Solvent Accessibility and Purifying Selection Within Proteins of *Escherichia coli* and *Salmonella enterica*

Carlos D. Bustamante, Jeffrey P. Townsend, and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University

The neutral theory of molecular evolution predicts that variation within species is inversely related to the strength of purifying selection, but the strength of purifying selection itself must be related to physical constraints imposed by protein folding and function. In this paper, we analyzed five enzymes for which polymorphic sequence variation within *Escherichia coli* and/or *Salmonella enterica* was available, along with a protein structure. Single and multivariate logistic regression models are presented that evaluate amino acid size, physicochemical properties, solvent accessibility, and secondary structure as predictors of polymorphism. A model that contains a positive coefficient of association between polymorphism and solvent accessibility and separate intercepts for each secondary-structure element is sufficient to explain the observed variation in polymorphism between sites. The model predicts an increase in the probability of amino acid polymorphism with increasing solvent accessibility for each protein regardless of physicochemical properties, secondary-structure element, or size of the amino acid. This result, when compared with the distribution of synonymous polymorphism, which shows no association with solvent accessibility, suggests a strong decrease in purifying selection with increasing solvent accessibility.

Introduction

The neutral theory of molecular evolution posits that the majority of evolution at the molecular level is due to the random fixation of mutations that do not affect fitness (Kimura 1983). Differences in rates of substitution and levels of heterozygosity among genes, or among different classes of sites within genes, are attributable to differences in the fraction of all mutations that are selectively neutral (mutations that are not selectively neutral are presumed to be deleterious owing to the rarity of beneficial mutations). Selection against these deleterious mutations is known as purifying selection and is acknowledged by most evolutionists as the predominant form of selection at the molecular level.

It is well known that the strength of purifying selection varies considerably between classes of DNA sites (e.g., between sites that alter amino acid sequence vs. those that do not). It has also been established through comparison of K_a/K_s (the ratio of divergence at amino acid replacement sites relative to divergence at synonymous sites) that purifying selection varies considerably between different proteins and, within the same protein, between different regions (Li 1997). Kimura's formulation of the mutation-drift hypothesis postulated that differences in the strength of purifying selection between different proteins are due to differences in functional constraint, such that genes that evolve quickly are more robust with respect to amino acid sequence than those that evolve slowly.

Understanding variability in substitution rates between different regions of proteins and between different classes of amino acid residues has been of considerable interest to molecular evolutionists. A growing literature in molecular phylogenetics has begun to address the question of how structural constraints relate to rate variation and thus to phylogenetic estimation (e.g., Naylor and Brown 1997). It has also been shown that location in the secondary structure and solvent accessibility systematically affect substitution rates in a wide range of protein families (Goldman, Thorne, and Jones 1998). The problem has also been of considerable interest to those who work on protein folding, since nonrandom substitution patterns can signal structural constraints as well as motifs important for optimizing protein-folding prediction (e.g., Koshi and Goldstein 1995; Overington et al. 1992).

The nature of structural factors determining levels of variation below the species level has not been examined. This is largely because the three-dimensional structures of the majority of proteins studied in population genetics are unknown. In this paper, we analyze five enzymes for which sequence variation among natural isolates of *Escherichia coli* and *Salmonella enterica* have been characterized and protein structures for *E. coli* forms of the enzymes are also known. For these five proteins, we find that solvent accessibility in the protein structure is a strong predictor of whether or not an amino acid will be polymorphic. This, of course, does not imply that any particular amino acid polymorphism is selectively neutral, only that purifying selection at the site is weak enough to allow the particular amino acid replacement to become polymorphic (Hartl et al. 2000). Here, we show that solvent accessibility is a better predictor of polymorphism for a given amino acid than its size, its physicochemical properties, or its location in the secondary structure of the protein.

Materials and Methods

Sequences and Structures

The enzymes analyzed in this study (which we will represent using the unitalicized gene symbols) are an-

Abbreviations: LRT, log-likelihood ratio test; MLE, maximum-likelihood estimate; SAS, solvent accessible surface.

Key words: purifying selection, polymorphism, solvent accessibility, neutral theory, *Escherichia coli*, *Salmonella enterica*, logistic regression.

Address for correspondence and reprints: Daniel L. Hartl, 16 Divinity Avenue, Cambridge, Massachusetts 02138. E-mail: dhartl@oeb.harvard.edu.

Mol. Biol. Evol. 17(2):301–308. 2000

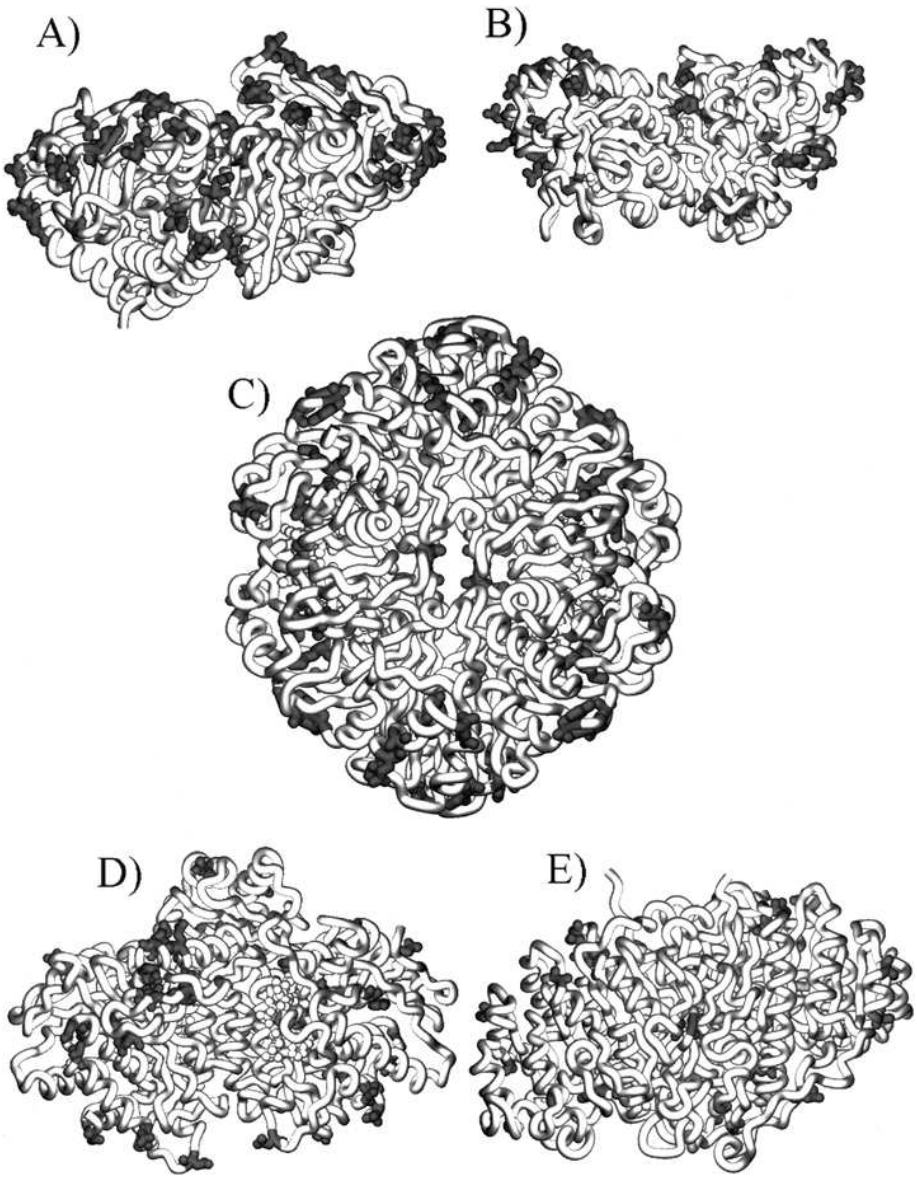

FIG. 1.—Backbone molecular structures of the enzymes used in this study. *A,* mdh = malate dehydrogenase. *B,* trpC = anthranilate isomerase. *C,* gapA = glyceraldehyde-3-phosphate dehydrogenase. *D,* icd = isocitrate dehydrogenase. *E,* phoA = alkaline phosphatase. The locations of variable residues in the α-carbon ribbon are colored dark gray, and their side chains are also shown.

thranilate isomerase (trpC), malate dehydrogenase (mdh), isocitrate dehydrogenase (icd), glyceraldehyde-3-phosphate dehydrogenase (gapA), and alkaline phosphatase (phoA). Diagrams of the folded backbones of these enzymes are shown in figure 1. Accession numbers for the sequences are found in the following: trpC (Milkman and Bridges 1993), mdh (Boyd et al. 1994; Pupo et al. 1997), icd (Wang, Whittam, and Selander 1997), gapA (Lawrence, Hartl, and Ochman 1991; Nelson, Whittam, and Selander 1991), and phoA (DuBose, Dykhuizen, and Hartl 1988). Sites that were variable within either species were considered polymorphic; sites that were identical within or between species were treated as invariant sites.

The structures used in this study were all determined through X-ray crystallography. For multimeric proteins, crystallographic transformations specified in the protein data bank (PDB) files were used to generate the functional multimeric molecule. The PDB files used in this study are from the following: trpC (Priestle et al. 1987), mdh (Hall, Levitt, and Banaszak 1992), icd (Stoddard, Dean, and Koshland 1993), gapA (Duee et al. 1996), and phoA (Wilmanns et al. 1992).

Solvent-Accessible Surface

A measure used throughout this paper is the exposed surface area of amino acid residues. This concept has been used extensively in structural biophysics to estimate the net gain in free energy due to protein folding as hydrophobic amino acid residues shed their ''water cages'' (Chothia 1974; Ooi et al. 1987) and is also used in energy refinement of protein structure (von Freyberg, Richmond, and Braun 1993). Consider the solvent-ac-

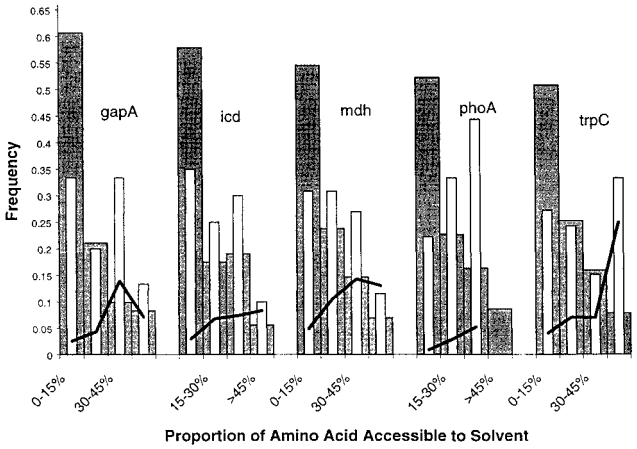

FIG. 2.—Frequency histograms of polymorphic and invariant residues for each enzyme. Gray bars represent the distribution of invariant residues, open bars represent the distribution of variable residues, and each solid line connects the fraction of residues within each solvent accessibility class that are polymorphic.

cessible surface area (SAS) of an atom, defined as the area on the surface of a sphere of radius R on each point of which the center of a solvent molecule can be placed in contact with the van der Waals sphere around the atom without penetrating any other atom in the molecule. The radius R is therefore given by the sum of the van der Waals radius of the atom and the chosen radius of the solvent molecule (Lee and Richards 1971). Finding the SAS of an amino acid is equivalent to rolling a water molecule (or another solvent molecule) over the van der Waals radii of the atoms in the amino acid as it is packed into the protein structure and calculating the surface area that the water molecule touches.

In our analysis, the SAS measure was used to estimate the proportion of each amino acid residue that is accessible to solvent. This was done by taking the ratio of SAS we calculated from the actual protein structure to that of the maximum exposed surface area in the fully extended conformation of the pentapeptide gly-gly-X-gly-gly, where X is the amino acid in question. We used two methods to estimate solvent accessibility that implemented in the package MOLMOL (Koradi, Billeter, and Wuthrich 1996) and Eisenhaber's ASC method (Eisenhaber and Argos 1993; Eisenhaber et al. 1995). The methods gave indistinguishable results. The distributions of solvent accessibility for polymorphic and invariant residues for each enzyme are indicated by the open and shaded bars, respectively, in figure 2. In this figure, the line segments connect the proportion of polymorphic amino acids observed in each category of solvent accessibility.

Logistic Regression, Confidence Intervals, and Estimation of K_a/K_s

Since we are interested in understanding how a set of predictor variables affect a dichotomous outcome variable (polymorphic or invariant), the logistic regression is an appropriate statistical model to employ. Specifically, letting P_i be the probability that the ith amino acid is polymorphic, the logistic regression models presented in this paper are nested within the model:

$$\ln\left[\frac{P_i}{1-P_i}\right] = \alpha + \beta_1 W_i + \beta_2 X_i + \beta_3 Y_i + \beta_4 Z_i + \beta_5 X_i W_i + \beta_6 X_i Z_i, \quad (1)$$

where α, β_1, β_2, β_3, and β_4 are the intercept and slopes for secondary-structure class, solvent accessibility, amino acid size, and physicochemical class, respectively, and W_i, X_i, Y_i, and Z_i are the values of secondary-structure class, solvent accessibility, size, and physicochemical class for the ith amino acid in the primary sequence of the solved structure. The parameters β_1 and β_4 allow for a unique intercept, and the parameters β_5 and β_6 allow for unique slopes for each secondary-structure element and physicochemical class of amino acid, respectively. For the sake of clarity, the parameter β_2 is hereinafter referred to as β_{sas}.

Maximum likelihood is the standard method used to estimate the slopes and intercepts for logistic regressions. Since the solutions to the derivative of the log-likelihood functions are not in closed form (Christensen 1997), we used Newton-Raphson iteration to obtain the estimates. Confidence intervals for the slopes and intercepts reported in this paper here are based on nonparametric bootstrapping of the data with 1,000 replicate data sets generated using a published algorithm for STATA (King, Tomz, and Wittenberg 1998). We report the 25th and 975th ranked estimates of the relevant parameter.

Multiple logistic regression models were explored to determine if including amino acid size (residue mass in daltons), physicochemical class, and secondary structure made a significant improvement on the reduced model with solvent accessibility alone. To assess improvement between nested models that differed in complexity, we used the difference in the log-likelihood of the hypotheses, which is approximately χ^2 distributed with degrees of freedom equal to the difference in degrees of freedom of the original models considered.

We estimated how K_a/K_s changes in trpC with solvent accessibility by using separate logistic regressions for replacement polymorphism versus synonymous polymorphism after classifying amino acids according to synonymy class (twofold redundant and fourfold redundant; amino acids that were neither twofold nor fourfold redundant were ignored). For each partition, we estimated K_a/K_s for a given value of solvent accessibility, X_o, as

$$\frac{K_a}{K_s}(X_o) = \left(\frac{P_a(X_o)}{P_s(X_o)}\right)\left(\frac{C}{1-C}\right), \quad (2)$$

where $P_a(X_o)$ is the probability of amino acid polymorphism per codon at X_o, and $P_s(X_o)$ is the probability of synonymous polymorphism per codon at X_o calculated from the logistic regression (eq. 1); C is the fraction of all single nucleotide changes that lead to a synonymous substitution assuming equal frequencies of nucleotide substitution. The quotient $C/(1-C)$ is a scaling coef-

Table 1
Maximum-Likelihood Estimates of Parameters for Logistic Regression of Polymorphism on Solvent Accessibility for Amino Acids Grouped by Protein

Protein	*N* Sequences (*Escherichia coli*, *Salmonella enterica*)	% Polymorphic	α (95% CI)	β_{sas} (95% CI)	LRT $Pr(\chi^2(1))$
mdh	19, 27	6.1	−3.59 (−4.46, −2.72)	3.85 (1.09, 6.62)	7.41 $P < 0.01$
trpC	25, 0	7.3	−3.46 (−4.14, −2.80)	3.81 (1.86, 5.75)	14.55 $P < 0.0001$
icd	17, 16	2.7	−4.54 (−5.68, −3.40)	4.08 (0.72, 7.44)	5.52 $P < 0.02$
gapA	10, 16	4.6	−3.69 (−4.56, −2.82)	3.20 (0.41, 5.98)	4.71 $P < 0.03$
phoA	8, 0	2.0	−4.36 (−5.47, −3.24)	2.18 (−1.40, 5.76)	1.35 NS
Combined	—	4.5	−3.87 (−4.26, −3.48)	3.53 (2.35, 4.71)	33.12 $P \ll 0.0001$

NOTE.—CI = confidence interval; LRT = log-likelihood ratio test.

ficient that allows us to generate a proxy for K_a/K_s from the ratio of the probability of replacement polymorphism to the probability of synonymous polymorphism. For twofold-redundant sites, $C = 1/9$ and $C/(1 - C) = 1/8$. For fourfold-redundant sites $C = 3/9$ and $C/(1 - C) = 1/2$. Confidence intervals for K_a/K_s were generated from replicate data sets generated through nonparametric bootstrapping.

Results

Figure 1 shows the α-carbon backbones of the molecular structures of the enzymes used in this study. Polymorphic residues are shaded, and their side chains in the canonical sequence are also shown. Figure 2 shows the distributions of solvent accessibility of invariant (shaded bars) and polymorphic (open bars) amino acids for each protein, as well as the fraction of all residues that are polymorphic within each solvent accessibility class. We found that the distribution of invariant sites for each gene was significantly skewed toward less solvent accessibility ($P < 0.001$ for all genes). The distribution of polymorphic sites showed no such skew, and there is a general trend toward increasing polymorphism with increasing solvent accessibility.

Table 1 summarizes the maximum-likelihood estimates of the parameters in the logistic regression model of polymorphism on solvent accessibility for amino acids grouped by protein. In the first column, we list the number of sequences from each species used in our study. The second column lists the proportion of sites that vary within each protein. The third and fourth columns give the maximum-likelihood estimates of α and β_{sas}, respectively. The fifth column gives the results of likelihood ratio tests (LRTs) of whether the model with $\beta_{sas} = \text{MLE}(\beta_{sas})$ fits the data significantly better than a model with $\beta_{sas} = 0$, where the test statistic is approximately distributed as χ^2 with one degree of freedom. For four out of the five genes, a model with increasing probability of polymorphism with solvent accessibility is a significantly better model, and the one protein (phoA) for which the test is not significant has the least polymorphism, thus compromising the power of the test. Using the analytical approximation of Whittemore (Whittemore 1981) it can be shown that for overall frequencies of polymorphism of 2%, 5%, and 10%, we have approximately 30%, 70%, and 90% power, respectively, to reject the null hypothesis that $\beta_{sas} = 0$ in favor of $\beta_{sas} = 3.5$ (the average β_{sas} for all of the genes).

An alternative explanation for the observed relationship between polymorphism and solvent accessibility is that it results secondarily from systematic differences in solvent accessibility of different classes of amino acids or elements of secondary structure. To test this possibility, we carried out separate logistic regressions for each of three major classes of amino acids (charged = Arg, Asp, His, Glu, Lys; hydrophobic = Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Val; uncharged = Asn, Gln, Ser, Thr, Trp, Tyr) and secondary-structural elements (alpha helices, beta sheets, and random coils/turns). The results are summarized in tables 2 and 3, respectively. Figure 3 shows a graph of the predicted probability of polymorphism and 95% confidence intervals for the logistic regression of all amino acids combined and the observed probability of polymorphism for each of the three physicochemical groups (fig. 3*A*) and for each of the elements of secondary structure considered (fig. 3*B*).

We also tested multiple logistic regression models that included the size, secondary structure, and/or physicochemical properties of the amino acid with and without solvent accessibility to see if they offered significant improvements over simpler models. The results relating to solvent accessibility are summarized in table 4. The first column gives the logistic regression null models, the second column gives the added parameter being tested, the third column gives the degrees of freedom in the comparison, the fourth column gives the results of the

Table 2
Maximum-Likelihood Estimates of Parameters for Logistic Regression of Polymorphism on Solvent Accessibility for Amino Acids Grouped by Physicochemical Property

Class	Residues	% Polymorphic	α (95% CI)	β_{sas} (95% CI)	LRT $Pr(\chi^2(1))$
Charged	624 (31.9%)	5.8	−4.14 (−5.19, −3.09)	3.99 (1.47, 6.52)	10.48 $P < 0.002$
Hydrophobic	465 (23.8%)	4.0	−3.76 (−4.27, −3.25)	4.02 (1.97, 6.07)	12.93 $P < 0.001$
Uncharged	866 (44.3%)	4.0	−4.04 (−4.81, −3.27)	3.70 (1.37, 6.02)	9.45 $P < 0.003$
Combined.	1,955	4.5	−3.87 (−4.26, −3.48)	3.53 (2.35, 4.71)	33.12 $P \ll 0.0001$

NOTE.—LRT = log-likelihood ratio test.

LRTs, and the fifth column gives the significance levels. It is clear from table 4 that all models that include solvent accessibility are significant improvements over those that do not, whereas including amino acid size and physicochemical class in a logistic regression does not yield a significant improvement over solvent accessibility alone. No model that excluded solvent accessibility presented a significant improvement over the null hypothesis that all β's are equal to zero, and all models that did include solvent accessibility presented significant improvements over the null hypothesis. We also note that the only improvement that can be made on a model of solvent accessibility alone is the addition of a different intercept for each element of secondary structure. Once this is done, adding a separate slope makes no additional improvement ($P > 0.32$).

Figure 4 shows the predicted K_a/K_s for trpC based on separate logistic regressions of amino acid and synonymous polymorphisms on solvent accessibility for twofold-redundant (fig. 4*A*) and fourfold-redundant (fig. 4*B*) amino acids. As expected, the probability of a synonymous polymorphism—averaging 9.6% in twofold-redundant and 24% in fourfold-redundant amino acids—shows no significant relationship with solvent accessibility ($P < 0.86$ and $P < 0.76$, respectively). In contrast, the maximum-likelihood estimates of α and β_{sas} for amino acid polymorphism are $\alpha = -4.1$ and $\beta_{sas} = 5.3$ for twofold-redundant amino acids, and $\alpha = -2.9$ and $\beta_{sas} = 2.9$ for fourfold-redundant amino acids. Both regressions fit the data significantly better than $\beta_{sas} = 0$ ($P < 0.05$). For both classes of amino acids, the K_a/K_s ratio increases dramatically as a function of solvent accessibility, suggesting a uniform relaxation in constraint and thus in the strength of purifying selection.

Discussion

Purifying selection is generally agreed to be the predominant form of natural selection involved in the patterning of most macromolecules most of the time. In a protein under predominantly purifying selection, less-constrained amino acid residues will tend to be more polymorphic and more-constrained amino acid residues will tend to be less polymorphic. In this paper, we investigated how well different properties of an amino acid—size, physicochemical class, and secondary-structure element—predict the relative likelihood that the site will be polymorphic within either *E. coli* or *S. enterica*. We find a strong positive relation between polymorphism and solvent accessibility, suggesting that amino acid sites that are more solvent-accessible are less likely to be constrained in identity. This finding is in accordance with Poisson random field analysis of polymorphic amino acids based on frequencies alone, which suggests that most polymorphic amino acids in the same proteins are slightly deleterious (Hartl et al. 2000), and with work done on multiple families of proteins showing that solvent accessibility and secondary structure impact

Table 3
Maximum-Likelihood Estimates of Parameters for Logistic Regression of Polymorphism on Solvent Accessibility for Amino Acids Grouped by Secondary Structure

Secondary Structure	Residues	% Polymorphic	α (95% CI)	β_{sas} (95% CI)	LRT $Pr(\chi^2(1))$
Coil	850 (43.48%)	4.0	−3.96 (−4.65, −3.26)	2.85 (0.94, 4.77)	8.38 $P < 0.004$
Helix	732 (37.44%)	5.7	−3.86 (−4.46, −3.26)	4.91 (3.03, 6.79)	26.34 $P < 0.0001$
Sheet	373 (19.08%)	2.9	−3.99 (−4.85, −3.13)	4.01 (0.26, 7.76)	4.01 $P < 0.05$
Combined	1955	4.5	−3.87 (−4.26, −3.48)	3.53 (2.35, 4.71)	33.12 $P \ll 0.0001$

NOTE.—LRT = log-likelihood ratio test.

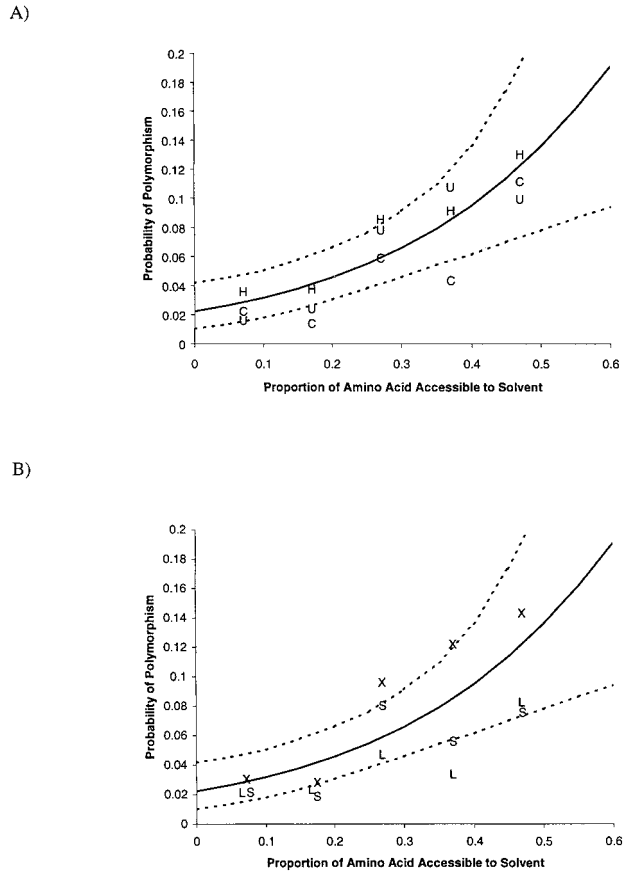

FIG. 3.—Comparison of observed probability of polymorphism and predicted values and 95% confidence intervals from the logistic regression for amino acids grouped by (*A*) physicochemical class (U = uncharged amino acids, C = charged amino acids, H = hydrophobic amino acids) and (*B*) secondary structure (X = helix, S = sheet, L = coil).

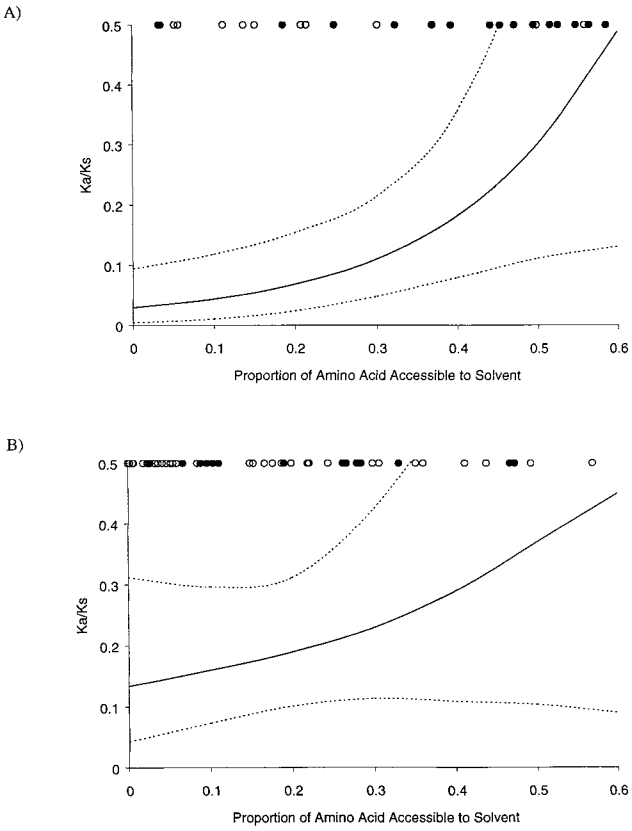

FIG. 4.—Estimated K_a/K_s ratio from logistic regressions of trpC for (*A*) twofold- and (*B*) fourfold-redundant amino acids. Closed circles represent the locations of amino acid polymorphisms, and open circles represent the locations of synonymous polymorphisms. The solid line represents the estimated K_a/K_s ratio, and the dashed lines are 95% confidence intervals estimated through nonparametric bootstrapping.

amino acid substitution rates (Goldman, Thorne, and Jones 1998).

It is suggested by the structures themselves (fig. 1) that there seems to be a concentration of polymorphic amino acid sites on the ''outside'' of each enzyme. These tend to be regions of relatively high solvent accessibility, and many of the polymorphic residues protrude from the structure in such a way that an amino acid replacement would not drastically alter hydrogen bonding or hydrophobic contacts made with other residues. That polymorphic residues tend to cluster on the outside of molecules is supported by histograms of solvent accessibility for invariant and polymorphic residues (fig. 2). Nevertheless, our analysis indicates that the ''outside''-''inside'' dichotomy is too simplistic. The probability of amino acid polymorphism increases as a continuous function of solvent accessibility.

The logistic regression analysis combines the intuitive appeal of ordinary least-squares regression with the ease of a likelihood framework for testing more complicated models. We found that all of the proteins surveyed showed strong effects of solvent accessibility on relative probability of polymorphism. This effect was significant for four of five proteins, and the one nonsignificant protein was also the least polymorphic so that

Table 4
Likelihood Ratio Tests for Nested Multivariate Logistic Regression Models of Polymorphism on Solvent Accessibility

Null Model	Added Parameter	df	LRT	Pr(χ^2, df)
Physicochemical class	+SAS	1	28.91	$P < 0.0001$
Size	+SAS	1	34.34	$P < 0.0001$
Secondary structure	+SAS	1	36.16	$P < 0.0001$
SAS	+Size	1	1.98	$P > 0.1593$
	+Physicochemical class	2	2.19	$P > 0.3350$
	+Secondary structure intercept	2	8.32	$P < 0.0156$
SAS, secondary structure intercept	+Secondary structure slope	2	2.27	$P > 0.3218$

NOTE.—LRT = log-likelihood ratio test.

the test had the least power. Unexpectedly, all five proteins had very similar regression coefficients, suggesting that lower solvent accessibility may be similarly associated with stronger selective constraints across a wide range of enzymes differing in myriad details of their individual structures.

We also investigated whether the effect of solvent accessibility reflects a shift in amino acid composition merely from areas of low solvent accessibility to areas of high solvent accessibility or from one element of secondary structure to another. For example, if hydrophobic residues tended to be concentrated in areas of low solvent accessibility and also tended to be monomorphic, but for charged amino acids the relations were the other way around, the overall correlation of polymorphism with solvent accessibility would be spurious. This is not the case. When we compare the estimates of the slope, β_{sas}, for each of the major classes of amino acids in tables 2 and 3, we note a striking similarity. There is also a good fit between the predicted probability of polymorphism from the combined logistic regression and the observed probability of polymorphism for amino acids grouped by physicochemical properties as hydrophobic (H), charged (C), and uncharged (U) (fig. 3*A*) and grouped by structural elements as helix (X), sheet (S), and coil (L) (fig. 3*B*). To address this issue formally, we also estimated multiple-regression models (table 4) that included size, secondary structure, and/or physicochemical class with and without solvent accessibility. Multiple-regression models that included solvent accessibility were significantly better at predicting probability of polymorphism than those that did not include it, and including amino acid size and/or physicochemical class in a multiple logistic regression made no significant improvement to a simpler model with solvent accessibility alone (table 4). The one improvement that could be made on the simplest model of solvent accessibility alone was to add an intercept term to account for differences in overall levels of polymorphism between elements of secondary structure. In short, the probability of polymorphism is more closely related to solvent accessibility than to amino acid identity, secondary structure, or size.

The logistic regression was also used in conjunction with data on synonymous polymorphism to estimate quantitatively the reduction in purifying selection with increasing solvent accessibility. When compared with the distribution of synonymous polymorphism, the increased probability of amino acid polymorphism with solvent accessibility (fig. 4) suggests strong purifying selection in areas of low solvent accessibility and weak purifying selection in areas of high solvent accessibility, irrespective of synonymy class. The reduction in purifying selection is so large that sites near the high end of the solvent accessibility range appear to be evolving at a rate 5–10 times as fast ($K_a/K_s \approx 0.5$ for both fourfold- and twofold-redundant sites) as those in areas of low solvent accessibility that are under very strong selection ($K_a/K_s \approx 0.1$ for fourfold-redundant sites, and $K_a/K_s < 0.05$ for twofold-redundant sites).

Although our results are based on only five proteins, they tentatively suggest that similar constraints may govern disparate enzymes independent of their function. This finding, if proven to be general, may be rationalized in a broader consideration of how enzymes are thought to function. For a particular enzyme, only a few key residues are directly involved in the catalytic function (i.e., those residues directly in the vicinity of the active site). The majority of other residues play a role in maintaining the correct three-dimensional structure of the protein so that the protein can perform its function (Pakula and Sauer 1989). Our results tentatively suggest that the majority of the sites that are allowed to vary within species are those sites that are less involved in the stabilizing of protein structure, since they are residues that are in close contact with solvent and thus do not form hydrogen bonds with other residues in the protein. The pervasive effect of solvent accessibility on polymorphism argues for a theory of universal structural constraint on amino acid evolution in enzymes and perhaps in other classes of protein structure.

Acknowledgments

The authors wish to thank the two anonymous reviewers, as well as Richard Lewontin, John Wakely, and Stephen Palumbi for thoughtful discussion. This work was supported by grants from the Howard Hughes Medical Institute (C.D.B.) and the U.S. National Institutes of Health (J.P.T., D.L.H.).

LITERATURE CITED

BOYD, E. F., K. NELSON, F. S. WANG, T. S. WHITTAM, and R. K. SELANDER. 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (mdh) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **91**:1280–1284.

CHOTHIA, C. 1974. Hydrophobic bonding and accessible surface area in proteins. Nature **248**:338–339.

CHRISTENSEN, R. 1997. Log-linear models and logistic regression. Springer, New York.

DUBOSE, R. F., D. E. DYKHUIZEN, and D. L. HARTL. 1988. Genetic exchange among natural isolates of bacteria: recombination within the phoA gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **85**:7036–7040.

DUEE, E., L. OLIVIER-DEYRIS, E. FANCHON, C. CORBIER, G. BRANLANT, and O. DIDEBERG. 1996. Comparison of the structures of wild-type and a N313T mutant of *Escherichia coli* glyceraldehyde 3-phosphate dehydrogenases: implication for NAD binding and cooperativity. J. Mol. Biol. **257**: 814–838.

EISENHABER, F., and P. ARGOS. 1993. Improved strategy in analytic surface calculations for molecular systems: handling of singularities and computational efficiency. J. Comput. Chem. **14**:1272–1280.

EISENHABER, F., P. LIJNZAAD, P. ARGOS, C. SANDER, and M. SCHARF. 1995. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comput. Chem. **16**:273–284.

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.

HALL, M. D., D. G. LEVITT, and L. J. BANASZAK. 1992. Crystal structure of *Escherichia coli* malate dehydrogenase. A complex of the apoenzyme and citrate at 1.87 A resolution. J. Mol. Biol. **226**:867–882.

HARTL, D., E. F. BOYD, C. D. BUSTAMANTE, and S. SAWYER. 2000. The glean machine: What can we learn from DNA sequence polymorphism? *In* S. SUHAI, ed. Genomics and proteomics. Plenum Press, New York (in press).

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KING, G., M. J. TOMZ, and J. WITTENBERG. 1998. Making the most of statistical analyses: improving interpretation and presentation. Am. J. Political Sci. (in press).

KORADI, R., M. BILLETER, and K. WUTHRICH. 1996. MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graph. **14**:51–55.

KOSHI, J. M., and R. A. GOLDSTEIN. 1995. Context-dependent optimal substitution matrices. Protein Eng. **8**:641–645.

LAWRENCE, J. G., D. L. HARTL, and H. OCHMAN. 1991. Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33**:241–250.

LEE, B., and F. M. RICHARDS. 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. **55**:379–400.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

MILKMAN, R., and M. M. BRIDGES. 1993. Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. Genetics **133**:455–468.

NAYLOR, G. J., and W. M. BROWN. 1997. Structural biology and phylogenetic estimation. Nature **388**:527–528.

NELSON, K., T. S. WHITTAM, and R. K. SELANDER. 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (gapA) in natural populations of Salmonella and *Escherichia coli.* Proc. Natl. Acad. Sci. USA **88**:6667–6671.

OOI, T., M. OOBATAKE, G. NEMETHY, and H. A. SCHERAGA. 1987. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc. Natl. Acad. Sci. USA **84**:3086–3090.

OVERINGTON, J., D. DONNELLY, M. S. JOHNSON, A. SALI, and T. L. BLUNDELL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. **1**:216–226.

PAKULA, A. A., and R. T. SAUER. 1989. Genetic analysis of protein stability and function. Annu. Rev. Genet. **23**:289–310.

PRIESTLE, J. P., M. G. GRUTTER, J. L. WHITE, M. G. VINCENT, M. KANIA, E. WILSON, T. S. JARDETZKY, K. KIRSCHNER, and J. N. JANSONIUS. 1987. Three-dimensional structure of the bifunctional enzyme N-(5′-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli.* Proc. Natl. Acad. Sci. USA **84**:5690–5694.

PUPO, G. M., D. K. KARAOLIS, R. LAN, and P. R. REEVES. 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect. Immun. **65**:2685–2692.

STODDARD, B. L., A. DEAN, and D. E. KOSHLAND JR. 1993. Structure of isocitrate dehydrogenase with isocitrate, nicotinamide adenine dinucleotide phosphate, and calcium at 2.5-A resolution: a pseudo-Michaelis ternary complex. Biochemistry **32**:9310–9316.

VON FREYBERG, B., T. J. RICHMOND, and W. BRAUN. 1993. Surface area included in energy refinement of proteins. A comparative study on atomic solvation parameters. J. Mol. Biol. **233**:275–292.

WANG, F. S., T. S. WHITTAM, and R. K. SELANDER. 1997. Evolutionary genetics of the isocitrate dehydrogenase gene (icd) in *Escherichia coli* and *Salmonella enterica.* J. Bacteriol. **179**:6551–6559.

WHITTEMORE, A. 1981. Sample size for logistic regression with small response probability. J. Am Stat. Assoc. **76**:27–32.

WILMANNS, M., J. P. PRIESTLE, T. NIERMANN, and J. N. JANSONIUS. 1992. Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 A resolution. J. Mol. Biol. **223**:477–507.

ANTONY DEAN, reviewing editor

Accepted November 1, 1999