PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Structure is three to ten times more conserved than sequence—A study of structural response in protein cores

Kristoffer Illergård,[1] David H. Ardell,[2,3] and Arne Elofsson[1]*

[1] Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
[2] Department of Natural Sciences, School of Natural Sciences, University of California, Merced, California 95344
[3] Linnaeus Centre for Bioinformatics, Uppsala University, SE-751 24 Uppsala, Sweden

ABSTRACT

Protein structures change during evolution in response to mutations. Here, we analyze the mapping between sequence and structure in a set of structurally aligned protein domains. To avoid artifacts, we restricted our attention only to the core components of these structures. We found that on average, using different measures of structural change, protein cores evolve linearly with evolutionary distance (amino acid substitutions per site). This is true irrespective of which measure of structural change we used, whether RMSD or discrete structural descriptors for secondary structure, accessibility, or contacts. This linear response allows us to quantify the claim that structure is more conserved than sequence. Using structural alphabets of similar cardinality to the sequence alphabet, structural cores evolve three to ten times slower than sequences. Although we observed an average linear response, we found a wide variance. Different domain families varied fivefold in structural response to evolution. An attempt to categorically analyze this variance among subgroups by structural and functional category revealed only one statistically significant trend. This trend can be explained by the fact that beta-sheets change faster than alpha-helices, most likely due to that they are shorter and that change occurs at the ends of the secondary structure elements.

Proteins 2009; 77:499–508.

Key words: protein structure; evolution; secondary structure; accessibility; residue contacts; RMSD.

INTRODUCTION

Evolutionary changes of individual protein domain primary structures that become fixed in populations are mainly replacements of single amino acid residues and short insertions or deletions. Since most three-dimensional structures of proteins are determined by their sequences[1] and solvent interactions, higher-order structure will also change in response to these changes. The extent of higher-order structural perturbation in response to sequence evolution will depend on the type and location of sequence changes. Some single mutations will completely disrupt structure, while others that conserve the physicochemical properties of the sequence will barely affect structure at all.[2]

Most amino acid substitutions are structurally conservative of the position of functional residues as well as the stability of the protein.[3] By comparing aligned proteins of known structures, ancient substitutions can be mapped to the structures. Sites in different local structural environment states exhibit different amino acid residue substitution patterns,[4–7] rates of substitution,[8,9] and polymorphic sequence variation.[10] However, how local structure evolves in response to evolutionary sequence changes and how the mapping between sequence and structure evolves requires further study.

One claim that is often used in discussions about proteins is that "structure is more conserved than sequence." This claim is supported by three observations: First, when comparing the root mean square deviation (RMSD) and sequence identity an exponential relation is found, see Figure 1(A).[11] Second, during evolution homologous protein sequences accumulate substitutions, and can diverge into the so-called twilight zone[12] of statistical sequence similarity, even while their structures retain detectable similarity.[13] Finally, due to physical constraints, there exist fewer distinct protein structures than there are distinct sequences, that is,

Additional Supporting Information may be found in the online version of this article.
This work was supported by grants from the Swedish Natural Sciences Research Council, SSF (the Foundation for Strategic Research) and the EU 6th Framework Program is gratefully acknowledged for support to the GeneFun project, contract No: LSHG-CT-2004-503567 and from the EMBRACE project, contract No: LSHG-CT-2004-512092.
*Correspondence to: Arne Elofsson, Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: arne@bioinfo.se.
Received 14 July 2008; Revised 10 March 2009; Accepted 19 March 2009
Published online 20 April 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22458

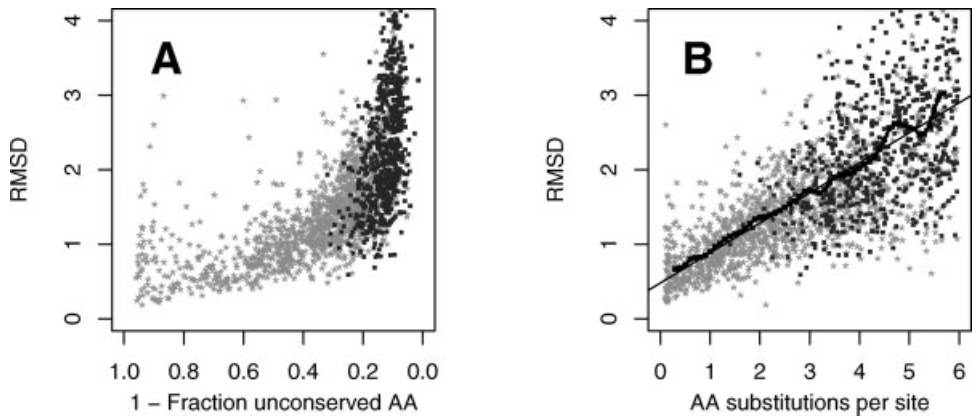

Figure 1

Structure versus sequence similarity plotted for homologous protein pairs. Domain pair from same SCOP family are plotted in black, others in grey. The line of best fit is plotted as a thin black line and a running average over 0.1 ED is plotted with a thick black line. RMSD is plotted against (**A**) sequence identity and (**B**) evolutionary distance (ED).

the space of possible folds is smaller than the space of possible sequences.[13]

However, several arguments may also be put forward to qualify this claim that "structure is more conserved than sequence." First, percent or proportion of identical residues, conventionally used to measure sequence divergence, is not an optimal measure of sequence conservation because it neglects counting hidden changes in the form of multiple, parallel, and back substitutions. In fact, it has been found that the pairwise structure-sequence relation within protein families is linear, when using measures based on global statistical significance[14] or when using length normalized BLAST bit-score and a score of structural divergence based on the Hausdorff metric.[15,16] Second, a change in sequence is discrete and represents a well-defined event, while a corresponding change in a native state of a flexible structure is continuous and depends on the particularities of the structural measure used. Therefore, comparing structural divergence and sequence divergence with their different dimensionalities, characteristics, and systematic sources of noise and measurement bias can be like comparing apples and oranges. Both sequence and structural alignments have uncertainties and these uncertainties must be managed to refine our understanding of the evolution of protein structure. Finally, one should really exhaustively or at least adequately sample all sequences of known structures in diverse proteomes to decide whether there really are more distinct sequences than distinct structures, that is, whether sequence space really is larger than structure space. Unfortunately, such an analysis is intractable for several reasons, for example, many protein families partly or completely consist of regions that do not fold into well-defined structures[17] and would have to be ignored, due to their under-representation (or absence) in structural databases. Another challenge comes from that although many substitutions are conservative, a few may completely ruin the stability of the protein structure or alter it dramatically.[18]

Here, we used a dataset of homologous pairs of protein sequences over a vast range of different sequence divergence and calculated how parts of those sequences map into the cores of their solved pairwise homologous protein domain structures. We then calculated several different summary measures of structural divergence and compared these against an unbiased estimator of their sequence divergence. As expected, the exponential relation between RMSD and sequence identity[11] transforms to a linear relation when using evolutionary divergence instead of percent identity. When we compared discrete measures of structural similarity that are comparable to our evolutionary distance estimator, we found that structure is, on average within the superposable core of domains, to be about three to ten times more conserved than sequence. Although we observed an average linear response of structural change with sequence change, we found a wide variance in relative rate. Different domain families varied fivefold in structural response to sequence evolution. A small part of the variation may be explained by the fact that secondary structure changes faster in all-β-sheet domains than in all-α-helix domains, but most of this variability remains unexplained.

MATERIAL AND METHODS

Dataset

SCOP 1.74 was used to select evolutionary-related protein domains. First, we excluded proteins diffracting to a

resolution less than 2.6 Å, or to which no local structure could be assigned. Thereafter, we created family and superfamily level subsets by randomly extracting one pair from each family/superfamily in SCOP. The family subset contains 1243 pairs that each belong to the same SCOP family and the second superfamily subset contains 656 pairs from the same superfamily but different families. We also created a third more redundant subset that contains 2567 pairs from 38 well-populated SCOP families. Here, to avoid bias, each protein is included at most once.

When comparing homologous protein domains, we needed alignments to infer corresponding homologous relationships among their residues. Here, we used pairwise structural alignments, based on STRUCTAL.[19] The alignments had to contain at least 50 aligned sites and yield an evolutionary distance between zero and six amino acid substitutions per site (see below).

We assumed that our alignments were accurate and divided them into two classes. The superposable regions, called the "core," were found within alignments by searching for aligned regions containing at least six consecutively aligned residues. The rest, referred to as "structurally divergent regions," were found to mostly contain divergent loops and regions where multiple insertions and deletions have occurred. In this study, the measures of sequence and structure similarity (see below) are only evaluated for the core regions, and thus the evolution of structurally divergent regions is completely ignored.

Classification of proteins

All pairs of protein domains were classified by their SCOP domains into structural classes ("All-α", "All-β", "α-β (a/b)," and "α-β(a+b)"),[20] fold descriptions ("layer," "sandwich," "barrel," and "bundle") and functional groups ("Metabolism," "Regulation," "Processes IC," "General," "Information") using SUPERFAMILY[21] annotations.

All proteins were further classified by SCOP domain architecture ("Single-Single," "Single-Multi," "Multi-Multi"). Additional classifications of the domains were based on lengths ("Longer than average," "shorter than average") and disulfide bonds ("No SSBonds-No SSbonds," "Has SSbonds-Has SSbonds," "Equal number of SSbonds," "Non-equal number of SSbonds").

Structural annotations

We used the core residues to superpose the two protein domains and calculate the RMSD between them. We also used STRIDE[22] to derive discrete structural annotations of secondary structure (SS) for each residue. The seven states in STRIDE were merged into three states: α-helices (G, H, and I), strand (E and S), and loops (T, S, and B). We obtained structural annotations of relative surface area (RSA) for each residue using Naccess 2.1.1,[23] which normalizes the accessible surface area of a residue by an extended Ala-X-Ala tri-peptide conformation. The probe size was set to 1.4 Å (the same radius as water). By assigning residues with RSA $\leq$ 25% as buried (B) and the rest as exposed (E), we obtained a binary two-state classification scheme of RSA.

We compared the SS and RSA structural annotations of corresponding positions in the two different structures, as given from the core regions of the structural alignments position-by-position. We provide the alignments used in this study with associated annotations in Supporting Information Table S1.

For comparison to the structural STRIDE-based alphabets, we divided the 20 amino acids into three alphabets of reduced cardinality (AA2, AA3, and AA6) by merging into different classes. In the alphabet AA2, the residues LIVFMWCPA are classified as hydrophobic and all others as polar, while in AA3 a third class consisting of P and G is created. In AA6, the following classification was used: [KR],[DE],[YWFH],[TSQN],[CVMLIA], and [PG] as has been proposed previously.[24,25]

We constructed binary two-dimensional contact maps for each protein. We inferred a contact if two β-carbon-atoms of side-chains (α-carbon for glycine) separated by more than two residues in the sequence were closer than 10 Å.

Evolutionary distance

Ideally, to correctly study the evolution of local structural environments, we would prefer to have used evolutionary sequence alignments, in which paired residues are homologous, that is, descend from the same residue in the ancestral protein. However, we used structural alignments here because these are superior for making alignments of distantly related proteins.[26] To reduce the possibility of inaccurate residue homology inference from the structural alignment, only core regions of structures from the same SCOP superfamily were considered, as described earlier.

The average number of substitutions per-site (evolutionary distance or ED) was calculated using maximum likelihood estimates with the JTT+Γ substitution model[27] allowing site-rate heterogeneity, computed using Tree-Puzzle v.5.2.[28] To work around an input requirement of 3 minimum sequences on input, we created input files to Tree-Puzzle in which each sequence appears in duplicate. Because we only calculate pairwise distances, this should have no effect on our results. We verified this by comparing the output of Tree-Puzzle for selected inputs in which we varied the number of redundant input sequences. As expected, we found the Tree-Puzzle pairwise distance calculations to be invariant to the presence and number of redundant input sequences.

Discrete, residue-based structural measures

We calculated fractional identities for different discrete descriptions of sequence and structure using exclusively core sites. In particular, we calculated the fractional identity of secondary structural states as:

$$\mathrm{ID_{SS}} = \frac{(\mathrm{HH} + \mathrm{SS} + \mathrm{LL})}{(\mathrm{HH} + \mathrm{SS} + \mathrm{LL} + \mathrm{HS} + \mathrm{SH} + \mathrm{HL} + \mathrm{LH} + \mathrm{SL} + \mathrm{LS})}$$

where among core sites, HH, SS, and LL are the number of coaligned helix (H), strand (S), and loop (L) states, respectively, HS and SH are sites with helix states aligned with strand states, HL and LH are sites with helix states aligned with loop states, and SL and LS are sites with strand states aligned with loop states.

We also studied changes in discrete representation of structure against evolutionary distance. For instance, the fraction of different aligned secondary structural states $\mathrm{DIFF_{SS}}$ can be calculated as $\mathrm{DIFF_{SS}} = 1 - \mathrm{ID_{SS}}$.

It is important to note that secondary structural identities are calculated on the basis of the same alignments and data as used for calculating evolutionary distances.

Contact similarity measure

We used contact maps to count the fraction of conserved and nonconserved contacts among aligned core residues. To be precise, suppose that there are $M \geq 0$ contacts between core residues in a protein X and $N \geq 0$ contacts in its homolog Y. Each contact involves two distinct residues, so it may be said that $2M$ residues are involved in contacts in X and $2N$ in Y. A core contact of X is conserved in Y if both of the core residues participating in that contact have homologous residues in Y engaged in the same contact. Suppose then that there are $C \leq \min(M,N)$ conserved contacts between X and Y, then the fraction of conserved contacts $\mathrm{ID_C}$ is defined as follows:

$$\mathrm{ID_C} = \frac{C}{(M + N - C)}$$

Linear regressions and detection of variation in structural response

The measures of sequence and structural similarity for all protein pairs were analyzed using R 2.2.1.[29] Linear regressions were obtained using the function lm ($Y \sim X$), with sequence similarity (DIFF or ED) as the X variable and different types of structural similarity measures as the Y variable. To measure the correlation between the variables R^2_{adj} was chosen over R^2 (Pearson correlation coefficient) as it should be more correct when subsamples are used. However, it should be noticed that the R^2_{adj} is virtually identical to R^2 for all data. Additionally, in the tables with regression data the 97.5 % confidence interval of the slope coefficients are shown.

We used the fact that an approximately equal number of values of protein pairs were found above and below the regression lines to detect differences in structural responses among subgroups of proteins. Here, "structural response" refers to the perturbation of structure in response to sequence evolution. We used the hypergeometric two-sided Fisher's exact test to test whether the same proportions of a subgroup and its complement were found above and below the line.

RESULTS

The aim of this study was to investigate how natural protein structures respond to evolutionary changes (substitutions) in the polypeptide sequence. For each of the 1899 structurally aligned protein domain pairs from the first two subsets of data, we compared several different measures of structural similarity against sequence similarity. Each pair of protein domains we compared is evolutionarily related and belongs to the same structural SCOP superfamily.[20] We examined pairs under a wide range of evolutionary distance, from zero up to six substitutions per site.

RMSD change linearly with evolutionary distance

The well-known exponential relation between RMSD and sequence identity, ID,[11] can also be seen in our data [Fig. 1(A)]. However, the exponential curve changes to a nearly linear function when adjusting for multiple substitutions per site, that is, when RMSD is plotted against ED [Fig. 1(B)]. The relationship between RMSD and ED can be approximated by the following equation, see Table I:

$$\mathrm{RMSD} = 0.5 + 0.4\,\mathrm{ED}$$

where ED is the evolutionary distance separating the two proteins. This means that two domains that have, on average, undergone one amino acid substitution per site each, after divergence are on average expected to differ by 1.3 Å RMSD. The equation also shows that two structures with identical polypeptide sequences are expected to vary by 0.5 Å RMSD, which is in agreement with earlier studies of structural effects of point-mutants[11] and different crystal forms.[30] It can also be noted that, although the average values change linearly with evolutionary distance, the variation among individual pairs is large, as indicated by the fact that the R^2_{adj} is only 0.48 (Table I). This means that pairs of protein at a distance of ED = 1 can differ structurally by anything between 0.5 and 2 Å RMSD, see Figure 1(B).

Table I
Estimates from linear Regression of Structural Similarity Measures Versus Evolutionary Distance (ED)

	Intercept	Slope	R^2_{adj}
$DIFF_{SS}$	0.032	0.035 ± 0.001	0.37
$DIFF_{RSA}$	0.060	0.041 ± 0.002	0.52
$DIFF_C$	0.117	0.065 ± 0.002	0.59
RMSD (Å)	0.48	0.40 ± 0.02	0.48

Linear divergence of discrete, residue-based structural measures

In addition to RMSD, three discrete residue-associated measures of structural divergence increase approximately linearly with evolutionary distance (Fig. 2). $DIFF_{SS}$ shows the most linear trend as well as an intercept closest to zero, while surface accessibility, $DIFF_{RSA}$, and a contact map measure, $DIFF_C$, show some signs of saturation and of intercepts further away from the origin (Table I).

The slopes of the regression lines provide various measures of how robust structures are to sequence change. These slopes, when drawn using discrete annotations of structural and sequence divergence, may be used as measures of the relative conservation of structure. From the regression with $DIFF_{SS}$, $DIFF_{RSA}$, and $DIFF_C$ to ED it can be seen that for each substitution, 0.035 discrete secondary structural states, 0.041 surface accessibility states, and 0.065 contacts change on average in response (Table I). This would infer the structure is 15 to 30 times more conserved than sequence.

These values refer to the ratios between changes in structural sequences with two and three discrete states, respectively, versus changes in the protein sequence with 20 different states. Therefore, these numbers are not correct as the number of possible state changes is dependent on the number of states describing the structure/sequence. However, if the amino acids are grouped by physicochemical similarity to a similar number of bins as the structural alphabets a fair comparison can be made. When this was done it was found that sequences still change more rapidly, between three and ten times slower than structure, that is, the "structural core is three to ten times more conserved than sequence" (Table II).

Large (unexplained) variation in structural response

Substitutions in amino acid residue sequences may be seen as discrete changes in a one-dimensional vector that represent well-defined events, while the corresponding changes in the structure are continuous and occur in three dimensions. Here, we have used four different structural measures of structural change. The first, RMSD, is a commonly used continuously varying geometric measure, while our three alternative measures, $DIFF_{SS}$, $DIFF_{RSA}$, and $DIFF_C$, are discrete residue-state-based measures respec-

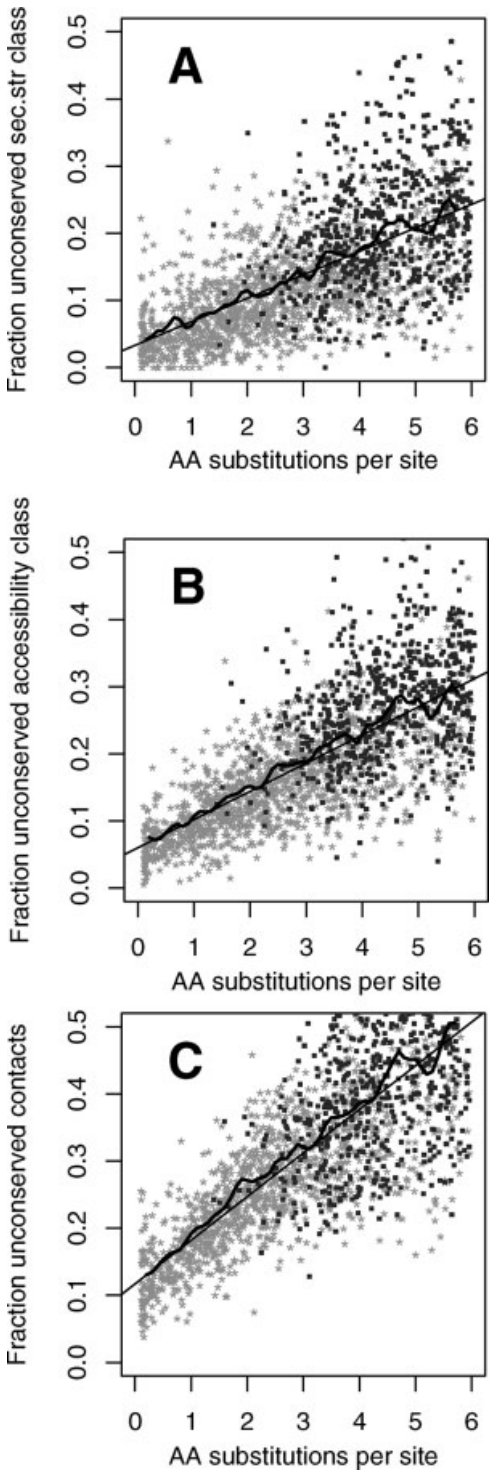

Figure 2
Structure versus sequence similarity plotted for homologous protein pairs. The plots are similar to Figure 1(B), but here the structural measure is in (**A**) fraction of unconserved secondary structure class, (**B**) fraction of unconserved accessibility class, and (**C**) fraction of unconserved residue contacts.

Table II
Estimates from Linear Regression of Structural Similarity Measures Versus Evolutionary Distance (ED) Using Reduced Alphabets

	Intercept	Slope
AA6 vs. $DIFF_{SS6}$	0.032	0.10 ± 0.01
AA vs. $DIFF_{SS}$	0.032	0.04 ± 0.002
AA6 vs. $DIFF_{SS}$	0.015	0.10 ± 0.01
AA3 vs. $DIFF_{SS}$	0.026	0.21 ± 0.01
AA2 vs. $DIFF_{SS}$	0.040	0.25 ± 0.02
AA vs. $DIFF_{RSA}$	0.048	0.03 ± 0.003
AA6 vs. $DIFF_{RSA}$	0.060	0.12 ± 0.01
AA3 vs. $DIFF_{RSA}$	0.055	0.23 ± 0.02
AA2 vs. $DIFF_{RSA}$	0.052	0.37 ± 0.03

tively characterizing conservation in secondary structure, relative surface accessibility, and residue contacts.

Two structural measures, RMSD and $DIFF_C$, correlate reasonably well (Table III). There exist several reasons why the different structural measures might not always agree, for example, a large insertion or deletion at the end of a protein domain may cause a large change in relative surface accessibility while having all secondary structure elements conserved. Furthermore, a conformational change between two domains may cause a large RMSD deviation while leaving most aligned secondary states conserved.

Two identical protein sequences may show substantial structural differences,[18] because of experimental errors, differing conditions of structure determination such as ligand- or ion-binding, loss of information from the transformation of continuously varying three-dimensional coordinates to discrete structural similarity measures, and the unavoidable fact that crystallized proteins represent only a snapshot of what are truly flexible structures in three dimensions. These sources of noise and bias might indeed be large relative to the "true" evolutionary changes we are trying to measure in this work.

One measure of noise in our data is an estimate of structural divergence for two identical sequences as given by the intercepts of our regression lines, which we have found to be about from 4 to 12% for the discrete structural measures, see Tables I and II.

Which factors determine the structural response?

As we note earlier, there is a high variability in structural responses for similar amounts of sequence change within domain cores, as indicated by the high R^2_{adj} values from the regressions (Tables I and II). A part of the variation is "noise" coming from, for example, non-sequence-based changes like experimental conditions as well as from sequence changes within the structurally variable regions. However, some part seems likely to be explained by differences in the "true evolutionary" changes of protein pairs, that is, some mutations are more disruptive than others.

Table III
Correlation Coefficients Between Four Different Structural Similiarity Measures

	$DIFF_{SS}$	$DIFF_{RSA}$	$DIFF_C$
RMSD	0.54	0.54	0.88
$DIFF_{SS}$		0.48	0.59
$DIFF_{RSA}$			0.63

$DIFF_{SS}$, fraction of unconserved secondary structure states; $DIFF_{RSA}$, relative surface accessibility state; $DIFF_C$, contacts; RMSD, root mean square deviation.

Table IV
Protein Subgroups Detected as Having Significantly High or Low Relative Divergence Measured by Different Structural Measures

	Percent above line (%)	Structural measure	P-*value*	Number of pairs
All-β	54	$DIFF_{SS}$	1e-6*	387
All-β	53	$DIFF_{RSA}$	1e-3**	387
Sandwich	55	$DIFF_{SS}$	1e-3**	246
Sandwich	55	$DIFF_{RSA}$	1e-3**	246
All-α	31	$DIFF_{SS}$	8e-4*	308
Bundle	30	$DIFF_{SS}$	3e-3**	125

The percentage of protein pairs of different subgroups that are found above the line of best fit, using the X-variable as evolutionary distance and the different Y-variables fraction of unconserved secondary structure states ($DIFF_{SS}$), relative surface accessibility states ($DIFF_{RSA}$), contacts ($DIFF_C$), and root mean square deviation (RMSD). Statistical significance from a two-sided Fischer's exact test is shown by *P*-values and asterisks, so that $^*P < 0.001$, $^{**}P < 0.01$.
The total number of pairs found in the specific subgroups are also shown.

To analyze this variability, all proteins were divided into subgroups based on their SCOP class, fold description, molecular function, domain architecture, lengths, and occurrence of disulfide bonds. Thereafter, it was examined whether any structural subgroups were over-or under-represented among pairs above the line of best fit using a hypergeometrical two-sided Fisher's exact test.

The only consistent trend we found was that for secondary structure similarity ($DIFF_{SS}$), the β-sheet containing subgroups are significantly over-represented in the high structural response group. Conversely, the α-helix subgroups are over-represented in the low structural response group (Table IV). The trend of higher degree of structural response in β-sheet containing proteins compared to α-helical proteins can also be seen, although weaker, when studying accessible surface area or contacts. From the estimated line of best fit for all-β and all-α it can be seen that with each amino acid substitution per site 0.039 of the secondary structure states are changed in the β-sheet proteins while only 0.024 in the α-helical proteins, Table V.

One likely explanation why β-sheets change faster than α-helices is that β sheets are shorter than α helices (median lengths 5 vs. 10 residues) and by the fact that practically all changes in secondary structural elements occur at their ends. When ungapped secondary structure units longer than eight sites in the pairwise alignments were removed from the all-β and all-α groups, the difference disappeared (not shown). This further supported the conclusion that the general shorter length of sheets com-

Table V
Estimates from Linear Regression of Fraction of Unconserved Secondary Structure States ($DIFF_{SS}$) Versus Evolutionary Distance (ED) for Four Different Protein Subgroups and Their Secondary Structure Content

	Intercept	Slope	Helix (%)	Sheet (%)
All-α	0.028	0.024 ± 0.004	69	4
All-β	0.037	0.039 ± 0.005	9	48
Bundle	0.047	0.015 ± 0.008	61	6
Sandwich	0.042	0.031 ± 0.006	11	47

pared to helices is responsible for most of the difference in their rates.

DISCUSSION

Our findings above lead us to conclude that protein structure changes approximately linearly with evolutionary distance within domain cores, when structure similarity is measured by RMSD or discrete measures of similarity. Furthermore, we have found structure to be from three to ten times more conserved than sequence when using discrete alphabets of similar size. Although the claim that structure is more conserved than sequence is quite old, we believe this is the first attempt to make a detailed quantification of this phenomenon. Our analysis of differences among subgroups by structural and functional categories only revealed one trend, which is that secondary structure changes faster in all-β-sheet domains than in all-α-helix domains.

Evolution of local structural environments is linear

Here, we found that local structural environment, as measured by, for example, discrete states of secondary structure and relative surface accessibility, change linearly, on average, in response to amino acid substitutions. In addition, the higher divergence in $DIFF_{SS}$ of proteins with high content of sheets and low divergence of proteins with high content of helices indicate that β-strands appear to evolve faster than α-helices. But the most likely explanation for this fact appears to derive from two additional trends we observed in our data: namely, first, that changes in strands and helices occur by lengthening and shortening at their ends in discrete multiple residue jumps, and second, that strands are, on average, shorter than helices. Therefore, a larger fraction of residues change state at strand-loop interfaces than at helix-loop interfaces. Length changes of secondary structural elements in this sense could occur either as insertions and deletions or as changes from coil to regular secondary structure or vice versa.

If helices and strands evolve by lengthening and shortening at their ends, this suggests that site-wise modeling of secondary structural evolution, such as with the quantity $DIFF_{SS}$, is incorrect, and instead, a duration-based model of structural change would be more appropriate. For example, a length-dependent random-walk model such as the single-step mutation (SSM) model, used to model the evolution of microsatellites, may be promising in future investigations of the evolution of protein structure. However, our initial attempts to fit such a model were inconclusive from lack of data.

Nevertheless, it is remarkable that a site-independent distance measure like $DIFF_{SS}$ does increase linearly with evolutionary distance anyway, as if these secondary structural elements do in fact evolve in a site-independent fashion on average.

Structure is indeed more conserved than sequence

Two previous studies have also found linear response of structural divergence with sequence divergence for

Table VI
Estimates from Linear Regression Values for Evolutionary Distance Versus Four Different Structural Similarity Measures for the Most Well-Populated Families

	RMSD		$DIFF_{SS}$		$DIFF_{RSA}$		$DIFF_C$		
Fam	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	No Proteins
a.1.1.2	0.734	0.312	0.056	0.010	0.050	0.033	0.158	0.062	198
a.2.11.1	0.459	0.210	0.004	0.026	0.093	0.016	0.052	0.047	27
a.25.1.1	0.365	0.233	0.027	0.004	0.144	0.003	0.102	0.036	26
a.133.1.2	0.964	0.045	0.048	0.015	0.084	0.024	0.198	0.009	62
b.1.1.1	0.647	0.311	0.066	0.038	0.068	0.030	0.185	0.045	202
b.1.1.2	0.915	0.267	0.104	0.034	0.103	0.021	0.182	0.050	234
b.1.18.2	0.227	0.486	0.099	0.031	0.009	0.069	0.078	0.098	34
b.60.1.1	0.576	0.425	0.047	0.026	0.072	0.046	0.147	0.066	51
c.1.8.1	0.285	0.497	0.044	0.038	0.055	0.043	0.083	0.080	80
c.1.8.3	0.338	0.498	0.043	0.039	0.078	0.031	0.133	0.071	68
c.37.1.8	0.868	0.217	0.065	0.023	0.092	0.030	0.193	0.039	107
c.94.1.1	0.657	0.491	0.047	0.040	0.065	0.045	0.092	0.085	74
d.19.1.1	0.586	0.322	0.030	0.025	0.062	0.048	0.119	0.057	93
d.93.1.1	0.986	0.103	0.065	0.022	0.144	0.009	0.192	0.029	36
d.117.1.1	0.636	0.241	0.070	0.009	0.081	0.024	0.155	0.035	38
d.144.1.7	0.933	0.354	0.066	0.020	0.138	0.024	0.190	0.045	140

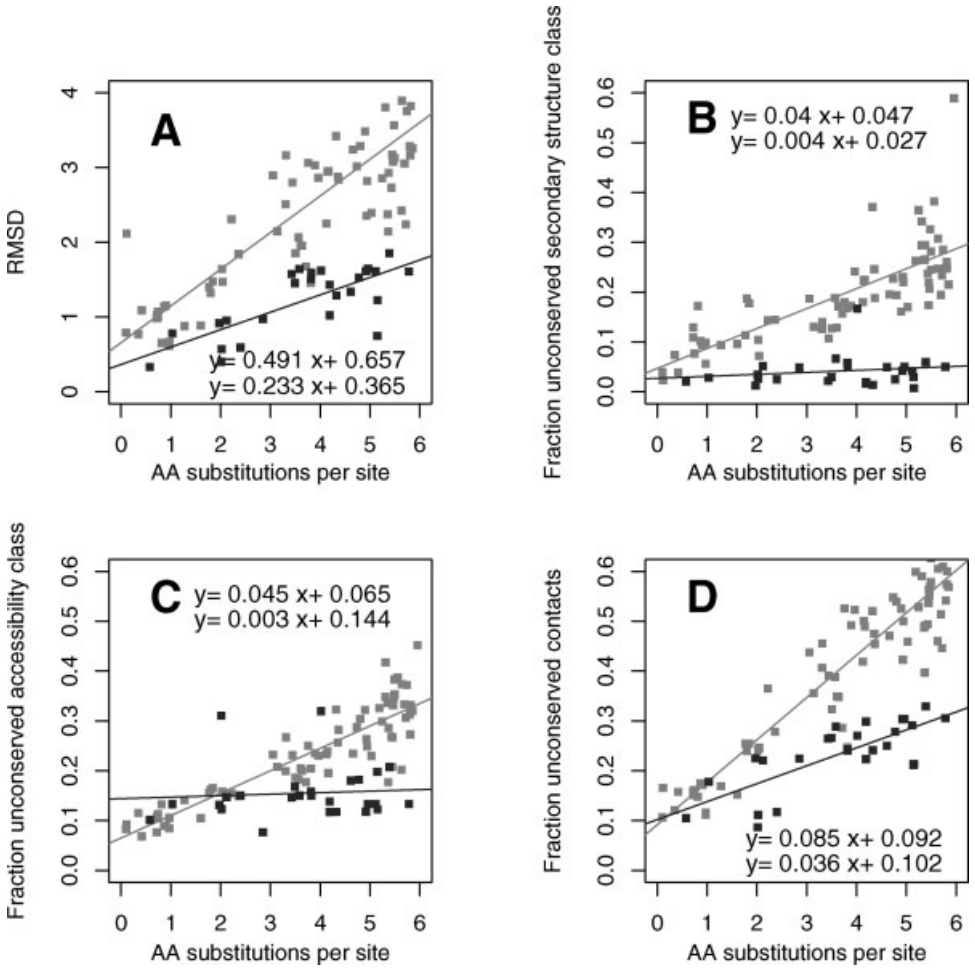

Figure 3

Structure and sequence similarity plotted for homologous protein pairs using four different sequence measures. Domain pairs from the phosphate binding protein-like family with SCOP code c.94.1.1 is in grey and pairs within ferritin family with SCOP code a.25.1.1 is in black. The sequence similarity measures is evolutionary distance (ED). The structural similarity measures are (**A**) RMSD, (**B**) fraction of unconserved secondary structure class, (**C**) fraction of unconserved accessibility class, and (**D**) fraction of unconserved residue contacts. The lines of best fit is plotted as thin lines and their coefficients are also seen within the four subfigures.

proteins within the same structural families.[14,16] In addition, these studies reveal that the linearity is valid at a residue scale in such a way that local structural environment as measured, for example, by discrete states of secondary structure and relative surface accessibility on average change linearly in response to amino acid replacements.

Although the protein structure-sequence relation have been studied many times,[31] the claim that structure is more conserved than sequence has not been thoroughly quantified and analyzed. Here, we have quantified this linearity in a dataset of structurally aligned protein domains, and found that structurally conserved domain cores are from three to ten times more conserved than sequence when using discrete alphabets of similar size. Although the average evolution of local structural environments is linear, at longer distances differences between pairs that are classified as belonging to same or different SCOP families (Figs. 1 and 2) can be seen. Additional analysis with other similarity measures would be needed to tell if inclusion of structural divergent regions, structural disordered regions, proteins with historically large structural changes and rearrangements, or use of other similarity measures will change the estimated structural response values radically.

Most of the variability in structural response cannot be explained by simple structural and functional descriptors

Although there is an approximately linear relationship between sequence and structure similarity on average, the response to different sequence changes varies heavily, at

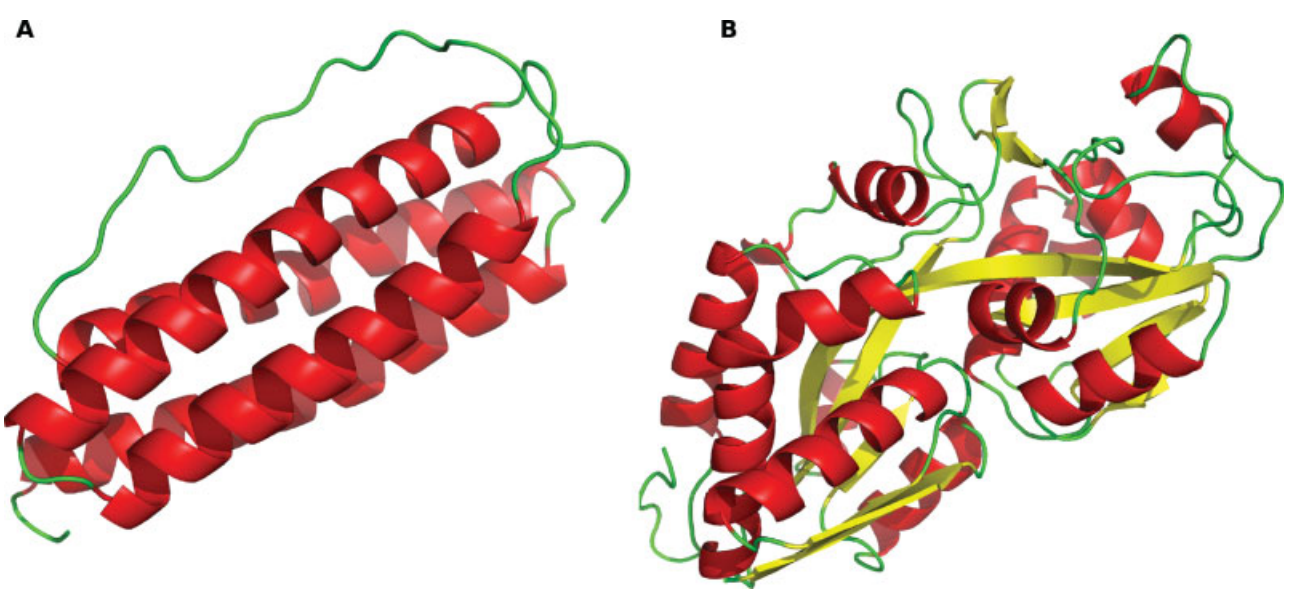

Figure 4

Structures of (A) Domain d1b71a1 belonging to the ferritin family (SCOP code a.25.1.1) have a high structural response and (B) Domain d1a40a belonging to phosphate binding protein-like family (SCOP code c.94.1.1) have a small structural response.

least in part probably because some mutations are more perturbative of structure than others. Among the 38 most well-populated families, relative structural divergence differs by a factor of five (Table VI), which is in line with previous estimates.[14,16] Unusually high and low structural response can be seen for the two families shown in Figure 3. Phosphate binding protein-like family with SCOP code c.94.1.1 (Fig. 4), containing both beta strands and helices in the core, have higher relative structural divergence then the Ferritin family with a α-helical core in a bundle with SCOP name a.25.1.1 (Figure 4). The structural responses vary by factors of 2.1, 11.1, 14.6, 2.4 for RMSD, $DIFF_{SS}$, $DIFF_{RSA}$, and $DIFF_{C}$, respectively.

The large variation in relative divergence within the structural core is puzzling. Our attempts to explain the variation by dividing domain pairs in different structural and functional groups, did not reveal any consistent trends, other than that β-sheets seemed to "evolve" faster than α-helices. However, a lot of parameters to investigate still remain both from structural, functional, and evolutionary perspectives. We believe from our results that it is clear that some mutations are more disruptive to the structure than others.

ACKNOWLEDGMENTS

The authors thank Gerard Kleywegt for discussion and encouragement at early stages of this project. Further, they thank Björn Wallner for help with initial scripts for working with pdb-files, Alexandra Jauhiainen for a valuable discussion and for suggestions regarding statistical analysis, and Åsa Björklund for proofreading.

REFERENCES

1. Anfinsen C. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Matthews B. Studies on protein stability with T4 lysozyme. Adv Protein Chem 1995;46:249–278.
3. Pal C, Papp B, Lercher M. An integrated view of protein evolution. Nat Rev Genet 2006;7:337–348.
4. Koshi J, Goldstein R. Context-dependent optimal substitution matrices. Protein Eng 1995;8:641–645.
5. Koshi J, Goldstein R. Correlating structure-dependent mutation matrices with physical-chemical properties. Pac Symp Biocomput 1996;488–499.
6. Koshi J, Goldstein R. Mutation matrices and physical-chemical properties: correlations and implications. Proteins 1997;27:336–344.
7. Koshi J, Goldstein R. Analyzing site heterogeneity during protein evolution. Pac Symp Biocomput 2001;191–202.
8. Dean A, Neuhauser C, Grenier E, Golding G. The pattern of amino acid replacements in alpha/beta-barrels. Mol Biol Evol 2002;19: 1846–1864.
9. Marsh L, Griffiths C. Protein structural influences in rhodopsin evolution. Mol Biol Evol 2005;22:894–904.
10. Bustamante C, Townsend J, Hartl D. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol 2000;17:301–308.
11. Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
12. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
13. Holm L, Sander C. Mapping the protein universe. Science 1996; 273:595–603.

14. Wood TC, Pearson WR. Evolution of protein sequences and structures. J Mol Biol 1999;291:977–995.
15. Panchenko AR, Madej T. Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol Biol 2005;5:10.
16. Panchenko AR, Wolf YI, Panchenko LA, Madej T. Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 2005;61:535–544.
17. Kamtekar S, Schiffer J, Xiong H, Babik J, Hecht M. Protein design by binary patterning of polar and nonpolar amino acids. Science 1993;262:1680–1685.
18. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 2008;71:891–902.
19. Subbiah S, Laurents D, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr Biol 1993;3:141–148.
20. Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
21. Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res 2007;35(Database issue):D308–D313.
22. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
23. Hubbard SJ, Thornton JM. NACCESS, Computer program, Department of Biochemistry and Molecular Biology, University College, London. 1993. Available at: http://wolf.bi.umist.ac.uk/unix/naccess.html.
24. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
25. Sandelin E. On hydrophobicity and conformational specificity in proteins. Biophys J 2004;86(1, Part 1):23–30.
26. Elofsson A. A study on protein sequence alignment quality. Proteins 2002;46:330–339.
27. Jones D, Taylor W, Thornton J. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
28. Schmidt H, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002;18:502–504.
29. R Development Core Team 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005.
30. Faber H, Matthews B. A mutant T4 lysozyme displays five different crystal conformations. Nature 1990;348:263–266.
31. Reeves G, Dallman T, Redfern O, Akpor A, Orengo C. Structural diversity of domain superfamilies in the CATH database. J Mol Biol 2006;360:725–741.