Article No. jmbi.1999.2972 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 291, 977–995

JMB

Evolution of Protein Sequences and Structures

Todd C. Wood and William R. Pearson*

Department of Biochemistry
University of Virginia
Charlottesville, VA
22908, USA

The relationship between sequence similarity and structural similarity has been examined in 36 protein families with five or more diverse members whose structures are known. The structural similarity within a family (as determined with the DALI structure comparison program) is linearly related to sequence similarity (as determined by a Smith-Waterman search of the protein sequences in the structure database). The correlation between structural similarity and sequence similarity is very high; 18 of the 36 families had linear correlation coefficients $r \geqslant 0.878$, and only nine had correlation coefficients $r \leqslant 0.815$. Inclusion of higher-order terms in the structure/sequence relationship improved the fit by less than 7% in 27 of the 36 families. Differences in sequence/structure correlations are distributed evenly among the four protein structural classes, α, β, α/β, and $\alpha + \beta$. While most protein families show high correlations between sequence similarity and structural similarity, the amount of structural change per sequence change, i.e. the structural mutation sensitivity, varies almost fourfold. Protein families with high and low structural mutation sensitivity are distributed evenly among protein structure classes. In addition, we did not detect strong correlations between structural mutation sensitivity and either protein family mutation rates or protein size. Our results are more consistent with models of protein structure that encode a protein family's fold throughout the protein sequence, and not just in a few critical residues.

Keywords: sequence similarity; structural similarity; DALI; Smith-Waterman; alignment

**Corresponding author*

Introduction

Two general models attempt to explain how the tertiary structure of a protein is encoded in its linear sequence of amino acids: (1) the local model, in which fold specificity is coded in just a few critical residues (10-20% of the sequence); and (2) the global model, in which the fold is formed by interactions involving the entire sequence (Lattman & Rose, 1993). The most obvious confirmation of the local model is the misfolding mutations associated with certain diseases, such as cystic fibrosis (Thomas *et al*., 1995). The global model is supported by numerous mutation studies which show that most mutations at any position in a protein sequence have no measurable impact on the protein function, and therefore the structure (Bowie *et al*., 1990; Lattman & Rose, 1993; Matthews, 1987).

The local model receives considerable support from examples of structurally similar proteins that do not share significant sequence similarity, e.g. actin and hexokinase (Kabsch & Holmes, 1995). Since actin and hexokinase share similarity of overall structure and ATP-binding sites but lack significant sequence similarity, they are frequently referred to as remote homologues. The structural similarity of remote homologues can be explained as the conservation of certain critical ''core'' folding residues, as the local model predicts.

If protein folding information is localized to critical residues, as the structures of remote homologues apparently imply, we would expect that the non-critical residues would be poorly constrained during sequence evolution. Such heterogeneity in functional constraint of amino acids would produce proteins with modest sequence similarity, but nearly identical structures. Alternatively, a strong, continuous correlation between sequence and structural similarity would imply that the folding information is distributed throughout the sequence, not localized to particular residues. In a

Abbreviation used: LCA, last common ancestor.

E-mail address of the corresponding author: wrp@virginia.edu

0022-2836/99/340977–19 $30.00/0

continuous correlation of sequence and structural similarity, each amino acid position contributes to the overall structural similarity.

Early studies by Chothia & Lesk (1986, 1987) showed a strongly non-linear relationship between sequence and structural similarity (Figure 1(a)). Very similar sequences showed modest structural differences, but structural differences increased dramatically as sequence identities dropped below 15-20 %. This observation supports the local model for protein folding; changes in sequences at 80-100 % identity have small effects on structure, but changes at 15-25 % identity, which are more likely to involve critical ''core'' residues, have a much larger effect. Recent studies have confirmed their findings with larger sets of protein structures (Flores *et al*., 1993; Russell *et al*., 1997). All these studies used the percent sequence identity and the root-mean-square difference (RMSD) of superimposed C^{α} atoms to measure sequence and structural similarity, respectively. The percent sequence identity and RMSD have shortcomings as measures of sequence and structural similarity (Brenner *et al*., 1998; Levitt & Gerstein, 1998), and thus a new evaluation of the correlation of sequence and structural similarity of homologous proteins is warranted.

To measure the correlation of sequence and structural similarity, we used modern database searching programs to estimate the significance of the sequence and structure similarity for 36 protein families (Table 1) with five or more known structures from sequences that are less than 80 % identical. We find that most of the evolutionary structural change in a protein family is linearly related to changes in sequence similarity, when plotted in terms of statistical significance or as RMSD *versus* percent identity. Although we detected significant non-linear components in the relationship between sequence and structural similarity, these additional components explained very little of the structural variance, supporting a largely global view of protein fold specificity. The slope of the linear fit of sequence/structure similarity defines how much the structure of a protein is expected to change with a given amount of sequence change. We call this quantity the structural mutation sensitivity and show that it differs among protein families and is not correlated with protein structural class or protein family mutation rate.

Results

Sequence and structural similarity

Although percent sequence identity is routinely used to quantify sequence similarity, it has been known for more than 20 years that similarity scores based solely on sequence identity perform poorly when compared to substitution matrices that recognized conservative substitutions with

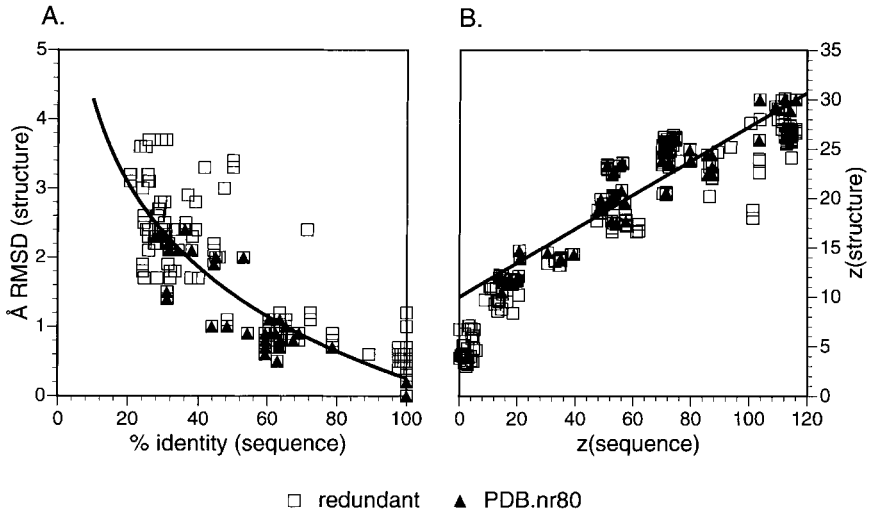

Figure 1. Sequence similarity and structural similarity from two perspectives (blue copper-binding proteins). A total of 346 homologous pairwise sequence alignments (ssearch3) and structure alignments (DALI) from queries of the DALI PDB database with 14 members of the blue copper-binding protein family (azurins, pseudoazurins, plastocyanins, etc.) are shown. Open squares report all the structural alignment pairs found in DALI for this family; filled triangles show only alignments from pdb.nr80 from structures determined to better than 2.2 Å that have statistically significant sequence and structural alignment scores. (a) The ssearch3 (Smith-Waterman) sequence alignment percent identity *versus* DALI average-carbon chain deviation (RMSD). (b) The statistical significance of the ssearch3 score (*z*-score, standard deviations above the mean) *versus* the statistical significance (*z*-score) reported by the DALI program. The RMSD *versus* percent identity relationship appears non-linear (Chothia & Lesk, 1986), while the relationships between the statistical significance of the structural and sequence statistical significance measures is linear, with a correlation coefficient of 0.93 for the filled triangles (line shown). The correlation coefficient for all the data points is 0.91.

Table 1. Protein families examined

Code	Description	Total	pdb.nr80	Class
SPR	Serine proteases (trypsins)	341	31	β
GLB	Globins	334	29	α
CAL	Calcium-binding EF hands	122	19	α
TOX	Snake neurotoxins	45	16	β
AAA	Alpha amylase	43	15	α/β
AZN	Blue copper-binding	104	14	β
PEP	Pepsins	74	13	β
THX	Thioredoxin/GST	85	11	α/β
PLI	Phospholipase A2	43	11	α
FAP	Fatty acid-binding protein	38	11	β
HOX	Homebox proteins	22	11	α
LPD	Lipoamide dehydrogenase	51	10	α/β
STX	Scorpion toxins	15	10	β
CYC	Cytochrome *c*	53	9	α
FXN	Ferredoxins	28	9	α + β
FRD	2Fe-2S ferredoxins	17	9	α + β
MIP	Macrophage inflammatory protein	60	7	α + β
ABP	Arabinose-binding protein	50	7	α/β
PER	Peroxidases	49	7	α
VCP	Viral coat proteins	49	7	β
PTI	Trypsin inhibitor	35	7	α + β
ADK	Adenylate kinase	21	7	α/β
LZM	C-type lysozymes	149	6	α + β
RIP	Ribosome-inactivating protein	20	6	α + β
DFR	DHFR	79	5	α/β
TPI	Triose phosphate isomerase	61	5	α/β
ADH	Alcohol dehydrogenase	56	5	α/β
SEL	Selectins	54	5	α + β
ISM	Peptidyl-prolyl isomerase	40	5	β
CPR	Cysteine proteases	38	5	α + β
ATH	Antithrombin	30	5	α + β
FVN	Flavodoxin	26	5	α/β
ANX	Annexin	18	5	α
PHC	Phosphocarrier protein	11	5	α + β
SOD	Superoxide dismutase	22	5	α + β
SUB	Subtilisin	44	5	α/β

similar biochemical properties (Pearson, 1995; Schwartz & Dayhoff, 1978); recently, shortcomings in the percent identity measure were demonstrated on a database of sequences whose structures are known (Levitt & Gerstein, 1998). Percent identity is a poor measure of sequence similarity in part because of its dependence on the length of the alignment (Sander & Schneider, 1991). While 30% identity is widely cited as a threshold of homology, short alignments often share 30% identity by chance and, for alignments of less than ten amino acid residues, 100% identity is not a significant similarity. A similar length dependence can be seen in structural comparisons using RMSD measures (Swindells, 1996).

Modern similarity searching programs, such as BLAST (Altschul *et al*., 1990, 1994, 1997), FASTA (Pearson & Lipman, 1988), and ssearch3 (Pearson, 1996) do not use percent identity, or even raw similarity scores, to characterize protein sequence similarity; they use bit scores, probabilities, or expectation values that reflect the statistical significance of the alignment score. The BLAST family of programs use a normalized bit score (Altschul *et al*., 1994) to characterize sequence similarity; FASTA and ssearch3 programs use a library sequence length-corrected *z*-value (Pearson, 1998). The probability of obtaining a sequence alignment bit score or *z*-value† by chance can be calculated using the extreme value distribution (Altschul *et al*., 1994; Altschul & Gish, 1996; Mott, 1992). We and others have shown that length-corrected *z*-values are the most effective at detecting distant evolutionary relationships (Brenner *et al*., 1998; Pearson, 1995, 1998).

The DALI structure comparison program also calculates a *z*-value to represent the significance of a structural similarity (Holm & Sander, 1993). In this case it is less clear how to transform the *z*-value into a probability of similarity by chance, but recent work by Levitt & Gerstein (1998) suggests that the extreme value distribution accurately describes structural similarity between non-homologous proteins as well as sequence similarity. Just as sequence *z*-values are more sensitive than percent identity, structural *z*-values can detect statistically significant structural similarities that are not readily identified using RMSD (Levitt & Gerstein, 1998).

† The *z*-value expresses a similarity score, s in terms of σ, the number of standard deviations from the mean similarity score μ, in a search of a database of unrelated sequences; $z = (s - \mu)\sigma$.

In addition to providing more sensitive measures of similarity, sequence and structure *z*-values can provide reliable information about the relationship between sequence similarity and structural similarity. Figure 1(a) shows a common representation of sequence/structure relationships by plotting the increase in the average C^{α} deviation between protein three-dimensional structures as their protein sequences diverge. Figure 1 shows the relationship between sequence and structure for the blue copper-binding proteins (azurins and plastocyanins), but very similar plots are seen for most protein families, including globins, serine proteases, and lysozymes (Chothia & Lesk, 1986). A common feature of these identity/RMSD plots is the dramatic increase in RMSD as comparisons are made between proteins that do not share significant structural similarity.

Figure 1(b) shows an alternate view of the sequence structure relationship, by comparing the *z*-value of the structural similarity shared by two proteins (as determined by the DALI structural alignment program) with the *z*-value of the proteins' sequence similarity (determined by the Smith-Waterman algorithm; Pearson, 1996; Smith & Waterman, 1981). In contrast to the curve seen in Figure 1(a), comparison of structural statistical significance with sequence statistical significance suggests a linear relationship between sequence similarity and structural similarity for this family ($r = 0.91$).

While structural and sequence *z*-values are less commonly used to measure similarity than percent identity and RMSD, they provide an accurate representation of these more familiar measures (Figure 2(a) and (c)). The *z*-values provide the additional advantage that one can know, in advance, whether the structural or sequence alignment is statistically significant. For the globin family and the pdb.nr80 database, sequence alignments with *z*-scores greater than nine are statistically significant ($p() < 0.01$), and the alignments that obtain that level of significance tend to align the entire globin sequence (Figure 2(b)). When the

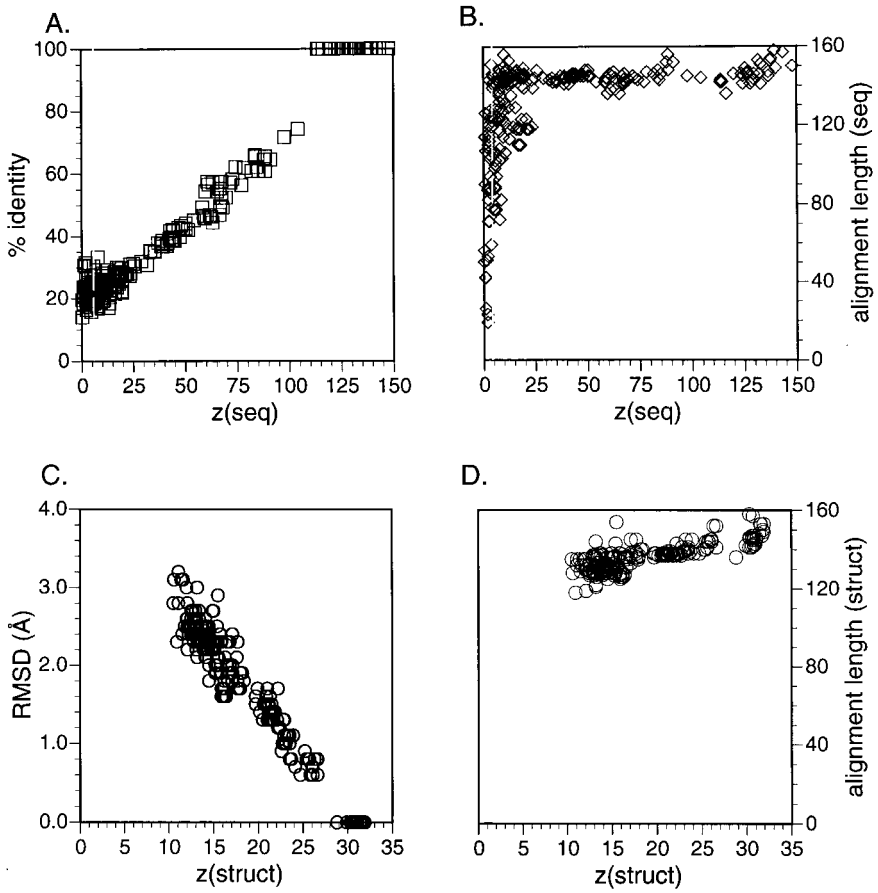

Figure 2. Percent identity and alignment length *versus* *z*-score. Alignments between 254 sequence/structure pairs from the globin family (pdb.nr100). (a) Sequence alignment percent identity *versus* sequence similarity *z*-score. (b) Sequence alignment length *versus* sequence similarity *z*-score. The alignment length is relatively constant near 145 residues, the average length of a globin family member, until the sequence alignment *z*-score drops below $z < 9$, the threshold for statistical significance for this library. (c) Structural alignment RMSD *versus* structural similarity *z*-score. (d) Structural alignment length *versus* structural similarity *z*-score. All the structural alignments are statistically significant and align the entire length of the structure.

two sequences do not share statistically significant similarity, the Smith-Waterman algorithm frequently aligns only a portion of the two sequences (Figure 2(b)), even a DALI structural alignment usually includes all the structure (Figure 2(d)).

Correlating sequence and structural similarity

The apparent biphasic relationship between structural change and sequence change (Figure 1(a)), supports the local model for protein folding. From the local perspective, sequence divergence from 0 %-50 % has relatively little effect on RMSD, while sequences that are more than 80 % divergent show large structural deviations. The *z*-score-based analysis (Figure 1(b)) supports the alternative global model; there is a linear correlation between sequence change and structural change. This apparent contradiction appears largely due to three problems that are encountered when comparing protein structures and sequences: (1) severe redundancy of protein structural data determined under different conditions, which can cause the same protein sequence to have considerable variation and thus obscure the structure/sequence relationship; (2) problems in aligning protein structures, even when the sequence similarity is quite high; and (3) problems with assigning an accurate percent identity in very distantly related sequences.

The effects of structure redundancy and alignment accuracy on the apparent sequence/structure relationship can be seen in Figures 1, 3 and 4. In contrast to the open squares in Figure 1, which show all the blue copper-binding protein structure/sequence relationships in the fully redundant PDB database (349 alignments), the filled triangles show the relationship between sequence and structure for those sequences and structures in the non-redundant pdb.nr80 that share statistically significant sequence and structural similarity (49 alignments). By focusing on sequences and structures that share significant similarity, it is more likely that the structural and sequence alignments are accurate. This constraint removes many of the points between 20 and 25 % sequence identity. By limiting our analysis to the far less redundant pdb.nr80 structures, we avoid analyzing structures that differ by as much as 1-2 Å RMSD, despite the fact that they are 99-100 % identical. When only the filled triangles are considered, the correlation coefficient for a linear regression of RMSD (structure) *versus* percent identity (sequence) increases slightly to $r = 0.93$ from $r = 0.91$ for the fully redundant data.

Figures 3 and 4 illustrate the effect of sequence/structure redundancy and structural resolution on the correlation of structure with sequence in the globin family. When every globin structure in the PDB is included in the analysis (Figure 3(a), 1813 alignments, see the legend), the linear correlation coefficient between *z*(structure) and *z*(sequence) is $r = 0.83$; when only alignments from pdb.nr80 between sequences that share statistically significant sequence similarity are included, the correlation increases to $r = 0.96$ (Figure 3(e)). When the comparisons are done using percent identity and RMSD, the correlation coefficient increases from $r = 0.81$ (Figure 4(a)) to $r = 0.92$ (Figure 4(e)). Comparison of the left and right panels in Figures 3 and 4 shows that removing lower-resolution structures and non-significant sequence alignments can both decrease and increase the correlation coefficient. However, as the redundancy of the structural data is reduced (Figures 3(a), (c), (e); 3(b), (d) and (f); 4(a), (c) and (e); or 4(b), (d) and (f)) the structure/sequence correlation consistently increases. This increase in structure/sequence correlation is expected; by reducing the number of different structures from identical (pdb.nr100), or nearly identical (pdb.nr80) sequences, the sequence-independent variation (different structures from the same sequence) is reduced.

The large non-sequence dependent structural variation appears to justify examination of structure/sequence relationships in a less redundant database, because there is just as much (actually slightly more) structural variation among sequences that are more than 97.5 % identical (153 sequences, RMSD ranges from 0 to 1.7 Å, *z*(structure) ranges from 13 to 21) as there is among sequences that do not share statistically significant sequence similarity (149 sequences, RMSD = 1.7-3.2 Å, *z*(structure) = 5.5-13; Figures 3(a) and (b) and 4(a) and (b)). While one might expect considerable variation among related proteins with very low sequence similarity, an equally large amount of variation among sequences that are less than 2.5 % different does not reflect genuine evolutionary differences. When sequences that are more than 80 % identical are removed from the set of alignments (Figure 3(f)), the amount of structural variation among the 17 globins aligned with themselves is 0 Å RMSD (*z*(structure) = 19-21). Diverse sequences show somewhat more structural variation; for the 20 weakest globin alignments with statistically significant sequence and structural alignment scores, RMSD ranges from 1.9-2.7 Å (*z*(structure) = 8.5-11). Thus, we focus on structure/sequence relationships among members of the less redundant pdb.nr80 dataset, because by reducing the extreme redundancy in the PDB database, we remove very large amounts of structural alignment ''noise'' that reduces our ability to explore sequence/structure relationships.

A more comprehensive evaluation of the relationship between sequence change and structural change is shown in Figure 5. Normalized *z*-score structure/sequence correlation coefficients were determined for each of the 36 protein families with five or more members in pdb.nr80 (Table 1). Half of the protein families had linear correlation coefficients greater than 0.878, implying that more than 75 % of the structural variance for these families could be explained by sequence variance. Table 2 summarizes the median, first, and third

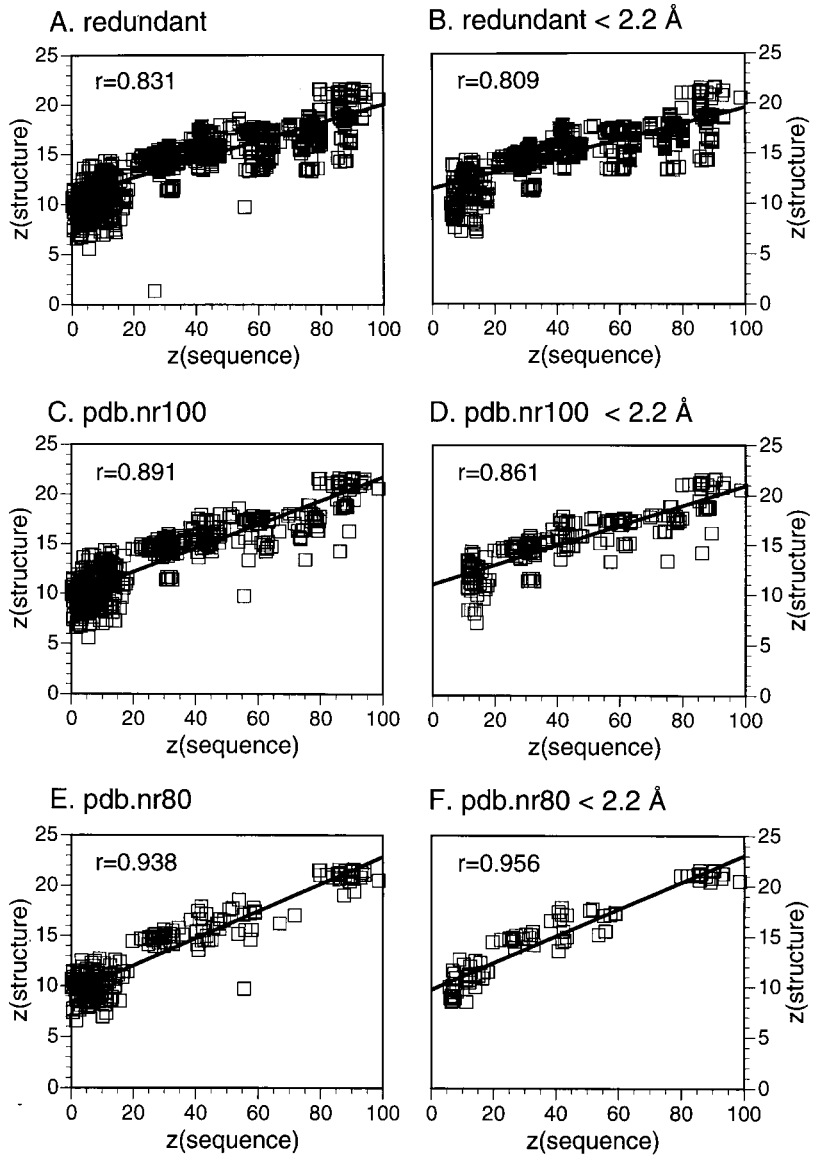

Figure 3. Sequence/structure correlations: redundancy and resolution. The relationship between sequence similarity and structural similarity is shown for six samples of the globin family structures. (a) Structural similarity (normalized DALI *z*-score) *versus* sequence similarity (normalized ssearch3 *z*-score) for all 334 globin structures in the fully redundant PDB database. A total of 1812 structural alignment scores reported by DALI are shown. (b) Structural similarities are shown only for the structures determined by crystallography to better than 2.2 Å resolution with statistically significant sequence and structural similarity (1185 alignments). Structural similarities are shown only for structures in the non-redundant (c) pdb.nr100 and (e) pdb.nr80 databases. Statistically significant similarities are shown for structures in the (d) pdb.nr100 and (f) pdb.nr80 databases that were determined to <2.2 Å resolution. Linear correlation coefficients for the sequence/structure regression lines are shown. (a) Excluded three DALI alignments between the globin structures 1myf and other globins, because two of the three alignments were not structurally significant although the sequence alignments showed >50% and >90% sequence identity. Examination of the structurally implied sequence alignment reported by DALI and the superimposed structures clearly show that the structural alignment was incomplete. Likewise, all three 1mbs alignments were excluded because of very poor structural alignments. The point in (a) where $z(\text{seq}) = 27$, $z(\text{struct}) = 1.4$ (1spg *versus* 1baba) is another example of a clearly incorrect DALI structural alignment, but other alignments with these two structures are correct. Removing this point improves the correlation coefficient to $r = 0.835$.

quartile linear correlation coefficients for both the *z*(structure)/*z*(sequence) and RMSD/percent identity data for the structures summarized in Figure 5, and also for the more redundant datasets and those containing lower resolution structures. The linear correlation coefficients increase slightly when RMSD and percent identity are plotted for the pdb.nr80, pdb.nr100, and redundant datasets, except for the median of the high-resolution, fully redundant set. Thus, when alignments between structures in pdb.nr80 are examined, more than 75% of the structural variance can be explained by sequence variation in 18 of the 36 families; and when alignments between all the redundant structures in the PDB are considered, 53 (*z*(structure)/*z*(sequence)) to 60% (RMSD/percent identity) of the structural variance is due to sequence variance in half the families. In contrast, linear correlation coefficients calculated for alignments between non-homologous proteins averaged 0.262. These high correlations within protein families imply that the structural variation of each family is largely the result of variation in the sequence.

The linear correlation coefficient between sequence and structural similarity for each family can be thought of as the amount of structural variation that is explained by sequence variation. Correlation coefficients depend on two factors, the slope of the structural-similarity/sequence-similarity relationship and the amount of residual noise that does not depend on sequence change. The top panel for each protein family in Figure 5

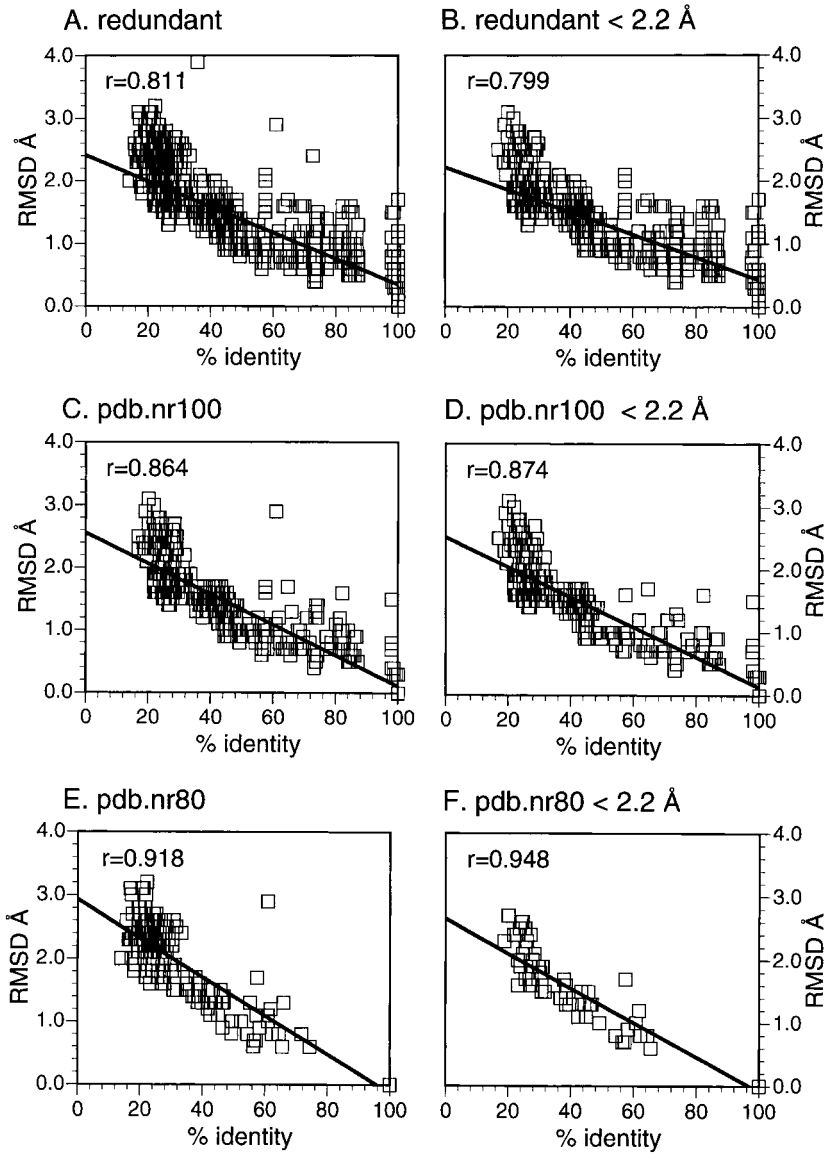

Figure 4. Sequence/structure correlations: percent identity and RMSD. The same pairwise comparisons shown in Figure 3 are plotted using percent sequence identity and DALI structural similarity reported as RMSD for comparisons of globin family members. (a) The fully redundant pdb. (b) Alignments involving structures determined to <2.2 Å. All or high-resolution structures from (c), (d) pdb.nr100 or (e), (f) pdb.nr80. Three structures are excluded as noted in Figure 3.

shows the structure/sequence correlation coefficient; the bottom panel shows the difference between each structure/sequence data point and the regression line (the residuals). In general, the families with the highest correlation coefficients have the least amount of scatter around the regression line (e.g. ADH, LZM, and PER) or they have a large number of data points with a high slope (AZN). We have the highest confidence in the structure/sequence relationship when the data samples a wide range of sequence similarity. For some families, most of the data points involve either very high or very low sequence similarity (e.g. ABP, DFR, and HOX); in these cases new structural data from sequences that are intermediate in similarity might reduce (or improve) the correlation. Nevertheless, currently available data show a very strong linear correlation between structural change and sequence change for most of the protein families examined.

Although most of the variation in structural similarity can be predicted by linear changes in sequence similarity, it appears that at low sequence similarity (<25 % identity), the slope of the structure/sequence relationship changes in some protein families (Figures 1(a), 3(f) and 4(f)). We examined these potential ''higher-order'' effects by fitting the data to a quadratic polynomial and restricted cubic spline polynomials (four knots). Linear, quadratic, and restricted cubic spline fits to the azurin blue copper-binding protein and globin families are shown in Figure 6. (These families were selected because they represent the amount of improvement obtained with a restricted cubic spline fit by half or three-quarters of the families.) Statistical analysis of the higher-order polynomial fits show that the coefficients of the higher-order terms were significantly different from zero, and thus that the higher-order terms provided an additional significant reduction in variance. How-

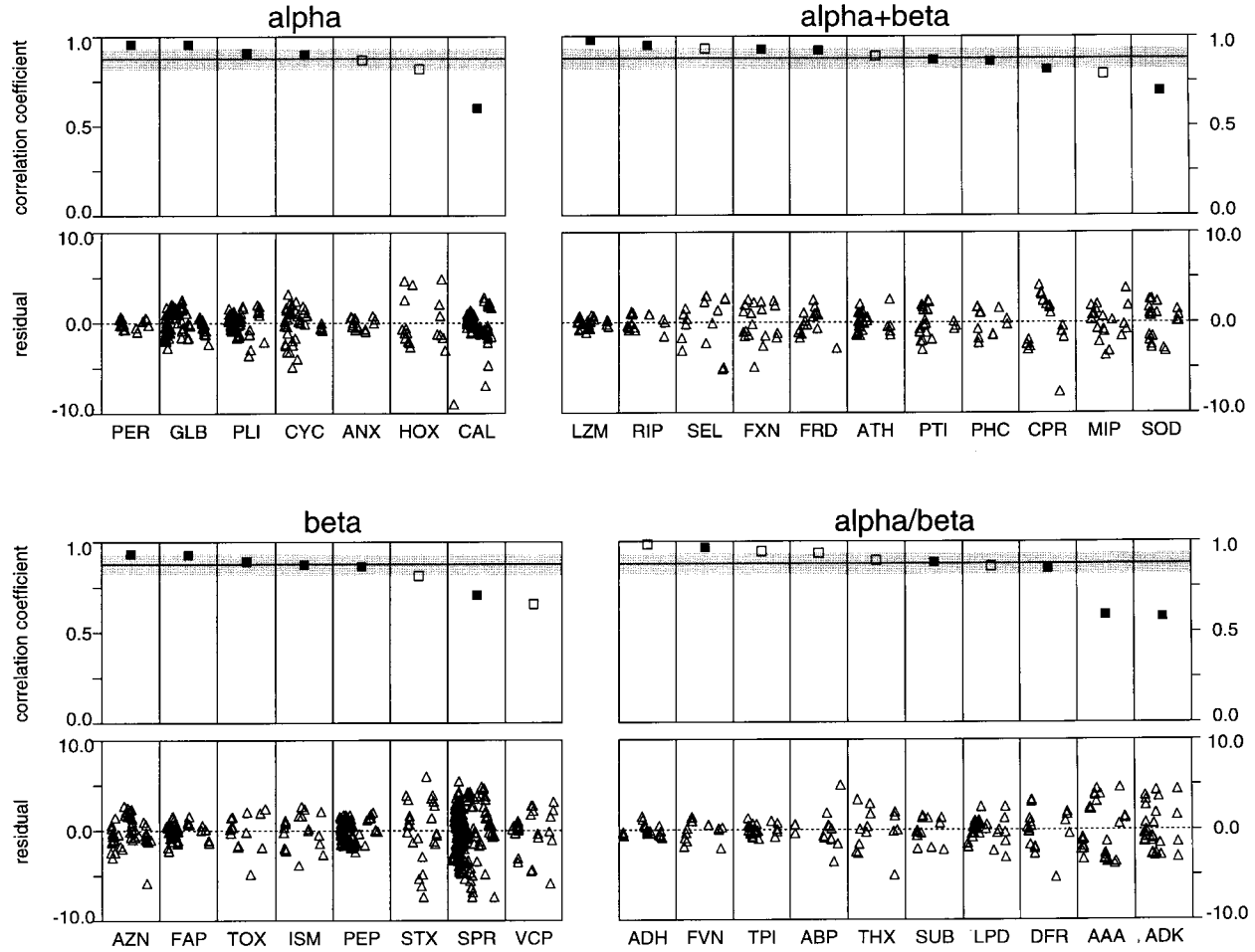

Figure 5. Sequence similarity and structural similarity is highly correlated. Linear correlation coefficients for comparisons of normalized ssearch3 *z*-scores with normalized DALI *z*-scores for 36 protein families with five or more members in pdb.nr80 are shown. Whenever possible, data from structures determined to less than 2.2 Å are shown (filled symbols) when more than ten structure/sequence pairs were available. For MIP, low-resolution pdb.nr100 data are shown. The other open symbols summarize low-resolution pdb.nr80 data. The horizontal line indicates the median correlation (0.878) from the least redundant, highest-resolution datasets; the shaded region indicates the upper (0.931) and lower (0.815) quartiles. The bottom half of each panel reports each of the residual distances between the least-squares regression line and the structure/sequence similarity pairs. Family codes are shown in Table 1.

ever, as Figure 6 shows, the effect of the additional terms is modest. To measure the magnitude of this effect, we calculate the ''adequacy'' of linear *versus* quadratic or restricted spline fits to the data (see Methods). For the data *z*(structure) *versus* *z*(sequence) relationships shown in Figure 5, the median adequacy for the linear *versus* quadratic fit was 0.983, indicating that 98.3 % of the structural variance that could be explained by the more complex fit could be explained by a linear relationship (i.e. the ratio of the adjusted r^2 of the linear model to the non-linear model was 0.983). The first and third adequacy quartiles were 0.996 and 0.955, indicating that for 25 % of the families, the quadratic fit provided almost no improvement, and for 75 % of the families, the improvement was less than 5 %. For the linear *versus* restricted spline fits, the median adequacy was 0.979; the first and third quartiles were 1.000 and 0.944. For the RMSD *versus* percent identity, linear/quadratic adequacy quartiles were 1.000, 0.988, and 0.951; linear/spline quartiles were 0.995, 0.971, and 0.935. Thus, although higher-order terms do significantly improve the structure/sequence fit, the additional terms account for less than a 5 % relative improvement in the variance in half the families, and less than 7 % in three-quarters of the families. Almost all the relationship between structural similarity and sequence similarity can be explained by a linear model.

The BLOSUM50 scoring matrix (Henikoff & Henikoff, 1992) was used to measure sequence similarity for the correlations in Figure 5, because it performs well in identifying distantly related sequences (Pearson, 1995). However, if protein fold specificity were coded by a small number of critical residues, we might expect that the BLOSUM80 matrix, which is derived from the most highly

Table 2. Sequence/structure linear correlation coefficients

	Median		First quartile		Third quartile	
	z(str)/ *z*(seq)	RMSD/ %ident.	*z*(str)/ *z*(seq)	RMSD/ %ident.	*z*(str)/ *z*(seq)	RMSD/ %ident.
Figure 5	0.878	0.916	0.931	0.953	0.815	0.839
pdb.nr80	0.862	0.908	0.916	0.952	0.798	0.836
pdb.nr100 < 2.2 Å	0.870	0.876	0.915	0.922	0.760	0.805
pdb.nr100	0.797	0.840	0.872	0.889	0.727	0.753
Redundant < 2.2 Å	0.841	0.803	0.885	0.905	0.657	0.707
Redundant	0.732	0.777	0.811	0.840	0.597	0.634

Linear correlation coefficients determined for the datasets shown for either the *z*(structure)/*z*(sequence) relationship or the RMSD/ percent identity relationship. The median correlation coefficient *r*, and correlation coefficients for the first quartile (75% of the families score worse) and third quartile (25% of the families score worse) are shown. Data from 36 families are shown for the fully redundant and pdb.nr100 datasets. pdb.nr80 is missing one family (MIP); pdb.nr100 < 2.2 Å is missing seven families and the fully redundant <2.2 Å set is missing two families.

conserved residues in protein blocks (Henikoff & Henikoff, 1992), might increase the correlation between sequence and structure. If certain highly conserved residues largely determine a protein's structure, the correlation between sequence and structural similarity should go up when BLOSUM62 or BLOSUM80 are used for sequence comparison. To test this, we repeated the previous experiments using the BLOSUM62 and BLOSUM80 matrices (Figure 7). The correlation coefficients for 24 of the 36 protein families (67%) decrease when conservative comparison matrices are used, but the decrease is never significant. For only 12 of the 36 protein families, the correlation coefficient calculated using the BLOSUM62 matrix is greater than the correlation with the BLOSUM50 matrix. This evidence further supports the global model; for most protein families, the most highly conserved residues do not account for more of the structural variability than ''average-conserved'' residues.

Structural mutation sensitivity differs among families

The slope calculated from a regression of structural *versus* sequence similarity provides further insight into the structural evolution of protein families. The slopes of these regression lines estimate how much a family's structure would be expected to change with a given amount of

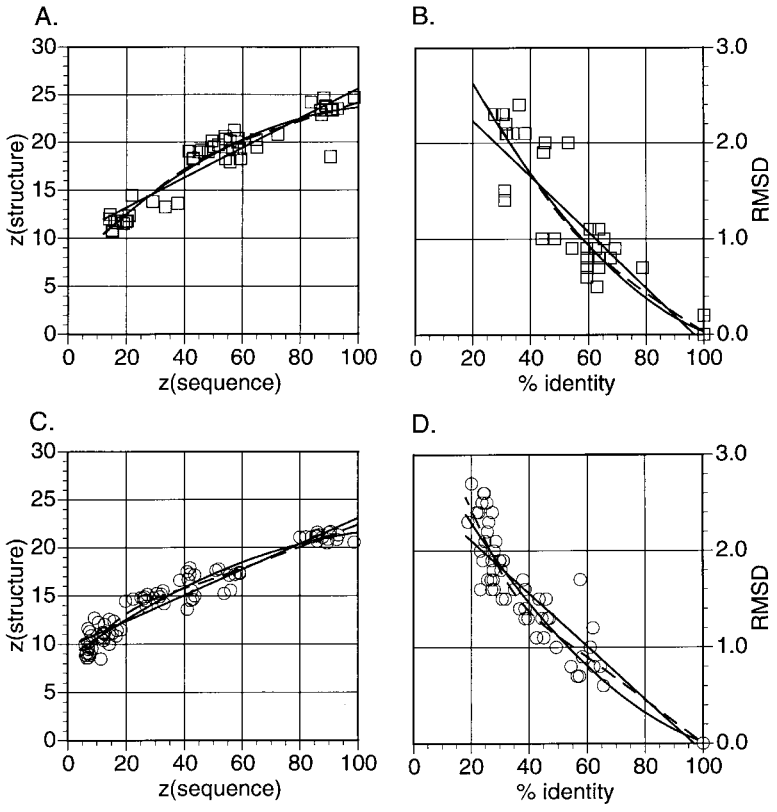

Figure 6. Higher-order relationships between sequence and structural similarity. Regressions using non-linear sequence/structure terms are shown for blue copper-binding proteins (a), (b), and globins (c), (d). Regression fits to a line, a quadratic equation, and a restricted cubic spline (four knots) for either (a), (c) *z*(structure)/*z*(sequence) data or (b), (d) RMSD/% identity data are shown. The azurin blue copper-binding family demonstrates the median improvement found after fitting the quadratic or spline functions; the globin family shows an amount of improvement achieved by only 25% of the protein families.

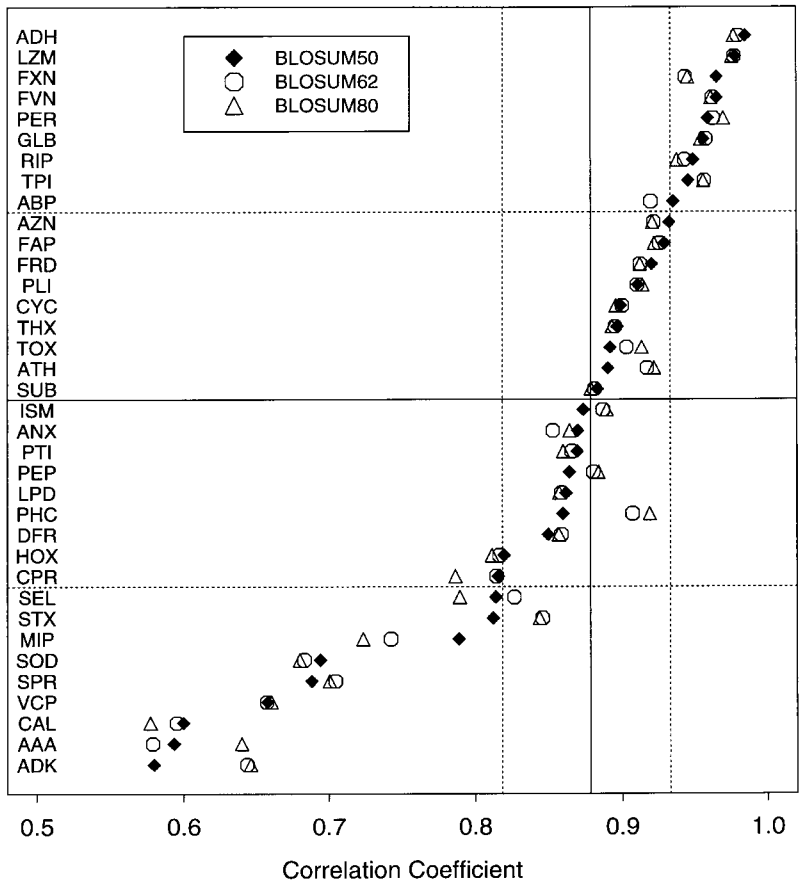

Figure 7. Structure/sequence correlations do not vary with different BLOSUM matrices. Sequence similarities were calculated with BLOSUM50 (◆), BLOSUM62 (○), and BLOSUM80 (△) matrices, and the correlation coefficient of structural similarity with sequence similarity is shown for least-redundant, high-resolution datasets as in Figure 5. Median (continuous lines) and upper and lower quartiles (broken lines) for the BLOSUM50 correlation coefficients are indicated.

sequence change, i.e. the structural mutation sensitivity. A small slope implies that the protein structure for that family is not very sensitive to sequence mutation; that is, the structure will change very little over the range of sequences in the family. In contrast, a large slope signifies protein structures that differ even with modest sequence changes. As seen in Figures 8 and 9, mutation sensitivity can vary widely among protein families.

Structural mutation sensitivity varies 3.9-fold (0.0545-0.213) among different protein families (Figures 8 and 9). The viral coat proteins (VCP, slope = 0.06, $r = 0.66$) might be expected to have a low structural mutation sensitivity because viral replication can be error-prone and substantial sequence divergence must be accommodated. In contrast, a protein family like the globins might be expected to have high structural mutation sensitivity, because of the large number of internal contacts that stabilize the packed helices. However, there does not appear to be an obvious relationship between structural mutation sensitivity and structural class (Figure 9). In the 36 protein families that we examined, proteins with similarly low structural mutation sensitivity are found in the α (annexins, ANX), β (viral coat proteins, VCP), α/β (adenylate kinase, ADK), and $\alpha + \beta$ (superoxide dismutase, SOD) structural classes. In general, protein families with lower structural mutation sensitivity have lower structure/sequence correlation coefficients, as expected from the definition of correlation coefficient. Less of the structural variance can be explained by sequence change when there is little change in structure with large changes in sequence. The largest structural mutation sensitivities are found in the β (scorpion toxins, STX, slope = 0.21) and $\alpha + \beta$ (ferredoxins, FXN, slope = 0.20) structural classes, while α and α/β proteins have a slightly lower range of structural mutation sensitivity (threefold). Nonetheless, it is striking that similar amounts of sequence change can cause dramatically different amounts of structural change. Since structural mutation sensitivity is not tightly associated with protein structural class, we considered two additional factors affecting mutation sensitivity: the average length of members of the family and the protein family divergence rate.

Differences in structural mutation sensitivity might simply be the result of size differences among the protein families. Single amino acid substitutions primarily influence the local conformation of a protein structure (Bowie *et al.*, 1990; Matsumura *et al.*, 1988; Sandberg & Terwilliger, 1989). In a large protein, that local change is small compared to the rest of the protein structure that stays the same. In a small protein, local changes

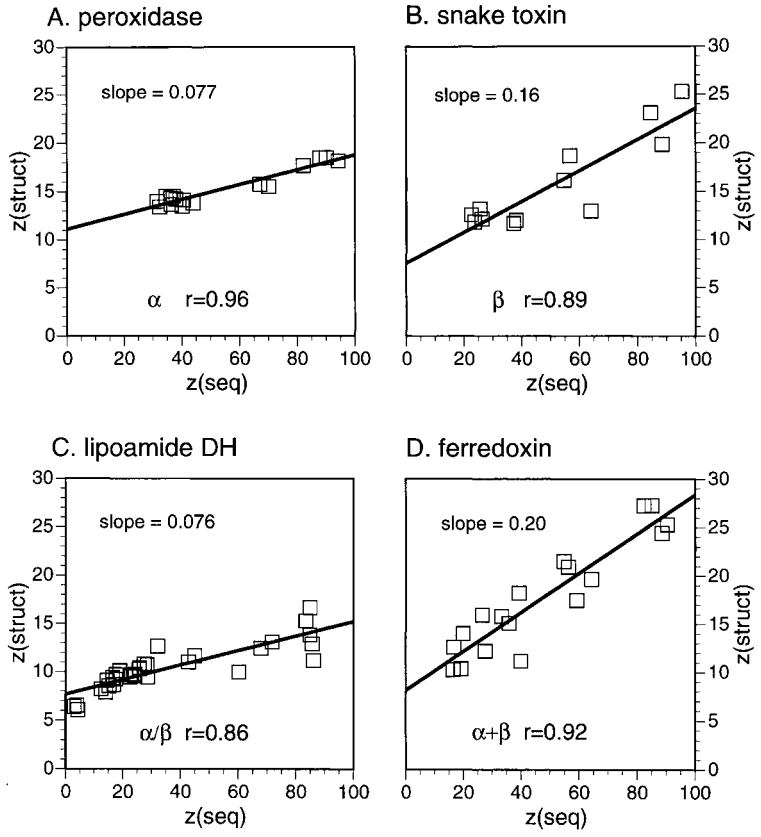

Figure 8. Structural sensitivity to sequence change varies among protein families. Length-normalized sequence-structure similarity relationships are shown for four protein families: (a) peroxidases (PER, α, 16 sequence/structure pairs); (b) snake neurotoxins (TOX, β, 12 pairs); (c) lipoamide dehydrogenase (LPD, α/β, 33 pairs); and (d) ferredoxins (FXN, α + β, 18 pairs). Linear correlation coefficients of the regression line for structure *versus* sequence are shown next to the structural class. The slope of the structure/sequence relationships (structural mutation sensitivity) is also shown.

should be much more noticeable in the molecular context. Thus, one might expect to find that the structural mutation sensitivity of large proteins is less than that of small proteins.

We compared the structural mutation sensitivity to the average protein size for each family (Figure 10(a)). As expected, families of large proteins have low mutation sensitivities, but surprisingly, small proteins have a wide range of structural mutation sensitivities. For example, calcium binding EF-hands (CAL), a protein family of length 130 residues, had a structural mutation sensitivity of 0.069, which is lower than that of α-amylases (AAA, structural mutation sensitivity 0.076), the longest protein (525 residues) in our dataset. Another small protein, MIP (average length 71 residues) had a structural mutation sensitivity of 0.092. Because we calculated the structural mutation sensitivity using sequence and structural similarity *z*-scores that were normalized for the query length, the mutation sensitivity/length relationship does not reflect a dependence of similarity on length. The variation in structural mutation sensitivity among small and medium length proteins indicates that protein size is not the only factor that determines structural mutation sensitivity.

Another possible influence on structural mutation sensitivity is a protein family's ''molecular clock'' or mutation rate. The low mutation sensitivity of the viral coat proteins suggests that mutation sensitivity may be directly related to mutation rate; proteins that evolve rapidly may have low mutation sensitivities to preserve their structure, and thereby their function. Conversely, proteins with high structural mutation sensitivity may not be able to tolerate a high rate of sequence mutation. Mutation rates can be investigated by examining average mutation rates for orthologous lineages.

Evolutionary mutation rates for the 16 families with independently derived dates for the last common ancestor (LCA) and clearly orthologous lineages were estimated by two different methods (Table 3). The method by Langley & Fitch (1974) derives the estimate from a least-squares regression fit of evolutionary distance and LCA dates. Xun Gu's method (Gu & Zhang, 1997) accounts for heterogeneity of mutation rate among the residues of the protein sequence. The ordering of the mutation rates is moderately consistent between the Langley-Fitch method and the Gu method (Table 3). Families that are faster (or slower) than the median with the Langley-Fitch are always faster (or slower) than the median with Gu's method. Complete consistency is not expected, since Gu's method is more rigorous than Langley and Fitch's method, but the limited consistency reassures us that the rate calculations are sensible. A comparison of the mutation rates to the structural mutation sensitivity (Figure 10(b)) shows that the proteins with the lowest mutation rates

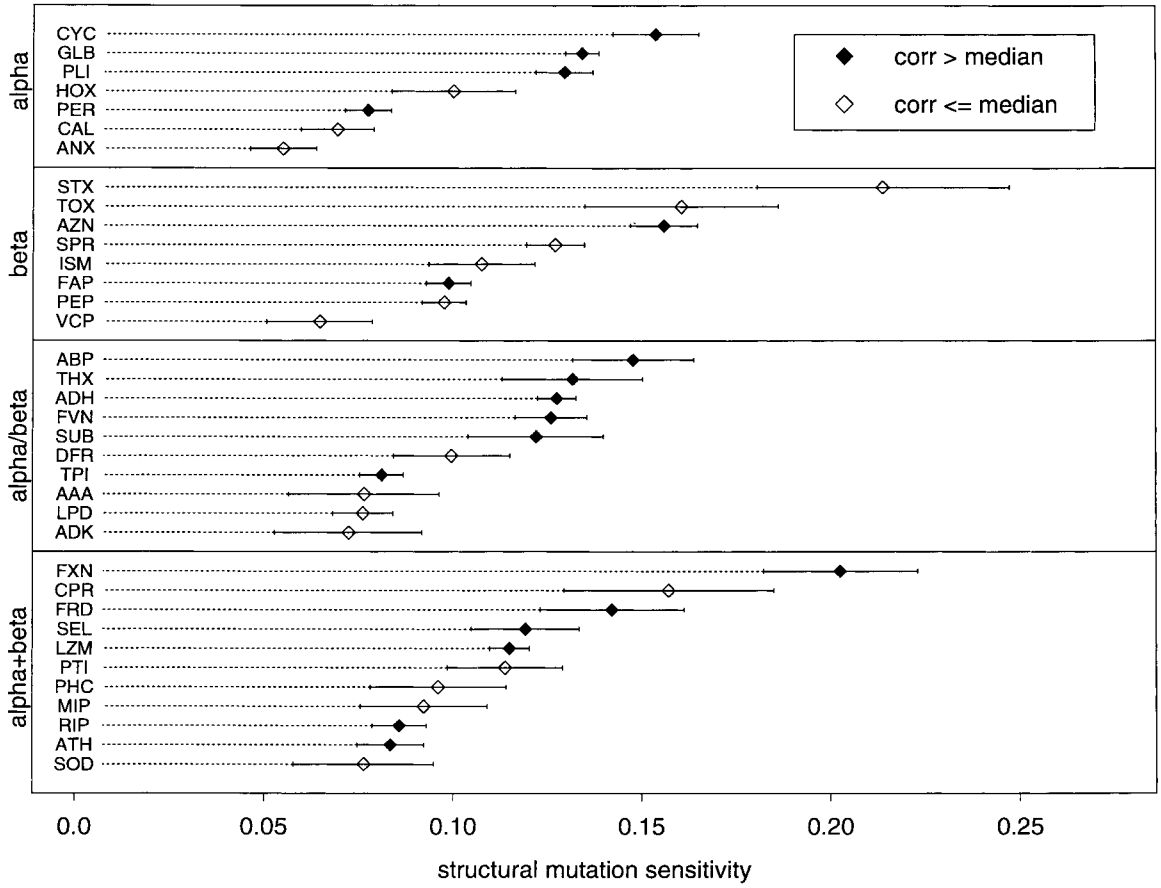

Figure 9. Structural mutation sensitivity does not depend on structural class. Structural mutation sensitivity (the slope of the sequence/structure relationship), and the standard error of the slope, is plotted for each of the 36 protein families. Filled symbols indicate families with a sequence/structure correlation greater than the median; open symbols show families with sequence/structure correlations less than or equal to the median.

(CAL, LPD, and SOD) have the lowest structural mutation sensitivity, but THX, with only a slightly higher mutation rate, has one of the highest structural mutation sensitivities, and ADK, with the highest mutation rate, has one of the lowest structural mutation sensitivities. In general, there appears to be no consistent relationship between structural mutation sensitivity and divergence rate, although generalizations must be tentative with such a small sample.

Table 3. Protein family divergence rates

Method	Langley-Fitch		Gu	
Fastest	GLBm	12.4	LZM	13.3
	GLBa	11.5	GLBa	11.6
	PEP	8.9	ATH	10.8
	GLBb	7.9	ADK	10.1
	ADK	7.1	GLBb	9.9
	LZM	6.0	PEP	9.5
	ATH	5.6	LPD	8.1
	PER	5.0	SPR	6.2
	SEL	4.1	GLBm	6.1
	CYC	3.3	DFR	5.7
	DFR	2.8	SEL	5.0
	ADH	2.6	TPI	4.2
	SPR	2.5	ADH	4.0
	THX	2.2	PER	3.8
	TPI	1.4	THX	2.9
	SOD	1.1	CYC	2.7
	LPD	1.0	SOD	2.7
Slowest	CAL	0.6	CAL	0.6

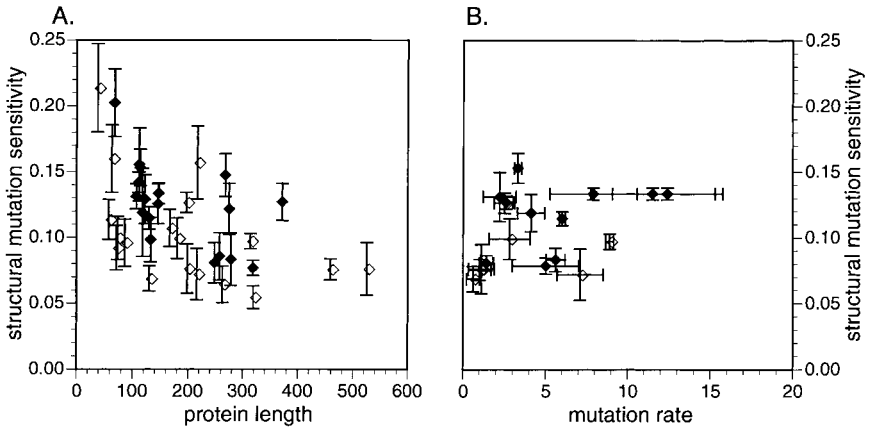

Figure 10. Structural mutation sensitivity does not depend on mutation rate or protein size. (a) The relationship between structural mutation rate and average protein family length. Filled symbols indicate families with sequence/structure correlation coefficients greater than the median; open symbols indicate families with correlation coefficients less than or equal to the median. Error bars report the standard error of the structural mutation sensitivity slope estimate. (b) The structural mutation sensitivities from Figure 4 are plotted against protein family mutation rates (Langley & Fitch, 1974) for 16 protein families for which a mutation rate could be estimated.

Discussion

We have examined the relationship between sequence change and structural change in 36 protein families with five or more diverse members whose structures are known. For most of the protein families that we examined changes in structural similarity are linearly dependent on changes in sequence similarity. In the globin family (Figure 2 and 3(d)), a change in a sequence z-score from $z = 15$ to 25 standard deviations above the mean (24.9 % identity at $z = 15$ to 30.3 % identity at $z = 25$) will change the structural z-score from $z = 11.7$ standard deviations (2.68 Å RMSD) to $z = 13.0$ (2.50 Å RMSD). Likewise, a change in sequence similarity from $z = 80$ (60.3 % identity) to $z = 90$ (65.7 % identity) changes the structural similarity from $z = 20.4$ (1.48 Å RMSD) to $z = 21.7$ (1.29 Å RMSD). Thus, for the globins, a ten standard deviation change from 24.9 % identity to 30.3 % identity has the same effect on structural similarity (Δ0.18 Å RMSD) as a sequence $\Delta z = 10$ change from 60.3 % identity to 65.7 % identity. This strong linear correlation is seen in at least three-quarters of the protein families (27 out of 36 families had $r \geqslant 0.815$, implying that almost 66 % of the variance in structural similarity could be accounted for by the change in sequence similarity).

Based on the results shown in Figures 3-6 and 8, we conclude that, on average, most sequence changes cause detectable structural changes and that the amount of structural change per sequence change (structural mutation sensitivity) is relatively constant within a protein family. While this observation is not completely unexpected (it is not surprising that changes in sequence cause changes in structure) our result is inconsistent with the widely held view (Chothia & Lesk, 1986, 1987), exemplified by the continuous line in Figure 1(a), that changes in protein sequence from 50-80 % identity are largely structurally neutral, but that below 30 % identity one sees changes in a core of highly conserved critical residues that determine the structure of a protein.

We believe that our conclusions differ from this conventional view for several reasons. First, we have focused on structure comparisons and sequence alignments between sequences that share statistically significant sequence similarity ($E < 0.001$). Sequences that do not share statistically significant similarity are difficult to align accurately (Figure 2(b)). Second, by working with a less redundant (pdb.nr80) set of protein structures, we reduce the structural variance for sequences that are 80-100 % identical. The major effect of using the less redundant dataset is to reduce dramatically the number of structures for a single sequence, which in turn reduces the amount of structural variation that is not the result of sequence change. For example, in the upper right-hand corner of Figure 2(a) and (b), there are several dozen structure/sequence pairs with structural z-scores that differ by more than ten standard deviations for sequences that are identical or nearly so. When the less redundant datasets are used, the amount of structural variation among these nearly identical sequences is reduced twofold and substantial structural noise is excluded.

Finally, the linearity of the structure/sequence relationship can be more readily detected in this study because we examine each protein family separately. Because of differences in structural mutation sensitivity, a composite structure/sequence plot that combined the points from all 36 protein families would show a relationship like the curved line in Figure 1(a). However, this curve would be simply the sum of the different linear structure/sequence relationships for each family.

Our results do not imply that all mutations have similar effects, even within a single protein family. However, virtually all mutations have measurable effects and, for most protein families, similar amounts of change in protein sequence appear to cause similar amounts of change in protein structure. The data are best for families with a large number of structures sampled across sequence *z*-scores from 10-100 or more, i.e. for families that have closely (>70 % identity), moderately (30-70 %), and distantly (15-30 %) related members. The strong linear correlation of sequence and structural similarity supports the global model for protein folding and suggests that, within a protein family, most changes in sequence are correlated with a constant, measurable change in structure. Although non-linear components of the sequence/structure relationship were detected, the adequacy of the linear component is typically greater than 95 %, supporting a mostly global model of protein folding.

Families with lower structure/sequence correlations must have other sources of apparent structural variation that are not accounted for by sequence change. For some families, ligand or ion-binding, structure-determination conditions, and random protein flexibility influence the structural variation of closely related proteins. Structural variation in the calmodulins is associated with the formation of protein complexes and with the binding of calcium and other ligands. For example, the structures of bovine recoverin with Ca^{2+} (1rec) and without (1iku) have an RMSD of 3.6 Å.

In another family where the structure/sequence relationship is weaker (serine proteases) the DALI program sometimes fails to align the structures properly and thus introduces artifactual structural variation. While DALI aligns only 117 residues of salmon elastase (1elt) and human plasminogen activator (1lmwB) to 1.7 Å RMSD, the structure alignment program SARF2 (Alexandrov, 1996) aligns 227 residues of the same protein pair to 1.69 Å. The shorter alignment of DALI is clearly erroneous and yields a much lower structural similarity score than the full alignment would. Unfortunately, SARF2 does not calculate *z*-scores.

We also examined the role of synthetic or mutant sequences in non-sequence dependent structural variation. Such sequences are excluded from pdb.nr80 because they are typically more than 80 % identical with wild-type sequences, but in the three largest protein families in pdb.nr100 96 of the 1097 serine proteases are non-wild-type sequences; globins contain 123 of 591, and lysozymes contain 126 of 233. Non-wild-type sequences did not contribute significantly to the other families. When the non-wild-type sequences and wild-type sequences were analyzed separately, the structure/sequence linear correlation coefficients differed by less than 10 % and structural mutation sensitivities by as much as 20 %. Thus, for these three families at least, it appears that non-wild-type structures show the same structure/sequence dependence seen in ''natural'' sequences.

Surprisingly, we also found that the structural mutation sensitivity can vary as much as fourfold among different protein families. Despite efforts to relate these differences to average structural class, mutation rate, and protein size, we were unable to account for the variation in mutation sensitivity. Differences in structural mutation sensitivity may reflect differences in the nature and extent of interactions between basic elements of secondary structure.

Our attempts to understand better the biophysical basis for differences in structural mutation sensitivity were hampered by the lack of structures from proteins that sample a broad range of sequence divergence within a protein family. Of the 1326 structure/sequence alignment pairs shown in Figure 5, half belong to the five largest families: serine proteases, pepsins, calmodulins, globins, and phospholipases A2. Thus, while we can conclude that different structural classes contain members with a wide range of structural mutation sensitivities, insufficient data are available to test whether non-homologus members of the same fold family have similar structural mutation sensitivities, or whether very distant branches of the same protein family can have detectably different structure/sequence relationships. To improve our understanding of intra-family structure/sequence relationships, we need more structures from sequences that are 30-50 % identical with sequences of known structure, in preference to structures from sequences that are more than 90 % identical or less than 20 % identical. Currently only two families in pdb.nr80, serine proteases and globins, have more than 20 structures, and only six families have more than 12 structures (Table 1). A five- to tenfold increase in the number of structures of intermediate sequence diversity (50 families with 12 or more diverse structures) should provide fundamental insights into the differences in protein folding interactions responsible for differences in structure/sequence relationships.

Methods

Sequence and structure comparisons

Sequence and structure similarity searches were performed on a database of 1770 (pdb.nr80) sequences for which structures have been determined. This database was produced from the 9039 sequences in the PDB (release 80) by a simple selection and database search process that identified related sequences in the fully redundant database ($E() < 10^{-4}$) that were 100 % identical (pdb.nr100) or more than 80 % identical (pdb.nr80). The Pdb.nr100 and pdb.nr80 databases are available from ftp.virginia.edu/pub/fasta. The non-redundant databases were constructed from the pdb_seqres.txt databases for PDB release 80,

obtained from ftp.pdb.bnl.gov/pub. Non-protein sequences were removed from pdb_seqres.txt to produce the redundant database pdb.r.

Non-redundant sequence databases were created as follows: a nascent database was created by placing the first sequence of pdb.r into a nascent database file. The remaining sequences of pdb.r were searched against the nascent database using ssearch3, one query sequence at a time. if the highest scoring sequence from the nascent database had a similarity score with an expectation value less than 10^{-4} and a percent identity below the identity threshold (100 % or 80 %), the query sequence was added to the nascent database. Expectation values were calculated using the Altschul-Gish scoring parameters (Altschul & Gish, 1996) based on a database size of 10,000 sequences. When every sequence from pdb.r had been compared to the nascent database, the nascent database was finished.

Protein family selection

To select a set of homologous proteins, we initially surveyed potential protein families in the PDB with a set of 46 families from the database PIR39b (Pearson, 1995). This dataset led to the identification of 18 protein families with known structures of high sequence diversity. An additional 26 protein families were found by comparing sequences from the SCOP superfamilies (Murzin *et al.*, 1995) to the pdb.r database using FASTA, for a total of 44 protein families that could be used in our study. Eight of these families were excluded because they had fewer than five members in pdb.nr80, leaving us with 36 protein families (Table 1). The PDB structure identifiers for the 342 structures in the 36 families from pdb.nr80 are listed in the Appendix.

DALI structural similarity searches (Holm & Sander, 1993) were conducted using the WWW interface at http://www2.ebi.ac.uk/dali/. Query structures for DALI searches were chosen from the pdb.nr80 database of structures with sequences that are $\geqslant 80$ % identical with any other query from that family. The average C^{α} root-mean-square deviation (RMSD), alignment length, and *z*-score reported by DALI were saved for later analysis. We tabulated the structural similarity scores reported by DALI in two different ways. In the simplest tabulation, we saved the DALI structural *z*-value, RMSD, and percent identity for the PDB structure identifier that matched the entries in pdb.nr80. In some cases, DALI might return higher structural similarity scores for a structure that is not in pdb.nr80, but which is encoded by a sequence that is 100 % identical with the pdb.nr80 structure. For a second tabulation of structure/sequence relationships, the best DALI similarity score was used from the set of DALI similarity scores from sequences that are 100 % identical to the pdb.nr80 entry found in the DALI search. We refer to these DALI similarity scores as DALI-selected scores. Results with the pdb.nr80 DALI score and the DALI-selected scores were indistinguishable.

For each protein pair reported in the DALI output, sequence similarity was calculated using the Smith-Waterman algorithm (Smith & Waterman, 1981) as implemented in the ssearch3 program of the FASTA3 package (Pearson, 1996). Smith-Waterman alignment scores were calculated using the BLOSUM50 scoring matrix (Henikoff & Henikoff, 1992) with a penalty of -12 for the first residue in a gap and -2 for each additional residue. The statistical significance of the sequence similarity scores was estimated using the statistical parameters (Pearson, 1998) for the query sequence from an ssearch3 database search of pdb.nr80. Expectation values of the sequence matches were calculated by sc_to_e, which converts a raw similarity score, sequence length, and ssearch3 statistical parameters to a *z*-score and statistical significance.

We did not use the manually assigned SCOP protein superfamilies (Murzin *et al.*, 1995) directly because more than half of the protein pairs that are assigned as homologous in SCOP lack both significant sequence and structural similarity (Levitt & Gerstein, 1998). With the exception of two families, the azurins/phycocyanins/plastocyanins (AZN) and the thioredoxins/glutathione transferases (THX), all the other proteins in our classification share either statistically significant sequence or structural similarity. The AZN and THX families contain proteins that are just at the borderline of significant structural similarity, but that share common functional groups or co-factors (Cu^{2+} for AZN and sulfur in THX). In most cases, our computationally based family assignments are very similar to those of SCOP, but two families (LZM and ADK) differ significantly from the SCOP classification. In both cases, our families are much smaller than the corresponding SCOP superfamily because of lack of statistical support for the homology inferences of SCOP (Wood, 1999). Because alignments are often inaccurate in the absence of significant similarity (Figure 2(b)), it is unlikely that the additional homology assignments in SCOP, which lacked significant structural similarity, would have been useful for this study.

Regression analysis

Linear and quadratic regression of DALI and ssearch3 structural and sequence similarities was done in S+ version 3.4 using the function 1m(). To minimize apparent structural variation that is unrelated to sequence differences, DALI results were filtered to eliminate sequence redundancy and low-resolution protein structures. When a DALI search returned results from several structures with the same (100 % identical) sequence, the highest DALI *z*-score was used. To filter redundant structures with more than 80 % sequence identity, those structures not found in pdb.nr80 were eliminated. Whenever a sufficient number of high-resol-

ution (solved by X-ray crystallography to better than 2.2 Å resolution), non-redundant (<80 % identical) structures were available, results from lower-resolution structures were excluded. Restricted cubic-spline polynomial fits to the structure/sequence-similarity relationships were done in S+ using the Hmisc library (Hmisc S-plus function library; programs available from http://lib.stat.cmu.edu/s/Harrell, Spline knots were located at the 0.05, 0.33, 0.67, and 0.95 quartiles of sequence similarity. The effect of additional terms in the regression of structure with sequence was quantified by using the adequacy ratio. $\text{Adequacy}_{\text{quadratic}} = \text{a}r^2_{\text{linear}}/\text{a}r^2_{\text{quadratic}}$, $\text{Adequacy}_{\text{spline}} = \text{a}r^2_{\text{linear}} \backslash i\ r^2_{\text{spline}}$, where a$r^2$ is the adjusted r^2 predicted by the linear, quadratic, or restricted cubic spline (four knots) relationship; ar^2 is the r^2 adjusted for the number of parameters used in the fit, and $\text{a}r^2 = 1 - (1 - r^2)(n - 1/(n - p)$, where n is the number of data points and p is the number of parameters in the regression fit ($p = 2$ for a linear fit).

To compare sequence/structure relationships among many different protein families, the DALI and ssearch3 z-scores were normalized so that protein families of different lengths could be compared. Because the similarity score is the sum of the individual similarities of each residue-pair in the alignment, longer protein families will have higher z-scores than shorter proteins at a given level of sequence identity. Thus, when comparing search results from families with proteins of different lengths, we normalize the z-score by dividing by the query sequence length (to produce an average z-score similarity per residue) and then multiply by 100. Since DALI similarities are also summed over the length of the alignment, a similar query length effect can be seen in DALI z-scores, and the same normalization was used. The result is a normalized z-score, which expresses the z-score that a match would get if it were calculated for a 100 residue query. Normalized z-scores do not show an obvious relationship between query length and z-score; homologous matches between small proteins receive z-scores just as high as homologous matches between large proteins.

Protein family divergence rates

Sequences used in mutation rate calculations were obtained from Swiss-Prot (Bairoch & Apweiler, 1996), and sequence alignments were constructed using CLUSTALW (Thompson *et al*., 1994). Mutation rates were calculated using the methods described by Langley & Fitch (1974) and Gu & Zhang (1997). Sequences were selected to maximize the number of fossil-derived last common ancestors. Trees were constructed using the kitsch program of the PHYLIP package (Felsenstein, 1989). Regression analysis using distances derived from the trees was done in S + . The rate estimate is made by dividing a mean distance by the LCA date for the appropriate taxonomic group(s).

Acknowledgments

This work was supported by a grant from the National Library of Medicine (LM04961). We thank Bob Kretsinger for his careful reading of the manuscript and for helpful suggestions and Frank Harrell for his statistical advice.

References

Alexandrov, N. N. (1996). SARFing the PDB. *Protein Eng.* **9**, 727-732.

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460-480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). A basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Bairoch, A. & Apweiler, R. (1996). The Swiss-Prot protein sequence data bank and its new supplement TrEMBL. *Nucl. Acids Res.* **24**, 21-25.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science,* **247**, 1306-1310.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399-405.

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

Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811-1826.

Gu, X. & Zhang, J. (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**, 1106-1113.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitutions matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.

Kabsch, W. & Holmes, K. C. (1995). The actin fold. *FASEB J.* **9**, 167-174.

Langley, C. H. & Fitch, W. M. (1974). An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**, 161-177.

Lattman, E. E. & Rose, G. D. (1993). Protein folding-what's the question? *Proc. Natl Acad. Sci. USA,* **90**, 439-441.

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA,* **95**, 5913-5920.

Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature,* **334**, 406-410.

Matthews, B. W. (1987). Genetic and structural analysis of the protein stability problem. *Biochemistry,* **26**, 6885-6888.

Mott, R. (1992). Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59-75.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1145-1160.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-258.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444-2448.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. & Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423-439.

Sandberg, W. S. & Terwilliger, T. C. (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science,* **245**, 54-57.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.

Schwartz, R. M. & Dayhoff, M. (1978). Matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure* (Dayhoff, M., ed.), vol. 5, suppl. 3, pp. 353-358, National Biomedical Research Foundation, Silver Spring, MD.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Swindells, M. B. (1996). Detecting structural similarities: a user's guide. *Methods Enzymol.* **266**, 643-653.

Thomas, P. J., Qu, B. H. & Pedersen, P. L. (1995). Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* **20**, 456-459.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Wood, T. C. (1999). *Theory and Application of Protein Homology,* PhD Thesis, University of Virginia, Charlottesville, VA.

Appendix

Table A1. PDB structures used

Name	Code	Structures
α-Amylase	AAA	1cdg 1cgt 2aaa 2exo 1xyza
		2taaa 1amy 1cyg 1ppi 1xas 1xys 1ciu 1bpla 1bpab 2 amg
Arabinose-binding protein	ABP	5abp 1pea
		3gbp (1dbp) 2liv 2lbp 1pnra
Alcohol DH	ADH	1cdoa
		1adg 1agna 1qora 1teha
Adenylate kinase	ADK	3adk 1akea 1uky 2ak3a 1ak2 1aky
		1ukd
Annexin	ANX	2ran 1aii
		1ala 1ain 1ann
Anti-thrombin	ATH	1hlea 1ovaa
		7apia 2acha 1anti
Azurin	AZN	2plt (3azua) 3pcy 7pcy 1aaj 1aiza 1paz 1pmy 1jer 1zia
		9pcy 1plb 1nin 1adwa
Calmodulin	CAL	1pal 1pvaa 2scpa 4cpv 4tnc 5pal 1omd 1rec 1rtp1 1tcob
		(1clb) 1ctr 2sas 1scmb 1scmc 1cnpa 1syma 2mysb 2mysc
Cysteine protease	CPR	1pe6 1aec (1ctea) 1ppo 1yal
Cytochrome *c*	CYC	(1csu) 1hrc 3c2c 1ccr 1cxc
		155c 1c2ra 1cry 1hroa
Dihydrofolate reductase	DFR	3dfr (3drca) 8dfr 1dyr
		(1dlr)
Fatty acid-binding protein	FAP	1crb 1hmr (1icn) 1mdc 1opba
		1alb 1cbq 1cbra 1ftpa 1pmpa 1eal
2Fe-2S ferredoxin	FRD	1frra 1frd 2pia 1dox 1roe 1doi
		1fxaa 1fxia 4fxc
Flavodoxin	FVN	3fx2 1flv 1ofv 2frc (1fla)
Ferredoxin	FXN	5fd1 1fdn 1fdx 1fxd 1blu
		1fxra 2fxb 1rof 1clf
Globin	GLB	(2spl) 3mba 3sdha 1ash 1ecn 1flp 1gdj 1hbg 1hsa 1hdsb 1lhs 1myt 1scta 1sctb 2lhb 1spga 1spgb
		2dhba 2dhbb 1fdhg 1hbha 1hbhb 1hdab 1hlm 1hlb 1itha 1bina (1outa) 1outb
Homebox	HOX	1enh 1fjla
		(1ahdp) 1aplc 1ftz 1hdp 1ocp 1pou 1ftt 1yrna 1vnd
Proline isomerase	ISM	2rmba 2rmca 1cyna 1lopa
		1clh
Lipoamide DH	LPD	1grg 1gera (1nhr)
		2tpra 3lada 1lvl 1ndaa (1tdf) 1ebda 1vdc

Lysozyme	LZM	3lhm 1alc 1lmn 1jug
		(2iffy) 2eql
Macrophage inflammatory protein	MIP	3il8 (1napa)
		1huma (1msga) 1plfa 1rhpa 1rtna
Pepsin	PEP	2psg 3app 3apre (3cms) 1htrb 1smra 1eaga
		2er0e 2ren 1lyba 1lbb 1smea 1jxra
Peroxidase	PER	1arp 1lgaa 1mnp 1qpaa
		(3ccp) 1apxa 1scha
Phosphocarrier protein	PHC	1ptf (1spha) 1pch
		1hdn 1zer
Phospholipase A2	PLI	3bb2 (3p2pa) 1poc 1poa 1ppa 1psj 1ae7
		1aypa (1clpa) 1pp2r 1buna
Pancreating trypsin inhibitor	PTI	2ptci 1aapa 1knt
		1dem 1dtk 1shp 1bunb
Ribosome inactivating protein	RIP	1abra 1ahc 1mrk
		1apa 1apga 1pafa
Selectin	SEL	1esl 1rdkl 1lit
		1msba 1hup
Superoxide dismutase	SOD	1mnga 3sdpa 1idsa 1isaa
		1abma
Serine protease	SPR	2ptn 2pkay 2sga 2tbs 2tbs 3gcta 3est 3sgbe 3rp2a (31pra) 1arb 1hpga 1hnee 1hyla 1sgt 1ton 1elt 1try 2sfa (1dst) 1fona (1hava)
		1bbre 1bbrh (2trm) 1hcga 1trna 1lmwb 1fuja 1pfxc 1rtfb 1pytd
Scorpion toxin	STX	1agt 1chl 1gps 1ica 1pnh 1sis 2crd 1mtx (1cmr) 1sco
Subtilisin	SUB	2pkc 2tece 2sece
		2sbt 1mpt
Thioredoxin	THX	(2tir) 3gsta 1glpa 1hna 1thx
		3trx (1agsa) 1gne 1gsq 1tof 1mek
Snake toxin	TOX	3ebx 1fas 1ntn
		1abta 1cdta 1cod 1cre 1cre 1cvo 1ctx 1cxn 1drs 1kbaa 1nea 1ntx 1tfs 1lsi
Triose phosphate isomerase	TPI	2ypia 3tima 1htia (1tmha) 1btma
Viral coat protein	VCP	2plv2 2r062 2rhn2 1bbt2 1mec2 1tmf2 1cov2

PDB structures in pdb.nr80. Each family is divided into two groups of structures. High-resolution (<2.2 Å) structures are shown above the line; lower-resolution structures are shown below the line. Structures from sequences designated as mutant in the PDB are enclosed in parentheses.

Edited by J. M. Thornton

(Received 22 January 1999; received in revised form 4 June 1999; accepted 28 June 1999)