

The Pattern of Amino Acid Replacements in α/β -Barrels

Antony M. Dean,* Claudia Neuhauser,† Elise Grenier,‡ and G. Brian Golding§

*The Biological Process Technology Institute and †Department of Ecology, Evolution, and Behavior, University of Minnesota; ‡Department of Biochemistry and Molecular Biology, FUHS/CMS, North Chicago; and §Department of Biology, McMaster University, Hamilton, Ontario

The determinants of site-to-site variability in the rate of amino acid replacement in α/β -barrel enzyme structures are investigated. Of 125 available α/β -barrel structures, only 25 meet a variety of phylogenetic and statistical criteria necessary to ensure sufficient data for reliable analysis. These 25 enzyme structures (from a wide variety of taxa with diverse lifestyles in diverse habitats) differ greatly in size, number, and topology of domains in addition to the α/β -barrel, quaternary structure, metabolic role, reaction catalyzed, presence of prosthetic groups, regulatory mechanisms, use of cofactors, and catalytic mechanisms. Yet, with the exception of ribulose-1,5-bisphosphate carboxylase, all structures have similar frequency distributions of amino acid replacement rates. Hence, site-specific variability in rates of evolution is largely independent of differences in biology, biochemistry, and molecular structure.

A correlation between site-specific rate variation and (1) distance from the active site, (2) solvent accessibility, and (3) treating glycines in unusual main-chain conformations as a separate class, explains approximately half the causal variation. Secondary structure exerts little influence on the pattern and distribution of replacements. Additional domains and subunits, side-chain hydrogen bonds, unusual side-chain rotamers, nonplanar peptide bonds, strained main-chain conformations, and buried hydrophilic-charged residues contribute little to variability among sites because they are rare. Nonlinear models do not improve the fits. In several enzymes, deviations from the typical pattern of replacements suggest the possible action of natural selection. A statistical analysis shows that, in all cases, much of the remaining unexplained variation is not attributable to chance and that other, as yet unidentified, causal relations must exist.

Introduction

Site-to-site variability in rates of evolution in molecular sequences presents both serious problems and wonderful opportunities for molecular evolutionists. Problems arise (examples discussed by Taneto, Takezaki, and Nei 1994; Yang 1994, 1996; Miyamoyo and Fitch 1996) when variability in rates affects the topologies of inferred phylogenies, causing incorrect classification, rejection of true null hypotheses, and (more insidiously) the generation of spurious evolutionary hypotheses. Yet opportunities also arise when variable rates are detected, for they suggest a glimpse at underlying evolutionary processes (Gu 1999, 2001) and constraints (Landgraf, Fischer, and Eisenberg 1999). For these and other reasons, the study of site-to-site variability in rates remains of central importance in molecular evolution (Thorne 2000).

A great deal of attention has been paid to the study of variable rates in DNA: between genes, between coding and noncoding sequences, between regulatory elements and their adjacent cistrons, and between nonsynonymous and synonymous substitutions within structural genes (Li 1997). Much discussion has centered on the role played by functional constraints in determining evolutionary rates, particularly with reference to the neutral theory (Kimura 1983). As previously noted (Dean and Golding 2000), many of these patterns are also consistent with Fisher's theory of evolution near fitness optima (Fisher 1930), wherein mutations of small

effect are more likely to be fitter than those of large effect—hence the former (e.g., synonymous substitutions) occur more frequently than the latter (e.g., nonsynonymous replacements).

Proteins, diverse in structure and in function, also form a natural arena in which to explore issues surrounding variability in evolutionary rates. After an early study demonstrating that amino acid replacement rates vary among sites in proteins (Uzzel and Corbin 1971), Kimura and Ohta (1973) established that sites lining the heme-binding pockets of hemoglobin evolve less rapidly than those on solvent accessible surfaces—a now oft cited example of functional constraint. In many recent studies high rates of amino acid replacement compared with the rates of silent substitution are taken as evidence of selection (methods and results reviewed by Yang and Bielawski 2000). A number of these studies reveal that rapidly evolving sites are localized within the three-dimensional structures of proteins, thereby providing additional insights into the mode of adaptive evolution (e.g., Hughes and Nei 1988; Bishop, Dean, and Mitchell-Olds 2000). Yet, between the extremes of casual inspection of protein structure and of rigorous application of statistical theory, there remains a vast gulf in our knowledge.

Incorporating protein structure into evolutionary models has recently become a focus of renewed interest (Thorne 2000). Bustamente, Townsend, and Hartl (2000) showed that polymorphic sites in several bacterial enzymes are far more likely to be on solvent accessible surfaces than in hydrophobic interiors, an observation entirely in accord with the observations of Kimura and Ohta (1973) and of Goldman and coworkers (Goldman, Thorne, and Jones 1998; Lio and Goldman 1999). The latter also attempted to extend the approach by incor-

Key words: amino acid replacement, evolution, rate, structure.

Address for correspondence and reprints: Antony M. Dean, The Biological Process Technology Institute, 240 Gortner Laboratories, 1479 Gortner Avenue, University of Minnesota, St. Paul, Minnesota 55108. E-mail: adean@biosci.umn.edu.

Mol. Biol. Evol. 19(11):1846–1864. 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

porating knowledge of secondary structures, but with mixed success. Atchley, Terhalle, and Dress (1999) and Atchley et al. (2000) used an information theoretic approach to analyze amino acid replacements in a DNA-binding helix-loop-helix domain of transcription factors. They found significant levels of covariation at surface sites that could not be ascribed to common phylogenetic history. Pollock, Taylor, and Goldman (1999) constructed models that explicitly incorporate covariation among sites and compared the fits with models that invoke no covariation.

In this article we explore two issues, building on an approach first used to analyze isocitrate dehydrogenase (Dean and Golding 2000). First, we determine when there is sufficient information in a protein phylogeny such that reliable inferences about the distribution of amino acid replacements can be made. Second, we explore the extent to which biological function affects the distribution of amino acid replacements within structures. To answer the first question we determine the proportion of variation in the number of amino acid replacements among sites that is attributable to causal effects. This is done using a new method that is independent of the structure of the underlying biological model, although it assumes an underlying Poisson process of amino acid replacement. To answer the second question we make a survey of proteins with an eightfold α/β -barrel, a motif that appears in a variety of structural and functional contexts.

The Model

We assume the following model for amino acid sequence evolution: among phylogenetically related amino acid sequences, different sites evolve independently along the branches of a given phylogenetic tree according to a Poisson process. Different sites are allowed to have different rates of evolution, and these rates may vary independently of others, both within and between branches. We do not use empirical correction matrices because these mask heterogeneity that is rightfully attributable to three-dimensional structural effects. When conditioned on phylogenetic history (tree topology, rates of evolution, and changes in rates of evolution), the total number of replacements at a given site for a sample of amino acid sequences is still Poisson distributed. We assume that at site i , the mean number of replacements over the entire history represented in the phylogenetic tree is μ_i .

Suppose we align a number of amino acid sequences, reconstruct their phylogeny, and infer the actual number of replacements per site. There are two distinct sources for variation in this data set: one is biological, namely the site-to-site variation that is caused by evolutionary forces that determine the rate of replacement, whereas the second is probabilistic, namely the inevitable variation that accompanies stochastic processes—in this case Poisson processes that determine the number of replacements per site, given the rate of replacement. Because both sources contribute to variation, it is important to determine what fraction of variation is due to

the stochastic process and what fraction is due to biological forces determining the site-to-site variation in replacement rates. Fortunately, in the case of Poisson noise, it is possible to tease apart these two sources of variation without recourse to a biological model of protein evolution, i.e., a model that seeks to explain the between-site variation in replacement rates.

Let Y_i be the random number counting the number of replacements at site i , $i = 1, 2, \dots, n$, accumulated throughout phylogenetic history. Each Y_i is Poisson distributed with mean μ_i for site i . Sites are independent of each other. Setting $\bar{\mu} = \sum_{i=1}^n \mu_i/n$, we can partition the variance in the number of replacements among all sites into two parts, the first ascribable to causal variation among sites and the second ascribable to residual stochastic error due to the Poisson process acting at each site:

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{\mu})^2 &= \sum_{i=1}^n (\mu_i - \bar{\mu})^2 + \sum_{i=1}^n (Y_i - \mu_i)^2 \\ &\quad + 2 \sum_{i=1}^n (Y_i - \mu_i)(\mu_i - \bar{\mu}) \end{aligned}$$

If we define $\sigma^2 = \sum_{i=1}^n (Y_i - \bar{\mu})^2/n$, $\sigma_{sites}^2 = \sum_{i=1}^n (\mu_i - \bar{\mu})^2/n$, and $\sigma_{error}^2 = E[\sum_{i=1}^n (Y_i - \mu_i)^2]/n$, then, after taking expectations and using $EY_i = \mu_i$,

$$\sigma^2 = \sigma_{sites}^2 + \sigma_{error}^2$$

Rearranging yields

$$\frac{\sigma_{sites}^2}{\sigma^2} = 1 - \frac{\sigma_{error}^2}{\sigma^2}$$

The distribution of replacements among sites (σ_{sites}^2) comes from the biological model. The distribution of replacements at each site is Poisson with mean equal to the variance ($\mu_i = \sigma_i^2$). Because sites are independent, it follows that $\bar{\mu} = \sigma_{error}^2$. Therefore,

$$\frac{\sigma_{sites}^2}{\sigma^2} = 1 - \frac{\bar{\mu}}{\sigma^2}$$

The ratio on the left-hand side is the proportion of the total variance among sites attributable to causal differences between sites. The ratio on the right-hand side is the proportion of the variance attributable to Poisson error. The expression on the right-hand side can be interpreted as a coefficient of determination, denoted by ρ^2 , an unbiased estimate of which (see Appendix at MBE web site: www.molbioevol.org) is given by

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

where $\bar{y} = \sum_{i=1}^n y_i/n$ and $s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n - 1)$ with y_i denoting the realized number of replacements at site i . For ease of reference, we will denote the Poisson estimated coefficient of determination, $\hat{\rho}^2$, by PECD. Hence, whenever errors are Poisson distributed, PECD can be computed from the sample mean and the sample variance of the data.

We also show (see appendix) that the variance of $\hat{\rho}^2$ is given (approximately) by

$$s_{\hat{\rho}^2}^2 = \left(\frac{\bar{y}}{s_y^2}\right)^2 \left(\frac{\text{Var}(\bar{y})}{\bar{y}^2} + \frac{\text{Var}(s_y^2)}{(s_y^2)^2} - 2 \frac{\text{Cov}(\bar{y}, s_y^2)}{\bar{y} \cdot s_y^2} \right) \quad (2)$$

upon substituting the exact solutions

$$\begin{aligned} \text{Var}(\bar{y}) &= \frac{\bar{y}}{n} \\ \text{Var}(s_y^2) &= \frac{\bar{y}}{n} + \frac{2}{n-1} \bar{y}^2 + \frac{(6n-8)}{n(n-1)} s_{sites}^2 \\ &\quad + \frac{4}{n-1} \bar{y} \cdot s_{sites}^2 + \frac{4}{n-1} \hat{\xi}_{sites}^3 \\ \text{Cov}(\bar{y}, s_y^2) &= \frac{\bar{y}}{n} + \frac{2}{n} s_{sites}^2 \\ \hat{\xi}_{sites}^3 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^3}{n-1} - \frac{(n-2)}{n} \bar{y} \\ &\quad - \frac{3(n-2)}{n} s_{sites}^2 \end{aligned}$$

where $s_{sites}^2 = s_y^2 - \bar{y}$ is the variation attributable to causal differences among sites, $\text{Cov}(\bar{y}, s_y^2)$ is the covariance of sample means and sample variances, and $\hat{\xi}_{sites}^3$ is the skewness attributable to causal differences among the n sites. The variance of s_{sites}^2 is estimated as

$$\begin{aligned} \text{Var}(s_{sites}^2) &= \text{Var}(s_y^2 - \bar{y}) = \frac{2}{n-1} \bar{y}^2 + \frac{2(n-2)}{n(n-1)} s_{sites}^2 \\ &\quad + \frac{4}{n-1} \bar{y} \cdot s_{sites}^2 + \frac{4}{n-1} \hat{\xi}_{sites}^3 \end{aligned} \quad (3)$$

Reducing Error

The variation due to Poisson error can be reduced by increasing sample size. The approximate increase in sampling size can be computed assuming constant rates of evolution. Suppose the total branch length in the tree is t and $\mu_i = \lambda_i t$, then the total variance is

$$\sigma^2 = \sigma_{sites}^2 + \sigma_{error}^2 = \sum_{i=1}^n \frac{(\lambda_i t - \bar{\lambda} t)^2}{n} + \bar{\lambda} t$$

where $\bar{\lambda} = \sum_{i=1}^n \lambda_i / n$. The proportion of variance attributable to the Poisson process is

$$\frac{\sigma_{error}^2}{\sigma^2} = \frac{\bar{\lambda} t}{(\bar{\lambda} t)^2 \sum_{i=1}^n (\lambda_i / \bar{\lambda} - 1)^2 / n + \bar{\lambda} t} = \frac{1}{\bar{\mu} (\sigma_{\lambda} / \bar{\lambda})^2 + 1}$$

where $(\sigma_{\lambda} / \bar{\lambda})^2$ is the coefficient of variation attributable to the sites. Rearranging yields

$$\bar{\mu} = \frac{(\sigma_{error}^2 / \sigma^2)^{-1} - 1}{(\sigma_{\lambda} / \bar{\lambda})^2}$$

The fold-increase in the data (η) necessary to reduce the proportion of the stochastic error from $\sigma_{error,old}^2 / \sigma_{old}^2$ to $\sigma_{error,new}^2 / \sigma_{new}^2$ is given by

$$\eta = \frac{\bar{\mu}_{new}}{\bar{\mu}_{old}} = \frac{(\sigma_{error,new}^2 / \sigma_{new}^2)^{-1} - 1}{(\sigma_{error,old}^2 / \sigma_{old}^2)^{-1} - 1} \quad (4)$$

To illustrate this, assume that $\bar{y}_{old} / s_{old}^2 = 0.5$, i.e., 50% of the variation is attributable to Poisson noise. To reduce the error from 0.5 to 0.05, the mean number of replacements must be increased by a factor of

$$\hat{\eta} = \frac{\bar{y}_{new}}{\bar{y}_{old}} = \frac{(0.05)^{-1} - 1}{(0.5)^{-1} - 1} = 19$$

with approximate variance

$$s_n^2 = \left(\frac{\sigma_{new}^2}{\sigma_{error,new}^2} - 1 \right) \left(\frac{\bar{y}_{old}}{s_{old}^2 - \bar{y}_{old}} \right)^2 \left(\frac{\text{var}(s_{sites}^2)}{(s_{sites}^2)^2} - \frac{3}{\bar{y}_{old} n} \right) \quad (5)$$

that is an underestimate because $\sigma_{error,new}^2 / \sigma_{new}^2$ is treated as fixed, whereas $\bar{y}_{new} / s_{new}^2$ is subject to (usually small) sampling errors.

Using PECD

The PECD $\hat{\rho}^2$ has three uses: (1) identifying data sets with sufficient causal variation to be worthy of analysis ($\hat{\rho}^2 \rightarrow 1$), (2) determining how much more data need be collected (η) to reduce the stochastic variation to some desirable limit, and (3) knowing when to stop tinkering with a regression model constructed from biological variables because its correlation coefficient (\hat{r}) has approached the theoretical limit ($\hat{r} \rightarrow \hat{\rho}$).

PECD helps identify those data sets most worthy of analysis. When all sites in a sequence evolve at the same rate, the expected distribution of replacements is Poisson, with the expected variance (σ^2) equal to the expected mean (μ). The ratio of the estimated variance (s_y^2) to the estimated mean (\bar{y}) weighted by the degrees of freedom, $(n-1)s_y^2/\bar{y}$, is approximately distributed as χ_{n-1}^2 and provides a convenient test for deviations from Poisson (Fisher 1948). Yet, significance alone is insufficient a criterion to pursue an analysis. For example, with $df = 400$ (common enough with molecular data) the variance need only be 14% larger than the mean to be significant, resulting in a PECD $1 - 1/1.14 = 0.12$ or 12% causal variation. Yet a data set with 12% causal variation is hardly worthy of detailed analysis when there may be others consisting of 90% causal variation.

Additional data must often be gathered in an effort to reduce stochastic noise to an acceptable level. There are two possible strategies: increasing sequence length or obtaining more sequences. The first is of limited use because interest often centers on sequences of defined length (e.g., a gene), and where lengthening is possible it does nothing to reduce the stochastic proportion of the estimated variance ($\bar{y}/s_y^2 = 1 - \hat{\rho}^2$), whereas its variance ($\text{Var}(\bar{y}/s_y^2) = \sigma_{\hat{\rho}^2}^2$) is reduced only in direct proportion to the increase in length. The second approach is far more efficient, even though the degrees of freedom remain unchanged. The stochastic portion of the variance decreases (roughly) as \bar{y} , whereas its variance decreases (roughly) as \bar{y}^3 . For this second approach, η provides a

Table 1
Theoretical, Simulated, and Estimated Coefficients of Determination from Simulated Protein Data ($n = 245$)

Ex-pected Mean	Theoretical ^b $\rho^2 \pm \sigma_{\rho^2}$	Simulated ^c $r^2 \pm \text{SD}$	Estimated ^d $\hat{\rho}^2 \pm s_{\hat{\rho}^2}(\pm \text{SD})$
0.853	0.382 ± 0.060	0.388 ± 0.050	$0.372 \pm 0.059 (\pm 0.061)$
1.706	0.553 ± 0.038	0.554 ± 0.041	$0.541 \pm 0.038 (\pm 0.037)$
3.412	0.712 ± 0.019	0.712 ± 0.030	$0.710 \pm 0.020 (\pm 0.019)$
6.823	0.832 ± 0.009	0.833 ± 0.018	$0.831 \pm 0.009 (\pm 0.009)$
13.650	0.908 ± 0.004	0.909 ± 0.011	$0.907 \pm 0.004 (\pm 0.004)$
27.290	0.952 ± 0.001	0.952 ± 0.006	$0.952 \pm 0.001 (\pm 0.001)$

^a Mean number of replacements per residue.

^b Calculated from the theoretical model, taking the inferred number of replacements at each site in triose phosphate isomerase as the Poisson expectation.

^c The mean and SD of 10,000 coefficients of determination obtained by regressing simulated data against their expectations.

^d The mean PECD and mean approximate SD together with its observed SD in parentheses.

useful, if approximate, gauge of the necessary increase in sample size—approximate only because the precise increase also depends on sequence relatedness and because very small initial samples necessarily produce large standard errors (table 2).

PECD also provides a means to ascertain when the coefficient of determination in the biological regression model (denoted by \hat{r}^2) needs further refinement ($\hat{r}^2 < \hat{\rho}^2$) and when a satisfactory explanation has been reached ($\hat{r}^2 \rightarrow \hat{\rho}^2$). Normalizing the coefficient of determination that comes from the biological model to the coefficient of determination that comes from partitioning the total variation into the variation due to the Poisson error and the variation that needs a causal explanation, provides a convenient measure of quality. We estimate the normalized coefficient of determination (NCD) as $\hat{q}^2 = \hat{r}^2/\hat{q}^2$, the proportion of variation that the model explains divided by the proportion of variation that needs an explanation. This preserves the intuitive scale from 0 to 1. As an example, compare the NCD $\hat{q}^2 = \hat{r}^2/\hat{q}^2 =$

$0.2/0.21 = 0.95$ to $\hat{q}^2 = \hat{r}^2/\hat{q}^2 = 0.4/0.95 = 0.42$. Even though the former $\hat{r}^2 = 0.2$ is much lower than the latter $\hat{r}^2 = 0.4$, efforts should be concentrated on improving the second model.

Simulations

To ascertain the reliability of our analysis when applied to proteins, we simulated the accumulation of amino acid replacements at 245 sites in the glycolytic enzyme triosephosphate isomerase (TIM). Analysis (see *Methods*) of 178 sequences from extant taxa allocates 90% of the site-to-site variability to causal effects. With 10% of the variation attributed to stochastic effects, the observed distribution provides a sufficiently robust estimate of the true underlying distribution for simulating Poisson scatter.

The mean and standard deviations (SD) of 2,000 replicate simulations (table 1), with the expected mean number of replacements per site varied between 0.853 and 27.29, reveal that the simulated $\hat{\rho}^2$ closely follows both the theoretical ρ^2 and the simulated r^2 . When there are few replacements, $\hat{\rho}^2$ underestimates r^2 . However, the bias is negligible compared with the SD values (table 1) and the 95% confidence intervals (fig. 1). Table 1 also shows that the approximate SD values of $\hat{\rho}^2$ ($s_{\hat{\rho}^2}$, from eq. 2) closely follow the empirically determined SD values and the approximate SD values of ρ^2 (σ_{ρ^2}). With many replacements, $s_{\hat{\rho}^2}$, SD, and σ_{ρ^2} are smaller than the empirically determined SD of \hat{r}^2 .

Table 2 presents the mean and SD values of 1,000 replicate simulations of η , the fold-increase in the sample size necessary to reduce the Poisson sampling variance to a predetermined fraction of the total (in this instance 0.05). The smaller the sample, the larger the fold-increase necessary and the less reliable its estimate. For very small samples, where the sampling variance accounts for more than 50% of the observed variance, the increases in sample size are large and poorly estimated. All realized SD values are slightly larger than

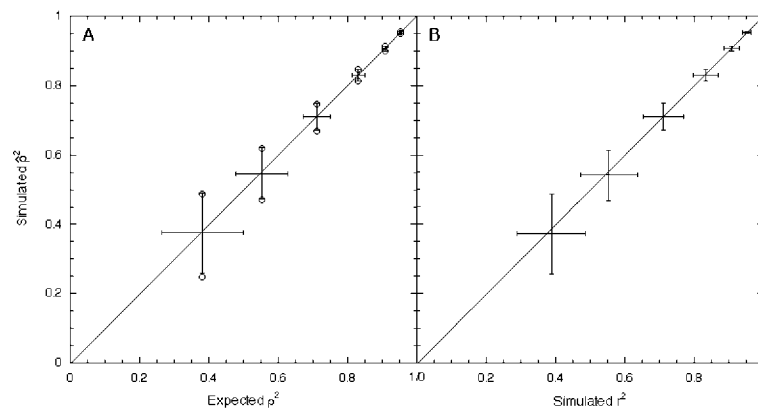


FIG. 1.—Simulations of the PECD using the numbers of amino acid replacements accumulated during the course of evolution in TIM as Poisson variates. (A) The PECD ($\hat{\rho}^2$) is an unbiased estimate of the true coefficient of determination (ρ^2) and that the approximate 95% confidence intervals (vertical lines) are not only similar to those of true coefficient of determination (horizontal lines) but are also close to those obtained empirically (circles) by excluding the upper and lower 2.5% of 2,000 replicate simulations. (B) The PECD provides an accurate estimate of the simulated coefficient of determination (r^2): the slight underestimate seen at low values is swamped by the range of the 95% confidence intervals, whereas at high values the underestimate is negligible. As judged by the 95% confidence intervals, the PECD provides a more reliable estimate of the proportion of variation attributable to causal effects than does the simulated coefficient of determination.

Table 2
Fold-Increase (η) in Sample Mean (\bar{y}) Necessary to Reduce Sampling Errors (\bar{y}/s^2_y) to Less Than 5% for Simulated Protein Data ($n = 245$)

Sample Mean \bar{y}	Sampling Error \bar{y}/s^2_y	Theoretical $\eta \pm \sigma_{\hat{\eta}}$	Simulated $(1 + \bar{y}_{new}/\bar{y}_{old}) \pm \text{SD}$	Estimated $\hat{\eta} \pm s_{\hat{\eta}} (\pm \text{SD})$
0.853	0.618	30.88 \pm 7.83	33.42 \pm 10.25	33.02 \pm 9.48 (± 9.15)
1.706	0.447	15.44 \pm 2.34	15.71 \pm 2.79	15.67 \pm 2.45 (± 2.47)
3.412	0.288	7.72 \pm 0.74	7.79 \pm 0.89	7.78 \pm 0.76 (± 0.76)
6.823	0.168	3.86 \pm 0.25	3.87 \pm 0.29	3.87 \pm 0.25 (± 0.26)
13.650	0.092	1.93 \pm 0.08	1.93 \pm 0.09	1.93 \pm 0.08 (± 0.08)

the predicted ones because they include variability associated with gathering the additional data (i.e., variability in $\sigma_{error.new}^2/\sigma_{new}^2$). This bias hardly matters when the target error is small ($\sigma_{error.new}^2/\sigma_{new}^2 < 0.05$), but it becomes increasingly important for larger target errors ($\sigma_{error.new}^2/\sigma_{new}^2 > 0.2$).

α/β -Barrel Proteins

α/β -Barrels (fig. 2) are large structures of at least 200 amino acids, with eight parallel β -strands forming a hub surrounded by a tire of eight α -helices (Branden and Tooze 1999). Connected in a repeating loop-strand-loop-helix motif, the β -strands and α -helices of different enzymes are readily superimposed but the loops are not, so variable are they in length and conformation. Sometimes the loops form entirely independent domains with their own hydrophobic cores (e.g., pyruvate kinase in fig. 3). Other structural variations on the α/β -barrel theme include: enolase which contains a single antiparallel β -strand, isocitrate lyase in which an α -helix from a second subunit completes the tire of the barrel, and quinolate phosphoribosyltransferase (fig. 3) which is a partial barrel.

α/β -barrels are common structures appearing in approximately 10% of unique enzymes in the Protein Data Bank (Gerlt 2000). Why so many enzymes should have this structure and why all have their active sites located at the carboxy termini of the β -strands are still matters of intense speculation. Those that are clearly homologous, as judged by shared vestigial sequence identities, clear-cut structural similarities, and related catalytic chemistries, can be classified into superfamilies (Babbitt and Gerlt 1997). Attempts to classify ever more dissim-

ilar superfamilies (Copley and Bork 2000) risk mistaking structural and functional constraints as evidence of homology.

α/β -barrel enzymes have diverse functions (table 3). At the organismal level they contribute to photosynthesis, respiration, cell growth, development, defense, and communication. They may be extracellular or confined to certain organelles. At the metabolic level they play roles in glycolysis and gluconeogenesis, CO_2 fixation, assorted biosyntheses and degradations, DNA repair, and bioluminescence. They carry out a diversity of biochemical transformations, including C–C bond formation, oxidations and reductions, hydrolyses and condensations using a variety of chemical mechanisms (Walsh 1979). Reactions may proceed concertedly (everything happens simultaneously) or sequentially (in a stepwise fashion). The transition states vary widely in chemical character, from enolates to radicals to oxocarbenium ions. These transition states may be stabilized directly by the protein or indirectly by way of divalent metals or pyridoxal phosphate. In the case of chitinase there is even the suggestion that the substrate itself directly assists catalysis (Terwisscha van Scheltinga et al. 1995). In some, the substrate becomes temporarily covalently attached to the enzyme or coenzyme (e.g., through a lysyl or pyridoxal phosphate Schiff-base, or the formation of acyl-enzyme intermediate). In others, the substrates are noncovalently bound throughout the reaction.

Methods

Structures, Sequences, and Alignments

Proteins containing α/β -barrels were identified using the SCOP classification (Lo Conte et al. 2000; <http://>

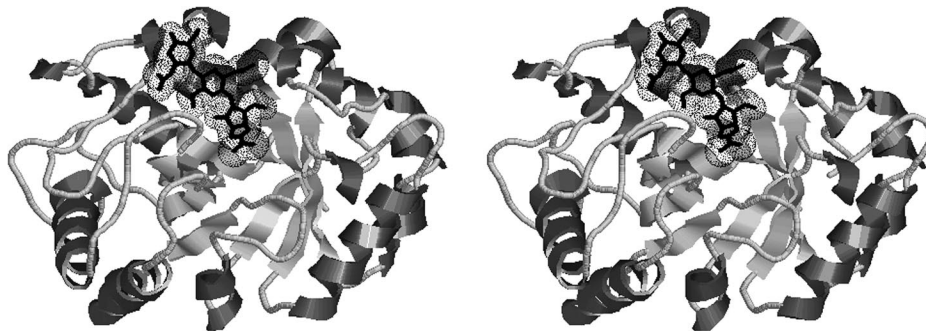


FIG. 2.—With a hub of eight parallel β -sheets surrounded by a tire of eight α -helices, monomeric plant acidic chitinase has a canonical α/β -barrel. The active site pocket, with the inhibitor allosamidin bound, lies at the carboxy termini of the β -sheets as it does in all α/β -barrel enzymes.

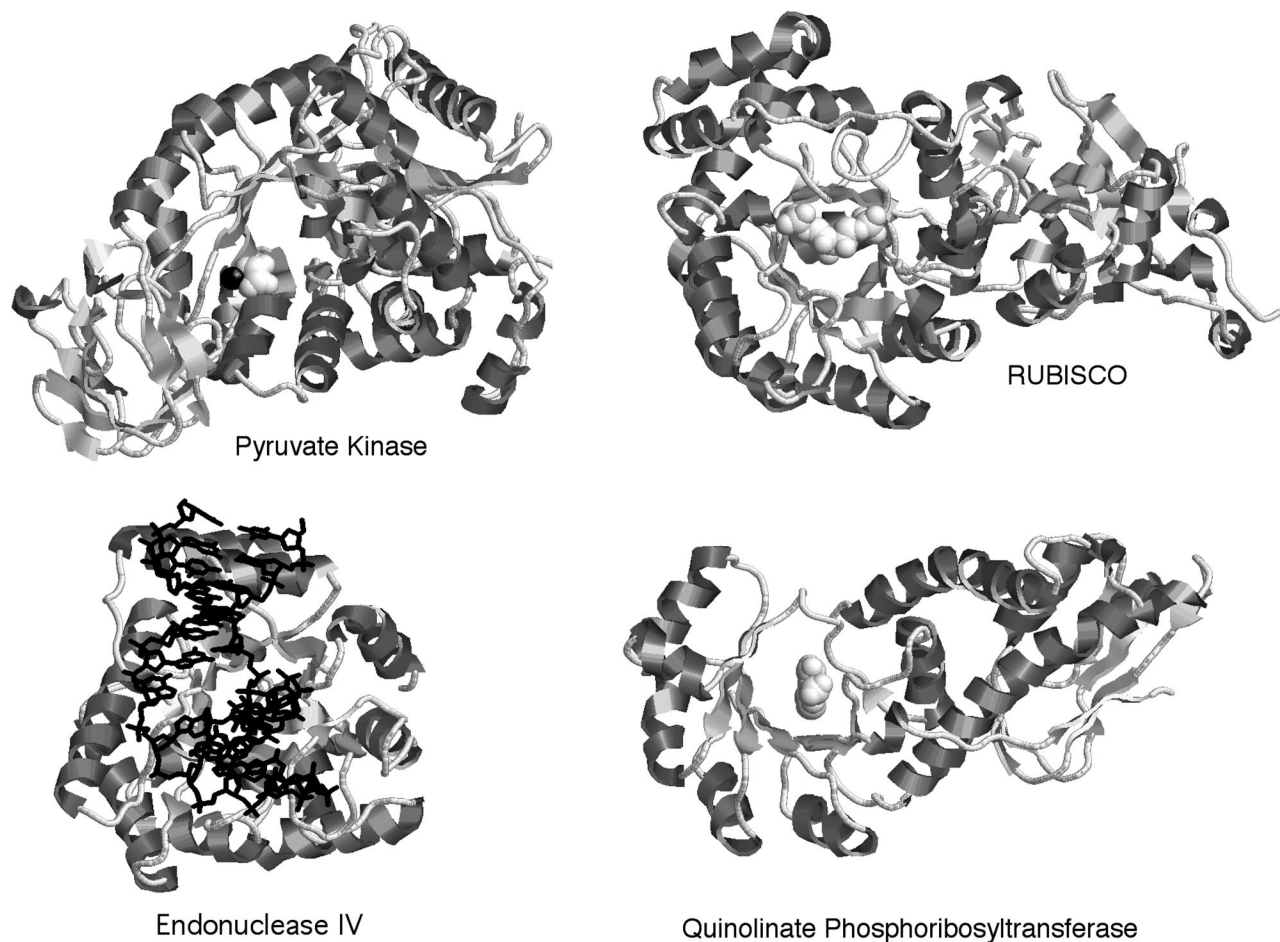


FIG. 3.—Subunits containing α/β -barrels (viewed end on) vary greatly, from the simple unembellished structure of endonuclease IV, to the large triple domain of pyruvate kinase, to the partial α/β -barrel of quinolate phosphoribosyltransferase. Substrates also vary greatly in size, from the tiny CO_2 of RUBISCO to the gigantic chromosomal DNA of endonuclease IV. DNA is represented by sticks, with other ligands, all bound in active sites, represented by van der Waals models.

scop.mrc-lmb.cam.ac.uk/scop/data/scop.1.html), the NCBI/Entrez database of structural alignments, and PUBMED searches using keywords “TIM barrel” or “alpha/beta barrel” (both accessible at <http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi>). Atomic coordinates for structures were downloaded from the Protein Data Bank (<http://www.rcsb.org/pdb/>). Where possible, structures with bound transition state analogues or covalently attached “suicide” substrates were chosen because they better represent enzymes undergoing catalysis. Otherwise, preference was given to structures with bound substrates, products, or inhibitors because these better represent Michaelis complexes.

Amino acid sequences homologous to those in the Protein Data Bank were identified using web-based gapped-BLAST (Altschul et al. 1997; <http://www.ncbi.nlm.nih.gov:80/BLAST>) and Neighbors (<http://www.ncbi.nlm.nih.gov:80/entrez>) algorithms. All mutant and chimeric sequences were discarded, as were short peptide fragments (<150 residues). Sequences were aligned using CLUSTALW (Thompson, Higgins, and Gibson 1994; Higgins, Thompson, and Gibson 1996; code and documentation available at <ftp://ftp.bio.indiana.edu/molbio/align/clustal/>). All sequences

less than 40% identical to a known structure were discarded. Only a single wild-type representative in clusters of sequences sharing more than 99% identity was retained. Partial sequences were removed, unless the missing sites comprised less than 5% of the sequence (commonly seen at the amino and carboxy termini). For many enzymes the structures from several different species are available. Structural alignments, obtained from web-based databases and programs (the MMDB database at <http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml>, VAST alignments at <http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>), CE (Shindyalov and Bourne 1998) at <http://cl.sdsc.edu/>, FSSP (Holm and Sander 1996) at <http://www2.ebi.ac.uk/dali/fssp/>, or locally with QUANTA (MSI, CA), were all similar and used as guides to adjust alignments with the SEQPUP editor (D. Gilbert 1999; <http://iubio.bio.indiana.edu/soft/molbio/seqpup/java>). Searches were completed on August 1, 2000.

Phylogenetic Reconstructions

Phylogenies were constructed using the Fitch-Margoliash least squares method (Fitch and Margoliash

Table 3
General Properties of Enzymes Analyzed

Enzyme	Metabolic Role(s)	Location(s)	Cofactors	Chemical Mechanism(s)	Key Intermediates or Transition States (or Both)	Allosteric Regulators	Modifications	Subunits	PDB
(De)carboxylases									
Ribulosebiphosphate carboxylase-oxygenase	Carbon dioxide fixation in photosynthesis, photorespiration	Chloroplast	Me ⁺⁺	CO ₂ or O ₂ fixed in ribulosebiphosphate then Aldol-like cleavage	Me ⁺⁺ stabilized enediolate?			8A,8B	1RCX
Phosphoenolpyruvate carboxylase	Photosynthesis (CAM and C4), TCA ana-pleurosis	Chloroplast	Me ⁺⁺	CO ₂ fixation into PEP with phosphate release	Me ⁺⁺ stabilized carbanion?	Aspartate, malate	Phosphorylated	4A	1FIY
Dehydratases									
Enolase	Glycolysis, gluconeogenesis		Me ⁺⁺	Dehydration	Me ⁺⁺ stabilized carbanion			2A	2ONE
Dehydrogenases									
Diketo-D-gluconic acid reductase	L-Ascorbate (vitamin C) biosynthesis			Acid catalysis with NADPH oxidation	Concerted using cationic intermediate			1A	1A80
3- α -hydroxysteroid dehydrogenase	Steroid inactivation			Acid catalysis with NADPH oxidation	Concerted using cationic intermediate			2A	1LW1
Aldose/aldehyde reductase	Various			Acid catalysis with NADPH oxidation	Concerted using cationic intermediate			1A	1EL3
Hydrolases									
Chitinase/lysozyme (family 18)	Chitin/murin hydrolysis β (1-4) endohydrolase	Extracellular		β -retaining double displacement mechanism via α -glycosyl-enzyme intermediate	Oxocarbenium ion-like Substrate assisted?			1A	1LLO
1,3-1,4 β -glucanase (family 17)	Plant endosperm cell wall β (1-4) endohydrolase	Extracellular		β -retaining double displacement mechanism via α -glycosyl-enzyme intermediate	Oxocarbenium ion-like		Glycosylated	2A	1AQ0
α -amylase (family 13)	Starch/glycogen metabolism, α (1-4) endohydrolase	Some Extracellular		α -retaining double displacement mechanism via β -glycosyl-enzyme intermediate	Oxocarbenium ion-like			1A	1AMY
Cyclodextrin glycotransferase (family 13)	Intramolecular α (1-4) transglycosylation			α -retaining double displacement mechanism via β -glycosyl-enzyme intermediate	Oxocarbenium ion-like			1A	1A47
Xylanase (family 10)	Xylan β (1-4) endohydrolase	Extracellular		β -retaining double displacement mechanism via β -glycosyl-enzyme intermediate	Oxocarbenium ion-like			1A	1B3Z
Isomerases									
Xylose isomerase	Aldose-ketose isomerization		Me ⁺⁺	Ring opens; 1,2 hydride shift; ring closure	Carbocation			4A	1BHW

Table 3
Continued.

Enzyme	Metabolic Role(s)	Location(s)	Cofactors	Chemical Mechanism(s)	Key Intermediates or Transition States (or Both)	Allosteric Regulators	Modifications	Subunits	PDB
Triose phosphate isomerase	Glycolysis, Gluconeogenesis, Calvin Cycle			Concerted acid-base catalyzed half reactions	Enediolate			2A	4TIM
Lyases									
Fructose 1,6 biphosphate aldolase (class 1)	Glycolysis, gluconeogenesis, Calvin cycle			Aldol cleavage, condensation	Lysyl Schiff-base			4A	1A5C
Fructose 1,6 biphosphate aldolase (class 2)	Glycolysis, gluconeogenesis, Calvin cycle		Me ⁺⁺	Aldol cleavage, condensation	Me ⁺⁺ stabilized enolate			2A	1B57
Tryptophan synthase (α -subunit)	Tryptophan biosynthesis			Aldol cleavage	Oxyanion			2A,2B	1BEU
5-aminolevulinat dehydratase	Porphyrin biosynthesis		Me ⁺⁺	Aldol condensation then cyclization by transaldimination	Lysyl Schiff-base gemdimine			8A	1B4E
Dihydrodipicolinate synthase	Lysine biosynthesis			Aldol condensation then cyclization by transaldimination	Lysyl Schiff-base	Lysine		4A	1DHP
Malate synthase G	Carbon fixation from acetate		Me ⁺⁺	Claisen condensation	Enolate			4A	1D8C
3-deoxy-D-arabinoheptulosonate	Aromatic amino acid biosynthesis		Me ⁺⁺	Aldollike condensation	Me ⁺⁺ stabilized enolate?	Phenylalanine, tyrosine, tryptophan		4A	1QR7
7-phosphate synthase	Carbon fixation from acetate								
Isocitrate lyase	Carbon fixation from acetate	Glyoxysome	Me ⁺⁺	Aldollike cleavage	Carbanion	PEP		4A	1dqu
Transferases									
Pyruvate kinase	Glycolysis		Me ⁺⁺	Transfer of phosphate from PEP to ADP	Me ⁺⁺ stabilized enolate	Alanine, fructose 1,6 biphosphate	Regulatory phosphorylation	4A	1A3W
tRNA-guanamine transferase	Substitution of wobble guanine with 7-aminomethyl-7-deazaguanine			β -Retaining double displacement mechanism using α -glycosyl-enzyme intermediate	Oxocarbenium ion-like			2A	1F3E
Quinolate phosphoribosyltransferase	NAD biosynthesis		Me ⁺⁺	NMN from quinolate and PRPP with anionic inversion from α to β	Oxycarbenium ion-like			2A	1QAP
Transaldolase	Pentose phosphate pathway, Calvin cycle			Aldol cleavage and condensation	Lysyl Schiff-base			2A	1UCW

1967), as implemented in PHYLIP (Felsenstein 1989; code and documentation available at <http://evolution.genetics.washington.edu/phylip.html>). Distances were calculated using a PAM250 matrix (Jones, Taylor, and Thornton 1992), with gaps treated as missing data. Branches longer than 0.3 were pruned, and clusters having five or more sequences were analyzed separately. For each cluster, the phylogeny was repeatedly reconstructed and pruned until all branch lengths were shorter than 0.3. Phylogenies were further pruned so that no more than 30% of their total lengths comprised branches longer than 0.2.

Inferred Amino Acid Replacements

For each Fitch-Margoliash tree the number of amino acid replacements per site was inferred by parsimony using PAUP* (Swofford 1998). Sites absent in the protein structure were discarded as were those present in the structure but absent in more than 10% of aligned homologous sequences.

Parsimony systematically underestimates the number of replacements per site. We therefore implemented a Jukes-Cantor-like correction to adjust for this bias. Following Gu and Zhang (1997), the probability of one or more amino acid replacements at site i on branch j (y_{ij}) when governed by a Poisson process is given by

$$P(y_{ij} > 0) = 1 - e^{-\lambda_i t_j}$$

where λ_i is the site-specific rate of amino acid replacement, and t_j is the length of branch j . Note that this correction assumes evolutionary rates are constant, in the sense that all rates are proportional to branch lengths t_j . The expected number of branches with replacements at site i (b_i) is obtained by summing over all m branches in the phylogeny,

$$b_i = \sum_{j=1}^m P(y_{ij} > 0) = \sum_{j=1}^m (1 - e^{-\lambda_i t_j})$$

with the expected number of branches receiving no replacements given by

$$m - b_i = \sum_{j=1}^m e^{-\lambda_i t_j} \quad (6)$$

On the left-hand side is the difference between the number of branches (m) in the phylogeny and the observed number of branches with at least one replacement (b_i). On the right-hand side, estimates of t_j are provided by the Fitch-Margoliash tree. There is no general analytic solution for λ_i (this equation being similar in form to Euler equations), and therefore we found its numerical value using the RootFind function in Mathematica (Wolfram Research, Inc., IL). The corrected number of replacements per site (y_i) is then estimated as

$$y_i = \lambda_i \sum_{j=1}^m t_j \quad (7)$$

Simulations (see *Results*) demonstrate that this correction accurately recovers the mean and variance in num-

ber of replacements when branch lengths are constrained as described above.

Protein Characterization

An SGI Indigo II (Mountain View, CA) running Quanta (MSI, CA) software was used to calculate H-bonds and ϕ and ψ angles and to define secondary structure from PDB files. Quanta was also used to identify those side-chains engaged in ionic interactions and those engaged in H-bonding, and whether the latter were side-chain to side-chain, side-chain to main-chain, interdomain or intersubunit interactions, whether they were donors or recipients, and whether the atoms involved were charged or polar (or both). The fraction of each amino acid side-chain exposed to the solvent was calculated using a 0.01-Å grid with a 1.4-Å radius probe (the diameter of water) in Quanta. The distance (Å) from the atom in each residue closest to the active site (taken to be an atom implicated in catalysis from mechanistic considerations and which, depending on context, may reside on a side-chain, in a prosthetic group, in a bound ligand or be a bound metal ion) was calculated from the x, y, z atomic coordinates using the calculator in JMP (SAS Institute Inc., NC). A similar calculation was performed with ligands bound in allosteric sites.

Regression Analyses

Linear least squares regression models of the form

$$y_i = a_0 + \sum_{k=1}^n a_k x_{ki}$$

were fitted to each data set, where y_i is the corrected number of replacements at site i , a_0 is the intercept, a_k are regression coefficients, x_{ki} are independent variables, and n is their number. Preliminary investigations suggested a minimal model of three terms that were highly significant across all enzymes:

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

where distance_i from the active site is measured in Å, access_i is the fraction of an individual amino acid side-chain exposed to solvent, and $\psi\phi\text{Gly}_i = 1$ if the bond angles at a Gly residue lie in regions where $\psi > -40^\circ$ or $\phi < -70^\circ$ (regions normally unoccupied by residues with side-chains) and $\psi\phi\text{Gly}_i = 0$ otherwise. Additional fits to the larger phylogenies were achieved using linear models supplemented with additional terms in a_i (for H-bonding, the class of residue occupied at a site—hydrophobic, aromatic, polar, charged—proximity to subunit and domain interfaces, to solvent filled cavities, and microcavities etc.) or with second- or third-order interaction terms (e.g., $\text{access} \times \text{distance} \times \psi\phi\text{Gly}$). A nonlinear model of the form

$$y_i = a_1 + \frac{a_2 \text{distance}_i^{b_2}}{c_2 + \text{distance}_i^{b_2}} + \frac{a_3 \text{access}_i^{b_3}}{c_3 + \text{access}_i^{b_3}} + a_4 \psi\phi\text{Gly}_i,$$

which allows sigmoidal fits, was also investigated for large data sets.

Results

Reliability of the Data

Of the approximately 125 α/β -barrel enzymes available in the Protein Data Bank, just 25 appear in table 4, an attrition required to isolate stringently reliable data sets. Enzymes in table 4 have the following attributes: (1) a Fitch-Margoliash phylogeny with an average of at least 1.5 amino acid replacements per site, (2) they are constructed from at least five sequences, (3) all sequences are less than 99% identical and, (4) each sequence is more than 40% identical to a sequence with a known structure with (5) no branch longer than 0.3 (mean number of replacements/site), and (6) have no more than 30% of the amino acid replacements assigned to branches of length greater than 0.2 (mean number of replacements/site). Criteria 1, 2, and 3 eliminate small phylogenies of few replacements, criterion 4 ensures all primary sequences are sufficiently similar such that major differences in protein structure are not likely to be present, and criteria 5 and 6 limit, but do not eliminate, the bias toward underestimating the number of replacements per site.

Parsimony estimates of the number of replacements per site are relatively insensitive to the precise topology of a phylogeny. With TIM, the correlation coefficient obtained with the Fitch-Margoliash tree and a randomly chosen maximum parsimony tree (one of 204 found using the heuristic search in PAUP*) is 0.997. This result is both typical and expected—variations in tree topology occur at nodes supported by few replacements, with the consequence that most replacements at most sites remain unaffected.

Corrections

Parsimony assigns no more than one replacement per site per branch. This causes the number of replacements to be systematically underestimated, particularly at rapidly evolving sites on long branches. The bias is inescapable, although pruning long branches (>0.3) and restricting those of modest length ($0.2 < t_i < 0.3$) to no more than 30% of the total length of the phylogeny minimizes serious underestimates. Nevertheless, the Jukes-Cantor-like correction produces a 15% increase in the mean number of replacements per site (table 4), with increases exceeding 30% at approximately 4% of sites. The need to implement the correction necessitated determining its accuracy.

We used computer simulations to assess the accuracy of the Jukes-Cantor-like correction (table 5). All simulations are based on the trees for enolase and TIM, using the observed branch lengths (t_j) and the observed distributions of rates of amino acid replacements (λ_i). In the simulations, each site i evolves at a constant rate, with the number of replacements on branch j drawn from a Poisson distribution with mean $\lambda_i t_j$. For each site, the total number of replacements (the “Poisson” data) and the number that would be inferred by parsimony (the “Parsimony” data) are recorded. The Jukes-Cantor-like correction is then applied to the Parsimony data to produce the “Corrected” data. In the first pair of sim-

ulations all sites evolve at the same constant rate ($\bar{\lambda}$ constant). As expected of Poisson processes, the mean and variance values are equal and all variability is stochastic ($\bar{y}/s_y^2 \approx 1$). In the second pair of simulations different sites evolve at different rates, with the λ_i distributions taken from the enolase and TIM phylogenies. Site-to-site differences produce an additional source of variation that inflates the variance relative to the mean. Now, only 10% of the variability is attributable to stochastic effects ($\bar{y}/s_y^2 \approx 0.1$).

These simulations show that, despite the severe pruning of trees, parsimony significantly underestimates both the mean and the variance in the numbers of replacements across sites (table 5). They also reveal that the stochastic portion of the variance (\bar{y}/s_y^2) is consistently overestimated. This is expected. Imagine an extreme case where parsimony inferences are made on a phylogeny with such long branches that virtually every site on every branch has at least one replacement. Then the variance in the number of replacements per site is severely underestimated ($s_y^2 \rightarrow 0$) with the consequence that the mean to variance ratio is overestimated ($\bar{y}/s_y^2 \rightarrow \infty$).

We conclude that parsimony seriously underestimates the number of replacements per site, even on severely pruned trees. Correcting this bias is essential for reliable analyses. Simulations show that the Jukes-Cantor-like correction accurately recovers the mean, the variance, and their ratio (table 5), with residual biases far smaller than the stochastic errors inherent to single replicates.

Reliability of the NCD ($r^2/\hat{\rho}^2$)

We used simulations to assess the accuracy of NCD. Data from TIM were used as expectations around which Poisson sampling effects were simulated. For each site in the sequence, the simulated number of replacements was drawn from a Poisson distribution whose expectation varied in proportion to the expected number of replacements per site (from 0.8 to 27). The NCD was determined using the regression coefficient of the simulated data against the true expectations and the PECD. Estimates of the NCD are highly unreliable (fig. 4A) below a mean of approximately two replacements per site and accurate above a mean of five. The same general trend is manifest in real data (fig. 4B) regressed against the minimal model (see *Regression Analyses* below), although here the observed NCDs are far lower. We conclude that simulations and real data indicate that reliable NCDs can only be obtained from large phylogenies.

Coefficients of Variation

When the rate of amino acid replacements at each site (λ_i) is constant, the variance in the number of replacements among sites that is attributable to the differences in rates among sites (σ_{sites}^2) increases as the square of the mean number of replacements per site ($\bar{\mu}$), viz.

Table 5
Mean and SD Values from 1,000 Iterations

Run	Enzyme	Data	Number of Replacements	Mean/Site $\bar{y} \pm \text{SD}$	Variance Enolaseance $s_y^2 \pm \text{SD}$	Mean/Variance $\bar{y}/s_y^2 \pm \text{SD}$
λ Constant	Enolase	Poisson	5,700 \pm 62	13.604 \pm 0.148	13.360 \pm 0.921	1.023 \pm 0.070
		Parsimony	5,269 \pm 54	12.575 \pm 0.130	10.568 \pm 0.713	1.195 \pm 0.070
		Corrected	5,676 \pm 63	13.546 \pm 0.150	14.104 \pm 0.982	0.965 \pm 0.066
	TIM	Poisson	3,036 \pm 44	12.340 \pm 0.178	12.339 \pm 1.129	1.008 \pm 0.091
		Parsimony	2,776 \pm 38	11.286 \pm 0.155	9.377 \pm 0.844	1.213 \pm 0.109
		Corrected	3,004 \pm 44	12.213 \pm 0.180	12.732 \pm 1.777	0.967 \pm 0.088
λ Variable	Enolase	Poisson	5,700 \pm 65	13.603 \pm 0.154	168.287 \pm 5.603	0.081 \pm 0.002
		Parsimony	4,942 \pm 53	11.796 \pm 0.126	110.731 \pm 3.180	0.107 \pm 0.002
		Corrected	5,592 \pm 67	13.346 \pm 0.160	158.986 \pm 5.606	0.084 \pm 0.002
	TIM	Poisson	3,054 \pm 49	12.416 \pm 0.201	109.338 \pm 5.351	0.114 \pm 0.005
		Parsimony	2,627 \pm 40	10.681 \pm 0.162	69.470 \pm 2.903	0.154 \pm 0.005
		Corrected	2,952 \pm 50	12.001 \pm 0.203	97.937 \pm 4.917	0.123 \pm 0.005

$$\sigma_{sites}^2 = \sum_{i=1}^n (\lambda_i T - \bar{\lambda} T)^2 = (\bar{\lambda} T)^2 \cdot \sum_{i=1}^n (\lambda_i / \bar{\lambda} - 1)^2$$

$$= \sigma^2 - \bar{\mu} = \bar{\mu}^2 \cdot (s_x / \bar{\lambda})^2$$

where $T = \sum_{j=1}^m t_j$ is the total length of the tree and $(s_x / \bar{\lambda})^2$, the coefficient of variation in rates, is defined as constant. A plot of $\log_{10}(s_y^2 - \bar{y})$ against $\log_{10}(\bar{y})$ is a straight line of slope = 2 with a y-axis intercept of $2 \log_{10}(s_x / \bar{\lambda})$. When the corrected data in table 4 are so plotted they yield a line with slope = 1.97 ± 0.06 and $s_x / \bar{\lambda} = 0.88 \pm 0.03$ (fig. 5).

Caution is warranted when interpreting this log-log

plot. Data inevitably cluster near the line when, as here, the range on the x-axis ($\log_{10}(\bar{y})$) is so very much greater than the range on the y-axis intercepts ($\log_{10}(s_x / \bar{\lambda})$). Indeed, the 95% confidence intervals indicate that a third of the enzymes deviate significantly from the line. Nevertheless, variability in the coefficients of variation is modest, with s/\bar{y} varying from 0.74 to 1.04 and with 80% of the data within ± 0.1 of the mean of 0.85. The obvious exception is ribulose-1,5-bisphosphate carboxylase (RUBISCO; not otherwise included in the analysis) with its dramatic deviation symptomatic of rates far more variable than is typical of α/β -barrel enzymes. These observations suggest that the distributions of amino acid replacements in many α/β -barrel enzymes are similar, though not identical.

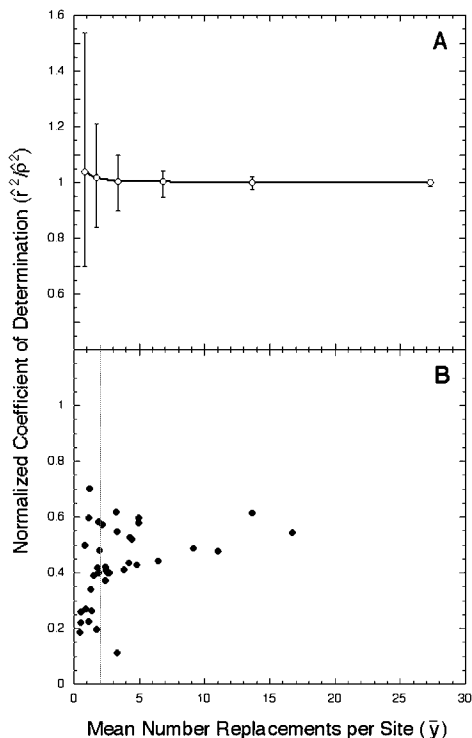


FIG. 4.—Estimates of the NCD (r^2/β^2) are unreliable when the mean number of replacements per site (\bar{y}) is less than 1.5. (A) Simulated data are regressed against the known underlying model. (B) Corrected data from 37 enzymes are regressed against the minimal regression model.

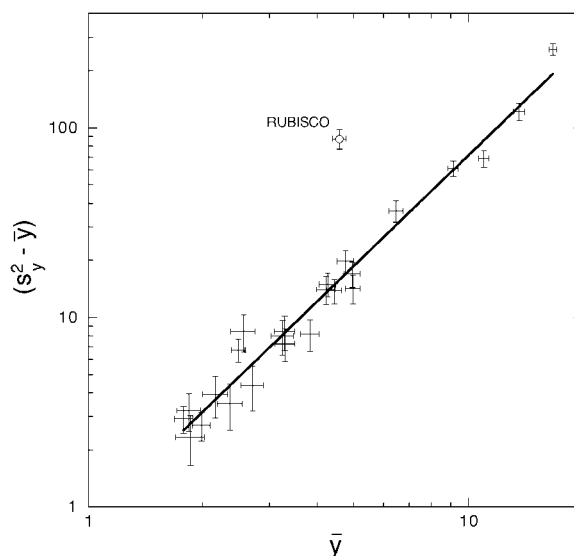


FIG. 5.—A simple Poisson process predicts that the proportion of variation attributable to causal effects ($s_y^2 - \bar{y}$) increases as the square of that attributable to stochastic effects (\bar{y}). With the notable exception of RUBISCO, data from the remaining 24 α/β -barrel enzymes yield a slope of 1.97 ± 0.06 , with relatively little scatter on this log-log plot. This suggests that protein evolution is dominated by a common set of rules, with variations on a general theme indicated by the 95% confidence intervals that do not overlap the fitted line.

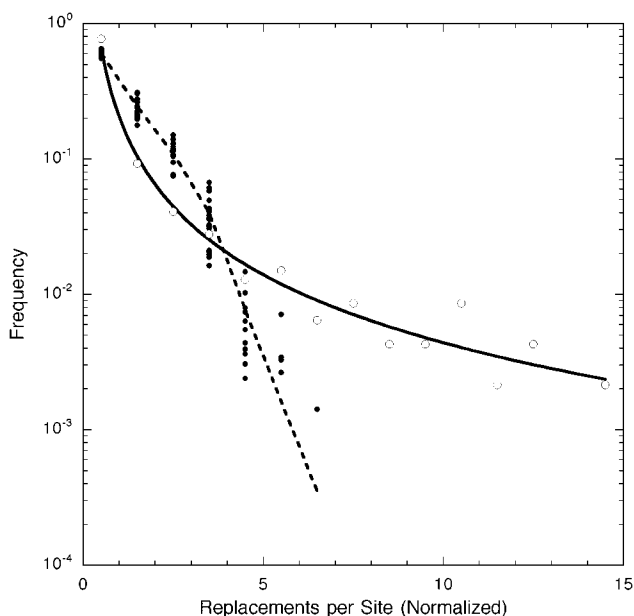


FIG. 6.—Frequency spectra of the number of replacements per site for 25 α/β -barrel enzymes after normalizing their mean values (\bar{y}) to 1. Twenty-four of the enzymes appear to share a common distribution or set of similar distributions (\bullet). RUBISCO (\circ) is exceptional in having so different a frequency distribution.

Frequency Distributions

To determine if α/β -barrel enzymes have similar frequency distributions of amino acid replacements per site we normalized our data such that each enzyme has an average of one amino acid replacement per site. This allows enzymes with few replacements per site to be directly compared with ones with many replacements per site. Of the 25 enzymes studied, 24 appear to share a common distribution (or a set of similar distributions) in which sites with many replacements are far less frequent than expected for a simple exponential decay (fig.

6). RUBISCO is the clear exception with sites having a moderate number of replacements being relatively rare, whereas those with many replacements are far more frequent than is typical.

Regression Analyses

The corrected number of replacements per site were regressed against distance from the catalytic center and solvent accessibility, while treating glycine residues with unusual ϕ and ψ angles as a separate class. For this minimal model the observed coefficients of determination average $\hat{r}^2 = 0.36 \pm 0.09$, whereas the PECDs average $\hat{\rho}^2 = 0.73 \pm 0.11$. This indicates that, on average, half the causal variation ($0.36/0.73 = 0.50$) is explained using just $df = 3$. Two outliers, RUBISCO and α -hydroxysteroid dehydrogenase, are discussed later.

We investigated additional models of amino acid replacement rates for four of the largest data sets (enolase, TIM, 5-aminolevulinatase dehydratase [5-ALDH] and class-I fructose 1,6 bisphosphate aldolase [F16BPI]) and where most variation is causal. Improvements in the fits obtained by adding additional terms to the minimal model are marginal, given the expenditure in degrees of freedom—the NCD increases by approximately 15% at a cost of over 25 df (table 6). Treating the two domains of each enolase monomer separately improves the NCD by only 2%. Distinguishing the amino terminus of each 5-ALDH monomer from the remaining α/β -barrel (the first 27 amino acids form an extend tail) produces no improvement in the fit. Including the x,y,z coordinates of the $C\alpha$ carbon atoms typically improves the NCD by approximately 5%, suggesting that weak directional gradients in the frequency of replacements across monomers are fairly common. Accounting for secondary structure (α -helix, β -sheet, β -bulge, turn, random coil) produces even less improvement. Hence, secondary

Table 6
Modifications to the Minimal Regression Model

VARIABLE	ADDITIONAL DF	NCD			
		Enolase	TIM	5-ALDH	F16BPI
Minimal Model ($df = 3$)	—	0.543	0.613	0.479	0.442
Domain (A-B)	1	0.566		0.479	
x,y,z	3	0.599	0.694	0.515	0.521
Secondary structure	4	0.552	0.643	0.531	0.483
HSBTE					
Rotamers	1	0.545			
Nonplanar peptide bonds	1	0.545			
Mainchain	1	0.546			
Buried hydrophyllic	1	0.545			
Sch-Sch H-bond	1	0.558			
Sch-Mch H-bond	1	0.545			
Mch-Mch H-bond	1	0.545			
AA-water-AA	1	0.545			
Intersubunit H-bond	1	0.552			
Sch H-bond to water	1	0.545			
Sch H-bond (pooled)	1	0.563			
AA	19	0.598	0.692	0.610	0.499
All	32(26)	0.671	(0.783)	(0.676)	(0.618)

NOTE.—The assumptions underlying use of F tests are violated, hence, statistical significance is not assessed.

structure exerts little influence on amino acid replacement rates.

Side-chain rotomers that clash with the main-chain and other side-chains are sufficiently rare that assigning them a unique class has a negligible effect on the fit. In any case, a structure with many errant side-chain rotomers is probably poorly determined and should not be analyzed. Nonplanar peptide bonds, unusual main-chain conformers (excluding Gly residues), and buried charged residues are each sufficiently rare that they too contribute little to the overall fit. Accounting for various H-bonding patterns in enolase (between side-chains, between side-chain and main-chain, between main-chains, and to water) improves fits marginally—pooling all types produces a 1.5% increase in the NCD. Including the 20 amino acids of the structure sequence improves the NCD by approximately 5%. The improvement is largely attributable to Arg, Asp, Glu, and Pro being significantly more conserved than that predicted by the minimal model, whereas Lys frequently occupies sites that rapidly evolve. Overall, improvements to the minimal model come at the expense of many degrees of freedom, suggesting that many variables, each of small effect, contribute to amino acid replacement rates. This does not exclude the possibility of another, as yet unidentified, variable having a major effect.

Introducing interaction terms (e.g., distance \times access) also produces small (<5%) improvements in the NCD at the expense of a large number of degrees of freedom. When the TIM data are fitted to the minimal model supplemented with x,y,z coordinates, the NCD rises from 0.69 to 0.72 when 10 second-order interaction terms are included (no interaction terms with glycine are included because of a lack of degrees of freedom) and then to 0.75 when all 26 third- and fourth-order interactions are included. Similar results are obtained when the same models are fitted to enolase data with the NCD rising from 0.60 to 0.61 and then to 0.63. We conclude that interactions between variables are of little consequence.

There is no theoretical reason to suppose that amino acid replacement rates should be a linear function of distance, access, or any other metric associated with protein structure. We therefore explored a nonlinear version of the minimal model based on the Hill equation of enzyme kinetics, an equation that displays a wide variety of behaviors from hyperbolic to sigmoid. For many data sets, fits do not converge. When fits did converge, increases in the NCDs were again marginal (not shown).

We conclude that distance from the active site, solvent accessibility, and glycine residues at constrained positions in the main-chain explain approximately 50% of the variability in rates of evolution not attributable to chance. Other variables, including secondary structure, and interactions account for only a small proportion of the observed variation.

Discussion

Standard regression investigates only deterministic dependencies among variables—stochastic errors are not

modeled. Consequently, the expected correlation coefficient of a “perfect” regression model must be less than ± 1 when stochastic errors are present. A model with a low correlation coefficient may indeed be of high quality if the remaining scatter is attributable to these errors. Improving such a model is impossible. This presents researchers with a very real quandary. Should they attempt to further refine a model and risk wasting precious time and resources analyzing random errors or should they abandon the model and risk missing some important effect? In the absence of a germane statistic, the decision must be based on other, subjective, criteria.

A solution in the special case of Poisson-distributed errors is possible. The reason is that the grand mean ($\bar{\mu}$) of summed Poisson distributions provides an estimate of the stochastic error that is entirely independent of the observed variance. This is not true for other distributions. For example, the grand mean of summed normal distributions provides no information about variances, whatever their source.

Each site in a protein accumulates amino acid replacements according to its own Poisson process. Though the rate at each site may vary independently of others, or coordinately with them (producing branch length effects), the overall process at each site is still Poisson when we condition on all historical contingencies. With evolution of a Poisson process, the proportion of site-to-site variation due to chance (\bar{y}/s_y^2) can be partitioned from that due to unspecified causal effects ($\hat{\rho}^2 = 1 - \bar{y}/s_y^2$).

This simple calculation allows us to concentrate on data with high information content. We decided that at least half the observed variation should be causal ($\hat{\rho}^2 > 0.5$) to warrant further analysis, a criterion that corresponds (roughly) to a phylogeny of 10 sequences averaging 1.75 replacements per site. Only 25 of 125 α/β -barrel structures in the Protein Data Bank, with phylogenies pruned of long branches and highly divergent sequences, satisfied this criterion on August 1, 2000 (table 4). Very large phylogenies (75 sequences averaging 10 replacements per site) are necessary if more than 90% of the variation is to be ascribed to causal effects. Few proteins of known structure are associated with such large phylogenies.

Despite pruning trees to remove branches longer than 0.3, parsimony underestimates both the mean and the variance in the number of amino acid replacements per site and overestimates their ratio. These biases are severe and a correction is essential to any reliable analysis (table 5). A Jukes-Cantor-like correction accurately recovers the mean, the variance, and their ratio, with remaining biases far below the stochastic noise inherent to the data. The only additional assumption needed for this correction is that changes in evolutionary rate affect all sites proportionally.

Approximately half of the causal variation is explained by just $df = 3$: distance from the active site, solvent accessibility, and glycines in unusual main-chain conformations (table 4). Like several recent analyses (Bustamante, Townsend, and Hartl 2000; Goldman, Thorne, and Jones 1998), we confirm that solvent ac-

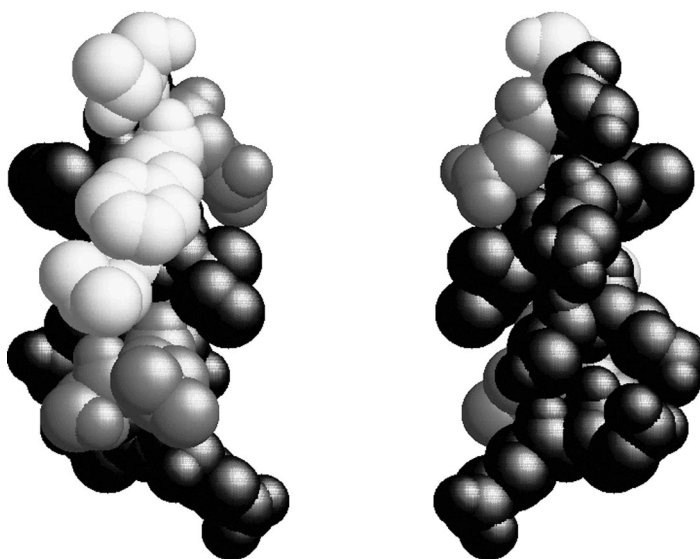


FIG. 7.—A van der Waals model of an α -helix from TIM that comprised residues 18–31: white, 0–10 replacements; gray, 11–20 replacements; black, 21–32 replacements. Typical of many amphipathic α -helices in α/β -barrels, the buried face (left) evolves far more slowly than the solvent-exposed face (right).

cessibility is a major determinant of amino acid replacement rates. We also show that distance from the active sites equals, and sometimes surpasses, solvent accessibility in importance (see the regression sums of squares in table 4). Glycines in unusual main-chain conformations make significant contributions to the regression sums of squares because they are highly conserved.

Asp, Glu, Arg, and Pro tend to be more conserved than expected, given their positions in crystal structures (data not shown). The first three are charged and their side-chains can H-bond to other polar and charged side-chains, but there is no correlation here: His occupies sites that evolve at expected rates, whereas Lys tends to occupy rapidly evolving sites. H-bonding among these and other polar side-chains does not contribute much to the fit (table 6). Asp sometimes plays a structural role in capping the dipole at the amino termini of α -helices but not so frequently to explain this level of conservation. Pro too is more conserved than most residues, perhaps because it often plays an important structural role by restricting acceptable main-chain conformations (its side-chain being covalently attached to the main-chain nitrogen). However, there is as yet no definitive method to predict, from structural data alone, which Pro residues will be conserved and which are free to evolve.

A surprising result of our analysis is that secondary structure has little predictive power regarding rates of evolution (table 6). TIM provides a typical example. Alone, secondary structure produces an NCD of $q^2 = \hat{r}^2/\hat{\beta}^2 = 0.12$, with helices evolving more rapidly than sheets and with turns and random coils having intermediate rates. When used to supplement the minimal model (distance, access, $\psi\phi$ Gly) secondary structure improves q^2 from 0.613 to 0.643, an increase of only 0.03. The difference (0.12 vs. 0.03) arises as a consequence of the construction of α/β -barrels (fig. 2). The sheets, which contain residues forming the active site, are buried in the hydrophobic core of the barrel. The helices,

farther from the active site and forming the perimeter of the barrel, have faces exposed to solvent (fig. 2). Hence, the helices of α/β -barrels evolve more rapidly than do sheets, not because they have any innate tendency to do so but because their position and exposure to solvent place them in regions where the functional and structural consequences of amino acid replacements are less severe. Indeed, sites in helices exposed to solvent evolve far more rapidly than those buried against the hydrophobic core (fig. 7). Secondary structure is of little consequence in determining rates of amino acid replacement.

The broad pattern of replacement rates is remarkably consistent among α/β -barrel structures (fig. 8). This is true, regardless of peptide length, primary, secondary, tertiary, and quaternary structures, presence of additional domains, catalytic chemistry used, the number and kinds of substrates used, kinetic behavior, mechanism of regulation, biochemical and physiological roles, and taxa analyzed (tables 3 and 4). The F16BP.I and F16BP.II are particularly interesting in this regard. Though they carry out precisely the same overall chemical reaction, cleaving the same C_6 substrate into the same two C_3 products, they display no evidence of homology and use unrelated chemistries. Catalysis in class-I enzymes (cyanobacteria, plants, and animals) proceeds through a lysyl Schiff-base, whereas class-II enzymes (cyanobacteria, bacteria, and fungi) use a divalent metal to stabilize the enolate intermediate (Walsh 1979). Their disparate chemistries and similar functions provide a striking example of convergent evolution at the biochemical level. Yet, despite the lack of homology and despite having distinct catalytic mechanisms, the frequency distributions and the patterns of amino acid replacements are similar.

Although the frequency distributions are similar among α/β -barrels (fig. 6) they are not identical—many confidence intervals fail to encompass the fitted line in

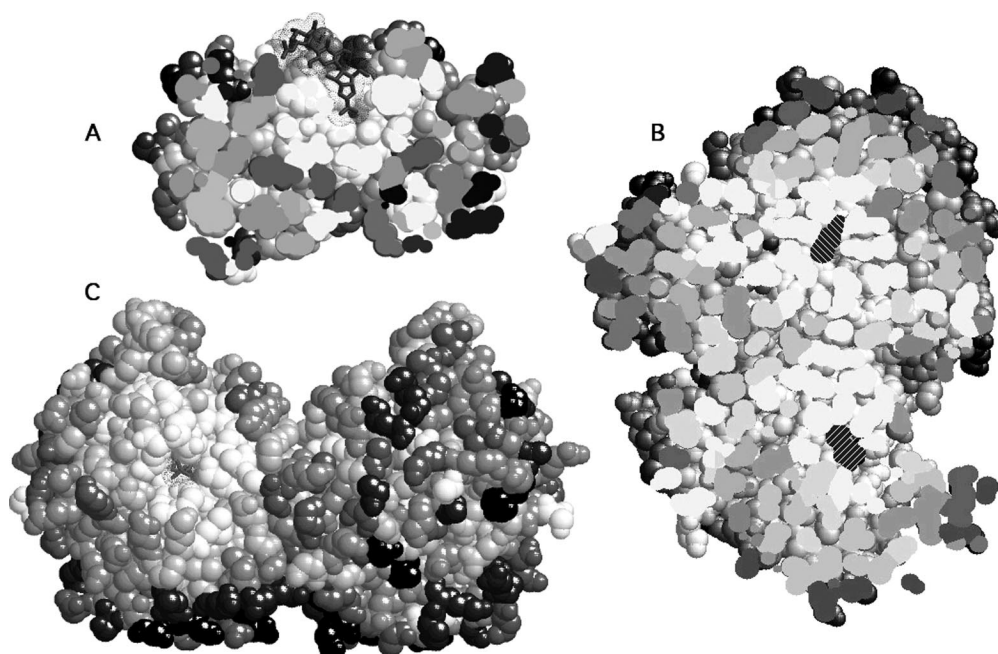


FIG. 8.—van der Waals representations of three α/β -barrel enzymes with amino acids shaded according to the number of replacements per site. Data were normalized to a mean of one replacement per site to facilitate comparisons among the three proteins. A cross-section through (A) monomeric plant acidic chitinase reveals the inhibitor allosamidin (sticks with dot surface) bound in a conserved active site (white residues), with rapidly evolving sites (black residues) located at the surface. A similar pattern is seen in (B) homodimeric enolase where residues at buried surfaces between the active sites (hatched substrates) evolve more slowly than do those exposed to the solvent. In (C) homodimeric TIM, surface residues surrounding the left-hand active site (stick substrate with dot surface) are conserved, whereas those on the opposite face of the enzyme, seen in the right-hand subunit, evolve rapidly. In all three examples, rapidly evolving residues lie farther from the active sites, regardless of whether they are (A and B) buried or (C) solvent exposed.

figure 6. Most deviations are minor and might arise from the many variations in size and shape among these enzymes. For example, the greater surface area–volume ratios in monomers might inflate the heterogeneity in rates compared with tetramers where much of the surface is buried at subunit interfaces.

RUBISCO is a notable exception to the above generalization. Its frequency spectrum of amino acid replacements differs dramatically from others in having far higher proportions of both slowly and rapidly evolving sites (fig. 6). Inspection of the structure reveals the typical overall pattern; a conserved active site surrounded by evolving sites, with the most rapidly evolving sites being remote and exposed to solvent. The structure offers no obvious explanation as to why the frequency spectrum should differ so markedly. It also does not offer any obvious insight into why the minimal regression model fits so poorly ($q^2 = \hat{r}^2/\hat{\rho}^2 = 0.18$; table 4), save that the greater site-to-site variability in rates, spread throughout the structure, reduces the correlation.

Analysis of the aldo-keto reductases also yields a poor fit to the minimal model ($q^2 = \hat{r}^2/\hat{\rho}^2 = 0.21$). Found in eukaryotes and prokaryotes, these enzymes belong to a diverse superfamily, sharing obvious sequence identity, a common structural fold, and a common catalytic mechanism but having widely different biological functions (Jez et al. 1997). Upon further investigation we discovered that the cause of the poor fit ($q^2 = \hat{r}^2/\hat{\rho}^2 = 0.113$; table 4) is attributable to a cluster of hydroxysteroid dehydrogenases (HSDs) and allied enzymes.

Once the HSDs are removed, the remaining aldo-keto reductases behave in a fashion typical of many other superfamilies ($q^2 = \hat{r}^2/\hat{\rho}^2 = 0.42$; table 4), with the most rapidly evolving sites scattered over the surface, well away from the active site (fig. 9A). In stark contrast, the most rapidly evolving sites in the HSDs cluster on either side of the substrate-binding cleft in the active site (fig. 9B). These replacements are concentrated in three loops that, when introduced into mammalian 3α -HSD from 20α -HSD, switch specificity from androgens to progestins (Ma and Penning 1999).

Though the acquisition of diverse physiological functions by HSDs (they are central to the metabolism of androgens, estrogens, glucocorticoids, mineralocorticoids, and progestins) is sufficient to account for the existence of adaptive replacements in the binding cleft, it is not sufficient to explain the high rate of replacement—43% of the replacements at these sites are attributable to replacements within each functional class (3α -HSD, 17β -HSD, 20α -HSD, dihydrodiol reductase, Δ^4 -3-ketosteroid 5β -reductase, and treating highly divergent 3α -HSD isozymes of human prostate and rat liver as different functionalities). The remaining aldo-keto reductases are also functionally diverse (e.g., biosynthesis of vitamin C in bacteria, production of sorbitol in mammals, mannitol biosynthesis in plants, xylose fermentation in yeasts, metabolism of neurotransmitter aldehydes, and possibly detoxifying assorted reactive carbonyls; Jez et al. 1997). Yet, replacements occur in these active sites at rates that are far lower than those in the

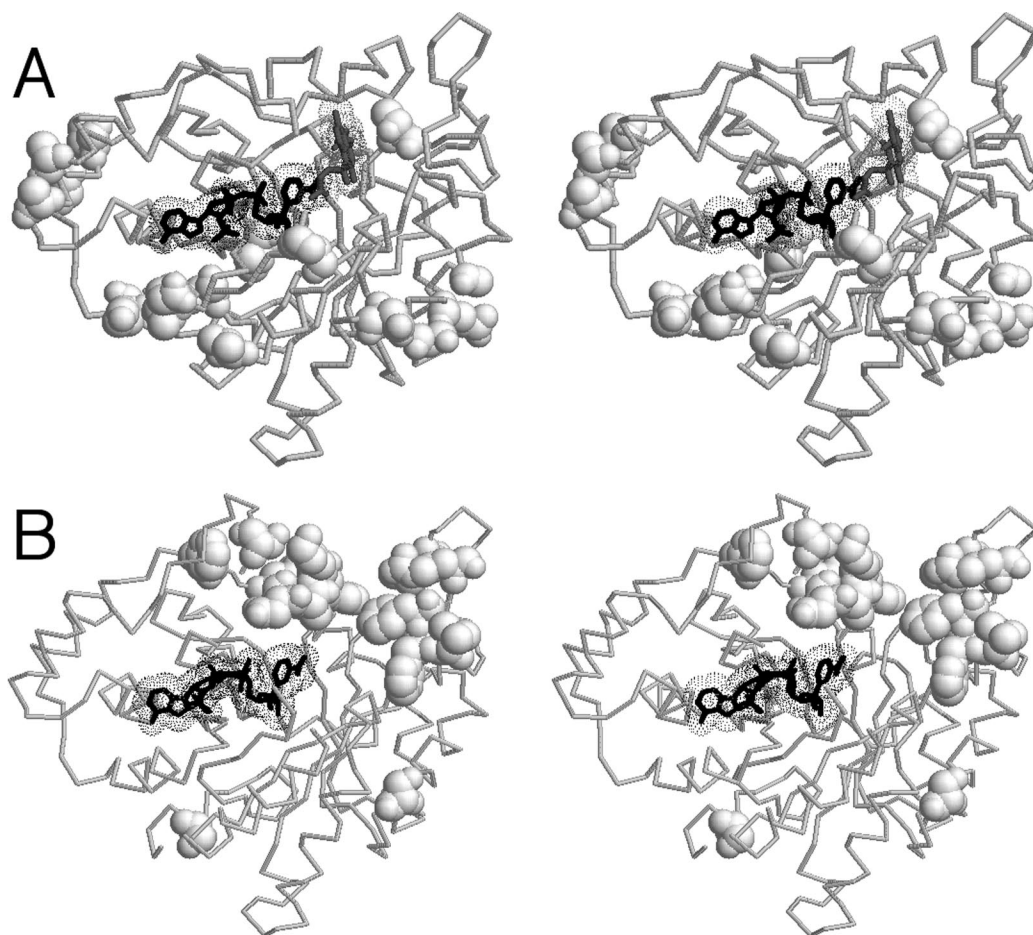


FIG. 9.—Cross-eyed stereo view of (A) aldose reductase showing the 15 most rapidly evolving sites (van der Waals surfaces) distributed over the surface and well outside the active site cleft in which are bound coenzyme (black sticks) and substrate (gray sticks). In the structurally and mechanistically related (B) HSDs the 15 most rapidly evolving sites cluster around the substrate-binding site.

HSDs and with only one rapidly evolving site anywhere near the substrate-binding cleft (fig. 9). Although a convincing explanation for this marked difference in the pattern of amino acid replacements among such similar homologues awaits thorough phylogenetic sampling, it is tempting to speculate that the HSDs, implicated in regulating sex hormone levels, may be subject to strong selection as life histories are fine tuned.

A reasonable fit ($q^2 = \hat{r}^2/\hat{\rho}^2 = 0.6$, table 4) is obtained when our simple model is applied to xylose isomerase. However, when residues with the 25 highest normalized deviations (observed/expected $- 1$) are plotted onto the protein structure, they form two contiguous bands, each flanking pairs of active sites in the tetramer (fig. 10). These “rings of fire” are not caused by the small number of expected replacements at sites near the active site—at only three of the 15 sites is the expected number of replacements less than one, and at other sites in the vicinity there is no tendency toward high normalized deviations. Similar patterns are not evident in other proteins—in pyruvate kinase, another large tetrameric enzyme, such sites are scattered haphazardly throughout the structure. The cause of these rings of fire remains a mystery, although their proximity to the active

sites is suggestive of mechanistic consequences subject to natural selection.

Conclusions

A simple statistical analysis of the distributions of amino acid replacements in α/β -barrel enzymes reveals that large phylogenies with many replacements are required to reduce the stochastic noise inherent to phylogenetic data to tolerable levels. Also, these large phylogenies should consist of many short branches so that corrections for multiple hits, essential to reliable analyses, remain but short extrapolations. Few proteins of known structure are associated with such phylogenies. Indeed, much literature illustrates the limited use of sophisticated statistical approaches when applied to marginal data.

The patterns and distributions of amino acid replacements among α/β -barrel enzymes are remarkably consistent, regardless of their diverse biochemical, metabolic, and biological roles. Indeed, fully half of the variation attributable to causal effects is explained by a simple regression model consisting of nothing more than solvent accessibility, distance from the active site, and

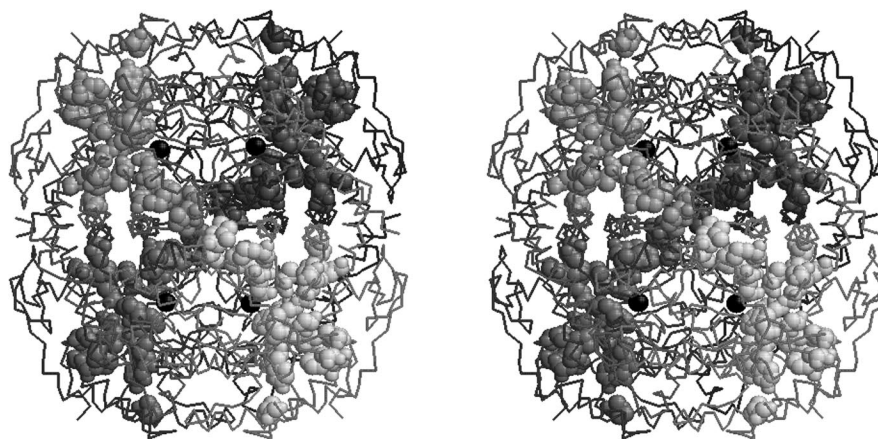


FIG. 10.—Cross-eyed stereo view of a homotetramer of xylose isomerase showing the contiguous rapidly evolving sites (van der Waals surfaces) that form bands on the sides of the active sites (black metal ion).

treating glycines occupying unusual main-chain conformations as a separate class. Other factors, notably secondary structure, exert little influence.

These results are general. The existence of additional domains in α/β -barrels have no obvious effect, but the simple model proves an equally good fit to isocitrate dehydrogenase, an enzyme that completely lacks an α/β -barrel (Dean and Golding 2000). On rare occasions when, as in the active site of HSDs, biological necessity disturbs the general pattern, a goodly portion of the causal variation in rates remains explainable by overall structural considerations. Nevertheless, our simple statistical analysis reveals that a considerable portion of the remaining unexplained variation is not attributable to chance. Other, as yet unidentified forces, must influence protein evolution.

Acknowledgments

The authors thank Charles Geyer, Glen Meeden, and Lauren Fisher for their thoughtful suggestions and encouragement.

LITERATURE CITED

- ALTSCHUL, S. F. L. M. T., A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- ATCHLEY, W. R., W. TERHALLE, and A. W. DRESS. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J. Mol. Evol.* **48**:501–516.
- ATCHLEY, W. R., K. R. WOLLENBERG, W. M. FITCH, W. TERHALLE, and A. W. DRESS. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* **17**:164–178.
- BABBITT, P. C., and J. A. GERLT. 1997. Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**:30591–30594.
- BISHOP, J. G., A. M. DEAN, and T. MITCHELL-OLDS. 2000. Rapid adaptive evolution in the active site of plant class I chitinases. *Proc. Natl. Acad. Sci. USA* **97**:5322–5327.
- BRANDEN, C., and J. TOOZE. 1999. Introduction to protein structure. 2nd edition. Garland Science, Ky.
- 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.
- COPLEY, R. R., and P. BORK. 2000. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**:627–640.
- DEAN, A. M., and G. B. GOLDING. 2000. Enzyme evolution explained (sort of). Pp. 1–12 in R. B. ALTMAN, A. K. DUNKER, L. HUNTER, K. LAUDERDALE, and T. E. KLEIN, eds. The Pacific symposium on bioinformatics 2000. World Scientific, Singapore.
- FELSENSTEIN, J. 1989. PHYLIP (phylogeny inference package). (Version 3.2). *Cladistics* **5**:164–166.
- FISHER, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.
- . 1948. Statistical methods for research workers. 10th edition. Oliver and Boyd, Edinburgh.
- FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science* **155**:279–284.
- GERLT, J. A. 2000. New wine from old barrels. *Nat. Struct. Biol.* **7**:171–173.
- GOLDMAN, N., J. L. THORNE, and D. T. JONES. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**:444–458.
- GU, X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* **16**:1664–1674.
- . 2001. Maximum likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18**:453–464.
- GU, X., and J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**:1106–1113.
- HIGGINS, D. G., J. D. THOMPSON, and T. J. GIBSON. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**:383–402.
- HOLM, L., and C. SANDER. 1996. Mapping the protein universe. *Science* **273**:595–602.
- HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**:167–170.
- JEZ, J. M., M. J. BENNETT, B. P. SCHLEGEL, M. LEWIS, and T. M. PENNING. 1997. Comparative anatomy of the aldo-keto reductase superfamily. *Biochem. J.* **326**:625–636.

- JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**:275–282.
- KENDALL, M., and A. STUART. 1977. *The advanced theory of statistics*. 4th edition, Vol. 1. MacMillan, New York.
- KIMURA, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, U.K.
- KIMURA, M., and T. OHTA. 1973. Mutation and evolution at the molecular level. *Genetics* **73**:19–35.
- LANDGRAF, R., D. FISCHER, and D. EISENBERG. 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **12**:943–951.
- LI, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, Mass.
- LIO, P., and N. GOLDMAN. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. *Mol. Biol. Evol.* **16**:1696–1710.
- LO CONTE, L., B. AILEY, T. J. HUBBARD, S. E. BRENNER, A. G. MURZIN, and C. CHOTHIA. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**:257–259.
- MA, H., and T. M. PENNING. 1999. Conversion of mammalian 3α -hydroxysteroid dehydrogenase to 20α -hydroxysteroid dehydrogenase using loop chimeras: changing specificity from androgens to progestins. *Proc. Natl. Acad. Sci. USA* **96**:11161–11166.
- MIYAMOYO, M. M., and W. M. FITCH. 1996. Constraints on protein evolution and the age of the eubacterial/eukaryotic split. *Syst. Biol.* **45**:568–575.
- POLLOCK, D., W. R. TAYLOR, and N. GOLDMAN. 1999. Co-evolving protein residues: maximum-likelihood identification and relationship to structure. *J. Mol. Biol.* **287**:187–198.
- SHINDYALOV, I. N., and P. E. BOURNE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**:739–747.
- Swofford, D. L. 1998. *Phylogenetic analysis using parsimony (* and other methods)*. Sinauer Associates, Sunderland, Mass.
- TANETO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, Neighbor-Joining, and maximum-parsimony methods when substitution rate varies with sites. *Mol. Biol. Evol.* **11**:261–277.
- TERWISSCHA VAN SCHELTINGA, A. C., S. ARMAND, K. H. KALK, A. ISOGAI, B. HENRISSAT, and B. W. DIJKSTRA. 1995. Stereochemistry of chitin hydrolysis by a plant chitinase/lysozyme and X-ray structure of a complex with allosamidin: evidence for substrate assisted catalysis. *Biochemistry* **34**:15619–15623.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- THORNE, J. L. 2000. Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* **10**:602–605.
- UZZEL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**:1089–1096.
- WALSH, C. 1979. *Enzymatic reaction mechanisms*. Freeman, NY.
- YANG, Z. 1994. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.
- . 1996. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites: approximate methods. *J. Mol. Evol.* **39**:306–314.
- YANG, Z., and J. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**:496–503.

WILLIAM MARTIN, reviewing editor

Accepted 4 June, 2002