has many differences in function and expression patterns. The *c-src*, *fyn*, *c-yes*, and *yrk* genes are expressed in a broad range of tissues, but *hck*, *blk*, *c-fgr*, *lck* and *lyn* genes are only expressed in haematopoietic cells. (Gu and Gu 2003) found evidence of rate-shifted sites between two subtrees of the *c-src* family. Our analysis may determine if there is a structure-based explanation for these rate shifts.

1.3 Methods

1.3.1 Phylogenetic Trees

Each candidate sequence was BLAST'ed against the non-redundant database. Only sequences with an E-value of 10^{-30} or less were accepted (an E-value of this level most likely implies functional equivalence). If any two sequences were more than 95% identical, only one representative was used.

Sequences were aligned with ClustalW (Chenna *et al.* 2003). Phylogenetic trees were initially generated with Mr. Bayes (Huelsenbeck and Ronquist 2001). Branch lengths longer than 0.3 were pruned from the Bayesian trees and the trees were re-generated until no long branches remained. The final trees were generated from this restricted set of sequences using proml in the Phylip package (Felsenstein 1989), with the slow option enabled. These trees were used to estimate the replacement rate at each site in the alignment using a maximum likelihood method (Fitch 1971) with a Jukes-Cantor-like correction (Dean *et al.* 2002). Each tree was partitioned into two or more subtrees, depending on how many species were available. Replacement rates at each site were also estimated for the subtrees.

1.3.2 The Linear Model

The linear model described the corrected replacement rate at each site as dependent on a number of biological factors. The parameters used in the linear model included solvent exposure of each residue, minimal distance to an active site, a binary variable for prolines or glycines with unusual side chain conformations, hydropathy of the residue, and a binary variable for membership in a turn. The model is of the form:

$$y_i = a_1 + a_2\text{access}_i + a_3\text{distance}_i + a_4\phi\psi\text{Gly/Pro}_i + a_5\text{hydro}_i$$

where y_i is the estimate of the replacement rate at site i, a_1 is the intercept, access_i is the fractional solvent exposure of the residue, distance_i is the distance of the residue from the active site, hydro_i is the Kyte-Doolittle hydropathy of the residue, and $\phi\psi\text{Gly/Pro}_i$ is a dummy variable for glycine residues in a range of torsion angles normally unoccupied by other amino acids, or proline residues, which normally occupy this range as well.

Solvent exposure values were obtained with DSSP (Kabsch and Sander 1983), and normalized by the solvent accessibility for the fully-exposed Gly-X-Gly tripeptide (Shrake and Rupley 1973). We calculated the minimal distance to an active site for each residue, with an active site being a bound ion, ligand, or specified residues in the PDBsum database (Laskowski *et al.* 1997). We did a second assigment of distances where residues specified by PROSITE (Hulo *et al.* 2004) were included. The set of distances that gave the best sum of squares for each pair of enzymes was chosen. Hydropathy values were the Kyte-Doolittle hydropathy for the residue. Turns were assigned based on the assessment of DSSP. Odd angles were defined as any proline residue, or a glycine with ψ less than -70 and ϕ greater than -40. The ψ and ϕ angles describe the rotation of bonds about the α-carbon of an amino acid, with reference to the carboxylic carbon and the nitrogen atom, respectively (see figure 1.1). Torsion angles in the selected range represent those that are unlikely to be found in amino acids that are not proline or glycine residues. The same biological factors were collected for the second structure in the cases where it was available.

For each tree, we did one linear regression using the replacement rate estimates from the entire tree, and repeated it using the replacement rates from each subtree. If multiple structures were available, we used one from each subtree and repeated the whole-tree and subtree analyses with the structural factors for the second structure.

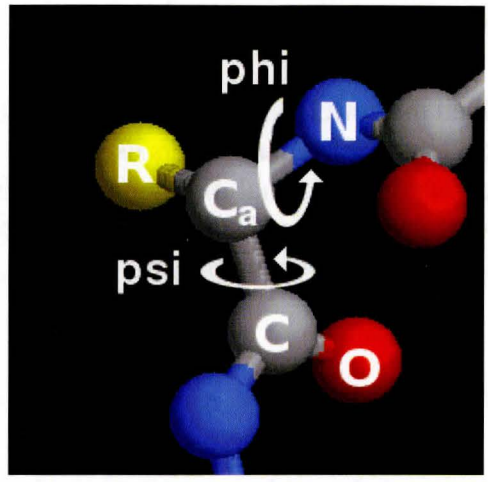

Figure 1.1: Torsion angles in the peptide backbone

1.4 Results

1.4.1 Data Quality

Dean *et al.* (2002) established a number of criteria that the phylogenies must meet in order to ensure that the data were statistically reliable. These critera are: 1) The parsimony tree has at least 1.5 replacements per site; 2) The tree contains at least 5 sequences; 3) All sequences are less than 99% identical; 4) Each sequence is at least 40% identical to a known structure; 5) No branch length is longer than 0.3; 6) No more than 30% of the replacements are assigned to branches longer than 0.2. Our data satisfies all of these criteria except for the sixth one (Table 1.2). ALAD has 35.9% of its replacements from branches longer than 0.2, and *c-src* has 26.2%.

Phylogenetic Trees

The details of the trees and subtrees are listed in Table 1.2. Each subtree automatically meets the above criteria 3, 4, and 5, as those were restrictions in the construction of the entire tree. The smallest subtree has only 13 species, but the average has 44 species. This is well above the minimum requirement of 5 for criteria 2. There are three subtrees for which criteria 1 is not met, even with the corrected

	Tree Length	Species	Length/ Species	Corrected Mean	Variance	Standard Deviation	Parsimony Mean	% Branches Over 0.2
ALAD	20.3162	93	0.2185	18.274	219.539	14.817	16.415	35.9
t1	10.7456	48	0.2239	10.213	71.958	8.483		
t2	9.4598	45	0.2102	8.180	58.127	7.624		
CaM	3.3582	63	0.0533	3.228	13.091	3.618	3.315	7.3
t1	1.8157	34	0.0534	1.789	7.198	2.683		
t2	1.5425	29	0.0532	1.470	2.735	1.654		
Enolase	3.1284	40	0.0782	3.174	14.822	3.850	2.920	0.0
t1	1.4774	13	0.1137	1.532	3.731	1.932		
t2	1.6092	27	0.0596	1.623	5.964	2.442		
ALDO	12.0534	91	0.1325	10.234	89.265	9.448	9.807	19.9
t1	6.5258	45	0.1450	5.526	26.873	5.184		
t2	5.3620	46	0.1166	4.731	28.081	5.299		
SRC	7.8241	64	0.1223	5.321	29.598	5.440	4.909	26.2
t1	1.3878	26	0.0534	0.764	1.662	1.289		
t2	6.3790	38	0.1679	4.551	21.488	4.636		
SOD	14.8074	117	0.1266	13.400	181.916	13.488	12.261	14.2
t1	7.7432	42	0.1844	6.863	40.522	6.366		
t2	3.2054	31	0.1034	3.111	16.196	4.025		
t3	3.5960	44	0.0817	3.750	22.521	4.746		
3αHSD	8.0332	63	0.1275	7.118	41.934	6.476	6.876	9.0
t1	2.6025	22	0.1183	2.334	6.396	2.529		
t2	3.0007	22	0.1364	2.684	8.511	2.917		
t3	1.9851	19	0.1045	1.867	5.050	2.247		
RUBISCO								
LSU	4.8294	265	0.0183	4.657	118.764	10.898	4.118	0.0
t1	3.3114	186	0.0178	3.147	62.146	7.883		
t2	1.4992	79	0.0190	1.506	12.475	3.532		
SSU	10.3015	127	0.0811	6.784	57.958	7.613	6.264	14.5
t1	7.3479	103	0.0713	5.336	43.416	6.589		
t2	2.2764	24	0.0949	1.056	1.865	1.366		

Table 1.2: Features of phylogenetic trees for enzymes used in this study.

replacement rates.

The phylogenetic trees are shown in Figures 1.2-1.10. The trees were preferrentially split along the longest branch in the phylogeny that split the tree in a balanced fashion.

The total branch lengths of the trees and subtrees are given in Table 1.2. Ideally, the total branch length of the entire tree would be about equally divided between the subtrees, but this is not always the case. SOD, *c-src*, and both RLSU and RSSU are particularly unbalanced. We divided the total whole and subtree lengths by the number of taxa to see if there were different overall rates for each subtree. Most enzymes do not show a drastic overall rate difference between subtrees, but enolase, *c-src* and SOD do have large differences in the branch length per species in their subtrees.

Dean *et al.* (2002) reported that the NCD is highly unstable below a mean of two replacements per site, and accurate above a mean of five. Three of the phylogenies (CaM, enolase, and RLSU) have a corrected mean of less than 5 replacements per site, and all of the phylogenies aside from ALAD have a subtree with a mean of less than 5 replacements per site. Further, 6 of the phylogenies contain at least one subtree with a mean of less than 2. This could have been avoided in one of the phylogenies, but a phylogeny would need to have a mean greater than 10 replacements per site in order to split it into two subtrees and ensure reliable NCDs in each.

The average length/ species is 0.10643, and most trees are close to this value. ALAD is by far the fastest-evolving tree overall, with 0.21845, and CaM (0.05331) and both subunits of RUBISCO (0.01834 and 0.08111) are the slowest. It is interesting to note that though they cover a similar taxonomic range, RSSU is evolving at 4.4 times the rate of the large subunit.

Table 1.3: Results of linear regressions.

Tree	Length	$\hat{\rho}^2$	$\hat{r}^2$	NCD	MSM	MSE	*acc*	*dist*	*hydro*	Gly/Pro
Enolase										
2ONE	435	0.7858	0.3871	0.4926	622.53	9.17	989.82	360.15	36.43	210.25
2ONE-t1	435	0.5895	0.2359	0.4002	95.47	2.88	173.78	43.29	23.91	35.52
2ONE-t2	435	0.7274	0.3478	0.4781	226.13	3.94	397.59	102.19	2.96	41.80
1PDZ	432	0.7846	0.3886	0.4953	618.53	9.11	1061.55	401.41	15.79	154.29
1PDZ-t1	432	0.5873	0.2134	0.3634	83.55	2.89	135.63	56.99	2.95	32.99
1PDZ-t2	432	0.7281	0.3446	0.4733	221.16	3.94	409.47	126.24	6.46	37.83
Fructose-1,6-Bisphosphatase										
1A5C	341	0.8854	0.3844	0.4342	2916.66	55.61	5089.88	1935.75	504.31	549.95
1A5C-t1	341	0.7944	0.3088	0.3887	705.38	18.80	1260.33	397.06	100.65	221.66
1A5C-t2	341	0.8313	0.2896	0.3484	674.16	19.69	944.23	698.33	123.29	14.34
1FBA	359	0.8818	0.3813	0.4324	3069.16	56.28	6025.88	2015.05	619.60	735.11
1FBA-t1	359	0.7863	0.3048	0.3876	727.28	18.74	1317.68	514.95	86.60	222.04
1FBA-t2	359	0.8275	0.3202	0.3869	828.57	19.88	1565.18	645.15	246.83	107.50
5-Aminolevulinate Dehydratase										
1B4E	322	0.9168	0.4003	0.4366	7052.23	133.32	19135.38	3001.57	2302.47	4082.32
1B4E-t1	322	0.8581	0.3565	0.4155	2058.39	46.89	5346.33	994.82	512.37	1259.99
1B4E-t2	322	0.8603	0.3062	0.3559	1428.97	40.85	4015.80	479.47	300.14	812.02

Continued on Next Page...

Table 1.3 – Continued

Tree	Length	$\hat{\rho}^2$	$\hat{r}^2$	NCD	MSM	MSE	*acc*	*dist*	*hydro*	Gly/Pro
3-α-Hydroxysteroid Dehydrogenase										
1LWI	304	0.8303	0.1690	0.2036	536.73	35.31	1111.44	324.80	202.53	437.63
1LWI-t1	304	0.6351	0.1236	0.1946	59.86	5.68	135.38	24.12	28.13	56.79
1LWI-t2	304	0.6888	0.0920	0.1336	59.91	7.91	95.28	43.77	13.78	72.31
1LWI-t3	304	0.6322	0.1576	0.2493	59.67	4.27	149.07	29.64	19.24	21.03
1PWM	315	0.8369	0.2073	0.2477	734.38	36.23	2311.40	820.82	497.14	468.84
1PWM-t1	315	0.6635	0.1112	0.1676	63.68	6.57	175.60	91.52	25.94	45.38
1PWM-t2	315	0.6921	0.1011	0.1461	72.08	8.27	192.15	111.80	72.82	44.63
1PWM-t3	315	0.6217	0.2130	0.3426	85.77	4.09	318.40	12.19	70.80	51.88
RUBISCO LSU										
8RUC	465	0.9608	0.1747	0.1818	2411.60	99.04	4963.39	1281.66	1796.32	621.96
8RUC-t1	465	0.9493	0.1552	0.1635	1120.98	53.05	2359.57	592.47	702.05	289.00
8RUC-t2	465	0.8792	0.1658	0.1886	240.41	10.51	472.02	122.88	246.64	62.46
1IR2	461	0.9611	0.1701	0.1770	2332.59	99.85	4736.01	1199.04	1832.17	612.89
1IR2-t1	461	0.9496	0.1506	0.1586	1084.23	53.64	2275.27	544.93	696.13	289.38
1IR2-t2	461	0.8802	0.1702	0.1934	243.56	10.41	469.14	115.61	286.59	56.86
RUBISCO SSU										
8RUC	122	0.8837	0.3454	0.3909	604.49	39.17	1162.96	717.19	102.45	40.04
8RUC-t1	122	0.8794	0.3067	0.3488	403.82	31.20	746.63	515.41	105.28	47.07
8RUC-t2	122	0.4523	0.2938	0.6495	17.76	1.46	36.87	12.20	0.00	0.12
1IR2	139	0.8825	0.3330	0.3773	642.85	38.44	1106.78	776.68	378.41	2.53
1IR2-t1	139	0.8849	0.2691	0.3041	399.46	32.38	665.26	493.59	266.30	2.50
1IR2-t2	139	0.4060	0.3045	0.7499	21.15	1.44	49.23	14.34	3.76	0.00

Continued on Next Page...

Table 1.3 – Continued

Tree	Length	$\hat{\rho}^2$	$\hat{r}^2$	NCD	MSM	MSE	*acc*	*dist*	*hydro*	Gly/Pro
Superoxide Dismutase										
1YAZ	152	0.9256	0.4586	0.4955	3151.27	101.22	4962.77	2268.45	897.30	1640.91
1YAZ-t1	152	0.8306	0.4512	0.5432	690.26	22.84	1091.38	555.51	339.40	251.03
1YAZ-t2	152	0.8094	0.3954	0.4885	240.42	10.00	345.91	190.76	11.57	104.86
1YAZ-t3	152	0.8289	0.2970	0.3583	242.31	15.61	535.49	80.58	134.17	99.53
1HL5	152	0.9252	0.4617	0.4990	3139.39	99.61	4315.30	1923.88	613.44	1518.12
1HL5-t1	152	0.8305	0.4147	0.4993	636.13	24.43	601.88	652.08	71.03	266.28
1HL5-t2	152	0.8093	0.4343	0.5367	264.56	9.38	381.28	158.44	20.14	75.31
1HL5-t3	152	0.8283	0.2876	0.3472	234.82	15.83	424.50	68.29	39.21	114.75
Calmodulin										
1CLM	143	0.7534	0.1405	0.1865	65.30	11.58	217.33	3.68	45.99	76.81
1CLM-t1	143	0.7514	0.1457	0.1939	37.24	6.33	136.72	31.73	50.60	38.47
1CLM-t2	143	0.4626	0.1150	0.2486	11.17	2.49	15.06	6.57	0.18	9.09
1CLL	143	0.7534	0.1418	0.1882	65.92	11.56	204.84	8.52	18.10	80.66
1CLL-t1	143	0.7514	0.1466	0.1951	37.45	6.32	136.47	40.66	27.03	42.39
1CLL-t2	143	0.4626	0.1117	0.2415	10.84	2.50	10.96	5.64	0.22	8.37
SRC Tyrosine Kinase										
2SRC	448	0.8213	0.3171	0.3861	1049.44	20.41	2232.79	643.73	9.38	253.57
2SRC-t1	448	0.7390	0.2040	0.2761	151.55	5.34	401.98	50.05	14.14	43.10
2SRC-t2	448	0.7043	0.2839	0.4031	374.72	8.54	625.22	348.63	2.82	53.91
1AD5	436	0.8246	0.2870	0.3481	935.95	21.58	1920.51	939.99	0.08	383.69
1AD5-t1	436	0.7405	0.1688	0.2280	121.71	5.56	278.19	93.58	0.34	57.44
1AD5-t2	436	0.7163	0.2855	0.3986	386.22	8.97	702.80	455.38	1.66	151.54

1.4.2 Regression Analyses

The results of the linear regression analyses for each structure and subtree are shown in Table 1.3. $\hat{\rho}^2$ ranges from 0.7534 (CaM) to 0.9608 (RLSU), with an average of 0.8624. This indicates that these data sets contain a large degree of causal rate variation, making them suitable for analysis. $\hat{\rho}^2$ almost always decreases for subtrees, as it decreases with the mean number of replacements and this always decreases for subtrees.

The 4-term linear regression model is significant at the 95% confidence level for all whole trees and subtrees used ($P(\hat{r}^2) < 0.05$). In all cases, the mean square model value (MSM) is much larger than the mean square error term (MSE), usually by an order of magnitude or more. This indicates that the 4-term model does manage to explain a significant amount of the variation in the replacement rates. The $\hat{r}^2$ value for the linear regressions ranges from 0.1405 (CaM) to 0.4586 (SOD), with an average of 0.2991 for the whole trees. It is also below 0.2 for 3αHSD and the RLSU, and above 0.3 for all other enzymes, including RSSU. The NCD ranges from 0.1818 (RLSU) to 0.4955 (SOD), with an average of 0.3564 for the whole trees. If the three lowest enzymes are excluded (RLSU (0.1818), CaM (0.1865), and 3αHSD (0.2036)), then the average NCD is 0.4393. Our basic model explains between 35%-43% of the causal rate variation in this set of enzymes.

Our model used the three terms found to be significant sources of rate variation in the last study (Accessibility, Distance, Gly/Pro), as well as one new one, Hydropathy. Membership in a turn was also tested, but was found to be a trivial source of variation in most of the enzymes. The criteria we use for a factor being a significant source of variation in the rates is if it is larger than the MSE, as all terms use one degree of freedom. If the sum of squares for a factor (SSF) is greater than the MSE, the factor is more useful than a random variable for each position. In most cases, the SSF is much larger than the MSE.

The Accessibility term is significant for all whole trees, and is the strongest source of variation for every enzyme. The Distance term is significant for all whole trees except CaM. The Gly/Pro term only applies to a few sites in each enzyme, but it is for the whole trees for all enzymes except RSSU. The new Hydropathy term is significant for 8 of the enzymes for the whole tree.

					REPLACEMENTS PER SITE				CORRELATION COEFFICIENTS						
					Parsimony		Corrected		Observed						
					Mean	Variance	Mean	Variance		Sums of squares				PECD	NCD
Enzyme		PDB	Sites	Taxa	$\bar{b}$	s_b^2	$\bar{y}$	s_y^2	$\hat{r}^2$	Distance	Access	Gly	Hydro	$\hat{\rho}^2$	$r^2/\hat{\rho}^2$
RUBISCO LSU	old	1BWV	466	160	4.118	64.134	4.589	91.884	0.171	1344.6	2102.7	627.3		0.951	0.180
	new	8RUC	466	265	4.457	103.590	4.657	118.764	0.175	1281.7	4963.4	622.0	1796.3	0.961	0.182
RUBISCO SSU	new		122	127	6.264	46.196	6.784	57.958	0.345	717.2	1163.0	40.0	102.5	0.884	0.391
Enolase	old	2ONE	419	104	14.302	172.428	16.756	275.746	0.511	10269.2	19293.3	5459.7		0.939	0.543
	new		435	40	2.920	11.731	3.169	14.738	0.387	360.2	989.8	210.3	36.4	0.786	0.493
3αHSD	old	1LWI	305	41	2.892	7.616	3.296	11.705	0.081	5.3	183.8	103.0		0.719	0.113
	new		304	63	6.876	34.598	7.502	45.725	0.169	324.8	1111.4	437.6	202.5	0.830	0.204
ALAD	old	1A5C	340	49	5.395	29.108	6.469	42.867	0.376	1119.6	1394.9	316.2		0.848	0.442
	new		341	91	9.807	70.124	10.695	90.497	0.384	1935.4	5089.9	550.0	504.3	0.885	0.434
ALDO	old	1B4E	318	46	9.535	51.321	11.039	79.906	0.413	3309.7	1388.8	1159.4		0.861	0.479
	new		322	93	16.415	159.138	18.324	219.662	0.400	3001.6	19135.4	437.6	2302.5	0.917	0.437
CaM	new	1CLL	143	63	3.315	12.623	3.609	16.499	0.141	3.7	217.3	76.8	46.0	0.753	0.187
SRC	new	2SRC	448	64	4.909	22.801	5.367	29.773	0.317	643.7	2232.8	253.6	9.4	0.821	0.386
SOD	new	1YAZ	152	117	12.261	139.260	13.516	182.771	0.459	2268.5	4962.8	1640.9	897.3	0.926	0.496

Table 1.4: Comparison of replacement rates and linear regressions for enzymes used in both this study and Dean *et. al* (2002).

To check how strongly the overall rate of evolution in the tree affects the fit of the model (and by extension, how much of the causal variation the model explains), we correlated the branch length per species in the trees with the $\hat{\rho}^2$, $\hat{r}^2$, and NCD. There was a strong and significant correlation between the branch length per species and the $\hat{r}^2$ value (r = 0.5414, P(0.5414) < 0.0001), and a slightly weaker one between the number of substitutions per species and the NCD (r = 0.4722, P(r) < 0.0001). The correlation between the number of substitutions per species and $\hat{\rho}^2$ was weak and not significant (r = 0.0500, P(r) = 0.7273). Not only is the model and the NCD more reliable when a phylogeny has a faster overall rate of evolution, but the amount of variation in rates that the model explains is higher as well.

1.4.3 Comparison With Previous Results

Five of the enzymes used in this study were also used in Dean *et. al* (2002). The results of the regression analyses from that study and the current ones are presented in Table 1.4. The same primary PDB structure files were used for all enzymes expect RUBISCO. We created a tree for the PDB file used by Dean *et al.* (2002) (1BWV from *Galdieria partita*), but the Bayesian trees we used in the first step produced longer branches than the parsimony method used earlier. This resulted in the *Galdieria* tree containing too many long branches, requiring pruning of so many sequences as to make the tree unuseable. The different tree construction method also affected enolase, as our tree has less than half the number of species that the tree from the earlier study does.

All of the previously-used enzymes have large differences in the number of species used in the final tree. In all cases aside from enolase, it is higher by a factor of at least 1.5. For enolase, the number of species is decreased by a factor of 2.6. There were 408 species in the enolase tree before pruning, so it has probably been more strongly affected by the use of more accurate Bayesian trees than the other protein sequences were. The mean and variance for both the parsimony and corrected replacement rates also differ. In all cases except RUBISCO, the larger phylogeny also has a higher mean replacement rate. This is expected, as a larger phylogeny would allow more opportunities for replacements to occur. However, this means that the sums of squares cannot be directly compared across studies (the total sum of squares for the linear regression analyses of the previous study were not provided, so the results cannot be normalized). The differences in means and variances seem to have an effect on the PECD ($\hat{\rho}^2$), as these all vary across the

two studies (they also tend to increase with phylogeny size). The $\hat{r}^2$ values for the linear regressions are close to those of the previous study for the RLSU, ALDO, and ALAD. For enolase and 3αHSD, they increase or decrease with the number of taxa. The NCDs follow a similar trend.

Though we cannot directly compare the sums of squares for the terms used in the regression, we can look at the relative values of them. Accessibility has the largest SSF for every enzyme but one in the previous study, and for every enzyme in the current one. The Accessibility SS makes up on average 48.5% of the SSM for the previous work, and 62.5% of the current one. The Distance term is slightly diminshed as a source of variation, making about 32.0% of the SSM for the previous work and 17.8% of this one. While the Gly term covers the remaining 19.4% of the last work, the remainder is about evenly divided between the Gly/Pro term (10.0%) and the new Hydro term (9.7%). This indicates that the Hydro term is a worthwhile addition, as it is about as significant source of variation as the Gly/Pro term is.

The new enzymes also have the Accessibility term explaining the majority of the variation (60.5% average) of the variation than the Distance term does (20.1%). However, the Hydropathy term explains about half as much variation as the Gly/Pro term for these enzymes (7.0% and 12.3%, respectively). These proportions are similar to the new values for the enzymes used in both studies.

1.4.4 The Fit is Not Improved at a Smaller Evolutionary Scale

In order to see if the structural linear regression model improves the fit of the model at a smaller evolutionary scale, we repeated the analysis with the phylogenetic tree divided into two or three smaller trees (Table 1.3). For 8 of the 9 enzymes used, we have structures from two different species, one in each subtree (except CaM, where both are in the same subtree). We predict that the causal variation will be better explained in the subtree that contains the target structure and less so in the other subtree. This will be reflected by an increase in the NCD for the subtree that contains the structure, relative to the NCD for the whole tree and the other subtree.

In general, we do not see this trend. Only 1 of the 9 enzymes (SOD) shows an NCD that increases more for the subtree containing the structure for both structures used, though 4 of the 9 enzymes show an increase in the subtree that has a higher number of substitutions per species. For the 8 enzymes that have a second structure available, the average difference in NCD between structures for the whole trees is

0.0138. Between the subtrees, there is an average difference of 0.0296, indicating a greater difference in the NCD at the subtree level. Though the difference is greater, there is no direction implied by the difference (*i.e.*, whether this is due to a relatively higher NCD for the subtrees containing their own structure when using the factors from that structure). For enolase, ALDO, RSSU, and SOD, one of the subtrees has a difference between the subtrees that is at least an order of magnitude greater than that between the whole trees. These higher differences correspond to a higher NCD in that subtree for the native structure.

In general, there is a high correspondence between the NCDs for the same subtree, across homologous enzymes. The correlation coefficient between whole-tree NCDs for the two enzymes is 0.9862, and for the subtrees it decreases only slightly to 0.9645. For comparison, the correlation coefficient between the two subtrees for a single enzyme is much lower, 0.5278. From this, we can conclude that the specific subtree has a greater effect on the NCD than the structural variant. We also took the average of all of the NCDs for subtrees containing their structure (*t1* for the main structure, *t2* for the alternative structure) and those subtrees which do not contain the structure (*t1* for the alternative structure, *t2* for the main structure). The mean NCD for the structure-containing subtrees is 0.3616, and for the non-structure-containing subtrees it is 0.3194. This indicates that to some degree, the NCD will be higher for the tree that is phylogenetically closer to the enzyme structure, though using a closer subtree does not increase the NCD over the whole tree.

The relative power of the structural factors changes somewhat with the partitioning. The Accessibility term is significant ($P < 0.05$) for all whole trees, and this does not change for the subtrees or across alternate structure. The Distance term is significant for all whole trees except CaM, and it is significant for all but one subtree among the other enzymes. The other two factors are less often significant across all subtrees. While the SSF varies substantially by subtree, it does not vary as much across homologous enzyme pairs. In the cases of ALDO and 3αHSD there is more of a difference in the relative SSF values across the enzyme pair, but there is no obvious reason why this should be so. While some of the enzyme pairs are fairly close phylogenetically, these enzymes are not the ones with the greatest branch length between them.

1.4.5 Rate Colourings

The replacement rates were normalized and mapped onto the tertiary structures for ALAD and ALDO. The replacement rates were normalized over the whole protein, and replaced the temperature column in the PDB structure file. The sites range from dark blue (highly conserved) through green, yellow, orange and red (rapidly evolving). This technique allows quick visualization of constraints on protein sites. A conserved core with fast-evolving sites scattered over the surface of the protein is a typical pattern. We may also use this technique to visualize sites which have altered constraints across subtrees.

ALAD has only one structure available. This is a long protein (322 aa) with a high mean number of replacements in both subtrees (10.2 and 8.1). Both subtrees have a decreased NCD, but *t2* shows a 4-fold greater decrease. The enzyme is an octamer. We show both sides of a single subunit for clarity. The extended tail wraps around a neighbouring subunit, so the interior side of it is shielded from solvent. The colouring for the whole tree shows the expected pattern, with the external residues generally evolving faster (Fig. 1.11c) and the surface shielded by other subunits (Fig. 1.11d), as well as the core (Fig. 1.11a, Fig. 1.11b), being more conserved. When the subtree rates are used, a number sites on the surface show a rate change (Fig. 1.11e, Fig. 1.11f, Fig. 1.11g, Fig. 1.11h). Some sites are faster in one subtree and not the other, and are evolving at a moderate speed over the whole tree. Others are evolving more slowly in both subtrees, but quickly over the whole tree. These are like the Type I and Type II sites described by Gu (1999), respectively. Comparing the rates for *t1* and *t2*, the *t1* rates seem to have a more reasonable distribution of fast sites on the unshielded surface, whereas *t2* has more hotter sites, but also more apparently conserved sites on the unshielded surface. Conserved sites would not be expected on a solvent-exposed surface, unless the sites were involved in a protein-protein interaction or some other ligand. This tree was divided into Zn^{2+} and Mg^{2+}-dependent forms of the enzyme, and some moderate structural rearrangements are not unlikely. Some degree of functional divergence was also possible for this enzyme, and it may be reflected in the different pattern of rates seen across the subtrees.

ALDO is another long protein (341 aa) with a relatively large mean number of replacements in both subtrees (5.5 and 4.7). ALDO had a relatively high RMS distance between the two structures, so it may be possible to detect rate distribution differences that are related to the structural differences. Specifically, *Plasmodium*

ALDO (1A5C) has a greater overall surface area exposed compared to the human form (Kim *et al.* 1998). Both subtrees experience a decrease in NCD, but it is greater for *t2* when the non-native structure is used to model the rates. Comparing the two structures with the rates for the whole trees (Fig. 1.12a, Fig. 1.12b), a characteristic pattern of replacement rates is seen across both structures. The conserved patch towards the middle of each subunit reflects the channel of the α/β barrel, which leads to the active site. The patch is clearly larger in 1A5C, which suggests that a greater protrusion of this channel is partly responsible for the increased surface area of this structure. Despite the significant structural differences, there is not a large difference in the NCD for the whole tree between these two structures (0.0018). Subtree 1 has an even smaller difference between subtrees (0.0011), and it shows roughly the same pattern of replacements scattered over the surface of the two structures, though with fewer very rapidly evolving sites (Fig. 1.12c, Fig. 1.12d). The conserved cleft is still more prevalent on 1A5C, but in both cases this does not seem to be a great enough difference to strongly influence the NCD. For subtree 2, the difference in NCDs is much larger (0.0385). The pattern of replacements has also changed somewhat (Fig. 1.12e, Fig. 1.12f). Though 1FBA shows more sites on the surface that appear to be conserved, 1A5C shows this pattern much more strongly. The conserved cleft has expanded to cover a much larger part of the exposed surface. It follows that the model, which is largely influenced by the solvent accessibility of residues, would provide a poorer fit with this combination of structure and subtree.

1.4.6 Atypical Replacement Patterns in the Large Subunit of RUBISCO

The previous work observed that the NCD was particularly low for RLSU, despite a fairly high $\hat{\rho}^2$. RLSU also had an abnormal distribution of replacements compared to the other α/β barrel proteins, in having a much higher variance. Specifically, while most sites are strongly conserved, quite a few had very high replacement rates. The authors did not propose any structure-based explanation for this deviance from the other proteins.

We see this pattern as well (Fig. 1.13). However, we have also included the small subunit of RUBISCO, which is nuclear-encoded. While the large subunit shows the same pattern of a number of sites with abnormally high replacement rates, the small subunit shows a pattern much more similar to the other enzymes

	Nuclear				Chloroplast			
	1st	2nd	3rd	Mean	1st	2nd	3rd	Mean
Chlamydomonas reinhardtii	64.7	47.9	86.1	66.3	44.4	37.4	19.4	33.7
Pinus thunbergii	54.0	47.0	53.6	51.5	45.8	38.5	31.7	38.7
Oryza sativa	58.4	46.4	61.3	55.4	48.5	39.9	33.7	40.7
Spinacia oleracea	52.5	41.5	42.3	45.4	47.7	39.1	31.6	39.5
Average	57.4	45.7	60.8	54.7	46.6	38.7	29.1	38.2
	RUBISCO SSU				RUBISCO LSU			
	1st	2nd	3rd	Mean	1st	2nd	3rd	Mean
Chlamydomonas reinhardtii	52.1	44.4	84.5	60.3	60.1	44.1	25.0	43.1
Pinus thunbergii	49.4	44.8	72.7	55.6	58.2	43.9	30.0	44.0
Oryza sativa	51.2	43.8	82.4	59.1	56.3	44.3	32.0	44.2
Spinacia oleracea	50.2	43.6	63.5	52.5	57.8	43.9	29.8	42.9
Average	50.7	44.2	75.8	56.9	58.1	44.1	29.2	43.6

Table 1.5: %G+C content of whole nuclear and chloroplast genomes, RUBISCO SSU and LSU genes.

in this study. We see this pattern is also reflected in a mapping of the rates onto the protein structure (Fig. 1.14). While both subunits display the fastest sites on the exterior of the protein, the large subunit is generally conserved with only a few rapidly-evolving sites. The small subunit shows more of a range of rates, and the distribution of these is not as skewed towards low values as it is with the large subunit. The different distribution between the large and small subunits suggests that the atypical distribution of replacements in the large subunit may be due to its location in the chloroplast genome, and not due to any structural constraints. While the mean replacement rate of the small subunit (6.73) is higher than that of the large subunit (4.67), its variance is much lower (57.86 vs 118.97), decreasing the $\hat{\rho}^2$ from 0.9608 for the large subunit to 0.8837 for the small subunit. The $\hat{r}^2$ value is also about twice as high for the small subunit (0.1747 vs. 0.3454), leading to an even larger difference in the NCDs (0.1818 vs 0.3909).

We analyzed the differences between the replacement patterns for the small and large subunits in more detail. Specifically, we looked at the profile of the numbers and types of different amino acids represented at the rapidly-evolving sites in each subunit, and we tested to see if the abnormal patterns were driven by a nucleotide

composition bias. The %G+C for the entire and chloroplast genomes of the four species, as well as the %G+C from the genes for the large and small subunits of RUBISCO are shown in Table 1.5. There is a pronounced difference in the %G+C content of the plant nucleus and the chloroplast genome for many species. We focused on a sample of four species from our tree, a chlorophyte (*Chlamydomonas reinhardtii*), a gymnosperm (*Pinus thunbergii*), a monocot (*Oryza sativa*), and a dicot (*Spinacia oleracea*). Among these, the overall nuclear-chloroplast difference in coding %G+C is greatest in Chlamydomonas (32.53%), and weakest in spinach (5.94%). Such a pronounced difference points to the possibility of the composition bias being strong enough to result in non-synonymous substitutions at some unconstrained sites.

For the whole genomes, the %G+C varies most between species at the third coding position, indicating that the bias would mostly result in synonymous substitutions. This pattern is also reflected for just the RUBISCO genes (except that there is little change in the composition of the large subunit at all, relative to the small subunit and genome-wide comparisions). This effectively rules out a change in composition bias in the chloroplast genome contributing to replacements in RLSU. The average 1st, 2nd, and 3rd codon positions have a %G+C of 58.07, 44.06, 29.04 for the large subunit, and 50.72, 44.12, 75.77 for the small subunit. While there is a noticeable difference in the composition of the first, and especially third position, there is virtually no difference in the second position. Further, the values for the first position are not changing much across species for the large subunit. Even if the composition bias may have favoured some amino acids over others initially, there is no evidence that it is currently driving nonsynonymous subsitutions in the large subunit.

We looked at the actual pattern of replacements in the hot sites for the large and small subunits. Hot sites were defined as those with a replacement rate greater than two standard deviations from the mean replacement rate in the protein. For the small subunit, 8.45% of the sites were above this cut-off, and 5.09% were above it for the large subunit. We found that the number of different types of amino acids represented at each hot site differed between the small and large subunits. The distribution of different amino acids also differs for the large and small subunits. We ordered the proportions of each amino acid represented in a hot site, and took an average of these across all sites. For the small subunit, the most-represented amino acid makes up an average of 33.77% of the sites in an alignment position, the second most-represented makes up 24.98%, the third makes up 16.54%, and there is a

gradual taper for the rest of the ranks. For the large subunit, the most-represented amino acid makes up a much greater proportion of the site at 63.82%. The second most-represented covers 23.97%, leaving little room for other amino acids to be represented. This describes the pattern that the large subunit hot sites display in the sequence alignment (not shown), which is an alternation of two amino acids throughout the whole tree. This pattern could indicate a high rate of change between two different amino acids throughout time. Alternately, it could represent high heterozygosity at these positions, which appears as replacements due to insufficient sampling of plastid genotypes for each species. The chloroplast genome is represented in much greater copy numbers than the nuclear genes, so the maintenance of heterozygosity is plausible in this case.

1.5 Discussion

Dean *et al.* (2002) found that a simple linear regression model with terms for solvent exposure, distance from the active site, and the presence of glycines or prolines explained half of the causal variation in a broad sample of α/β barrel proteins. In light of observations that replacement rates change along branches as well as across sites, we wanted to see if restricting our analysis to smaller lineages could account for some of the remaining variation. If replacement rate heterogeneity is influenced by phylogenetically local structural adaptations. we predict that a subtree will have a greater NCD when a structure native to that subtree is used to model the rates than if a non-native structure is used. We instead found that the NCD generally decreased for subtrees compared to the whole tree. Among subtrees, there was a slight increase in the NCD when a native structure was used relative to a non-native structure, but the specific subtree generally had a much greater effect on the NCD.

The previous study set standards for data quality that may have limited the power of this study. Specifically, the requirements that branch lengths all be shorter than 0.3 may have truncated some data sets such that upholding this criteria in conjunction with the requirement that each subtree have a mean of at least two replacements per site was not possible. Dean *et al.* (2002) used parsimony trees for their analysis, which generally underestimates the number of replacements. In the interest of improving tree quality, we used a Bayesian tree initially, and pruned sequences accoring to that tree. The Bayesian trees had longer branches overall than Neighbour-Joining trees made from the same sequences (data not shown).

Neighbour-Joining trees also typically have higher replacement estimates than parsimony trees. While the Bayesian tree is more accurate, it is also probably incompatible with the branch length maxima established for parsimony trees.

Only one enzyme (ALAD) has a mean of more than 5 replacements for all subtrees, and that tree exceeds the 30% allotment of branches over 0.2. The enzymes that did have reliable NCDs (mean greater than 2 replacements) for all subtrees are ALDO, SOD, and 3αHSD. Among these, only SOD follows the expected pattern of NCD elevation for the subtree when modelled by its native enzyme, though ALAD and ALDO also have relatively higher NCDs for subtrees modelled by their native enzymes. This suggests our basic results would not change even if our trees had higher means.

Since this study sought changes in the fit of the linear model across different lineages, the requirement of high sequence identity (40% in Dean *et al.* (2002), we used a BLAST E-value above 10^{-30}) probably limited the power of the method as well. This cut-off was chosen to minimize structural differences between homologs, but structural differences were exactly what are required to see differences in the fit of the linear model. However, this may be an inherent limitation, since structure tends to change much more slowly than sequence. If sequences with greater structural differences were chosen, the probability of multiple replacements at each protein site becomes greater, and rate estimates become more inaccurate. However, there is probably some conjunction of identity requirements and phylogenetic method will maximize power of this analysis without compromising the accuracy of the rate estimates.

Despite different phylogeny sizes and the addition of a new term to the linear model, our results for the whole tree regressions are fairly similar to those of Dean *et al.* (2002) for the old enzymes. It is presumed that where the number of sequences used changes between this study and Dean *et al.* (2002), these sequences are more or less equally distributed over the whole tree, not all added to one new cluster. Thus, we can compare the changes in $\hat{\rho}^2$, $\hat{r}^2$, and the NCD with those for the subtree analysis to gauge how much of an effect non-lineage-dependent changes in phylogeny size has on these values. The RLSU is the only enzyme for which the mean and $\hat{\rho}^2$ seem largely unaffected by the change in phylogeny size. However, RLSU has both the largest phylogeny in both studies and the lowest replacement rate per species, indicating that it is probably very conserved and evolutionarily stable. For all of the other enzymes used in both studies, the number of taxa change by a factor of 1.5-2.6, and the mean number of replacements follows a similar pattern.

The $\hat{\rho}^2$ also increases with number of taxa, but not as drastically. The NCD changes even less drastically, and hardly at all for the enzymes where the uncorrected mean is greater than 5 for both studies. This supports the simulation studies of Dean *et al.* (2002). The notable decrease in $\hat{\rho}^2$ for the enzymes with a mean below 5 in one study and above it in another (enolase and 3αHSD) suggests that more of the variation in rates appears stochastic over smaller phylogenies with fewer replacements. We see this pattern reflected in the subtree analyses as well.

Though the models have differences across the two studies, these are somewhat minor. Our model included a term for hydropathy, and included prolines in the binary variable for glycine residues in unusual main chain conformations. Minor differences in the assignment of distance from the active site were also likely. Despite these differences, there is a clear trend to the fluctuation in NCDs that follows the change in the number of taxa across studies, so it is reasonable to use the differences in NCD between studies as a guide to how much NCDs fluctuate when the phylogenetic differences are not restricted to distinct lineages. For most of the enzymes used in both studies, the difference in NCD due to lineage-specific effects (non-overlapping subtrees) is greater than the difference in NCD between studies (overlapping subtrees). This difference is greater even for the two enzymes that have a mean of more than two replacements in each subtree, so this result is not likely just due to fluctuation in NCD based on the size of the data set.

Some enzymes had a subtree with a much greater difference in NCD between structures compared to the whole tree difference in NCD between structures (enolase, 3αHSD, RSSU, SOD). For these enzymes, there was a strong and significant correlation (r = -0.6504, P(r) = 0.0161) between the difference between structures (for whole and subtrees) and the PECD ($\hat{\rho}^2$). This correlation is much greater than that for the enzymes which do not have a subtree with a much greater difference (*c-src*, 3αHSD, RLSU, CaM, r = -0.2242, P(r) = 0.4616). In the cases where one subtree shows a much greater NCD for one structure, it is always higher for the native structure. This pattern indicates that the subtree which is better modelled by its native enzyme also has a greater proportion of Poisson rate variation.

The enzymes with a large difference between structures for a subtree do not fit any obvious pattern. The subtrees are not universally those expected to be undergoing functional divergence, and some of the subtrees that were expected to be divergent do not show great differences. These enzymes do not have the greatest RMS distances between structures, nor are there any obvious differences in the sums of squares for individual model terms in all cases.

We propose that for these subtrees where more of the variation is stochastic, the NCD is higher because the causal variation is simpler in form, and is more fully described by our model. There are other types of causal rate variation that is not included in our model, and for the subtrees with the higher PECD, it is likely that other constraints are at work. These constraints are probably represented in the structure for that subtree, but not in a form that our model draws out. Without this additional information about causal rate variation, the two structures appear to provide roughly equivalent NCDs. Some other terms not yet tested that could be relevant when looking for lineage-dependent changes in constraint include sites of protein-protein interactions and the number of other proteins that the protein interacts with (Lopez, Casane and Philippe 2002). Another likely cause of deterministic but unaccounted for rate variation is the effects of a fluctuating neutral space model (Takahata 1987), wherein certain replacements in a protein results in a new landscape of neutral or deleterious replacements in neighbouring residues. Such effects are not suitable for inclusion in a simple linear model, but they do appear to have an effect on site-wise rate heterogeneity.

In almost all cases, the NCD was greater for the whole tree than for either subtree. This result indicates that the subtree approach is not useful for explaining the remaining rate variation. This pattern suggests that the effects of simple structural parameters are stronger over longer evolutionary times. At smaller scales, more of the variation in the rates can occur randomly, but over longer scales, the difference between structurally constrained and unconstrained sites becomes more apparent. This is supported by the observation that the variance in rates is larger for all whole trees relative to the subtrees. The approach described in this paper does not focus on rate shifts that occur at specific sites, and may be too general to detect subtle changes that occur over relatively close evolutionary times. A model that describes and models the differences in the rates and all factors between sites may perform better. A difference-based model would also allow the inclusion of differences in the organisms, such as the temperature they live at, or relative exposure to various substances such as acid or free radicals.

We tested two new terms for addition to the linear regression model: the hydropathy of the residue, and membership of the residue in a turn. Turn membership was not a significant source of variation for almost all enzymes (data not shown), but the Kyte-Doolittle hydropathy value was. The Hydropathy term may seem redundant when the Accessibility term already explains such a large part of the variation, but the Hydropathy value conveys additional information. It is the only term in our

model that allows some indication of amino acid identity other than Gly or Pro, but it does so in a manner that requires only one degree of freedom. It also indicates charged residues.

The coefficients of the Hydropathy term were all positive where they were significant. The Kyte-Doolittle hydropathy scale runs from 4.5 for the most hydrophobic residues to -4.5 for the most hydrophilic/ charged residues. This suggests either an elevated replacement rate with hydrophobic residues (which would contradict the strongly conservative effect that being buried has on residues), or a decrease in the replacement rate for charged residues. As the Kyte-Doolittle scale is weighted towards negative values (hydrophilic and charged residues), the latter explanation seems more likely. Dean and Golding (2000) did not find a factor for hydrophobic, hydrophilic, charged, and Pro or Gly sufficiently significant for the four degrees of freedom required. However, in an additional analysis of enolase, Dean *et al.* (2002) found that Arg, Asp and Glu were the most conserved of the 20 individual amino acids, indicating some conservative role for charged amino acids. The hydropathy scale consumes only a single degree of freedom, and allows for a gradient of hydrophilicity or charge.

The large subunit of RUBISCO displayed an atypical distribution of rates in both Dean *et al.* (2002) and in this work. We included the nuclear-encoded small subunit in this study to see if coding location had some effect on the pattern of replacement rates. The large subunit is generally highly conserved, but some sites have very high replacement rates. This leads to an abnormally low mean-to-variance ratio as compared to the other enzymes in this study. We found that the small subunit had a more typical mean-to-variance ratio, suggesting that the encoding of the large subunit in the chloroplast genome was responsible for the atypical pattern.

The chloroplast genome has a much different genomic and population structure than the plant nuclear genes. It is a small, circular chromosome, with an interspersed repeat and two single-copy regions. Each plant cell has between 50-150 chloroplasts, and these segregate in a more or less random fashion upon cell division, and they are strictly maternally inherited for most species. Each chloroplast contains many copies of its genome, on the order of about 100. There is some evidence that there is recombination between these, particularly in the interspersed repeat regions (Maier *et al.* 1995). The chloroplast genome-encoded gene used in this study showed a pattern of replacements that differed from the other proteins studied, in that most sites were strongly conserved, and a few had a very high

replacement rate. Closer analysis has shown that most of these sites are fluctuating between two amino acids throughout the tree. The great copy number of the chloroplast genome, coupled with some gene conversion biased in favour of the wild type (Birky CW and Walsh 1992) is probably sufficient to prevent most new replacements, as even if a new neutral mutation arose, the chances of it coming to prominence through drift are much lower than they would be for a nuclear gene. For the hot sites, we propose that there is a certain level of heterozygosity that is more or less permanently maintained, which is possible in a very large population. It is likely that any chloroplast would be genotypically uniform, but any cell may have a collection of plastids with different genotypes. Since dividing cells would inherit between 25 and 75 chloroplasts, it is reasonable that each cell inherits plastids with a variety of genotypes. Whatever the cause of the atypical pattern, it is likely that the population structure of the organelle genomes will create some deviation from the patterns that would be expected for a diploid nuclear population, which may make it unsuitable for analysis by this method. Additionally, this method is currently most suitable for globular proteins, and most organelle-encoded proteins are membrane-embedded.

1.6 Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (Canada) Research grants to GBG.

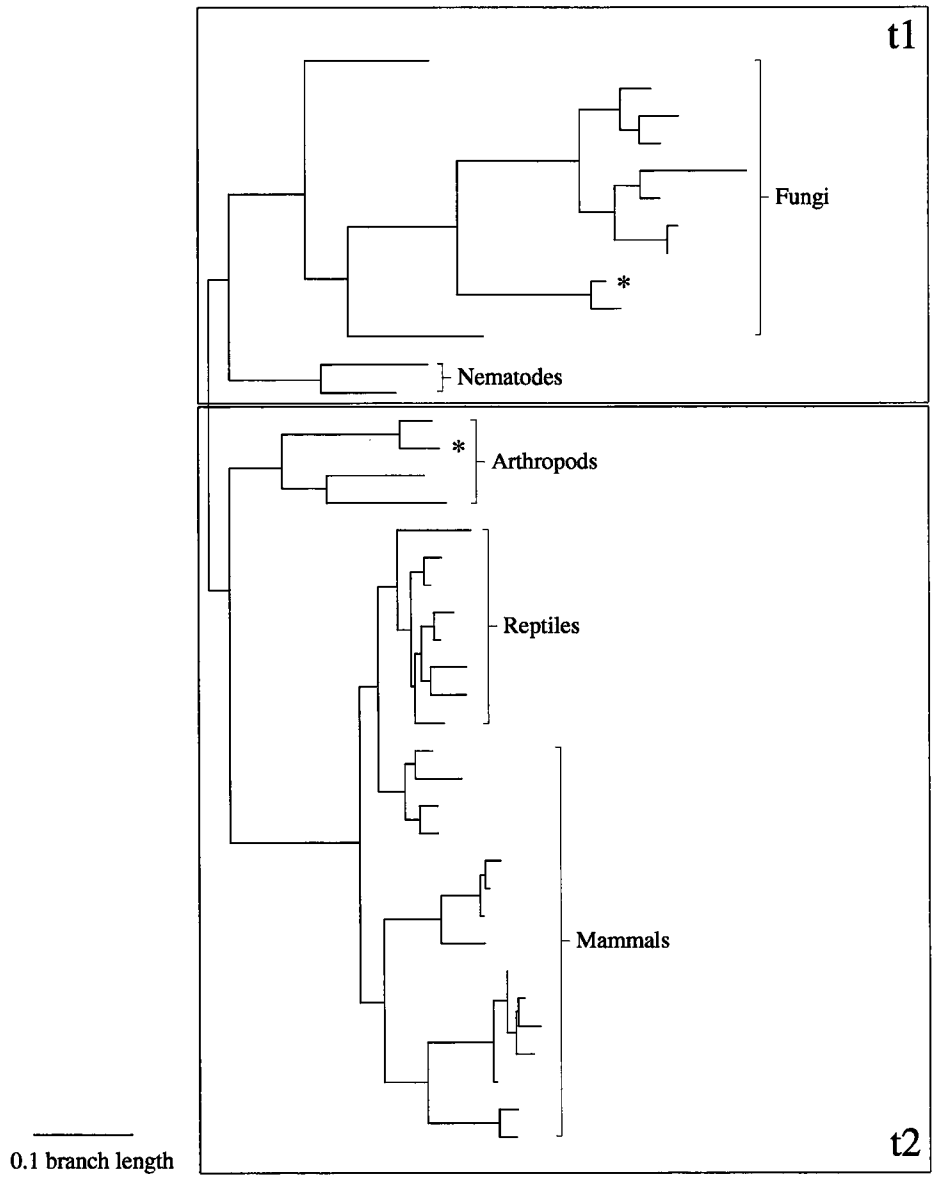

Figure 1.2: Phylogenetic tree for Enolase. The asterisks indicate sequences with structures which were used for this study.

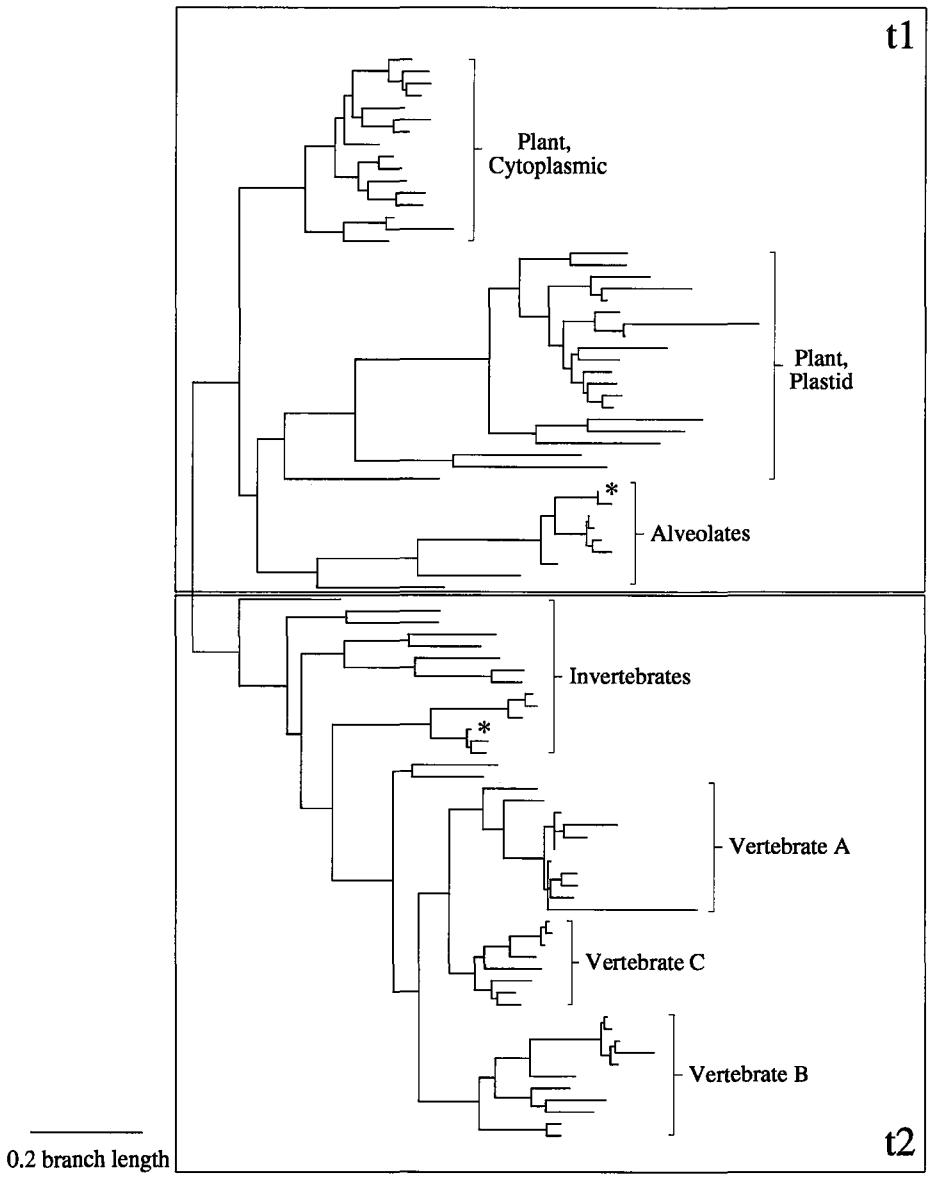

Figure 1.3: Phylogenetic tree for Fructose-1,6-bisphosphate aldolase (Class I). The asterisks indicate sequences with structures which were used for this study.

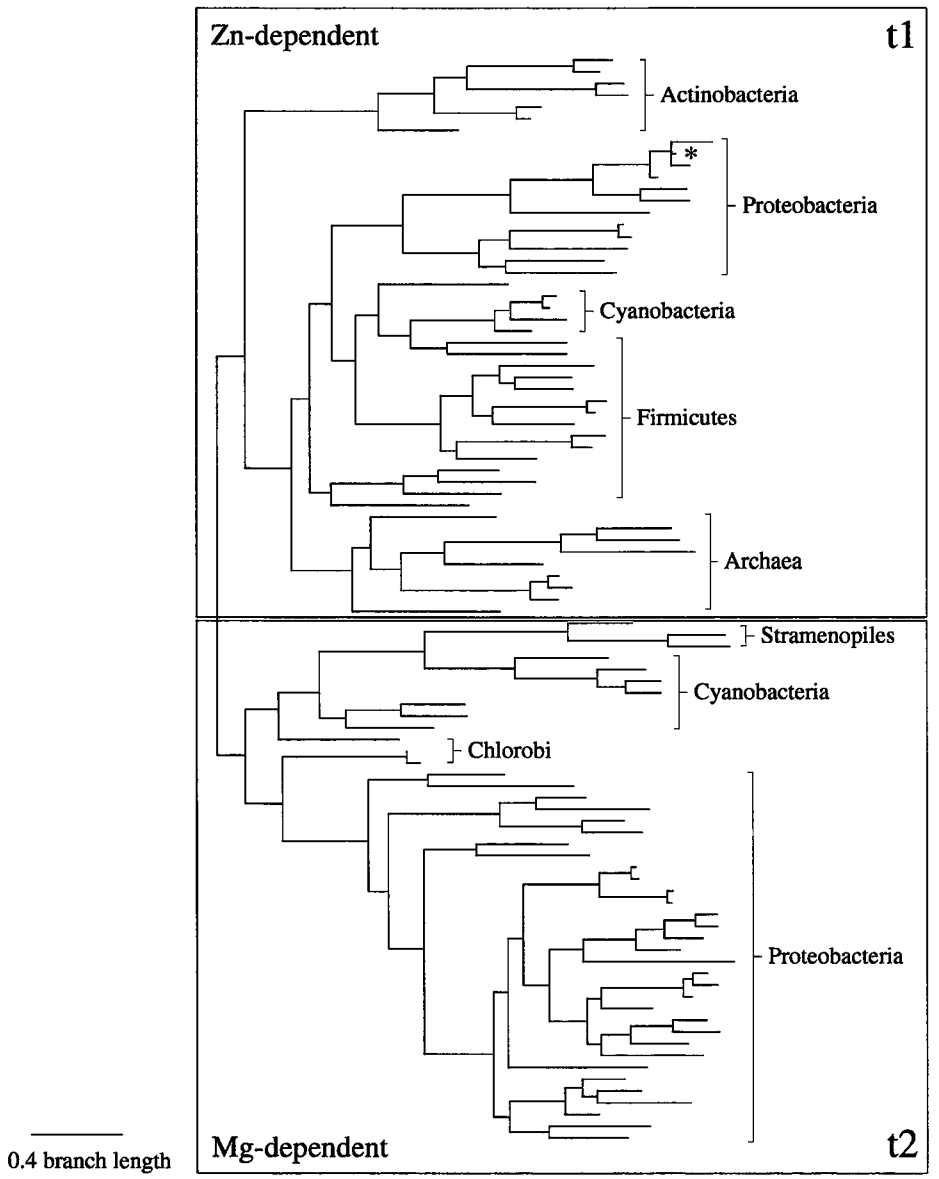

Figure 1.4: Phylogenetic tree for 5-Aminolevulinate Dehydratase. The asterisk indicates the sequence of the structure which was used for this study.

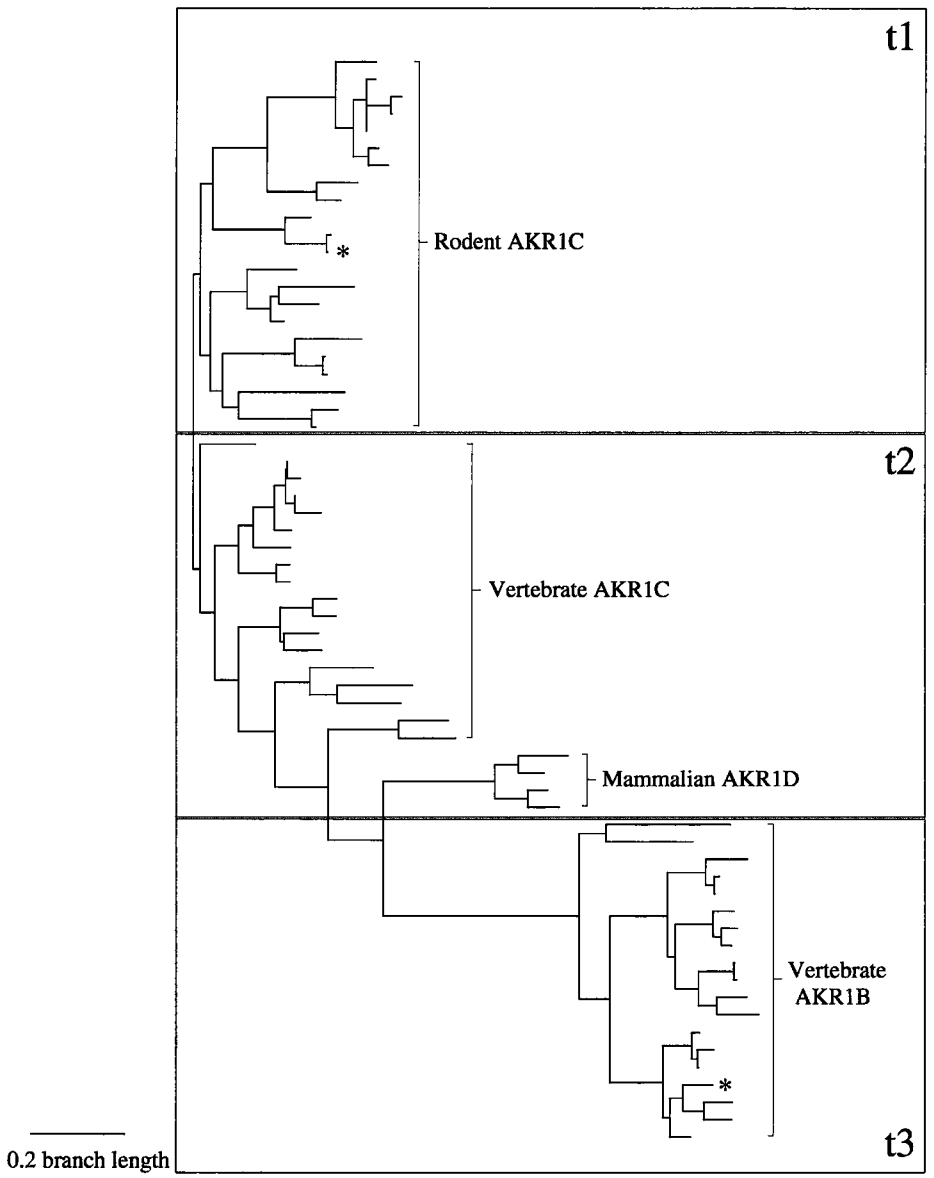

Figure 1.5: Phylogenetic tree for 3-α-hydroxysteroid dehydrogenase. The asterisks indicate sequences with structures which were used for this study.

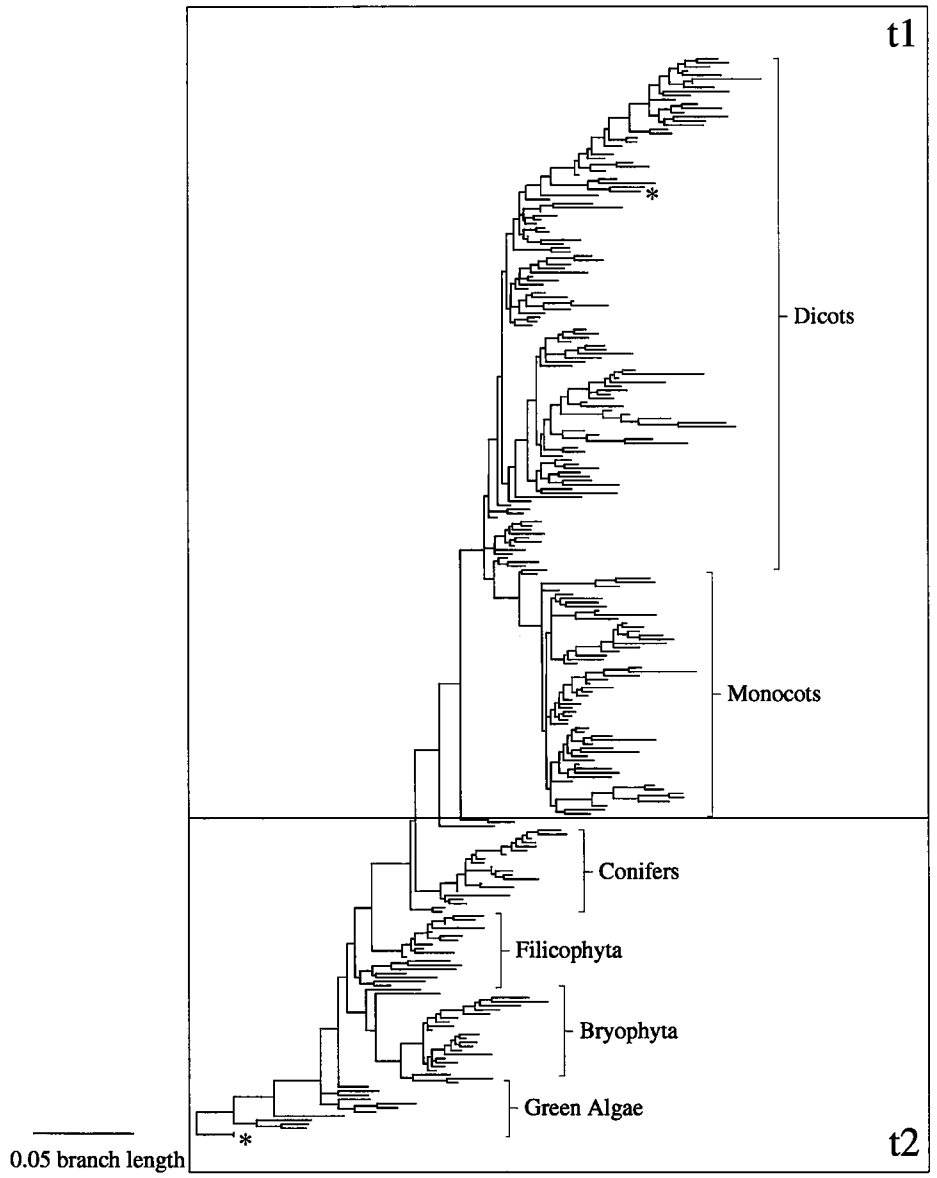

Figure 1.6: Phylogenetic tree for the large subunit of RUBISCO. The asterisks indicate sequences with structures which were used for this study.

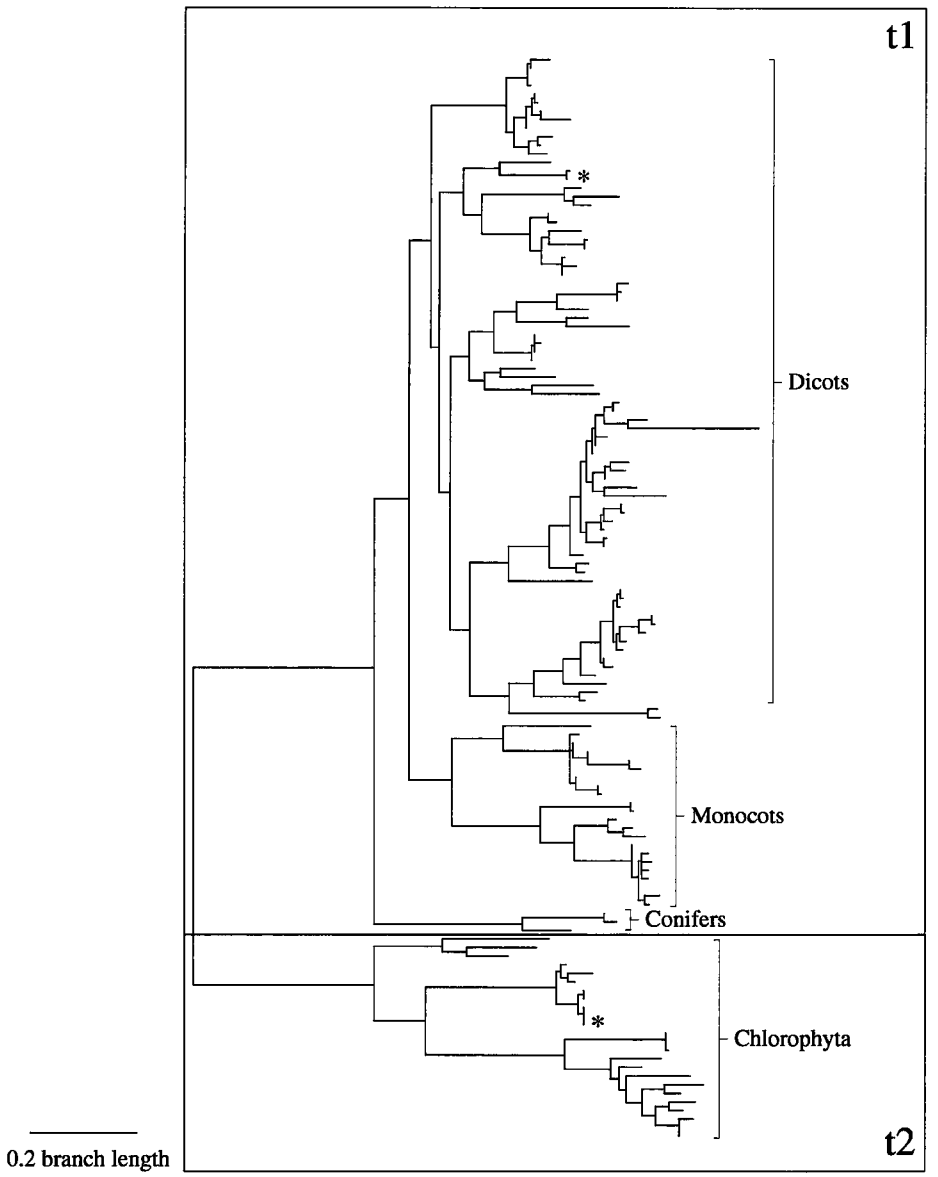

Figure 1.7: Phylogenetic tree for the small subunit of RUBISCO. The asterisks indicate sequences with structures which were used for this study.

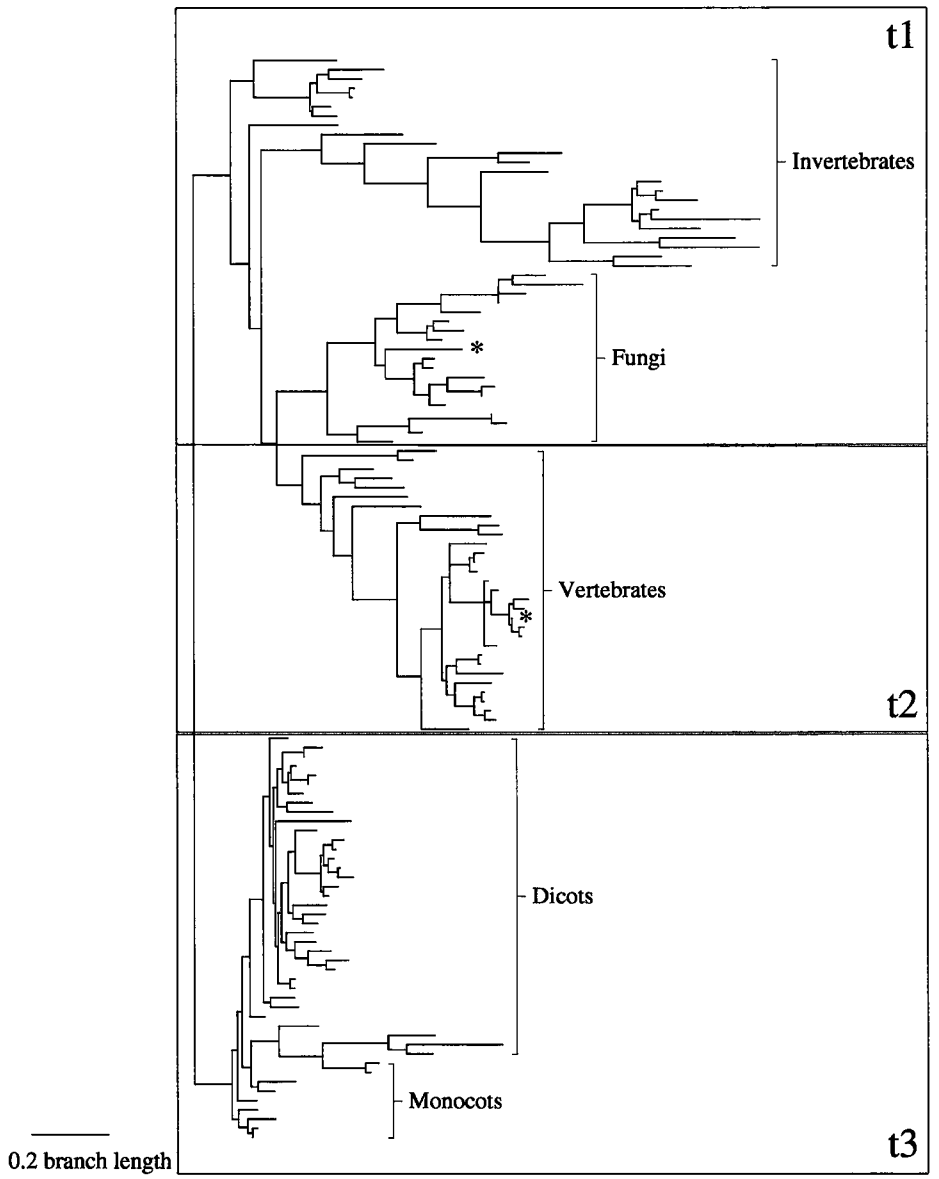

Figure 1.8: Phylogenetic tree for Superoxide Dismutase. The asterisks indicate sequences with structures which were used for this study.

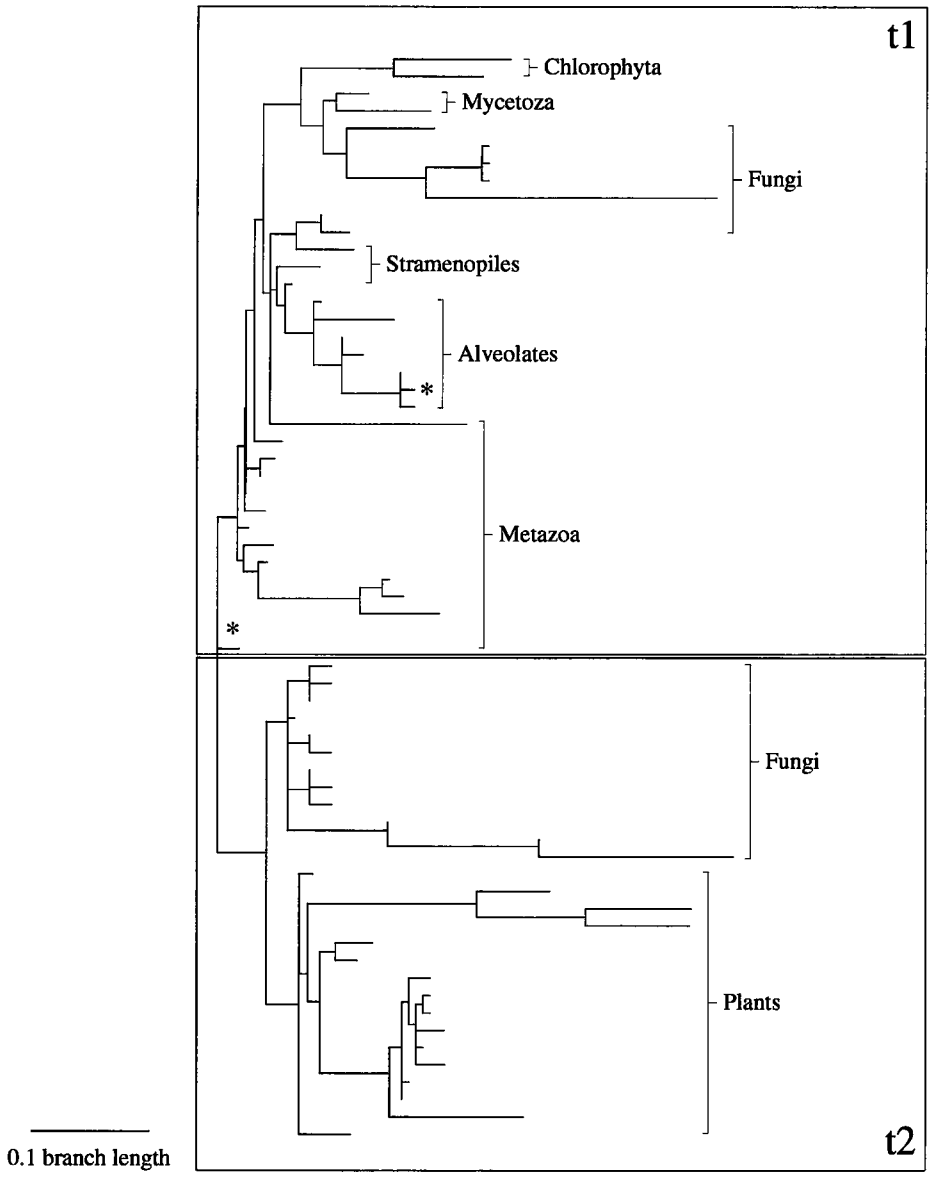

Figure 1.9: Phylogenetic tree for Calmodulin. The asterisks indicate sequences with structures which were used for this study.

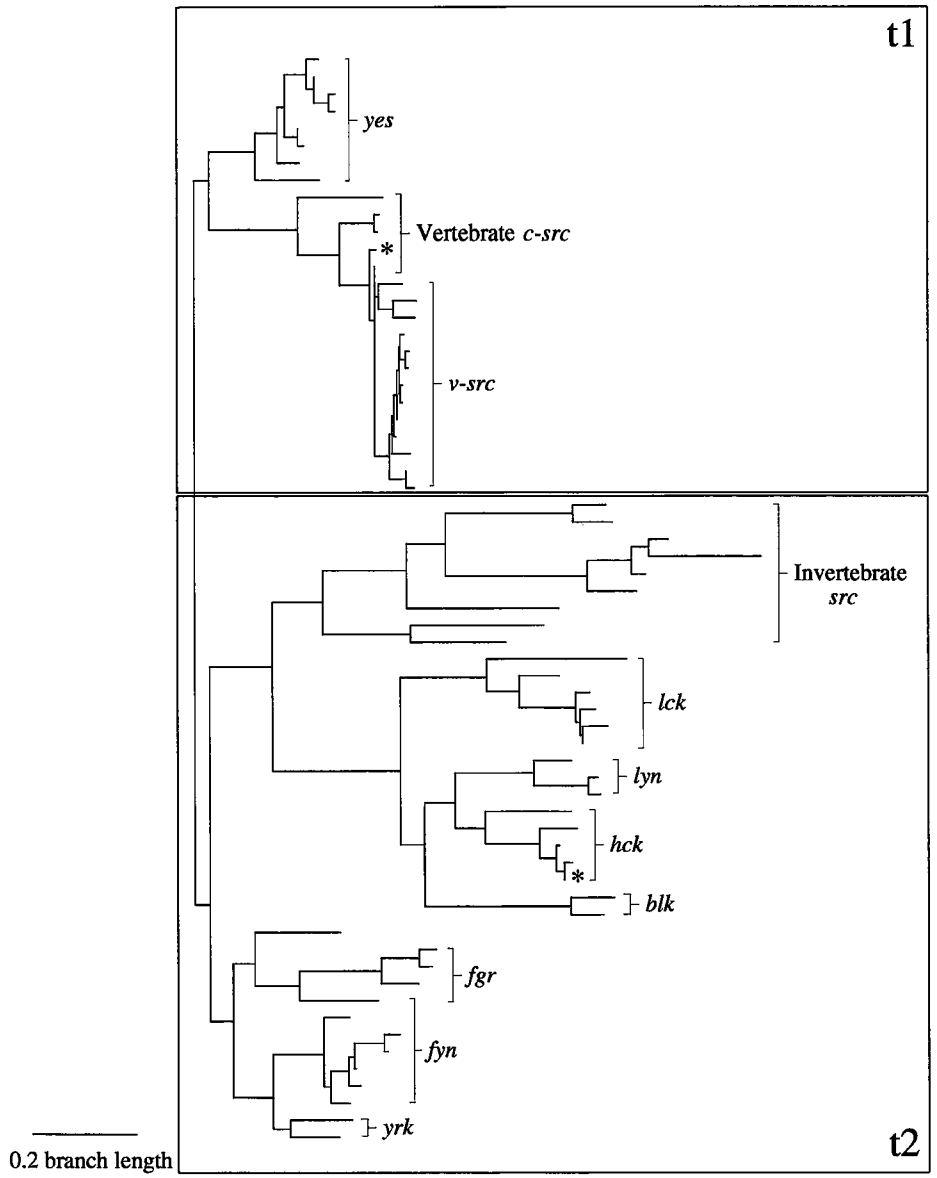

Figure 1.10: Phylogenetic tree for SRC Tyrosine Kinase. The asterisks indicate sequences with structures which were used for this study.

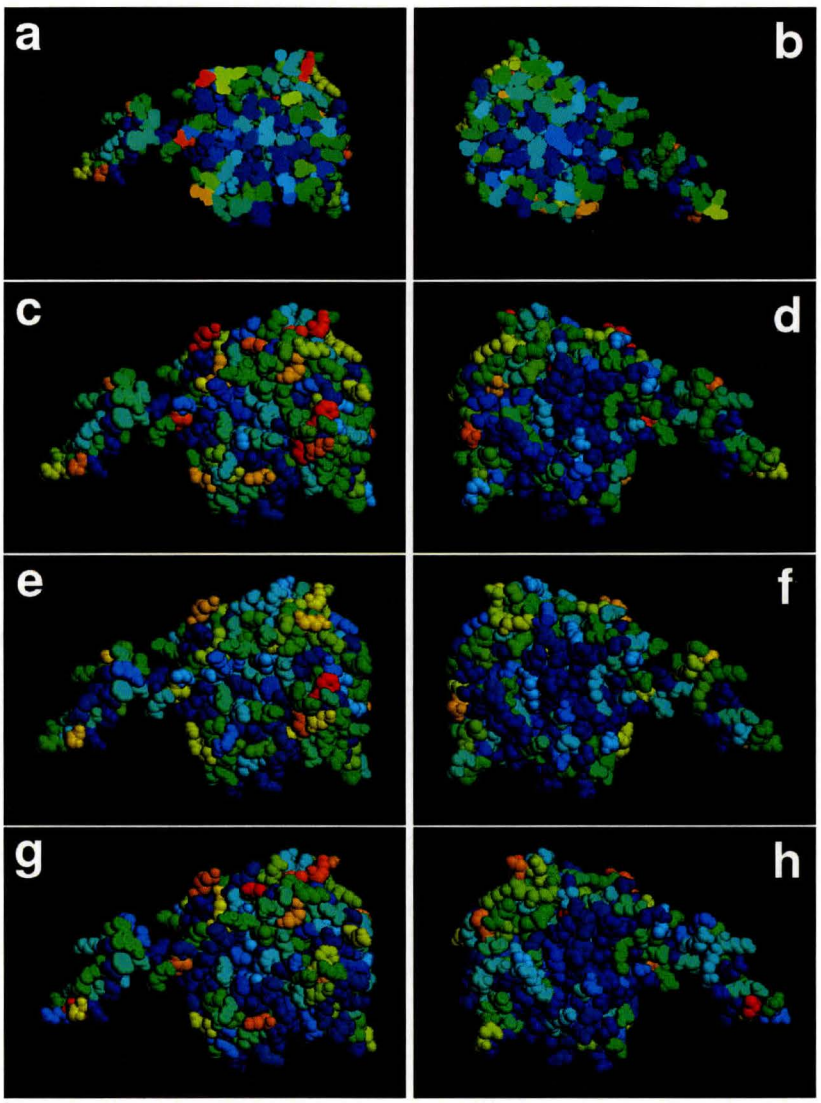

Figure 1.11: Rate-coloured structure for 5-Aminolevulinate Dehydratase monomer, *Escherichia coli* (1B4E). The left side shows the more solvent-exposed surface, the right side shows the surface that is buried in the octamer. The top row (a, b) shows a cross-section with rates from the whole tree, showing conserved interior. The second row (c, d) shows rates from the whole trees, third row (e, f) is for subtree 1, and bottom row (g, h) is for subtree 2. See figure 1.4 for phylogeny.

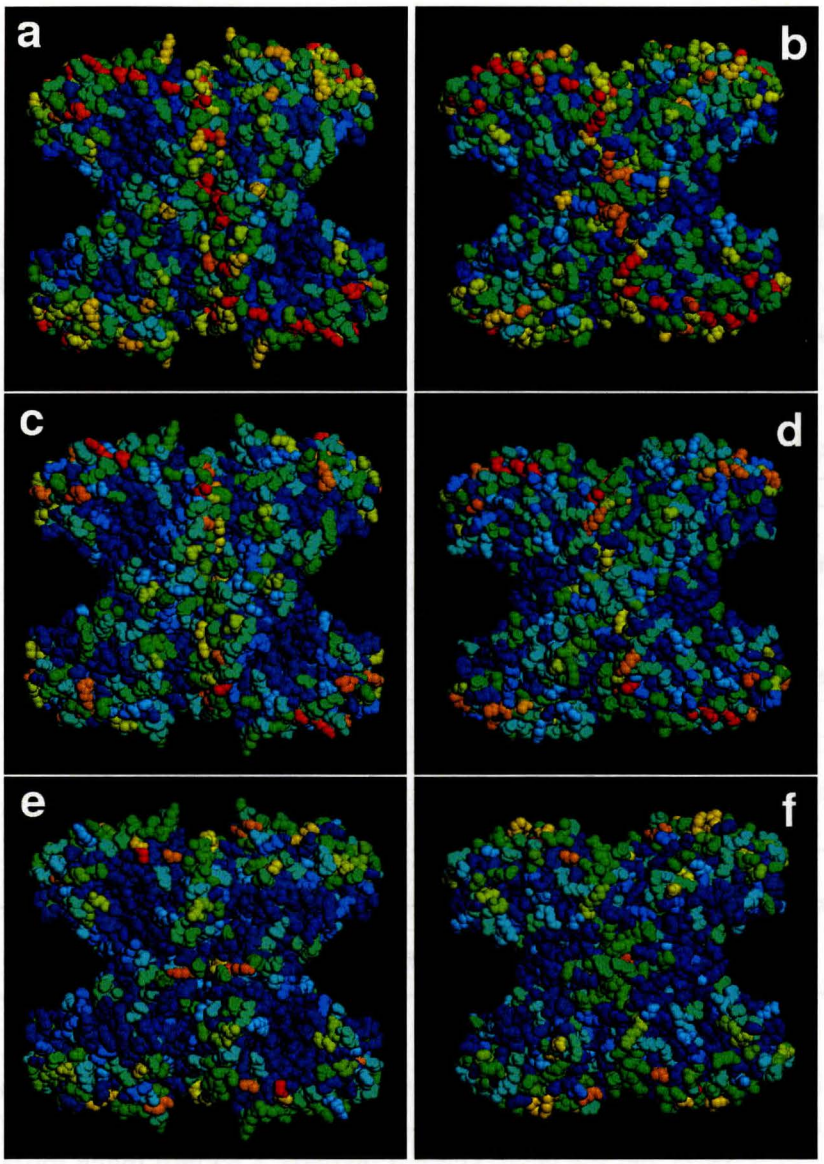

Figure 1.12: Rate-coloured structures for Fructose-1,6-bisphosphate aldolase(class I). *Plasmodium falciparum* (1A5C) structures left, *Drosophila melanogaster* (1FBA) structures right. Top row (a, b) rates are for the whole tree, second row (c, d) is for subtree 1, and third row (e, f) is for subtree 2. See figure 1.3 for phylogeny.

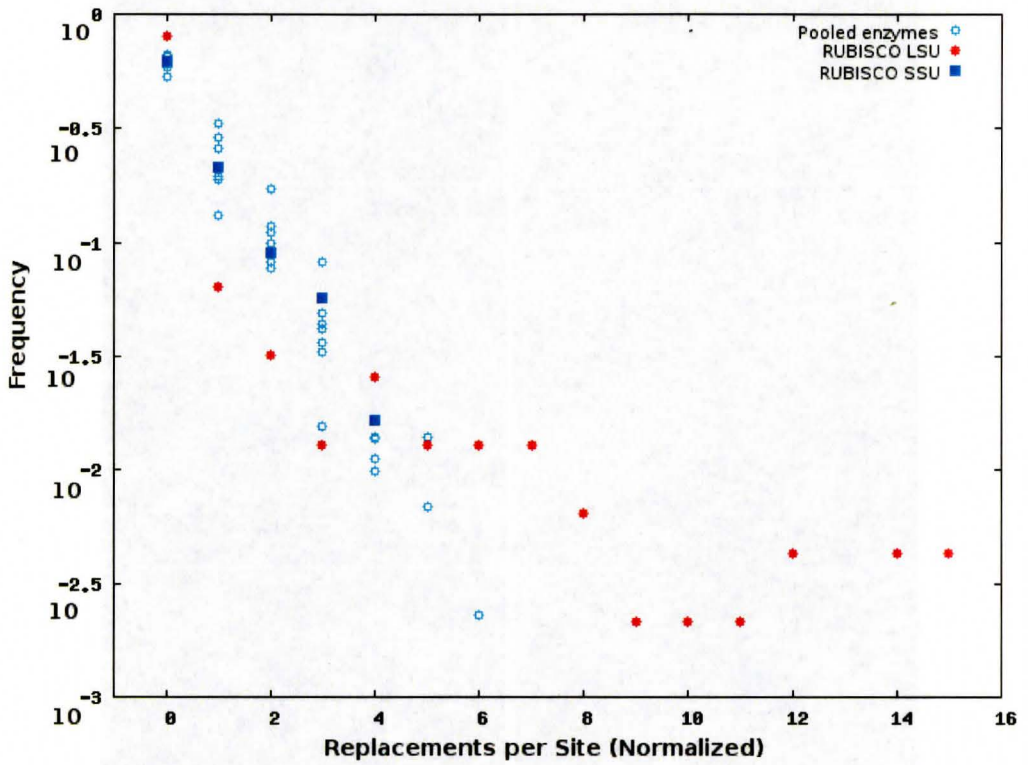

Figure 1.13: Distribution of normalized replacement rates for enzymes used in this study. The large subunit of RUBISCO displays the same atypical pattern that it did in Dean *et al.* (2002), but the small subunit displays a pattern much more like that of the other enzymes.

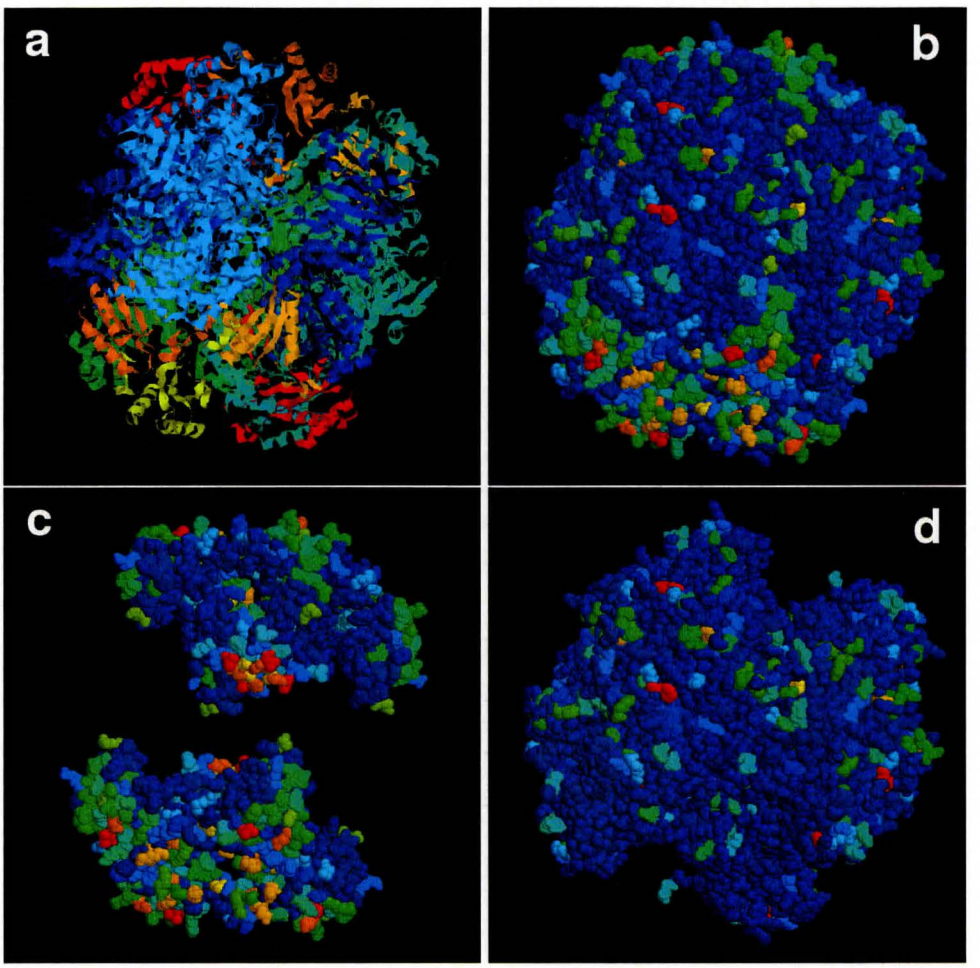

Figure 1.14: Schema and rate-coloured structures for Ribulose-1,5-bisphosphate carboxylase/oxygenase, *Chlamydomonas reinhardtii* (1IR2). (a) Shows arrangement of small (yellow, orange, red) and large subunits (green and blue). (b) Shows rate colouring for all subunits. (c) Shows rates for only the small subunit, and (d) shows rates for the large subunit. See figures 1.6 and 1.7 for phylogenies.

Chapter 2

Patterns of Amino Acid Replacement at Protein-Protein Interfaces

2.1 Abstract

Some of the variation in replacement rates that occurs across protein sites corresponds to structural factors of the protein site, particularly the solvent exposure of the site. Physically unconstrained sites tend to evolve more rapidly, but constraints can change over the evolutionary history of a protein. Such changes would be expected to affect the evolutionary rate of the site in turn. A set of four proteins that have changed their quaternary structure at some point are examined to see how the change in constraint on the new interface sites affects the replacement rate, and its relationship with other structural factors. The existing and new interfaces are also compared in order to see exactly how protein interfaces change over time. We find that the unique interfaces are as conserved as the shared ones, and they exhibit a different relationship between replacement rates and indicators of constraint than the shared interfaces or other protein sites. We also find that the unique interfaces display characteristic amino acid preferences that may identify interfaces which are still in the process of stabilizing.

2.2 Introduction

The neutral theory of evolution (Kimura 1989) states that the majority of fixed mutations have little or no effects on the fitness of the organism. While most nucleotide sites may be evolving in a more random, rather than positively selected manner, protein sequences typically have stronger functional constraints and so are often subject to negative selection at many sites. The strength of these constraints typically varies throughout a protein sequence, frequenly in a very general manner that can be correlated with physical features of the protein such as the solvent exposure of a site, its distance from an active site in the protein, or special folding constraints (Dean and Golding 2000; Dean *et al.* 2002). These constraints can aslo change throughout time, based on changes in the organism's environment or lifestyle (Golding and Dean 1998; Susko *et al.* 2002; Gaucher *et al.* 2002; Fay and Wu 2003), new functional roles as a result of gene duplication (Gu 1999; Gu 2003; Susko *et al.* 2002; Knudsen *et al.* 2003), or due to random changes in neighbouring residues that may allow or restrict new replacements (Lopez, Casane and Philippe 2002; Takahata 1987).

Dean and Golding (2000) developed a linear model that described the sitewise replacement rate heterogeneity in a protein as a product of the solvent exposure of the site, its distance from the active site, and identity as a glycine in a main chain conformation that could not be found in an amino acid with a longer sidechain. In a subsequent study Dean *et al.* (2002), half of the causal replacement rate variation in a variety of proteins was explained using this model. We expanded on this work in the previous chapter, and tried to use the model to detect changes in the replacement rate at a site throughout time by dividing a large phylogeny into two smaller phylogenies. We used a representative structure from each subtree to provide structural factors. There was weak evidence for local adpatation (demonstrated as a relatively better fit of the model when a structure native to the subtree was used), but generally the model performed better for the larger phylogeny than for either of the smaller ones. However, the points of division for the phylogenies were fairly arbitrary, and for most of the proteins there was little or no evidence of functional divergence between the two structures. In this chapter, we use the same linear model and divided phylogeny approach with sites of known functional divergence. We compare the differences in constraint that arise when a protein changes its quaternary structure at some point in its evolutionary history. In order to ensure the strongest phylogenetic signal possible, our work is restricted to proteins that form permanent associations. For the sake of simplicity, we only consider homomeric proteins.

Though proteins frequently change their quaternary structure in a transient manner, permanent changes in the number of subunits in a protein are rare. In *Escherichia coli*, only 19.4% of proteins occur as monomers, and 79% have higher-order quaternary structures (Goodsell and Olson 2000). These figures suggest that such associations are favourable, though most proteins are likely already in a stable quaternary structure. (Goodsell and Olson 2000) presented a list of compelling arguments in favour of larger proteins obtained through an increase in the number of subunits. Larger proteins have greater stability due to a greater number of stabilizing contacts (many of which are weak, and so allow greater flexibility that is often necessary for catalysis), and require less solvent. An increase in active sites through an increase in the number of subunits means that diffused substrates will have an increased chance of useful contact, are easier to regulate, and are more densely encoded in the genome. Symmetric protein contacts are additionally easier to evolve, as a single interface-stabilizing mutation will result in two new points of contact (Goodsell and Olson 2000). In cases where a new gene arises from duplication, it might be expected that the new gene would maintain the quaternary structure of the old gene as well. Cases where a protein has different quaternary structures throughout a phylogeny are likely to be experiencing different constraints. Such differences in constraint include a requirement to diffuse quickly or remain stable at low concentrations (for monomers), or a requirement for even greater stability due to life in extreme environments (Goodsell and Olson 2000). Alternately, the fixation of the quaternary structure could have occurred after some speciation or duplication event, or through drift in one lineage. There is no general rule for the form of an interface, though they can be classified into three general types that correspond to the manner in which the interface evolved (Xu, Tsai and Nussinov 1998). One type involves the initial exchange of one domain between subunits, in which case the individual subunits undergo significant conformational change upon association. A second type requires association for proper folding; it has no stable smaller subunit. The simplest case involves mutation in the surface residues of existing stable smaller units (Xu, Tsai and Nussinov 1998). This latter type is the general form that all of the larger structures used in this study take, as the smaller homologs demonstrate the existence of a stable subunit.

Though there has been little work focusing on the specifics of how new interfaces evolve, the specific composition of existing interfaces and the evolutionary constraints acting upon them has been covered in some detail. Protein-protein interfaces have evaded simple and general characterization based on single parameters such as hydrophobicity relative to other surface regions, shape, electrostatic

interactions, and flexibility, though taken together, these factors can predict interfaces with some accuracy (Jones and Thornton 1997; Liang *et al.* 2004). However, no single factor dominates for all proteins, which seem to maintain a large diversity of interfaces (Larsen, Olson and Goodsell 1998; Xu, Tsai and Nussinov 1998; Nooren and Thornton 2003). When proteins are divided into specific interface categories (based on permanence of interface, and heterogeneity or homogeneity of the involved subunits), clearer patterns in composition differences can be found, and these differences can be used to predict the same interface categories with slightly better accuracy than structural features (Ofran and Rost 2003b; Ofran and Rost 2003a). Glaser *et al.* (2001) also found differences in contact preferences between smaller and larger interfaces. Mutation of each residue in an interface to alanine provided values for the free energy contribution of various residues, identifying the few amino acids that generally contribute the most energetically to inter-protein associations (Bogan and Thorn 1998). Bogan and Thorn (1998) found a pattern to the distribution of hot spots in interfaces, and described a general interface anatomy where a core of hot spot residues were surrounded by a ring of more hydrophobic residues that exclude solvent. However, their data set was relatively small and did not necessarily represent a general structure for all interfaces.

Perhaps because surface residues are typically experiencing little constraint, they tend to have relatively rapid evolutionary rates. Previous works have found a positive correlation between replacement rates and the degree of solvent exposure of a protein site (Mizuguchi and Blundell 2000; Bustamante, Townsend and Hartl 2000; Goldman, Thorne and Jones 1998; Thompson and Goldstein 1996b; Thompson and Goldstein 1996a; Dean and Golding 2000; Dean *et al.* 2002). This relationship may be due to the greater number of interactions that buried residues must maintain with other amino acids for the structural integrity of the protein. Exposed sites would have fewer of these specific interactions, and so could be expected to experience less constraint in terms of specific amino acid identities. Residues that participate in interfaces experience greater constraint, and so tend to be more conserved. This feature has been studied and exploited in a number of ways. Evolutionary metrics detect greater relative conservation of interface residues than of other surface residues, though surface sites involved in ligand binding or active sites generally carry stronger conservation signals (Elcock and McCammon 2001; Glaser *et al.* 2003). Though the strength of constraint may not be as strong for interfaces, the entire sequence of a protein involved in some sort of complex is overall more conserved, indicating that the conservertaive effect probably extends beyond the sites involved in the interface (Teichmann 2002). Furthermore, obligate interfaces

tend to have lower replacement rates than transient interfaces (Landgraf, Xenarios and Eisenberg 2001). The constraint imposed by interface membership is also likely to be due to the requirement for specific interactions with other residues. This constraint may be of the same nature as that imposed by burial in the protein core, or it may have a somewhat different character. Most of the above studies examined the entire interface, defined as either contiguous surface patches, or residues within a certain distance from a residue on another protein chain. However, the existence of hot spot residues revealed by alanine scanning mutations (Bogan and Thorn 1998) suggests that some residues play more important roles in the interfaces than others. Hu *et al.* (2000) set out to identify these residues by finding structurally conserved residues in the interface. With smaller data sets, these residues corresponded fairly well to experimentally-validated hotspots, but some residues that were conserved did not make large energetic contributions. In later works (Ma *et al.* 2003; Halperin, Wolfson and Nussinov 2004), the correlation between hot spots and structurally conserved residues improved. It was also found that both hot spots and structurally conserved residues had greater packing density than other interface sites, and the conserved residues that were not hot spots likely facilitated this closer packing around energetically important residues (Halperin, Wolfson and Nussinov 2004). This latter study used the largest data set, but does not contradict the picture put forth by Bogan and Thorn (1998). It presented a picture of interfaces as being held together by a few critical residues which limit flexibility in the remainder of the enzyme. These important contacts were proposed to be restricted to a maximal density in the interface so as to allow the enzyme enough conformational freedom to carry out its catalytic function.

There are some general rules of interface stabilization that emerge from these studies, though many nuances and idiosyncracies remain in individual interfaces. These differences may exist because only a few sites actually play important roles, or because there are many different ways to form an interface. Another possibility is that interfaces change their character over time. We have collected a set of homologous protein pairs with different quaternary structures. Typically, one interface is conserved between the pair, and another interface is unique to one member. In these cases, it is very likely that the unique interface is newer, and has not had as long to stabilize as the shared interface has. Our dataset allows us to contrast the differences between shared and unique interfaces. This offers a snapshot of an evolving interface, which may shed some light on the kind of changes or constraints that are at work in an early interface. We use the linear model and subtree approach that was used in the previous chapter to examine differences in replacement rates,

structural features, and amino acid composition that occur between related structures. We find that differences in rates can be detected across subtrees between the sites that are under different constraints, and these differences in constraint manifest as changes in the relationship between the replacement rates and structural factors. We also find that the two different types of interfaces are both conserved, but they display differences in character and amino acid preferences.

2.2.1 Enzymes Studied

We found four pairs of enzymes that were suitable for study. Relevant details of these enzyme pairs are given in Table 2.1. The enzymes collected for this study were identified by Enzyme Commission number (a categorization based on catalytic function), which serves as an *a priori* assurance of functional similarity. Homology was assured by PRSS (Probability of Random Shuffle) scores (Pearson 2000). The P-values for the PRSS scores are generally low, ranging from 0.00078 and 0.0024 for the most distantly-related Purine nucleoside phosphorylase enzyme pair, to 4.784e-25 and 4.863e-27 for the Alcohol dehydrogenase pair. For three of the four enzymes, one member of the pair is a thermophile. While the additional stability requirements of high-temperature environments would clearly favour greater-numbered quaternary structures, it is possible that the evolutionary and compositional features of these proteins are biased by the thermostability requirements as well. However, the enzyme pairs in which one member is thermophilic have greater sequence similarity, suggesting that such drastic selective pressure is required for proteins to change their quaternary structure over a relatively short evolutionary period. The Root Mean Square (RMS) distances between α-carbons of structures provided in the table were obtained from with the Swiss-PdbViewer (Guex and Peitsch 1997).

Enzyme	EC number	PDB	Sub-units	Species	Length	% Identity	RMS (Å)	PRSS P-val	Interfaces M	Y	B	Folds (CATH topology)
Alcohol Dehydrogenase	1.1.1.1	1MP0	2	*Homo sapiens*	374	24.74	1.73	4.784e-25	6	39	35	Quinone Oxidoreductase; Rossman fold
		1H2B	4	*Aeropyrum pernix*	360			4.863e-27	39	6	35	
Triose Phosphate Isomerase	5.3.1.1	1AMK	2	*Leishmania mexicana*	252	20.23	1.58	1.135e-05	3	22	35	TIM Barrel
		1HG3	4	*Pyrococcus woesei*	226			2.654e-06	22	3	35	
Inorganic Pyro-phosphatase	3.6.1.1	1WGI	2	*Saccharomyces cerevisiae*	287	17.48	1.38	4.793e-10	3	40	14	Inorganic Pyro-phosphatase
		1QEZ	6	*Sulfolobus acidocaldarius*	174			1.357e-10	40	3	14	
Purine Nucleoside Phosphorylase	2.4.2.1	1V2H	3	*Homo sapiens*	254	13.89	1.58	0.00078	10	38	41	Rossmann fold
		1ECP	6	*Escherichia coli*	289			0.0024	38	10	41	

Table 2.1: Enzymes used in this study.

Alcohol Dehydrogenase

Alcohol dehydrogenase is an oxidoreductase with broad distribution and variable substrate specificity and cofactor requirements. There are three classes of ADHs. Our tree (Fig. 2.1) includes only type I ADHs, ranging through bacteria, archaea, and eukaryotes. Our structures include a tetrameric ADH from *Aeropyrum pernix* (4ADH, a hyperthermophile:Guy, Isupov and Littlechild 2003), and a dimeric glutathione-dependent formaldehyde dehydrogenase (2ADH) from *Homo sapiens* (Sanghani *et al.* 2003). Both are Zn^{2+}-dependent, and have a 3-Layer ($\alpha\beta\alpha$) sandwich, nucleotide-binding Rossman fold. They have similar tertiary structures (RMS distance of 1.58 Å) and similar sequences (24.74% sequence identity). The tetramer is composed of a dimer of the dimeric form.

The dimeric Class III ADH (2ADH) is widely expressed in animal tissues. It binds NAD(H), and oxidizes a variety of substrates, preferrentially long-chain carboxylic acids. It plays an important role in formaldehyde detoxification (Sanghani *et al.* 2003). The tetrameric Class I ADH also binds an NAD(H) cofactor, is inhibited by octanoic acid, and prefers cyclic structures (Guy, Isupov and Littlechild 2003). Though increased quaternary structure is a common adaptation to high-temperature environment, the Class I ADH is also tetrameric in many non-thermophilic bacterial species. Upon binding the cofactor, the Class I ADH undergoes a conformation change, which does not occur in the Class III dimer (Sanghani *et al.* 2003; Guy, Isupov and Littlechild 2003). The *Aeropyrum pernix* tetramer shares 70% sequence identity with two other non-thermophilic species. The difference in thermostability of the various tetramers is due to the enhanced stability of the subunit interfaces (Guy, Isupov and Littlechild 2003). The two ADH structures are the most closely related of all enzymes used in this study, though there are some clear differences in structure and functional constraints.

Triose Phosphate Isomerase

Triose phopsphate isomerase (TIM) is a glycolytic α/β-barrel enzyme with wide distribution. TIM catalyzes the interconversion of dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate with a diffusion-limited rate. Our phylogeny includes species from archaea, bacteria, and eukaryotes (Fig. 2.2). The enzyme is a dimer in most organisms (Williams *et al.* 1999), but a few thermophilic archaea have tetrameric forms. Our structures include a dimer from a eukaryote, *Leishmania*

mexicana (2TIM, (Williams *et al.* 1999)) and a tetramer from an archaea, *Pyrococcus woesei* (4TIM:Walden *et al.* 2001). These two forms have the same general tertiary structure (RMS distance 1.58 Å) and a sequence identity of 20.32%. The tetramer is a dimer of the dimeric form.

The *Pyrococcus woesei* tetramer has experienced pruning of several helix and loop regions, making it more compact relative to other TIMs. These pruned regions were fairly flexible in dimeric structures, so the pruning has led to enhanced stability of the protein. It may be expected that the tetrameric TIM would have many further adaptations to promote the thermostability of the enzyme, the *Leishmania mexicana* dimer experienced an increase in thermal stability from 56°C to 83°C with a single point mutation (Williams *et al.* 1999). An increase in ionic interactions is also expected for thermostable enzymes, but the tetramer interface of *Pyrococcus woesei* TIM is stabilized by mostly hydrophobic interactions (Walden *et al.* 2001). Though there are structural differences between the two TIMs, these differences are not entirely along the expected mesophilic versus thermophilic axes.

Inorganic Pyrophosphatase

Inorganic pyrophosphatase (IPPase) is an enzyme that catalyzes the irreversible hydrolysis of the phosphoanhydride bond in inorganic pyrophosphate. Since build-up of inorganic pyrophosphate can be toxic, this function is essential for the continuation of processes that use nucleotide triphosphates. The enzyme is generally dimeric in eukaryotes, and hexameric in bacteria and archaea. Our phylogeny for this enzyme encompasses this range (Fig. 2.3). We use a dimer from *Saccharomyces cerevisiae* (2IPPase:Heikinheimo *et al.* 1996) and a hexamer from *Sulfolobus acidocaldarius* (6IPPase, a thermophile:Leppanen *et al.* 1999). The hexamer is formed from a trimer of the dimeric unit.

Both the active site and the catalytic mechanism are well-conserved across the dimeric and hexameric forms. Both structures contain the Inorganic pyrophosphatase fold, which consists of a twisted five-stranded barrel. The eukaryotic IPPases are generally longer due to extentions on both terminii. This is the case with our two structures as well: the fungal dimer is 287 amino acids in length, and the archaeal hexamer is only 174 residues long. Some of these extra residues make up the dimer interface, which is smaller in the hexameric protein. The truncation of the hexameric form also means that some of the active site residues participate in the dimer interface, but they do not in the eukaryotic dimer. The hexameric forms

generally have a more flexible active site, but that is not the case for the *Sulfolobus* hexamer. Both forms require divalent cations, but the dimer binds two Mn^{2+} per subunit, and the hexamer binds a single Mg^{2+} per subunit. In the hexamer, this ion may be lost at low temperatures, causing the enzyme to deactivate (Leppanen *et al.* 1999).

Purine Nucleoside Phosphorylase

Purine nucleoside phosphorylase (PNP) is part of the purine salvage pathway. It converts purine ribonucleosides into the free base and ribose-1-phosphate. The enzyme can either be trimeric or hexameric. The two forms are encoded by two different genes, and have different substrate specificities and response to inhibitors (Mao *et al.* 1997). We use a trimeric form from *Bos taurus* (3PNP:de Azevedo WF *et al.* 2003), and a hexameric form from *Escherichia coli* (6PNP:Mao *et al.* 1997). Representatives of the hexameric form are found in bacteria, archaea, and some eukaryotes, though these are not featured in our tree (Fig. 2.4). The trimeric form is found over the same phylogenetic range. The amino acid similarity is not high between the two forms (13.89% identity), but the PRSS score between our two structures is significant at the 99.9% level (P(score) = 0.00078). Similarly, the overall topology of the monomer and the active site location are shared between the two forms, but the actual residues in the active site are very different (Mao *et al.* 1997). The subunit interfaces are also quite different. Though some of the same positions are involved, the trimer has a disc-like, cyclical arrangement of subunits, and the hexamer is a disc formed from a trimer of dimers (Fig. 2.8a and 2.8b). In the hexamer, the active site is located at the dimer interface, but the trimer interface has shifted slightly so that this is not the case.

There are also some functional differences between the two forms, despite the structural similarity of the monomers. The hexameric active site is larger and more accessible than the trimer. It will also accept a greater variety of substrates. Both forms take a (2′-deoxy)purine ribonucleoside. The hexamer will take both adenine and guanine/ hypoxanthine, but the trimeric form will only take guanine or hypoxanthine, and the hexamer will also accept substrate with modified ribose groups. The greater substrate specificity protects the trimer from inhibitors that affect the hexamer, such as formycin A (Mao *et al.* 1997).

2.3 Methods

Candidate proteins were initially found by searching the entire PDB database for proteins with the same Enzyme Commission (EC) number but differing numbers of protein subunits. Differences were first detected by the number of chains in a PDB file, or by the quaternary structure as provided in the BIOMOLECULE entry. Each of these candidate proteins were then manually inspected. In order to survive the next round of screening, proteins required confirmation of different quaternary structures through either available literature, or supporting Assumed Biological Molecule coordinates. Heteromeric proteins were rejected. Where the option was available, structures binding an inhibitor were chosen in order to more closely capture the active conformation of the protein. The protein sequences were then tested for homology using PRSS (Pearson 2000). A P-value of less than 0.10 (90% confidence level) was required for the assumption of homology.

Each structure surviving these requirements were BLAST'ed for similar sequences, using a maximum E-value of 10^{-30}. All of the unique sequences obtained for each protein were grouped and aligned using ClustalW (Chenna *et al.* 2003). Puzzle (v 5.0) (Schmidt *et al.* 2002) was used on a neighbor-joining tree to estimate the Γ distribution parameter α (exact, 8 Γ categories + 1 invariant, WAG 2000 model). The alignment was then bootstrapped (100 replicates), and distance matrices for each replicate were obtained with Protdist (Felsenstein 1989) (JTT model, Γ + invariant). Neighbor-joining trees were obtained for each replicate. These trees were then evaulated by proml (phylip 3.6b (Felsenstein 1989), 6 categories). The tree with the highest likelihood was chosen for further analysis. Any branch lengths longer than 1.5 were pruned to avoid excessive multiple hits due to long branches, with an exception allowed for the branch that joined the two subtrees. The trees were sub-divided on long branches on the assumption that all sequences in the subtrees had the same quaternary structure as the structure sequence within the cluster. This assumption was tested whenever possible by checking the quaternary structure of other structures found in each subtree. The process of alignment and bootstrapping was repeated on the subtrees, and these trees were used in a likelihood method (Fitch 1971) to estimate the number of replacements at each site in the protein. The likelihood counts were normalized on a scale from 0-999 and used in the temperature column of the PDB file for visual inspection of replacement rates. The rate estimates were also used for a variety of statistical anaylses.

2.3.1 Identifying Residues Involved in the Protein-Protein Interface

Residues were assumed to be involved in a protein-protein interface if they were within 5.2Å of any atom on a different protein chain. The distance of 5.2Å represents the length of two hydrogen bonds to the oxygen atom of a water molecule. Distances were calculated from the PDB files for each protein structure.

The sequences of all the available structures for a protein were aligned with ClustalW, and the sites involved in interfaces were compared across the different quaternary structures. Sites in the homologs were classifed accoring to participation in interfaces. The sites were classified as: Not participating in an interface (NINT); Participating in an interface both in its own structure and aligned with a position in the homolog which participates in an interface (BINT); Participating in an interface in its own structure, but not aligned to a residue in an interface in the homolog (MINT); Not participating in an interface in its own structure, but aligned to a residue which particpates in an interface in the homolog (YINT).

2.3.2 Statistical Analyses

The rate estimates from the subtree of each homolog were used in a series of ANOVAs to explore the relationships between replacement rates and interface categories. Other structural and functional constraints and various interactions were also considered for comparison. The degree of solvent exposure at a site is strongly correlated with the replacement rate. Any effect that the interface membership may have on replacement rates may not be independent of these effects. The Kyte-Doolittle hydropathy of each residue is similar to the solvent exposure in these respects. The solvent exposure was determined with DSSP (Kabsch and Sander 1983), and normalized by the value for the fully-extended Gly-X-Gly tripeptide (Shrake and Rupley 1973). ANOVAs were also used to explore how the percentage of each amino acid at an alignment position is affected by interface membership. We also took the average values of the normalized replacement rate, composition of the same amino acid at the corresponding site, and solvent exposure at each site, weighted by the proportion of the specific amino acid at the site for direct comparison between interface categories.

2.4 Results

2.4.1 Phylogenetic Trees

The quality of the underlying phylogenies (Figures 2.1-2.4, Table 2.2) will affect the reliability of the rate estimates which underlie subsequent analyses. Dean *et al.* (2002) listed a number of criteria which phylogenies must meet to be suitable for analysis. These are as follows: 1) A phylogeny with mean of at least 1.5 replacements per site; 2) The tree includes at least five sequences; 3) All sequences are less than 99% identical; 4) Each sequence shares at least 40% identity with a sequence of known structure; 5) No branch is longer than 0.3 (the mean number of replacements per site); 6) No more than 30% of the branches are longer than 0.2. For this study, we also required that the quaternary structure be consistent within each subtree. The first three criteria ensure that the data set is large enough to be robust, and the fourth ensures that sequences are similar enough to avoid large differences in protein structure. The last two minimize the number of multiple replacements at each protein site, which would lead to underestimating the replacement rate.

The phenomenon we investigate in this study is somewhat unusual, and very few enzymes were initial candidates. It was necessary to increase the long branch cutoff (criteria 5) to 1.5. All trees meet criteria 1, 2, 3, and 4 by construction. Additionally, initially choosing the enzymes by EC number further ensures similar function between the two proteins. Most trees meet criteria 6. The 2IPPase tree marginally exceeds 30% of branches over 0.2.

The subtrees were chosen on the basis of a long branch between them (to ensure consistent quaternary structure of subtrees), so the distribution of taxa among them could not be changed. As a result, some trees have a very unbalanced number of taxa in the subtrees and total tree length (particularly TIM). There are also large differences in the branch length per species (approximately, the speed of the molecular clock) for TIM and ADH. The branch length per species is greater for the larger protein in TIM and ADH. The mean number of replacements also varies, particularly in TIM, PNP, and IPPase, but the mean roughly follows the overall size of the tree (total branch length). However, the mean is large enough to ensure a robust data set in most cases.

	Tree Length	Species	Length/ Species	Sequence length	Corrected Mean	Variance	% Branches over 0.2
2TIM	48.26	228	0.2117	252	58.09	1504.55	17.66
4TIM	9.71	20	0.4854	226	15.90	125.98	29.79
2ADH	26.82	257	0.1044	374	27.04	418.59	19.03
4ADH	43.75	177	0.2472	360	20.40	225.40	5.09
2IPP	8.21	28	0.2932	287	4.19	8.29	32.08
6IPP	21.95	74	0.2966	174	24.16	300.12	19.29
3PNP	20.74	99	0.2095	254	46.81	647.47	15.90
6PNP	16.74	74	0.2263	289	5.93	17.95	19.31

Table 2.2: Features of the phylogenetic trees for the enzymes used in this study.

2.4.2 Rate Colourings

The replacement rates for each subtree were normalized to a scale of 0-999 and used in the temperature column of the PDB file to visually explore how the rates vary by position in the three-dimensional structure of the protein. The residues involved in the interfaces for each protein are shown below.

TIM is one of the three enzymes in this study that has a fairly simple change in quaternary structure. The tetramer is a dimer of the dimeric form (Figures 2.5a and 2.5b), and only three residues that are involved in the dimer interface are exclusive to the interface of the dimeric form. The dimer interface (comprised of 35 residues) is more extensive than the tetrameric interface (22 residues). The residues involved in the dimer interface appear fairly structurally conserved among both forms, and show a similar distribution of rates (Figures 2.5c and 2.5d). However, the residues of the tetramer interface show some rearrangement between the two forms, and a different rate distribution (Figures 2.5e and 2.5f). The tetrameric form is from a hyperthermophile, which may explain the apparent compression of the residues in the tetramer interface for the tetrameric form. These residues are slightly more conserved in 4TIM, which may indicate constraint from involvement in an interface. Additionally, the residues involved in the tetrameric interface are on the surface of the dimeric form, which could be expected to reduce the constraint on these sites even further.

ADH is similar to TIM in that the larger form is a hyperthermophilic, and a

simple dimer of dimers (Figures 2.6a and 2.6b). Only six residues are involved in the dimeric interface of 2ADH that are not used in 4ADH. The dimer and tetramer interfaces are of similar sizes, with 35 residues in the dimer interface common to both forms, and 39 residues in the tetrameric interface. The dimeric interface shows a bit of tertiary structure change between the two forms, and a very slight decrease in overall conservation for 4ADH (Figures 2.6c and 2.6d). The tetrameric interface is more structurally condensed for 4ADH (as it is for 4TIM), but it shows no obvious differences in conservation compared to 2ADH (Figures 2.6e and 2.6f). These residues are exposed in 2ADH, so it is likely that there is some other difference in constraint for the more conserved residues of the 4TIM tetramer interface. The difference may be due to the more extensive subtree of 4ADH relative to that of 4TIM. The subtree for 4TIM is restricted to archaea, whereas the subtree for 4ADH covers a more extensive phylogenetic range, comparable to that of 2ADH. It is possible that there is an accompanying broader set of constraints acting on the tetramer interface residues of 4ADH.

IPPase is another enzyme for which the two structures are fairly close phylogenetically and where the larger homolog is built up from the smaller. 6IPPase is made of a trimer of 2IPPase. The dimer interface is much smaller than the hexamer interface, at only 14 residues common between forms and an extra three that are unique to 2IPPase. The hexameric interface uses 40 residues, some of which are relatively buried in 2IPPase (Figures 2.7a and 2.7b). Both sets of interface residues are generally more conserved in 2IPPase, though there is a pattern in the hexamer interface residues that is consistent with the solvent exposure of those residues (the more conserved ones are also buried). For 6IPPase, the dimeric interface is marginally more conserved than the hexameric interface, and the hexamer residues may have experienced more structural change than the dimeric interface residues (Figures 2.7c, 2.7d, 2.7e and 2.7f). The overall greater rates in 6IPPase may also be due to the broader phylogenetic range that the hexamer subtree spans (Fig. 2.3).

PNP has a more complicated structural relationship between the two forms. Though there is some overlap in the specific residues used (41 are common to both forms) they are used differently, and the trimeric form has more unique residues involved (39 versus 10 unique residues for the hexamer). 3PNP has a single type of asymmetric interface (Fig. 2.8a), whereas 6PNP has two different symmetric interfaces, one which forms a dimer, and one that joins three dimers radially to form a hexamer (Fig. 2.8b). Figure 2.8c-f shows the relative positions of the residues involved in each interface on the monomers. The monomers are positioned with

the rotational axis of the protein pointing down. The trimer interface residues are shifted to one side of the monomer relative to the hexameric interface. (Figures 2.8g and 2.8f) show the monomers rotated 90 degrees towards the viewer, with the rotational axis of the protein facing forwards. There is a clear difference in the rate distribution between the two forms. The trimer interface (Fig. 2.8g) is generally conserved, with the residues that are unique to the hexamer interface evolving more quickly. The hexamer interface has one interface that is clearly conserved, and one that is evolving much more quickly (Fig. 2.8h). 6IPPase and 3IPPase cover similar phylogenetic ranges (Fig. 2.4), but the active site of 6IPPase is located at the conserved dimeric interface. The active site of 3IPPase is not located at an interface. This difference likely explains why there is a discrepancy in rates between the two interfaces of 6IPPase, but not why its trimeric interface is evolving so much more quickly.

2.4.3 Statistical Analyses

To test whether interface membership had an effect on the evolutionary rates at a protein site, the replacement rate was modelled as a product of interface membership and other factors already known to affect rates. We explored the relationships between replacement rates, interface categories, and other structural factors (solvent exposure and hydropathy) in a series of ANOVAs and correlations. In the most general anaylsis, the rates are modelled as a product of these factors and various interactions of the factors. We also treated residues separately by interface class to see if any of the factors had different effects in interfaces. Since specific amino acids may play more or less important roles in interfaces, we also modelled the percentage of each amino acid at each position in the alignment as a function of interface membership, rates, structural factors, and corresponding values at the aligned site in the other structure. Finally, we compared the averages of the replacement rates, conservation of the specific amino acid across structures, and the solvent exposure for each amino acid, weighted by the composition of that amino acid at the alignment position.

Initially, the mean replacement rates and solvent exposures were compared for each interface category to see if general differences existed between the sites in each group. The means for each category are given in Table 2.3. The replacement rates were normalized for each protein such that the average site had ten replacements, and the solvent exposures were normalized by the maximal exposure of that amino

	mean % exposure	mean rate
NINT	20.86	9.93
YINT	24.90	11.87
BINT	13.31	9.50
MINT	18.61	9.47

Table 2.3: Mean solvent exposure and normalized replacement rate for the four interface categories.

acid in a chain flanked by two Gly residues. MINT and BINT sites are generally on the surface of the protein, but should be relatively shielded from solvent by contact with other interface residues. YINT sites, however, are aligned with MINT sites and so should also be on a protein surface. These sites are not shielded in an interface, and so might be expected to be more exposed in general. This pattern is reflected in the average solvent exposure values. YINT sites have the greatest overall exposure, with NINT sites being slightly less exposed (but presumably with greater variation among sites). The MINT sites are more shielded than the NINT sites, but the BINT sites are even more drastically shielded. This differences between MINT and BINT sites may be due to a greater maturity of BINT sites, as the BINT sites have likely been shielded by their interface inclusion for a longer time. There is a general correlation between the degree of solvent exposure of a residue and the replacement rate at that site, so the average replacement rates might be expected to follow the same pattern. The same general ranking between categories is observed, but there is a greater difference between YINT and NINT sites than there is between NINT and MINT / BINT sites. There is virtually no difference between the average rate for MINT and BINT sites. This result might suggest that the overall degree of constraint is similar for residues participating in an interface. Conversely, the sites that are aligned to interface sites and which are not participating in an interface are relatively unconstrained, which may be due to their expected location on protein surfaces. Additionally, since MINT sites are more exposed than BINT sites but are similarly conserved, the depressed replacement rate is probably not solely due to the increased burial of the interface residues.

The replacement rate at a protein site is affected by factors such as the solvent exposure and the hydropathy of the current residue. In order to establish to what degree the rate is affected by the constraints imposed by interface membership, we

Factor	2TIM	4TIM	2ADH	4ADH	2IPP	6IPP	3PNP	6PNP	Total
acc1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	8
rate2	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.153	0.007	7
hydro1	0.051	0.277	<0.001	0.117	0.012	0.225	0.096	0.353	4
YINT	0.060	0.301	0.659	0.489	0.705	0.073	0.006	0.383	3
BINT	0.010	0.017	0.367	0.025	0.529	0.807	0.474	0.048	4
MINT	0.490	0.299	0.691	0.046	0.055	0.429	0.007	0.035	4

Table 2.4: P-values for the ANOVA of replacement rates. *acc1* is the solvent accessibility of the site, *rate2* is the replacement rate at the corresponding site in the other member of the enzyme pair, *hydro1* is the hydropathy of the residue at the site.

modelled the replacement rate with an ANOVA. We initially used a simple model, consisting of the replacement rate at the aligned site in the corresponding protein, the solvent exposure and hydropathy of the site, as well as a binary variable representing inclusion in the YINT, BINT, or MINT interface categories. The P-values for these factors are shown in Table 2.4. We used the 90% confidence level as a general cut-off for these analyses. The first ANOVA confirms that the solvent exposure has a strong effect on the replacement rate for all proteins in this study. The replacement rate at the corresponding site is the next most commonly significant term, with 7 of 8 significant cases. This suggests that overall, similar constraints are acting on each site across the two subtrees. The hydropathy was only significant for 4 of 8 cases, all of which are the smaller members of the protein pairs. For the interface categories, YINT is a significant influence on the rates for 3 enzymes, and both MINT and BINT are significant for 4 enzymes. Each category tended to be significant for different enzymes, though both BINT and MINT are significant for 4ADH and 6PNP.

The initial ANOVA was repeated with a number of interaction terms added to investigate whether the effect of the interface categories on rates is influenced by the solvent exposure or hydropathy of the sites. The P-values for each term are shown in Table 2.5. The solvent accessibility is significant for only 5 enzymes, but this is likely because some of the variation in rates that the accessibility term was accounting for previously has been shared with the interaction terms. The interactions of the interface categories with the solvent exposure accounts for three significant cases, though two of these are in enzymes for which the solvent exposure is still significant. In general, the interaction of the solvent exposure with the interface categories does not account for more of the rate variation, suggesting that

	2TIM	4TIM	2ADH	4ADH	2IPP	6IPP	3PNP	6PNP	Total
rate2	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.073	0.064	8
acc1	<0.001	0.228	0.846	0.004	<0.001	0.002	0.806	0.028	5
hydro1	0.047	0.971	0.542	0.226	0.120	0.697	0.691	0.165	1
YINT	0.328	0.999	0.557	0.367	0.247	0.371	0.244	0.904	0
BINT	0.013	0.090	0.807	0.058	0.391	0.293	0.472	0.064	4
MINT	0.204	0.735	0.397	0.082	0.968	0.160	0.933	0.056	2
Y**acc2*	0.972	0.010	0.405	0.908	0.622	0.534	0.815	0.991	1
M**acc1*	0.231	0.731	0.321	0.803	0.137	0.472	0.164	0.847	0
B**a1***a2*	0.001	0.602	0.231	0.712	0.062	0.457	0.462	0.718	2
Y**hydro2*	0.505	0.843	0.753	0.235	0.530	0.088	0.992	0.446	1
M**hydro1*	0.219	0.701	0.950	0.885	0.885	0.503	0.919	0.408	0
B**h1***h2*	0.620	0.611	0.044	0.068	0.709	0.079	0.803	0.716	3

Table 2.5: The P-values for ANOVA of factors influencing sitewise replacement rates, including interactions between terms. *acc2* and *hydro2* are the solvent exposure and hydropathy of the aligned site in the other member of the protein pair, respectively.

the effect that interface membership has on replacement rates is largely independent of the degree of burial of the residues. The hydropathy is significant for only a single enzyme with the interaction terms added, though the interaction of BINT sites with the hydropathy across both sites is significant for three enzymes (2ADH, 4ADH, 6IPPase). Two of these are enzymes for which the solitary BINT term was not significant. The BINT category is the only category which did not lose significance with the inclusion of interface categories. The YINT category is no longer significant alone, though it regains significance for 4TIM in concert with the solvent exposure at that site. The MINT term loses significance for two enzymes, and gains nothing from interaction with the accessibility or hydropathy terms.

Since the number of sites in each interface category is relatively low compared to the total number of sites, it is possible that the power of the first replacement ANOVA was low for interface-based rate differences, particularly when considering the interactions with solvent accessibility and hydropathy. In order to detect more clearly if interface membership changes the effect that solvent accessibility and hydropathy have on replacement rate sites, we divided the sites by interface category (NINT, YINT, BINT, and MINT) and performed separate ANOVAs for each. The P-values for these ANOVAs are given in (Table 2.6). We modelled the rates as a factor of the rate at the corresponding site, the solvent exposure, the hydropathy, and the interaction of the solvent accessibility and that of the corresponding site. The interface categories are all smaller data sets than the non-interface category,

	rate2	*acc1*	*a1*a2*	*hydro1*	Total
2TIM	>0.001	>0.001	>0.001	0.230	3
M	n/a	n/a	n/a	n/a	
Y	0.949	>0.001	0.563	0.012	2
B	0.009	0.003	0.048	0.108	3
4TIM	>0.001	>0.001	0.004	0.663	3
M	0.364	0.373	0.779	0.867	0
Y	n/a	n/a	n/a	n/a	
B	0.002	0.241	0.760	0.003	3
2ADH	>0.001	>0.001	0.279	>0.001	3
M	0.550	0.397	0.253	0.106	0
Y	0.037	>0.001	0.096	0.230	3
B	0.035	0.085	0.824	0.798	3
4ADH	>0.001	>0.001	0.389	0.167	2
M	0.052	0.140	0.649	0.744	1
Y	0.618	0.641	0.612	0.696	0
B	0.036	0.283	0.925	0.397	1
2IPP	>0.001	>0.001	0.003	0.055	4
M	n/a	n/a	n/a	n/a	
Y	0.604	0.097	0.868	0.719	1
B	0.963	0.177	0.494	0.396	0
6IPP	>0.001	>0.001	0.958	0.139	2
M	0.396	0.091	0.555	0.898	1
Y	n/a	n/a	n/a	n/a	
B	0.891	0.970	0.810	0.901	0
3PNP	0.010	>0.001	0.045	0.797	3
M	0.816	0.214	0.285	0.698	0
Y	0.466	>0.001	0.323	0.004	2
B	0.290	0.001	0.230	0.017	2
6PNP	0.012	>0.001	0.239	0.601	2
M	0.794	0.445	0.140	0.271	0
Y	0.849	0.460	0.711	0.282	0
B	0.984	0.226	0.926	0.502	0
NINT(8)	8	8	4	2	22
MINT(6)	1	1	0	0	2
YINT(6)	1	4	1	2	8
BINT(8)	4	3	1	2	10

Table 2.6: P-values for replacement rate ANOVA, with separate analyses for each interface category.

so the power of each analysis is reduced. However, the interface categories are all approximately the same size, so comparisons can readily be made between them. For the sites that are not involved in any interface, the corresponding rate and the solvent exposure are strong determinants of the rate. In the previous analysis of all sites, the corresponding rate was not a significant factor for PNP, though it is significant when only the non-interface sites are considered. This suggests that there is a stronger disjunction in replacement rates between the interface sites for this enzyme. While the replacement rate at the corresponding site is only significant for 4 of 8 sets of BINT sites (ADH and TIM), it is only significant for a single enzyme for both MINT and YINT sites. This result is hardly surprising, as the BINT sites share the constraint of participating in an interface, whereas the MINT and YINT sites do not share this constraint across proteins. The solvent accessibility is significant for 4 of 6 YINT cases, and 3 of 8 BINT cases, suggesting that the rates of YINT sites are still largely affected by the exposure of the residue, and BINT sites are only to a moderate degree. The solvent exposure is significant for only one MINT site, indicating that the rates at these sites are largely independent of the degree of solvent exposure. The same pattern for MINT sites carries through to the hydropathy and the interaction of the solvent accessibilities of both sites. The interaction of the solvent exposure at both sites was only significant for half of the sets of NINT sites, and for only a single enzyme for YINT and BINT sites. The hydropathy term did not display a pattern that differed for interface residues as the solvent exposure did, having two significant cases for each of the NINT, YINT, and BINT sites. Overall, this ANOVA suggests that the unique interfaces (MINT sites) are less likely to have replacement rates affected by typical indicators of constraint as compared to the common interfaces (BINT sites).

The factors that differed most strongly in explanatory power between interface and non-interface sites were the solvent accessibility of each residue, and the rate at the corresponding site. To explore this change in more detail, we performed a number of correlations, with sites grouped by interface categories.

In general, the replacement rate and solvent accessibility at a site are strongly correlated (Table 2.7). For the NINT sites, this relationship is supported. The correlation is significant at the 90% level for all sets of NINT residues, and the correlation is fairly strong ($r = 0.4348$). Fewer cases are significant for all of the interface categories. However, the decrease in number of significant cases is not noteworthy, because the cases where the YINT and MINT correlations are not significant are those where each set of residues is very small, and the power of the analysis

	MINT	YINT	BINT	NINT
2TIM	0.5326	0.6096*	0.4532*	0.4287*
	0.6425	0.0026	0.0063	>0.0001
4TIM	0.2745	-0.7148	0.2742	0.2965*
	0.2164	0.4930	0.1110	0.0001
2ADH	-0.0470	0.5673*	0.4827*	0.5265*
	0.9296	0.0002	0.0033	>0.0001
4ADH	0.4806*	0.4274	0.3092*	0.4686*
	0.0020	0.3979	0.0707	>0.0001
2IPP	-0.9970*	0.4149*	0.5501*	0.3943*
	0.0493	0.0078	0.0415	>0.0001
6IPP	0.4661*	0.2311	0.1199	0.4509*
	0.0024	0.8516	0.6831	>0.0001
3PNP	-0.3350	0.6788*	0.3568*	0.4134*
	0.3440	>0.0001	0.0238	>0.0001
6PNP	0.3751*	0.2574	0.3082*	0.4998*
	0.0203	0.4727	0.0500	>0.0001
mean of sig	0.0812	0.5677	0.4100	0.4348

Table 2.7: Correlation of replacement rate and solvent accessibility, and P-value of correlation. The correlations that are significant at the 90% level are starred.

is too low. The average strength of the correlation is lower for BINT sites (r = 0.4100). The significant correlation for 2IPPase is based on only three sites, so it is unlikely that this correlation is meaningful. With this value removed, the average correlation coefficient for the MINT sites is 0.4406, which is also similar to that of NINT sites. The average correlation coefficient for YINT sites is higher than that of the NINT sites, at 0.5677. As with the accessibility term in the previous ANOVA, YINT residues are more likely to be on the surface, and the greater overall exposure leading to a stronger relationship. There is clearly a difference in the relationship between replacement rates and solvent accessibility for the two forms of IPPase. The NINT sites give the most realistic estimate the general relationship between replacement rates and solvent accessibility in each protein. It is interesting to note that the correlation coefficient for the NINT sites often varies quite a bit between homologs, particularly TIM. The correlation for BINT sites is not significant for two enzymes (4TIM, 6IPPase), though the correlation is significant and stronger than average for the smaller partners of both of these enzymes. This suggests a possible change in constraint across structures for these two enzymes, despite the shared requirement of forming a stable interface. Both of these enzymes are from thermophiles, so the change in constraint may be due to the special requirements of thermostable interfaces. This explanation is supported for 4TIM, for which the correlation is also insignificant. 6IPPase, however, has a significant correlation for its unique interface sites, which suggests that the thermostability demands are not the cause of the disjunction between replacement rates and solvent accessibility for its BINT sites. Overall, the expected relationship between the replacement rate and the solvent exposure is supported for both types of interfaces. The relationship is even stronger for YINT sites, likely because they have a higher overall degree of solvent exposure, and replacement rates offering more data for correlation than sites where both sets of values are lower.

The replacement rate at the corresponding site was a strong determinant of replacement rates in the ANOVAs, though it was less likely to be significant for the interface sites than for the NINT sites. This effect was stronger for the MINT and YINT sites than for the BINT sites, suggesting that the constraints had changed across homologs for these sites. To investigate this in more detail, we correlated the rates across sites by interface category (Table 2.8). The pattern seen in the ANOVA is very pronounced here. The correlation is significant and fairly strong for all NINT sites (average r = 0.4033), significant for only half of the BINT cases, and only one MINT case (YINT sites were not included because they are redundant where sites are compared across homologs). Where the BINT cases are signifi-

	MINT	BINT	NINT
2TIM	0.9619	0.5599*	0.2839*
	0.1762	0.0005	0.0003
4TIM	0.1852		
	0.4092		
2ADH	-0.2202	0.4598*	0.5460*
	0.6750	0.0054	>0.0001
4ADH	0.3707*		
	0.0202		
2IPP	-0.7410	0.1322	0.5157*
	0.4687	0.6524	>0.0001
6IPP	0.1881		
	0.2452		
3PNP	0.0761	-0.0846	0.2674*
	0.8345	0.6037	0.0010
6PNP	0.2496		
	0.1307		

Table 2.8: P-values for correlation of rate with rate at corresponding site.

	MINT	YINT	BINT	NINT
2TIM	-0.9449	-0.3742*	-0.1738	-0.4871*
	0.2123	0.0863	0.3180	>0.0001
4TIM	-0.6565*	-0.9921*	-0.2496	-0.6033*
	0.0009	0.0803	0.1482	>0.0001
2ADH	-0.4251	-0.2555	-0.5053*	-0.4838*
	0.4007	0.1164	0.0020	>0.0001
4ADH	-0.2674*	-0.6065	-0.4813*	-0.5513*
	0.0998	0.2018	0.0034	>0.0001
2IPP	0.4046	-0.3943*	-0.1761	-0.4813*
	0.7348	0.0118	0.5470	>0.0001
6IPP	-0.4807*	0.0778	-0.6582*	-0.4597*
	0.0017	0.9504	0.0105	>0.0001
3PNP	-0.3910	-0.4614*	-0.4438*	-0.4924*
	0.2639	0.0031	0.0041	>0.0001
6PNP	-0.2550	-0.0282	-0.3189*	-0.5485*
	0.1223	0.9384	0.0421	>0.0001
mean of sig	-0.4682	-0.5555	-0.4815	-0.5134

Table 2.9: P-values for correlation of hydropathy and solvent accessibility at each site.

cantly correlated, the value is fairly strong (TIM, r = 0.5599, and ADH, r = 0.4598). These two enzymes also have the most similar shared interface. For the single case where the rates between MINT sites are correlated, the strength of the correlation is low relative to those of NINT and BINT sites (r = 0.3707). The previous correlation between replacement rates and solvent accessibility suggested that there were differences in constraint between the BINT sites of TIM and IPPase. There was more evidence for a thermostability-based difference for TIM than for IPPase. The correlation between rates at BINT sites is significant for TIM, but not for IPPase, which further suggests some other difference in constraint between the shared interface sites of this enzyme.

The ANOVA results for the hydropathy were ambiguous. It had an equal (but low) propensity to be a significant cause of variation in replacement rates for NINT, YINT, and BINT sites. It also followed different patterns than the solvent accessibility term. It might be expected that the hydropathy and solvent accessibility have similar effects on the replacement rates, as both terms generally describe the interactions of the residue with water. We correlated the solvent accessibility and the hydropathy of each residue, divided by interface categories to explore patterns between these two factors more closely (Table 2.9). As with the other factors, the correlations are always significant for the NINT sites. The correlation is less likely to be significant for the BINT sites (5 of 8), the YINT sites (4 of 8), and even less so for the MINT sites (3 of 8). Again, four of the MINT and YINT sets are too small to have sufficient power for a significant correlation to be likely. Among the significant correlations, the average values follow a pattern similar to that of the correlations between replacement rate and solvent accessibility. All of the correlations are negative here, as the hydrophilic values of the hydropathy scale are negative. The YINT sites have the strongest average correlation (r = -0.5555), followed by NINT sites (r = -0.5134), then BINT and MINT sites (-0.4815 and -0.4682 respectively). This suggests that MINT and BINT sites are more likely to violate the simple assumption that more hydrophilic residues will be more exposed to solvent, as would occur if charged residues are used to form salt bridges at buried interfaces between subunits. IPPase has an atypical pattern for its BINT sites, in that there is a significant correlation between hydropathy and solvent accessibility for the hexamer, but not for the dimer. This effect is more pronounced considering that the YINT sites of 2IPPase and the aligned MINT sites of 6IPPase have similar and significant correlation coefficients. This suggests that the two forms of IPPase make different use of the residues that are involved in a common interface, to a greater degree than the other enzymes do.

The varying relationship between the hydropathy and solvent exposure of residues suggests that there may be some constraints on interface residues that are more specific than general hydrophobic or hydrophilic effects. To further investigate the specific useage patterns of the various amino acids in common and unique interfaces, we carried out another ANOVA on the composition of each amino acid at alignment positions for each subtree. The ANOVA explored which structural factors and interface categories affected the composition of each amino acid at protein sites. For the factors that were generally relevant, weighted averages were obtained by each amino acid and interface category to examine category-based differences.

In the first ANOVA, the percentage of each amino acid at a site was modelled as a product of the replacement rate, solvent exposure, percentage of that amino acid at the corresponding site, and interface membership. Table 2.10 shows the number of enzymes (of 8) which had a P-value less than 0.10 for each factor. The factors all varied in their explanatory power by specific amino acid type more than they did by factor. This is not surprising, as some of the rarer amino acids would not have enough data to show clear trends. However, even among the more abundant amino acids there is variation in the relevance of the various factors. A factor was considered generally relevant for an amino acid if it was significant for at least 3 of 8 cases. The factor that is most frequently significant is the amino acid at the corresponding site, followed by the solvent exposure and the replacement rate at the site. The proportion of the corresponding amino acid is more generally significant for Val, Phe, Met, Gly, and Lys, while it is barely significant for Trp and Tyr. These amino acids are relatively scarce (Table 2.11), though Met, His, and Gln are less abundant than Tyr but are more commonly affected by the accompanying composition. The solvent exposure follows a reasonably predictable pattern of influence that generally follows the hydropathy of the amino acids. The amino acids at either end of the scale are generally more affected by the solvent exposure. Arg is a notable exception to this pattern. The replacement rate affected only a few amino acids strongly; Gly, Ile, and Gln. Gly tends to occupy the most slowly-evolving sites, and Gln is generally found at the fastest ones (Table 2.14). In comparison, the interface membership factors are generally much weaker sources of variation in amino acid composition. The interface category which is generally the most relevant is the BINT sites. However, there are slightly more BINT sites than YINT or MINT sites, so this increase may simply represent an increase in the power of the ANOVA. The amino acids that are generally most affected by interface status are Met, Arg, Asn, Lys, and Val. However, Val is only generally affected by BINT membership, and this represents a decrease in the composition (Table 2.11). While most amino acid compositions are much more generally influenced by the non-interface factors, a few are comparably or more influenced by the interface factors. These include Trp, Tyr, Arg, Met, Asn, and His.

The ANOVA indicated that the effects of the interface status of residues did not play a very strong role in specific amino acid composition. However, the number of sites included in an interface are generally low relative to the entire protein, and certain interface categories did have a minor effect on the proportion of certain amino acids. The compositions of each amino acid by interface category are given in Table 2.11 (Fig. 2.9). The ANOVA found the compositions of Met, Arg, Asn,

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R	160
rate1	5	4	3	0	3	2	1	7	2	3	1	0	3	2	1	5	2	2	4	2	52
acc1	8	8	7	3	4	2	5	6	0	3	1	0	3	1	7	4	8	3	8	1	82
AA2	6	8	5	8	4	8	7	8	6	4	0	2	6	6	6	4	6	6	8	4	112
MINT	1	1	0	0	0	2	0	2	0	0	2	2	2	2	0	2	0	1	3	4	24
YINT	0	1	0	0	0	3	1	2	0	0	1	0	1	1	2	2	2	3	2	3	24
BINT	1	5	1	1	0	3	1	0	2	1	2	1	2	2	0	2	1	3	2	0	30

Table 2.10: Number of enzymes (of 8) that had a P-value < 0.10 for each factor in ANOVAs modelling % composition of each amino acid. *AA2* is the composition of the same amino acid at the corresponding site.

and Lys to be influenced by interface membership, as well as Trp, Try, and His to a lesser degree. The composition of Met is greatly elevated in BINT sites. Asn is more abundant in YINT, BINT, and MINT sites than in NINT sites, particularly for the latter two. Arg is very abundant in MINT and YINT sites. Lys, however, is not over-represented in YINT or MINT sites, but is under-represented in BINT sites. The values in Fig. 2.9 show other patterns that were not found significant by the ANOVA but still seem noteworthy. YINT sites are generally low in the more hydrophobic residues, and somewhat more abundant in the hydrophilic residues, which is what would be expected for surface residues. BINT and MINT sites are both less abundant in Ile and Val compared to NINT sites, and both have elevated proportions of Tyr, Pro, His, and Asn. Phe is most abundant in BINT sites, and is slightly elevated in MINT sites. Both also have marginally more Trp, though the proportion of this amino acid is generally very low. This suggests a general preference for amino acids with an aromatic character in both types of interfaces. The BINT and MINT sites seem to have different preferences for charged residues, however. While both sets of sites are relatively low in Glu, BINT sites are much lower in Lys, Asp, and particularly Arg, which are abundant in MINT sites. BINT sites also show a preference for Met, Thr, and Ser that is not shared by MINT sites. Both MINT and BINT sites have relatively low proportions of Ala, though this is a typically common amino acid which may just be avoided in favour of amino acids that are more specifically useful in interface formation.

It was shown above that BINT and particularly MINT sites have a weaker relationship between the hydropathy and solvent exposure of a residue compared to NINT sites, whereas YINT sites had a stronger relationship. This relationship is shown by individual amino acid in (Table 2.12, Fig. 2.10). The stronger correlation between the hydropathy and solvent exposure of YINT sites is likely due to the

	% Composition									
	I	V	L	F	C	M	A	G	T	S
All	7.13	9.47	7.69	3.46	1.76	2.39	9.78	9.16	4.97	4.91
N	8.00	10.80	7.99	3.13	1.99	2.04	10.42	9.40	5.04	4.73
Y	3.66	6.83	5.02	2.75	1.86	1.38	11.13	10.50	3.92	4.75
B	5.90	5.22	8.12	5.44	0.76	5.06	6.89	9.07	6.01	6.10
M	5.07	7.26	7.20	3.98	1.21	2.27	7.44	5.97	3.81	4.85
	W	Y	P	H	E	Q	D	N	K	R
All	0.93	2.78	4.57	2.34	7.05	2.44	5.43	3.68	6.27	3.78
N	0.84	2.30	4.27	1.86	7.27	2.26	5.29	3.19	6.44	2.74
Y	0.95	2.83	4.61	1.90	8.12	4.25	6.18	4.07	7.12	8.17
B	1.25	4.27	5.63	4.14	5.87	2.70	4.56	4.97	4.59	3.44
M	1.19	4.53	5.54	4.07	5.93	1.75	7.28	5.46	6.48	8.72

Table 2.11: Composition of each amino acid by interface category.

	Solvent exposure									
	I	V	L	F	C	M	A	G	T	S
All	7.71	8.83	10.36	9.63	8.47	12.87	14.53	22.75	19.58	22.62
N	7.22	7.81	10.08	8.92	8.24	16.24	14.17	24.35	20.62	24.77
Y	13.22	13.38	15.11	7.92	10.81	13.76	20.09	20.61	31.90	30.13
B	6.56	10.12	7.54	13.36	5.14	5.63	12.30	18.15	8.21	12.97
M	12.34	16.18	14.56	7.68	11.32	11.31	13.69	15.81	22.73	16.05
	W	Y	P	H	E	Q	D	N	K	R
All	11.25	11.48	23.91	22.72	35.61	31.65	32.48	26.73	38.83	25.72
N	13.94	12.53	24.40	27.11	38.49	35.63	35.08	32.31	41.48	28.16
Y	7.80	10.59	38.49	29.79	37.31	32.41	29.62	23.57	42.18	26.19
B	6.28	12.19	15.56	12.95	20.83	15.31	19.27	11.64	24.40	21.22
M	5.90	6.41	21.60	17.56	25.69	24.81	31.57	22.46	28.34	21.46

Table 2.12: Mean solvent exposure of sites weighted by % composition of amino acid at each protein site.

	% Composition of amino acid at corresponding site									
	I	V	L	F	C	M	A	G	T	S
All	18.18	22.13	15.76	10.78	29.91	11.46	19.17	44.50	12.44	9.71
N	20.99	23.97	15.32	13.78	30.11	9.23	22.20	46.77	12.51	10.47
Y	6.50	13.60	17.04	1.71	32.91	1.26	8.86	23.80	5.07	5.19
B	9.28	16.56	19.82	8.74	0.80	21.34	8.79	48.07	18.21	11.09
M	4.72	12.88	11.95	1.19	50.98	0.76	13.30	42.12	5.25	5.10
	W	Y	P	H	E	Q	D	N	K	R
All	1.80	13.40	21.84	16.30	19.59	8.21	23.44	13.12	19.65	10.70
N	2.68	15.68	25.65	20.12	21.25	5.89	28.28	12.44	21.76	4.94
Y	0.42	23.12	22.54	1.33	10.05	1.48	14.09	14.53	8.07	18.07
B	0.14	1.67	7.40	21.26	20.62	27.82	12.46	16.37	24.75	16.81
M	0.33	14.51	18.85	0.63	13.72	3.62	12.01	10.90	8.89	15.54

Table 2.13: Composition of same amino acid at corresponding site, weighted by % composition of amino acid at each protein site.

	Normalized replacement rate									
	I	V	L	F	C	M	A	G	T	S
All	10.12	10.00	9.67	8.80	6.67	9.56	10.13	5.30	10.77	11.22
N	9.84	9.76	9.23	8.29	6.07	9.96	9.83	5.29	10.60	10.96
Y	13.12	11.36	12.86	11.06	9.41	12.73	12.24	6.44	12.57	14.86
B	11.12	12.12	10.89	9.17	11.55	7.84	10.14	4.36	10.60	10.83
M	9.98	9.49	9.50	9.88	6.11	10.57	10.59	5.57	11.34	10.66
	W	Y	P	H	E	Q	D	N	K	R
All	5.99	8.23	7.99	9.53	12.24	14.13	10.85	11.12	14.10	10.87
N	5.96	8.15	7.80	10.03	12.64	14.65	10.70	12.22	14.04	12.01
Y	5.40	9.46	11.42	12.00	14.64	14.41	11.54	10.87	17.17	10.12
B	6.45	8.82	7.72	8.42	8.43	10.67	11.49	8.10	13.58	11.09
M	5.91	6.92	6.78	8.13	10.64	15.88	10.58	10.08	11.78	8.36

Table 2.14: Replacement rate at each site, weighted by % composition of amino acid at that site. Replacement rates are normalized to an average of 10 replacements per site.

higher proportions of hydrophilic amino acids like Glu, Gln, Asp, Asn, and Arg relative to NINT sites, even though these residues do not have greater solvent exposure on average. The weaker relationship between hydropathy and solvent exposure for MINT and BINT sites can be seen in this graph, with more pronounced differences in the more hydrophilic amino acids. Residues at BINT sites are generally more buried than they are for any other category. Hydrophilic residues at MINT sites are generally more buried than they are at NINT sites, though not to the same extent that BINT sites are. The only amino acid for which BINT sites were more exposed is Phe, which is also relatively abundant for BINT sites. This pattern of exposure suggests that the BINT interfaces are generally more hydrophobic than the MINT interfaces. The charged and hydrophilic residues that are used are more likely to be buried, possibly as salt bridges in BINT sites.

While relative abundance of the various amino acids is informative, it does not indicate differences in constraint as much as measures of conservation do. We used two different measures of conservation. The first is the proportion of the same amino acid at the corresponding site in the homologous protein, which indicates how well that specific amino acid is conserved across the two subtrees. The second is the weighted mean replacement rate for that amino acid, which is taken only over the native subtree. This indicates how conserved positions rich in the specific amino acid tend to be, though this can differ across subtrees. When looking for amino acids that play important roles in maintenance of BINT interfaces, it would be expected that these amino acids would be at generally slowly-evolving positions, and also that they would be conserved across subtrees as they are expected to be under similar constraints in both structures. Amino acids that are important to MINT sites, however, should be conserved in their native subtree, but not necessarily across subtrees, as the functional constraints on these positions are expected to differ. The weighted means averages by amino acid are shown for these two measures in (Tables 2.13 and 2.14, Figures 2.11 and 2.12). The replacement rates were normalized for each site prior to taking the weighted average by setting the mean number of replacements at each site equal to 10.

For BINT sites, the amino acids Met, Gly, Thr, Gln, Asn, Lys are more conserved across structures relative to the other interface categories. Of those, Met, Gly, Gln, and Asn are also at slowly-evolving sites. His and Glu are also at slower sites, and are as conserved across structures as they are for NINT sites. The amino acids Met, Phe, Thr, Ser, Tyr, Pro, His, and Asn were more abundant at BINT sites. Of these, Met, His, and Asn are conserved and slowly-evolving, indicating that

they are fairly important in shared interfaces. Thr, Ser, Tyr, and Pro have comparable rates to those of NINT sites. While Thr and Ser are reasonably conserved across structures, Try and Pro are poorly conserved for BINT sites. This suggests that Thr and Ser may play minor roles in BINT interfaces, while Try and Pro are either less important, or are used differently for each structure. The remaining amino acids that are conserved or slowly-evolving at BINT sites are Gly, Gln, Glu, and Lys. Gly and Gln were of moderate abundance in BINT sites, while Glu and Lys were under-represented. This suggests that they may play important roles, though Glu and Lys may be restricted in quantity.

For MINT sites, Tyr, His, Glu, Asn, Lys, and particularly Arg-rich sites are evolving more slowly compared to NINT sites. Sites rich in Glu and Asn are even slower for BINT sites, but Arg and Tyr are slower for MINT sites. None of these amino acids except Arg are also conserved across structures. However, this does not indicate anything about the utility of these residues at MINT sites, as the constraints are expected to differ across subtrees for these sites. Tyr, His, Asn, and Arg were also more abundant in MINT sites. As with the BINT sites, the proportion of Glu was relatively low in MINT sites, though this does not necessarily mean that Glu does not play a role in MINT sites. Pro was also more abundant in MINT sites, but it displays no greater conservation relative to NINT sites (though it is generally conserved). Asp was also somewhat elevated in MINT sites, but similarly showed no rate difference relative to MINT sites.

When determining which amino acids play important roles in maintaining interfaces, abundance alone is not a sufficient indicator. Metrics of conservation can indicate which sites are under constraint, and thus important, as opposed to those amino acids which are merely tolerated. The unique nature of interfaces may also result in some atypical uses of various amino acids. For these reasons, we considered four different metrics in looking for amino acid preferences of common and unique interfaces: The average percent composition of each amino acid for the four interface categories; The average degree of solvent exposure for each amino acid; The average normalized replacement rate for the subtree; The percent composition of the same amino acid at the corresponding site in the partner subtree. The percent compositions were taken over the full alignment for each subtree, so they more accurately reflect the large-scale preferences and reduce any bias that the specific structures may have introduced. The other metrics were also weighted by the composition of the amino acid under consideration. Taken together, these measures show some commonalities and some differences in the use of various amino acids

in the different interfaces. A generally important amino acid could be expected to have a relatively elevated composition, be occupying slowly-evolving sites, be conserved across subtrees for BINT sites, and atypical in its exposure to solvent (which may indicate burial in an interface). Two amino acids, His and Asn, generally meet these criteria for both BINT and MINT sites. Met is clearly preferred in BINT sites, as is Gln and Thr, but to a lesser degree (they are conserved but not particularly abundant). Arg is an important residue for MINT sites, along with Tyr and Lys. BINT sites display conservation and relative burial of Glu, though it is diminished in both BINT and MINT sites. Phe and Ser were somewhat enriched in BINT sites, though neither were unusually conserved. For BINT sites, Cys seems particularly unimportant, being diminished and generally unconserved. Ile and Val seem similarly unimportant for both BINT and MINT sites.

There was a concern that the heavy thermophile membership of our dataset would lead to a strong bias in amino acid composition that was due to thermostability constraints rather than interface constraints. However, such a bias would be mediated by the following: 1) Not all of the species in the subtree of the thermophilic structures are thermophiles; 2) The BINT category would be equally represented by mesophiles and thermophiles even if the subtree for the thermophilic structure consisted entirely of thermophiles. To estimate the degree of potential bias, we took the average compositions of each amino acid for only the thermophilic structures and their mesophilic homologs (Table 2.15). There are 101 thermophilic MINT sites, 84 of each mesophilic and thermophilic BINT sites, and 12 mesophilic MINT sites. The compositions of the mesophilic and thermophilic BINT groups are similar, which is the strongest argument against a systematic thermophilic bias. In particular, the residues which are more abundant in MINT than in BINT sites (Val, Tyr, Asp, Asn, Lys, Arg) are not universally favoured by the thermophilic MINT and BINT sites. There are only 12 mesophilic MINT sites, but Arg is much more abundant in the mesophilic MINT sites than in the thermophilic MINT sites. Thermophiles are expected to prefer charged residues Glu, Lys, Asn, and Arg (Bogin *et al.* 2002), and have relatively less Ser, Thr, and Gln (Haney *et al.* 1999). While Glu is more abundant in the thermophilic interfaces, the other three are not, nor are Ser, Thr, and Gln relatively depleted.

	I	V	L	F	C	M	A	G	T	S
MesoB	5.92	4.53	8.14	6.26	0.57	4.17	6.04	9.78	5.57	5.88
ThermoB	6.68	7.21	10.57	3.98	0.86	2.35	6.85	8.05	6.04	5.60
MesoM	5.16	7.20	3.19	2.13	0.33	0.63	1.08	3.27	3.33	1.48
ThermoM	4.18	6.42	8.58	3.58	1.63	2.81	8.73	5.99	3.32	4.46
	W	Y	P	H	E	Q	D	N	K	R
MesoB	2.94	2.98	4.52	4.41	5.87	2.62	3.45	5.34	7.16	3.82
ThermoB	0.24	4.70	6.43	5.73	6.53	3.90	2.87	3.83	3.84	3.73
MesoM	8.23	3.85	9.27	8.29	0.64	0.68	7.30	2.54	8.75	22.73
ThermoM	0.66	4.54	5.28	2.89	6.14	1.95	7.37	6.31	7.12	8.05

Table 2.15: % compositions of each amino acid, partitioned by interface category (BINT or MINT) and thermophilic structures or mesophilic homologs thereof.

2.5 Discussion

Constraints that affect the evolutionary rate at protein sites vary both by site (Dean and Golding 2000; Dean *et al.* 2002; Robinson *et al.* 2003; Saunders and Baker 2002; Shi, Blundell and Mizuguchi 2001; Mizuguchi and Blundell 2000; Bustamante, Townsend and Hartl 2000; Goldman, Thorne and Jones 1998; Thompson and Goldstein 1996a) and over time at a specific protein site (Lopez, Casane and Philippe 2002; Susko *et al.* 2002; Yang, Swanson and Vacquier 2000; Gaucher *et al.* 2002; Knudsen *et al.* 2003; Gu 2003). The causes of these site-specific rate changes may be due to fairly obvious changes in the environment or function of the protein (Golding and Dean 1998; Susko *et al.* 2002; Yang, Swanson and Vacquier 2000; Gaucher *et al.* 2002; Gu 2003), or new structural constraints imposed by changes in neighbouring residues (Lopez, Casane and Philippe 2002; Knudsen and Miyamoto 2001). Previous work (Dean and Golding 2000; Dean *et al.* 2002) showed that half of the causal sitewise rate variation in a variety of proteins could be explained by a linear model that used general structural terms (solvent accessibility, distance from an active site, identity of Gly or Pro residues). In the previous chapter, we tried to detect general changes in constraint across two halves of a phylogeny by comparing the performance of this model when structural parameters were provided by structures from each subtree. We found a weak relative effect, but the model generally performed better at the larger scale where more of the rate variation was

deterministic. The enzymes used in the previous study did not universally display any drastic functional differences, so whatever local phylogenetic signals that did exist were likely weak. Here we use a data set with known functional differences between subtrees to see if changes in the relationship between sitewise replacement rates and structural factors are more obvious when a stronger change in constraint is present. Our data set consists of four enzyme pairs that have changed their quaternary structure at some point. We compare how the rate and structure relationships change across sites which have and have not experienced a change in constraint with respect to interface participation. We further describe some of the differences between the interfaces which are common to both structures in a homologous pair and those interfaces which are unique to one structure.

Overall, there were differences in the replacement rate, solvent accessibility, and amino acid compositions and the relationships between these factors between the various interface categories. The protein sites were partitioned into four categories. NINT sites are those which do not participate in any interface. BINT sites are those which participate in an interface in both members of the enzyme pair. MINT sites are those which only participate in an interface for one member, and YINT sites are those which are aligned to the MINT sites in the homologous structure. Generally, one might expect differences in constraints between the sites which are actively participating in an interface (BINT and MINT) and those which are not (NINT and YINT). However, MINT and YINT sites are homologous, and so may be expected to share some similarities. Since MINT sites are only forming functional interfaces in one subtree, they might be expected to show some differences from BINT sites as well. We found generally that MINT sites are similar to BINT sites, but with some key differences.

We found a lower average replacement rate for BINT and MINT sites as compared to NINT sites, and a higher than average rate for YINT sites. The higher rate of YINT sites is likely due to their greater average solvent exposure. YINT sites are more exposed to solvent, and so are likely experiencing reduced constraint because they would be maintaining fewer specific contacts. This reduced constraint would lead to an elevated replacement rate. MINT sites were more buried on average than NINT sites, but BINT sites were even further buried. This suggests that the conservation of MINT sites is not solely due to the expected effects of burial. It also suggests that the MINT sites may have been surface residues more recently. MINT sites had a weaker and less often significant correlation between the solvent exposure and the hydropathy of the residue at a site, which further supports more recent

solvent exposure for these residues. Alternately, the weaker correlation suggests an atypical use of charged residues in unique interfaces, likely in the formation of buried salt bridges.

Interface membership did have an effect on replacement rates, though it was not as strong as the effects of solvent accessibility or the replacement rate at the aligned site in the homologous structure. This effect was not stronger for MINT or YINT sites when considered with the solvent accessibility or the hydropathy of the residues at the site, though BINT sites gained some explanatory power when coupled with the hydropathies of the residues at both sites. The solvent exposure was less likely to have an effect on the replacement rate of MINT sites than of YINT or BINT sites. The replacement rate at the homologous site was more often a significant determinant of replacement rates for BINT sites than for MINT or YINT sites. Generally, there was a greater difference between the replacement rates of MINT and YINT sites than between BINT or NINT sites, and this difference was not due to differences in solvent exposure between these sites in the homologous structures.

The compositions and use patterns of specific amino acids are also somewhat affected by interface category. BINT and MINT sites shared a common preference for His and Asn, which were both more buried on average than they were in NINT or YINT sites, though both were even more buried in BINT sites. BINT sites preferred Met and Thr, while Gln and Glu were not abundant but were conserved and buried relative to other interface categories. MINT sites expressed a strong preference for Arg, as well as for Tyr and Lys. These amino acids were also relatively buried, but not as drastically as the amino acids preferred by BINT sites. Despite the fact that BINT and MINT sites are more buried overall, the representation of the more hydrophobic amino acids was relatively low. Though both types of interfaces made relatively high use of moderately polar amino acids (Tyr, Pro, His), the most important roles in these interfaces seem to be played by the polar and charged residues.

The greater conservation of interface residues has been demonstrated at length (Elcock and McCammon 2001; Landgraf, Xenarios and Eisenberg 2001; Glaser *et al.* 2003; Halperin, Wolfson and Nussinov 2004; Ma *et al.* 2003; Hu *et al.* 2000; Teichmann 2002), so our finding of slower rates in sites which participate in interfaces is not surprising. What is noteworthy is the differences between the solvent exposure and residue hydropathies, and their relationships to the replacement rates between the BINT and MINT interface categories. The differences in amino acid

compositions may be stochastic, as the data set is not large. However, the compositions are taken over full phylogenetic trees rather than just pairs of sequences, which provides a more robust sample. All of the enzymes studied have obligate subunit associations, so no pressure to maintain transient stability and tolerance of exposure to solvent currently exists for the MINT interfaces. Three of our enzyme pairs contain one member from a thermophilic species, which could introduce bias, but this effect is mediated in several ways. Though it is impossible to know in most cases which quaternary structure is ancestral, it is likely that the higher-order one is more recent. It follows that the MINT sites have spent less time participating in an interface relative to BINT sites, and the observed differences are a reflection of this novelty.

Interfaces can vary widely in their specific physical features (Larsen, Olson and Goodsell 1998), though there are some generalities. Ofran and Rost (2003b) found that interfaces could be identified by sequence alone if they were treated as belonging to a number of different functional categories based on symmetry and duration of contact. This suggests that some sets of amino acids are preferred for specific roles. Though hydrophobicity alone is not a reliable indicator of interface residues (Jones and Thornton 1997; Liang *et al.* 2004), larger interfaces tend to be richer in hydrophobic residues than smaller interfaces are (Glaser *et al.* 2001). Interfaces have been commonly described as having a general structure of a core of conserved strongly-binding amino acids, surrounded by a ring of less critical and more hydrophobic residues that exclude solvent from the interface (Bogan and Thorn 1998; Hu *et al.* 2000; Ma *et al.* 2003; Halperin, Wolfson and Nussinov 2004). The strongly-binding hot spot residues are limited in density in an interface, as the strong contacts rigidify the interface and limit the flexibility of the enzyme (Hu *et al.* 2000; Ma *et al.* 2003). The conserved hot spot residues also have a greater packing density around them, which further contributes to the rigidity of these sites (Halperin, Wolfson and Nussinov 2004). The multiple and weaker hydrophobic contacts offer greater flexibility, but must be present in greater numbers to maintain the stability of the interface (Goodsell and Olson 2000). Though interfaces are generally more conserved than other surface residues, they are not as conserved as active sites (Glaser *et al.* 2003; Elcock and McCammon 2001; Landgraf, Xenarios and Eisenberg 2001). The difficulty in predicting interfaces indicates that there are a variety of ways to form them, so some degree of flux can be expected. Larger interfaces would generally take a longer to evolve, and the replacements that are successful in the context of an existing stable interface are likely different from those that are successful in stablizing a new interface. The conserved hot spot

residues tend to be more polar (Bogan and Thorn 1998; Hu *et al.* 2000), and may already be present on the surface of a protein. Once several of these strong contacts are established and the nearby environment becomes shielded from the solvent, hydrophobic replacements are more likely to be successful.

Our results support this gradual development of an interface that is more hydrophobic and is stabilized by more hydrophobic contacts than strong hot spots. Though the common interfaces were not larger overall than the unique interfaces, they were more shielded from solvent. The common interfaces consist of more hydrophobic residues than the unique interfaces do. The hydrophobic residues (Ile, Val, Leu, Phe, Cys, and Met) together make up 30.5% of BINT sites, compared to 26.99% of MINT sites. The hot spot residues which make the strongest energetic contributions to interfaces are Trp, Tyr, Arg, Asp, His, and Pro. While NINT sites are generally very low in these amino acids (17.3% total), BINT sites are more enriched (23.29%), but not to the extent that MINT sites are (31.33%). The residues that are more abundant in BINT sites relative to MINT sites (Met, Gly, Thr, Ser) do not form strong hydrogen bonds, but are rather moderately polar and more tolerant of buried environments than the more hydrophilic residues that are abundant in MINT sites (Asp, Asn, Lys, Arg). Hu *et al.* (2000) found that structurally conserved sites were not always hot spot residues, indicating that residues that are not energetically significant may still play important roles. These conserved but non-hot spot residues were His, Asn, Gln, Thr, Ser, Phe, and Met, which are exactly those which are conserved or abundant in BINT sites. While His and Asn are also abundant and relatively conserved in MINT sites, both of these residues are more exposed in MINT sites than they are in BINT sites. Additionally, Tyr and Arg were more conserved in MINT sites than they were in BINT sites. This suggests that as the interface matures and develops weaker contacts, the hot spots play a less crucial role.

The differences between the MINT and YINT sites demonstrates that differences in constraint do manifest as changes in replacement rate when the functional differences are sufficiently pronounced. The relationships between the replacement rate and structural factors (solvent exposure and hydropathy of the residue) were stronger for those sites not involved in an interface in one subtree, and weaker for the sites that were involved in any interface. The specific cause of the constraint could not be detected by the changes in these relationships, but a disjoint in the rate between sites was significant relative to all other sites in the protein. The ANOVA was not an ideal method of detecting these differences, as the number of

sites participating in interfaces is likely too small to have a significant overall effect on the rates. While the pressure for thermostability was generally minimized by using phylogenies rather than single sequences, the tree for 4TIM was comprised entirely of thermophiles. This additional change in constraint was also detected by a breakdown in the relationship between the replacement rate and solvent accessibility. This suggests that a variety of strong changes in functional constraint may be detected by comparing the strength of this relationship across subtrees.

2.6 Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (Canada) Research grants to GBG. Thanks to Dr. Jonathon Stone for suggestions regarding ANOVA.

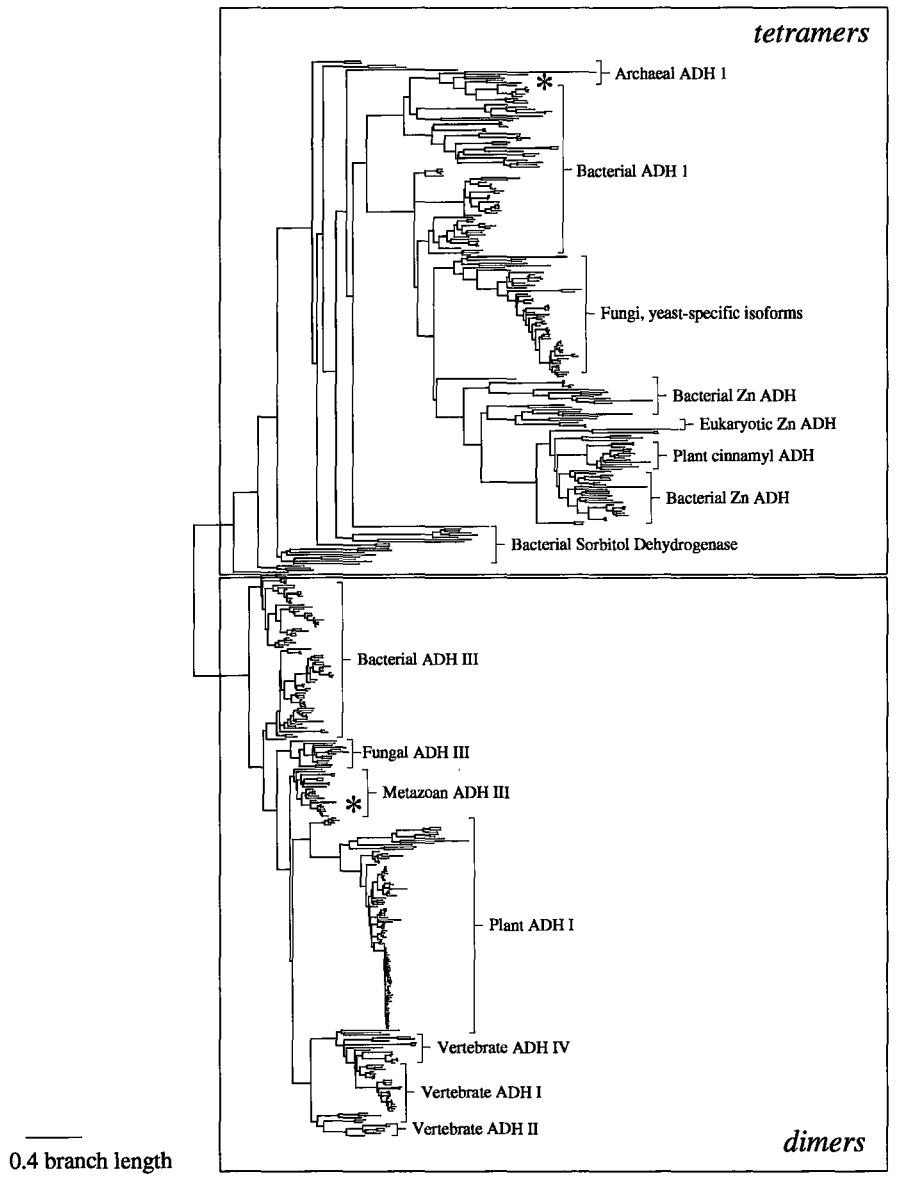

Figure 2.1: Phylogenetic tree for Alcohol Dehydrogenase. The asterisks indicate sequences with structures which were used for this study.

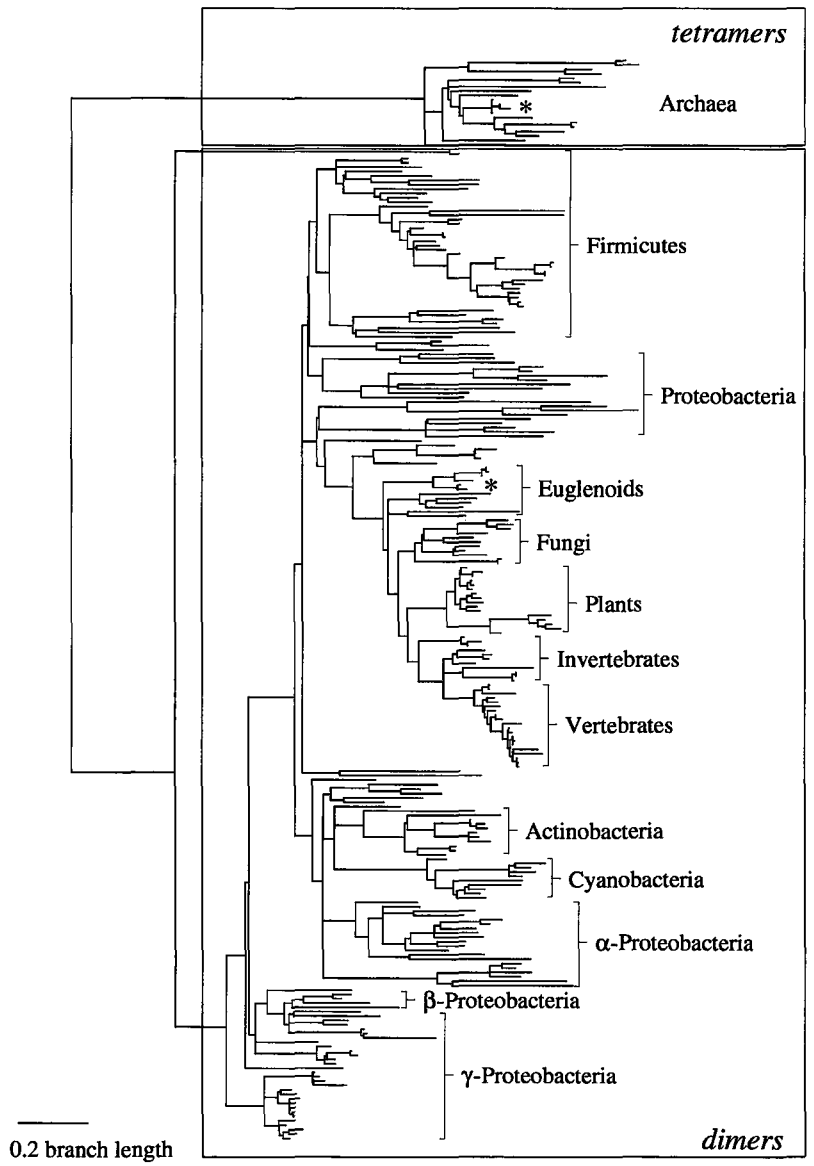

Figure 2.2: Phylogenetic tree for Triose Phosphate Isomerase. The asterisks indicate sequences with structures which were used for this study.

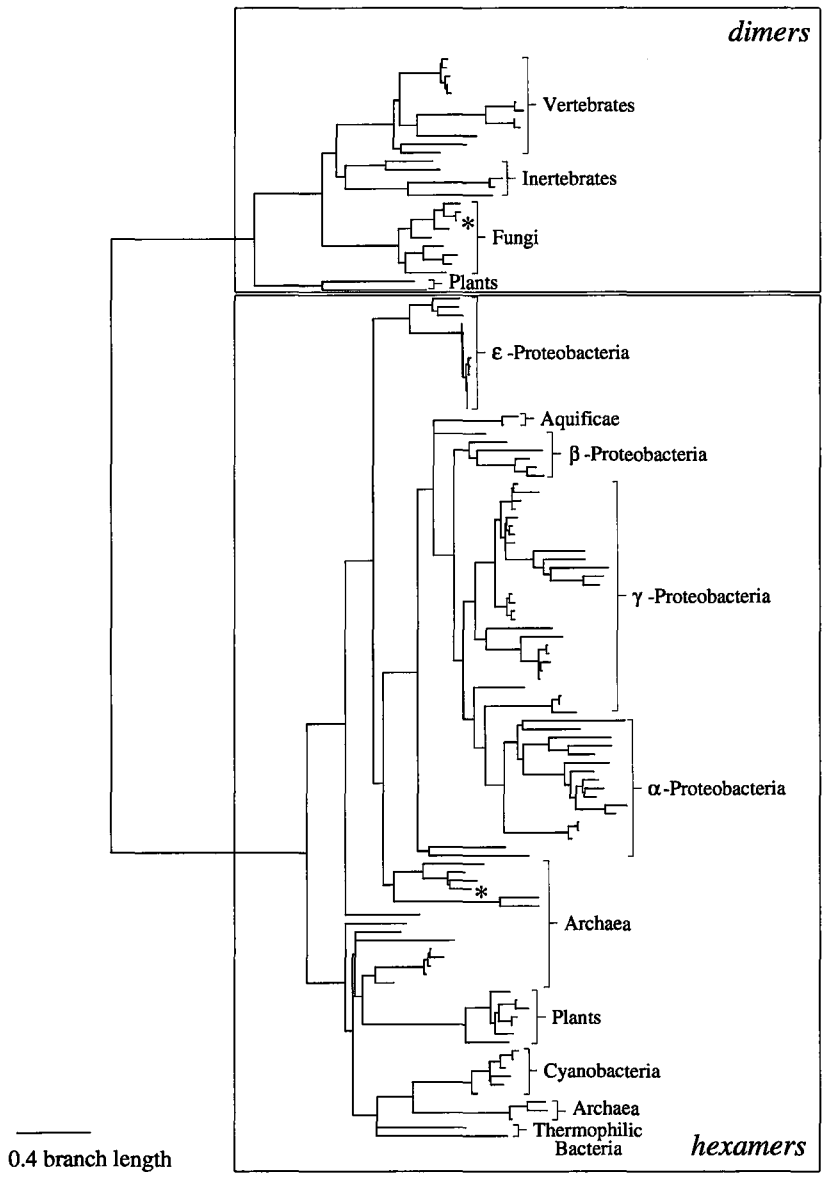

Figure 2.3: Phylogenetic tree for Inorganic Pyrophosphatase. The asterisks indicate sequences with structures which were used for this study.

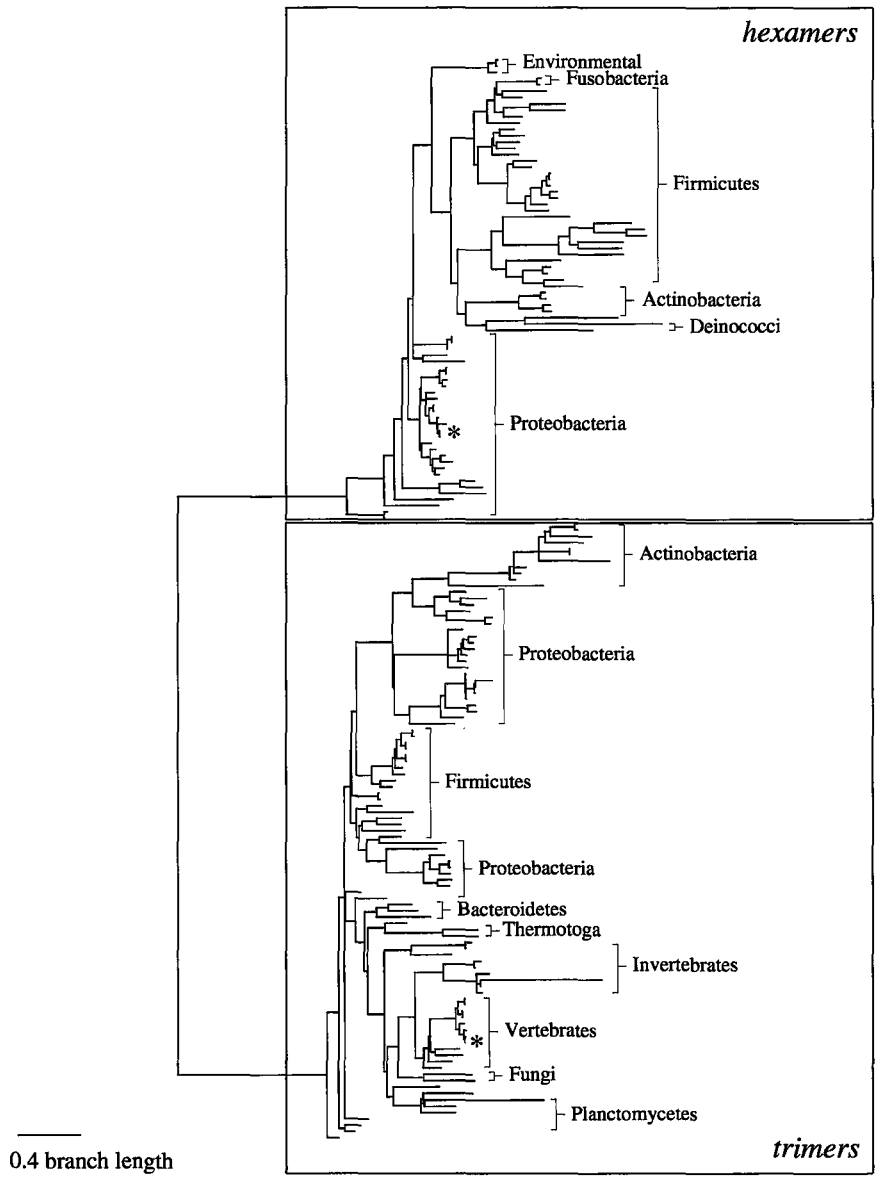

Figure 2.4: Phylogenetic tree for Purine Nucleoside Phosphorylase. The asterisks indicate sequences with structures which were used for this study.

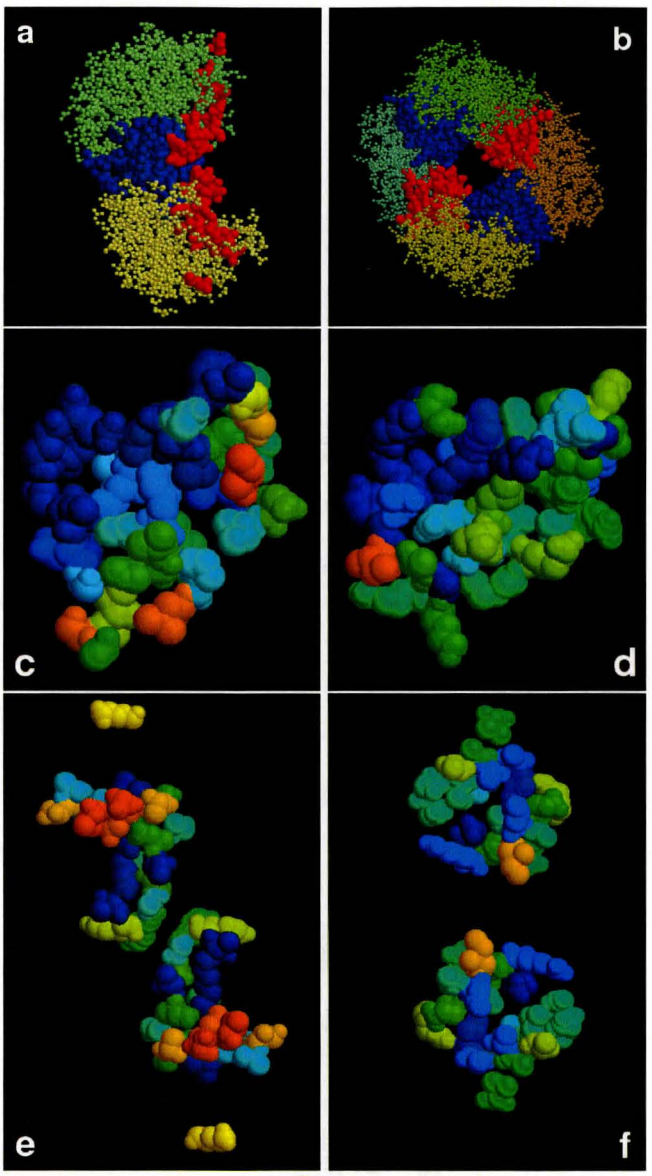

Figure 2.5: Schema and rate-coloured structures for Triose Phosphate Isomerase. The dimer is shown in (a) and the tetramer is shown in (b). The rate-coloured dimeric interface sites are shown for the dimer in (c), and for the tetramer in (d). (e) and (f) show the tetramer interface sites, for the dimer and tetramer respectively. See figure 2.2 for phylogenies.

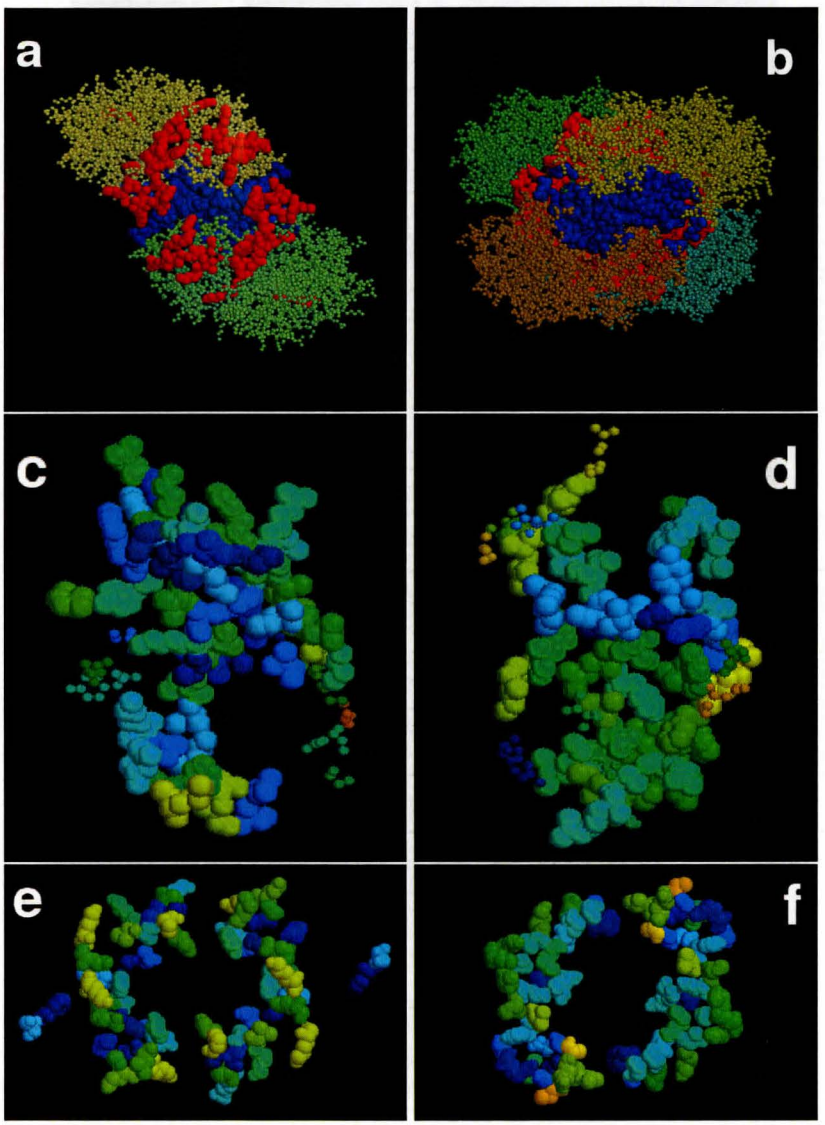

Figure 2.6: Schema and rate-coloured structures for Alcohol Dehydrogenase. The dimer is shown in (a) and the tetramer is shown in (b). The rate-coloured dimeric interface sites are shown for the dimer in (c), and for the tetramer in (d). Small residues indicate YINT sites for each enzyme. (e) and (f) show the tetramer interface sites, for the dimer and tetramer respectively. See figure 2.1 for phylogenies.

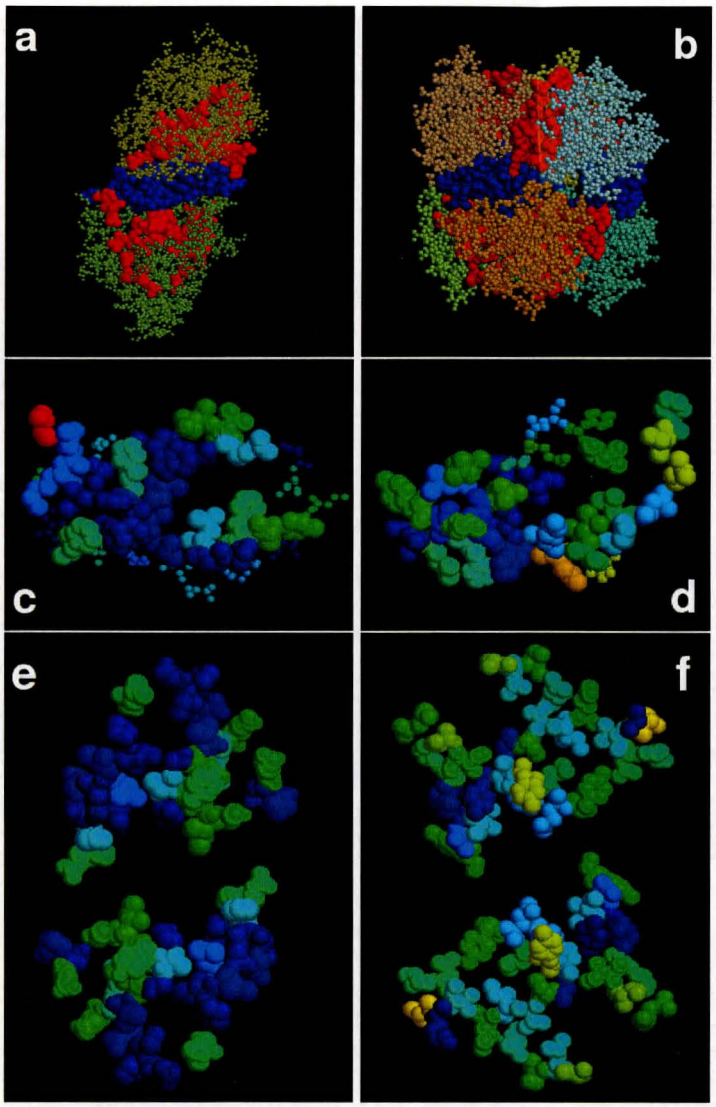

Figure 2.7: Schema and rate-coloured structures for Inorganic Pyrophosphatase. The dimer is shown in (a) and the hexamer is shown in (b). The rate-coloured dimeric interface sites are shown for the dimer in (c), and for the hexamer in (d). Small residues indicate YINT sites for each enzyme. (e) and (f) show the hexamer interface sites, for the dimer and hexamer respectively. See figure 2.3 for phylogenies.

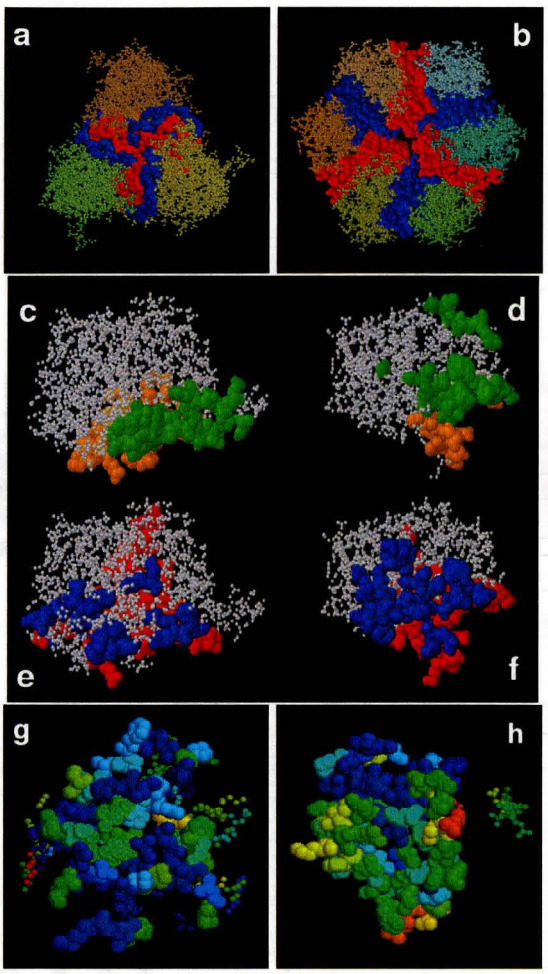

Figure 2.8: Schema and rate-coloured structures for Purine Nucleoside Phosphorylase. (a) and (b) show the full trimer and hexamer, respectively. The monomers with interface-participating residues are shown in (c) and (e) for the trimer, and (d) and (f) for the hexamer, with the rotational axis of the proteins at the bottom. Green and orange residues (c, d) show the positions of the residues in both sides of the trimer interface sites, and the blue and red residues show the location of sites participating in the hexameric interfaces (e, f). The rate-coloured interface residues are shown for the trimer in (g), and for the hexamer in (h). In (g) and (h), the rotational axis of the proteins is in the horizontal plane of the page, facing the viewer. Smaller residues indicate YINT sites. See figure 2.4 for phylogenies.

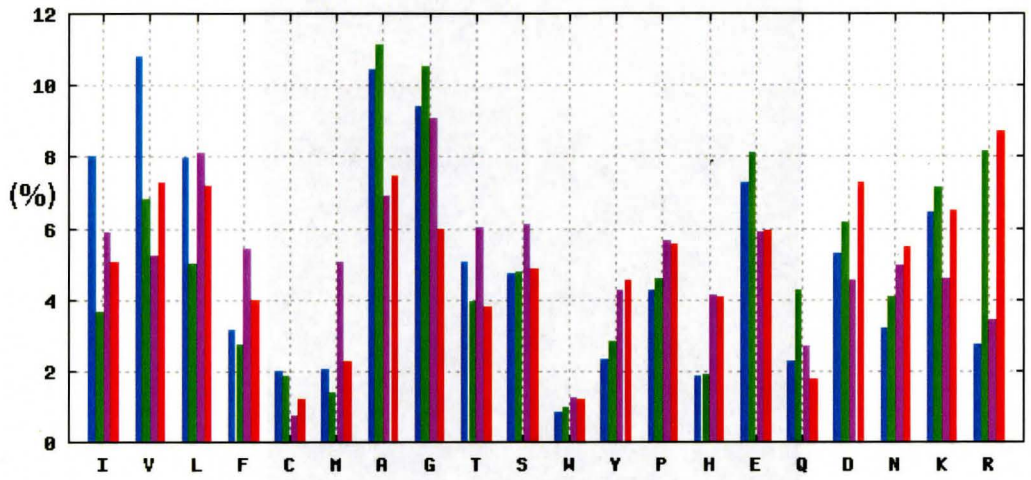

Figure 2.9: Average composition of amino acids by interface category. Blue bars indicate NINT sites, green bars are YINT sites, purple bars are BINT sites, and red bars are MINT sites.

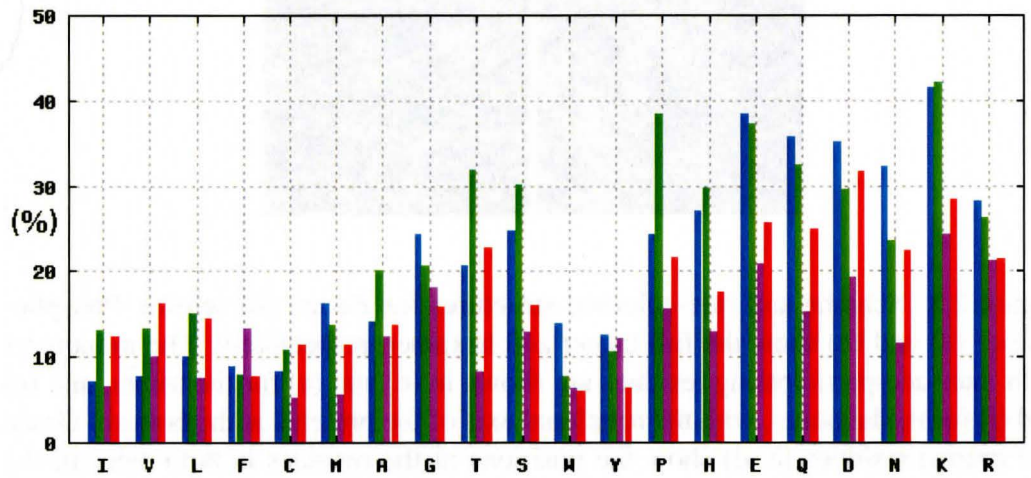

Figure 2.10: Average degree of solvent exposure for each amino acid, weighted by the composition of the amino acid at the alignment site.

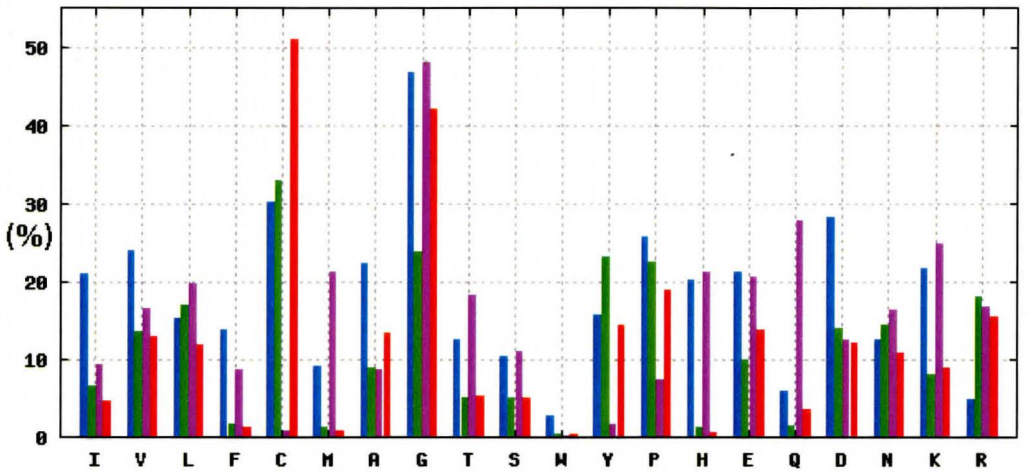

Figure 2.11: Composition of the same amino acid at the aligned site in the corresponding subtree, weighted by proportion of amino acid at site.

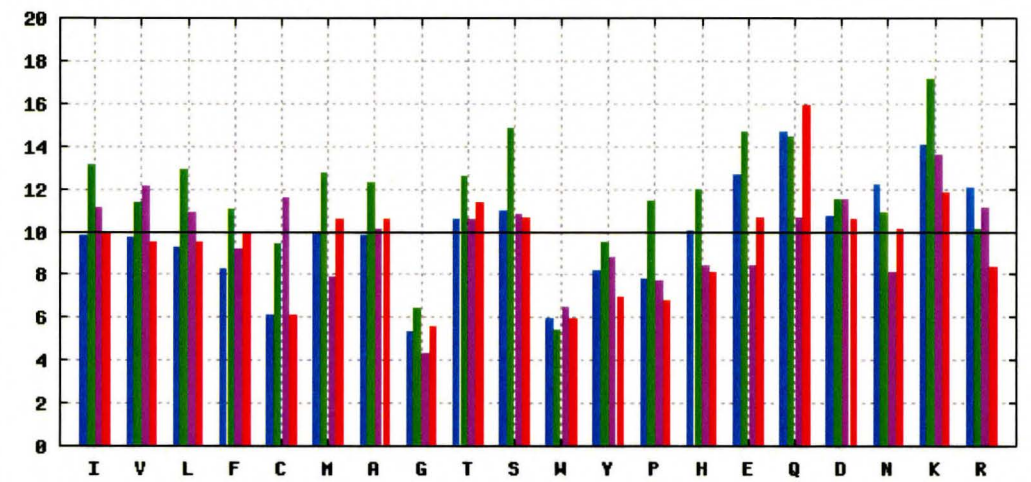

Figure 2.12: Average normalized number of replacements per site, weighted by proportion of amino acid at site. The mean number of replacements at a protein site has been set to 10.

Part II

CONCLUSION

Protein sequences do not evolve at a constant rate across all sites, nor do those sites maintain the same rate throughout time. While some of this variation can be attributed to the underlying Poisson mutation process, most sites are under some degree of constraint, which affects the rate at that site. Additionally, the constraints acting on a protein site may change at any point in time. When the causal variation in rates is associated with simple physicochemical indicators, it can be largely explained with a linear model. The simple model used in this study includes terms for the solvent exposure of each site, the distance of the residue from the active site, the hydropathy of the residue, and an amino acid such as glycine or proline which play important torsional roles. In a survey of α/β barrel proteins, Dean *et al.* (2002) explained half the causal rate variation using a variant of this model. In this work, we explored changes in the relationship between replacement rates and general physicochemical indicators of constraint between different lineages with differences in the degree of rate divergence expected between these lineages.

In the first chapter, we explored the changes in the NCDs of a number of enzymes with varying degrees of expected divergence across sub-phylogenies for these enzymes. The NCD is calculated from the $\hat{r}^2$ value from the multiple regression divided by the expected proportion of causal rate variation, and so offers an indication of how much variation remains to be explained (as opposed to the variation which is stochastic). We found that local adaptation can be detected by changes in the NCD. In certain subtrees where the proportion of Poisson rate variation was greater, the NCD was also greater when the native structure was used, indicating that the deterministic variation in that subtree was of a simpler nature which was more fully explained by the simple model. For the other subtree which had more causal rate variation, the NCD was lower and therefore more of the causal rate variation was due to factors which were not included in the simple model. This result suggests that changes in the NCD can indicate where changes in constraint that are not simply due to general structural features occur by the greater proportion of unexplained causal variation. Though changes in the NCD were affected more strongly by the specific subtree and the evolutionary patterns in that subtree, a relative positive effect of the native structure was apparent. A difference-based multiple regression model, where changes between structures are emphasized may more effectively highlight changes in constraint across the subtrees.

In the previous work (Dean *et al.* 2002), the large subunit of RUBISCO was found to have a very unusual distribution of replacement rates. We found that this pattern was due to its location in the chloroplast genome. The nuclear-encoded

small subunit displays a pattern of replacements like that of the other enzymes studied. Additionally, the sites in the large subunit with unusually high replacement rates displayed a different pattern than rapidly-evolving sites in the small subunit. The fast sites in the large subunit typically fluctuated between two amino acids, suggesting possible polymorphisms which may be maintained due to the unusual population structure of the chloroplast genome.

In the second chapter, we examined a set of proteins that had changed their quaternary structure at some point in the phylogeny. The sites that were involved in the new interface were expected to display differences in replacement rates and constraint relationships when compared to their homologous sites which were not participating in an interface. The new interface sites were more conserved and had a weaker relationship between replacement rates and structural factors than the homologous sites did, and displayed more similarity to the interface sites that were shared between structures. The new interface sites also differed from the shared interface sites in ways which suggest the mechanism by which new interfaces evolve. The shared interfaces are more shielded from solvent, and make greater use of small hydrophobic residues than of the strong 'hot spot' residues which have strong binding energy, but which also confer rigidity upon the structure. It is thus likely that new interfaces are established based on only a few strong contacts. Over time, abundant and weak hydrophobic contacts can evolve in the newly-shielded environment, leading to a more flexible interface.

Generally, the relationship between replacement rates and physicochemical indicators of constraint does change across lineages in a way that indicates changes in this constraint. If replacement rate heterogeneity is viewed naively, so much information is present that it is difficult to know which patterns represent unusual adaptations and which are more simply explained by general structural factors. The analysis used in this work may point to lineage-specific changes in constraint which are not simply due to general structural features, and allow researchers to focus their attention on outliers which may represent unusual and interesting adaptations.

Part III

REFERENCES

Bibliography

Andersson, I. (1996). Large structures at high resolution: the 1.6 A crystal structure of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase complexed with 2-carboxyarabinitol bisphosphate. *J Mol Biol. 259*, 160–174.

Bennett, M. J., B. P. Schlegel, J. M. Jez, T. M. Penning, and M. Lewis (1996). Structure of 3 alpha-hydroxysteroid/dihydrodiol dehydrogenase complexed with NADP+. *Biochemistry. 35*, 10702–10711.

Birky CW, J. r. and J. B. Walsh (1992). Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. *Genetics. 130*, 677–683.

Bogan, A. A. and K. S. Thorn (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol. 280*, 1–9.

Bogin, O., I. Levin, Y. Hacham, S. Tel-Or, M. Peretz, F. Frolow, and Y. Burstein (2002). Structural basis for the enhanced thermal stability of alcohol dehydrogenase mutants from the mesophilic bacterium *Clostridium beijerinckii*: contribution of salt bridging. *Protein Sci. 11*, 2561–2574.

Bustamante, C. D., J. P. Townsend, and D. L. Hartl (2000). Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol. 17*, 301–308.

Chattopadhyaya, R., W. E. Meador, A. R. Means, and F. A. Quiocho (1992). Calmodulin structure refined at 1.7 A resolution. *J Mol Biol. 228*, 1177–1192.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res. 31*, 3497–3500.

de Azevedo WF, J. r., F. Canduri, D. M. dos Santos, J. H. Pereira, M. V.

Bertacine Dias, R. G. Silva, M. A. Mendes, L. A. Basso, M. S. Palma, and D. S. Santos (2003). Crystal structure of human PNP complexed with guanine. *Biochem Biophys Res Commun. 312*, 767–772.

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

Dean, A. M., C. Neuhauser, E. Grenier, and G. B. Golding (2002). The pattern of amino acid replacements in alpha/beta-barrels. *Mol Biol Evol. 19*, 1846–1864.

Duquerroy, S., C. Camus, and J. Janin (1995). X-ray structure and catalytic mechanism of lobster enolase. *Biochemistry. 34*, 12513–12523.

El-Kabbani, O., C. Darmanin, T. R. Schneider, I. Hazemann, F. Ruiz, M. Oka, A. Joachimiak, C. Schulze-Briese, T. Tomizaki, A. Mitschler, and A. Podjarny (2004). Ultrahigh resolution drug design. II. Atomic resolution structures of human aldose reductase holoenzyme complexed with Fidarestat and Minalrestat: implications for the binding of cyclic imide inhibitors. *Proteins. 55*, 805–813.

Elcock, A. H. and J. A. McCammon (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A. 98*, 2990–2994.

Erskine, P. T., E. Norton, J. B. Cooper, R. Lambert, A. Coker, G. Lewis, P. Spencer, M. Sarwar, S. P. Wood, M. J. Warren, and P. M. Shoolingin-Jordan (1999). X-ray structure of 5-aminolevulinic acid dehydratase from *Escherichia coli* complexed with the inhibitor levulinic acid at 2.0 A resolution. *Biochemistry. 38*, 4266–4276.

Fay, J. C. and C. I. Wu (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet. 4*, 213–235.

Felsenstein, J. (1989). PHYLIP (phylogeny inference package). (Version 3.2). *Cladistics. 5*, 164–166.

Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool. 20*, 406–416.

Fitch, W. M. (1976). The molecular evolution of cytochrome c in eukaryotes. *J Mol Evol. 8*, 13–40.

Gaucher, E. A., X. Gu, M. M. Miyamoto, and S. A. Benner (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends*

Biochem Sci. 27, 315–321.

Glaser, F., T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics. 19*, 163–164.

Glaser, F., D. M. Steinberg, I. A. Vakser, and N. Ben-Tal (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins. 43*, 89–102.

Golding, G. B. and A. M. Dean (1998). The structural basis of molecular adaptation. *Mol Biol Evol. 15*, 355–369.

Goldman, N., J. L. Thorne, and D. T. Jones (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics. 149*, 445–458.

Goodsell, D. S. and A. J. Olson (2000). Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct. 29*, 105–153.

Gu, J. and X. Gu (2003). Natural history and functional divergence of protein tyrosine kinases. *Gene. 317*, 49–57.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol. 16*, 1664–1674.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol. 18*, 453–464.

Gu, X. (2003). Functional divergence in protein (family) sequence evolution. *Genetica. 118*, 133–141.

Guex, N. and M. Peitsch (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis. 18*, 2714–2723.

Guy, J. E., M. N. Isupov, and J. A. Littlechild (2003). The structure of an alcohol dehydrogenase from the hyperthermophilic archaeon *Aeropyrum pernix. J Mol Biol. 331*, 1041–1051.

Halperin, I., H. Wolfson, and R. Nussinov (2004). Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure. 12*, 1027–1038.

Haney, P. J., J. H. Badger, G. L. Buldak, C. I. Reich, C. R. Woese, and G. J. Olsen (1999). Thermal adaptation analyzed by comparison of protein sequences

from mesophilic and extremely thermophilic Methanococcus species. *Proc Natl Acad Sci U S A. 96*, 3578–3583.

Hart, P. J., M. M. Balbirnie, N. L. Ogihara, A. M. Nersissian, M. S. Weiss, J. S. Valentine, and D. Eisenberg (1999). A structure-based mechanism for copper-zinc superoxide dismutase. *Biochemistry. 38*, 2167–2178.

Heikinheimo, P., J. Lehtonen, A. Baykov, R. Lahti, B. S. Cooperman, and A. Goldman (1996). The structural basis for pyrophosphatase catalysis. *Structure. 4*, 1491–1508.

Hester, G., O. Brenner-Holzach, F. A. Rossi, M. Struck-Donatz, K. H. Winterhalter, J. D. Smit, and K. Piontek (1991). The crystal structure of fructose-1,6-bisphosphate aldolase from *Drosophila melanogaster* at 2.5 A resolution. *FEBS Lett. 292*, 237–242.

Hu, Z., B. Ma, H. Wolfson, and R. Nussinov (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins. 39*, 331–342.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics. 17*, 754–755.

Hulo, N., C. J. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch (2004). Recent improvements to the PROSITE database. *Nucleic Acids Res. 32-7*, D134–D137.

Jones, S. and J. M. Thornton (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol. 272*, 121–132.

Kabsch, W. and C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers. 22*, 2577–2637.

Kim, H., U. Certa, H. Dobeli, P. Jakob, and W. G. Hol (1998). Crystal structure of fructose-1,6-bisphosphate aldolase from the human malaria parasite *Plasmodium falciparum. Biochemistry. 37*, 4388–4396.

Kimura, M. (1989). The neutral theory of molecular evolution and the world view of the neutralists. *Genome. 31*, 24–31.

Knudsen, B. and M. M. Miyamoto (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A. 98*, 14512–14517.

Knudsen, B., M. M. Miyamoto, P. J. Laipis, and D. N. Silverman (2003). Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. *Genetics. 164*, 1261–1269.

Landgraf, R., I. Xenarios, and D. Eisenberg (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol. 307*, 1487–1502.

Larsen, T. A., A. J. Olson, and D. S. Goodsell (1998). Morphology of protein-protein interfaces. *Structure. 6*, 421–427.

Laskowski, R. A., E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones, and J. M. Thornton (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci. 22*, 488–490.

Leppanen, V. M., H. Nummelin, T. Hansen, R. Lahti, G. Schafer, and A. Goldman (1999). *Sulfolobus acidocaldarius* inorganic pyrophosphatase: structure, thermostability, and effect of metal ion in an archael pyrophosphatase. *Protein Sci. 8*, 1218–1231.

Liang, S., J. Zhang, S. Zhang, and H. Guo (2004). Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. *Proteins. 57*, 548–557.

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

Ma, B., T. Elkayam, H. Wolfson, and R. Nussinov (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A. 100*, 5772–5777.

Maier, R. M., K. Neckermann, G. L. Igloi, and H. Kossel (1995). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J Mol Biol. 251*, 614–628.

Mao, C., W. J. Cook, M. Zhou, G. W. Koszalka, T. A. Krenitsky, and S. E. Ealick (1997). The crystal structure of *Escherichia coli* purine nucleoside phosphorylase: a comparison with the human enzyme reveals a conserved topology. *Structure. 5*, 1373–1383.

Miyamoto, M. M. and W. M. Fitch (1995). Testing the covarion hypothesis of molecular evolution. *Mol Biol Evol. 12*, 503–513.

Mizohata, E., H. Matsumura, Y. Okano, M. Kumei, H. Takuma, J. Onodera, K. Kato, N. Shibata, T. Inoue, A. Yokota, and Y. Kai (2002). Crystal structure of activated ribulose-1,5-bisphosphate carboxylase/oxygenase from green alga *Chlamydomonas reinhardtii* complexed with 2-carboxyarabinitol-1,5-bisphosphate. *J Mol Biol. 316*, 679–691.

Mizuguchi, K. and T. Blundell (2000). Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics. 16*, 1111–1119.

Nooren, I. M. and J. M. Thornton (2003). Diversity of protein-protein interactions. *EMBO J. 22*, 3486–3492.

Ofran, Y. and B. Rost (2003a). Analysing six types of protein-protein interfaces. *J Mol Biol. 325*, 377–387.

Ofran, Y. and B. Rost (2003b). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett. 544*, 236–239.

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol. 132*, 185–219.

Rao, S. T., S. Wu, K. A. Satyshur, K. Y. Ling, C. Kung, and M. Sundaralingam (1993). Structure of *Paramecium tetraurelia* calmodulin at 1.8 A resolution. *Protein Sci. 2*, 436–447.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol. 20*, 1692–1704.

Sanghani, P. C., H. Robinson, R. Bennett-Lovsey, T. D. Hurley, and W. F. Bosron (2003). Structure-function relationships in human Class III alcohol dehydrogenase (formaldehyde dehydrogenase). *Chem Biol Interact. 143*, 195–200.

Saunders, C. T. and D. Baker (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol. 322*, 891–901.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics. 18*, 502–504.

Shi, J., T. L. Blundell, and K. Mizuguchi (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol. 310*, 243–257.

Shrake, A. and J. A. Rupley (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol. 79*, 351–371.

Sicheri, F., I. Moarefi, and J. Kuriyan (1997). Crystal structure of the Src family tyrosine kinase Hck. *Nature. 385*, 602–609.

Strange, R. W., S. Antonyuk, M. A. Hough, P. A. Doucette, J. A. Rodriguez, P. J. Hart, L. J. Hayward, J. S. Valentine, and S. S. Hasnain (2003). The structure of holo and metal-deficient wild-type human Cu, Zn superoxide dismutase and its relevance to familial amyotrophic lateral sclerosis. *J Mol Biol. 328*, 877–891.

Susko, E., Y. Inagaki, C. Field, M. E. Holder, and A. J. Roger (2002). Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol Biol Evol. 19*, 1514–1523.

Takahata, N. (1987). On the overdispersed molecular clock. *Genetics. 116*, 169–179.

Teichmann, S. A. (2002). The constraints protein-protein interactions place on sequence divergence. *J Mol Biol. 324*, 399–407.

Thompson, M. J. and R. A. Goldstein (1996a). Constructing amino acid residue substitution classes maximally indicative of local protein structure. *Proteins. 25*, 28–37.

Thompson, M. J. and R. A. Goldstein (1996b). Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins. 25*, 38–47.

Wako, H. and T. L. Blundell (1994). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J Mol Biol. 238*, 682–692.

Walden, H., G. S. Bell, R. J. Russell, B. Siebers, R. Hensel, and G. L. Taylor (2001). Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase. *J Mol Biol. 306*, 745–757.

Williams, J. C., J. P. Zeelen, G. Neubauer, G. Vriend, J. Backmann, P. A. Michels, A. M. Lambeir, and R. K. Wierenga (1999). Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Eng. 12*, 243–250.

Xu, D., C. J. Tsai, and R. Nussinov (1998). Mechanism and evolution of protein dimerization. *Protein Sci. 7*, 533–544.

Xu, W., A. Doshi, M. Lei, M. J. Eck, and S. C. Harrison (1999). Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell. 3*, 629–638.

Yang, Z., W. J. Swanson, and V. D. Vacquier (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol. 17*, 1446–1455.

Zhang, E., J. M. Brewer, W. Minor, L. A. Carreira, and L. Lebioda (1997). Mechanism of enolase: the crystal structure of asymmetric dimer enolase-2-phospho-D-glycerate/enolase-phosphoenolpyruvate at 2.0 A resolution. *Biochemistry. 36*, 12526–12534.