doi:10.1006/jmbi.2000.3550 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **297**, 233–249

JMB

Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure and Function through Traditional and Probabilistic Scores

Cyrus A. Wilson[1], Julia Kreychman[1] and Mark Gerstein[1,2]*

[1]*Department of Molecular Biophysics and Biochemistry*

[2]*Department of Computer Science, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, CT 06520, USA*

**Corresponding author*

Measuring in a quantitative, statistical sense the degree to which structural and functional information can be ''transferred'' between pairs of related protein sequences at various levels of similarity is an essential prerequisite for robust genome annotation. To this end, we performed pairwise sequence, structure and function comparisons on ~30,000 pairs of protein domains with known structure and function. Our domain pairs, which are constructed according to the SCOP fold classification, range in similarity from just sharing a fold, to being nearly identical. Our results show that traditional scores for sequence and structure similarity have the same basic exponential relationship as observed previously, with structural divergence, measured in RMS, being exponentially related to sequence divergence, measured in percent identity. However, as the scale of our survey is much larger than any previous investigations, our results have greater statistical weight and precision. We have been able to express the relationship of sequence and structure similarity using more ''modern scores,'' such as Smith-Waterman alignment scores and probabilistic *P*-values for both sequence and structure comparison. These modern scores address some of the problems with traditional scores, such as determining a conserved core and correcting for length dependency; they enable us to phrase the sequence-structure relationship in more precise and accurate terms. We found that the basic exponential sequence-structure relationship is very general: the same essential relationship is found in the different secondary-structure classes and is evident in all the scoring schemes. To relate function to sequence and structure we assigned various levels of functional similarity to the domain pairs, based on a simple functional classification scheme. This scheme was constructed by combining and augmenting annotations in the enzyme and fly functional classifications and comparing subsets of these to the *Escherichia coli* and yeast classifications. We found sigmoidal relationships between similarity in function and sequence, with clear thresholds for different levels of functional conservation. For pairs of domains that share the same fold, precise function appears to be conserved down to ~40 % sequence identity, whereas broad functional class is conserved to ~25 %. Interestingly, percent identity is more effective at quantifying functional conservation than the more modern scores (e.g. *P*-values). Results of all the pairwise comparisons and our combined functional classification scheme for protein structures can be accessed from a web database at http://bioinfo.mbb.yale.edu/align

Keywords: bioinformatics; sequence similarity; percent identity; structure similarity; functional classification

Abbreviations used: EC, Enzyme Commission; EST, expressed sequence tags; SCOP, structural classification of proteins; GO, Gene Ontology Project.

E-mail address of the corresponding author: Mark.Gerstein@yale.edu

0022-2836/00/010233–17 $35.00/0

Introduction

The problem of genome annotation

Perhaps the most valuable information to be gained from a genome analysis is functional annotation of all the gene products. Unfortunately, of all the proteins whose sequences are known, functions have been experimentally determined for only a very small number (Andrade & Sander, 1997). Given the current size and accessibility of sequence and structure data, homologs of a newly sequenced gene's product can be identified *via* database searches, and probable structure and function assigned to the gene product (Bork *et al*., 1998). This is based on the concept that sequence similarity implies structural and functional similarity. However, structural and functional annotations should be transferred with caution. If a protein is assigned an incorrect function in a database, the error could carry over to other proteins for which structure or function is inferred by homology to the errant protein (Brenner, 1999; Karp, 1996, 1998a). In large databases such an error can propagate out of control, presenting a serious quality control issue as we move to larger genomes from multicellular organisms.

Benchmarking fold and function recognition

Here, we used manually curated structural and functional classifications as standards in analyzing to what degree annotations of a protein's structure and function can be transferred to a similar sequence. The knowledge gained from the study can be used to establish confidence levels for structure and function prediction, improving our understanding of how long it will take to annotate accurately an entire genome.

Our simultaneous analysis of relationships between sequence and structure, sequence and function, and structure and function (Figure 1) may provide insight into paradigms for functional prediction other than that based alone on sequence similarity (Enright *et al*., 1999).

Past results

Sequence-structure

The transfer of structural annotation is well characterized. Chothia & Lesk (1986, 1987) found that structural divergence, when expressed in terms of the RMS separation of matching alpha carbon atoms, was an exponential function of sequence divergence, expressed in terms of the fraction of residues that differed between sequences. The reliability of structural annotation transferred by homology, then, depends on the sequence identity of the homologous proteins (Chothia & Lesk, 1986). Flores *et al*. (1993), Russell & Barton (1994), and Russell *et al*. (1997) observed the same general trend, and also characterized the conservation of structural features other than the C^{α} backbone, such as secondary structure, accessibility and torsion angles. A paper by Wood & Pearson (1999) re-expressed the sequence-structure relationship in terms of statistically based ''Z-scores'' and found that this relationship had a simple linear form in terms of these scores. They also noted that protein families differed in detail in the slope of this linear relationship.

Others have focused on the limits of sequence comparison, specifically around the ''twilight zone,'' the region of sequence similarity that does not reliably imply structural homology (Doolittle, 1987), and on establishing cut-offs for significant sequence similarity. Using the SCOP structural classification (Murzin *et al*., 1995), Brenner *et al*. (1998) benchmarked the effectiveness of the popular FASTA and BLASTP programs and their probabilistic scoring schemes (i.e. the *e*-value) (Pearson & Lipman, 1988; Pearson, 1996; Altschul *et al*., 1990, 1994; Karlin & Altschul, 1993). They found that in making fold assignments, the FASTA *e*-value closely tracked the number of false positives, i.e. the error rate, and that at a conservative *e*-value cut-off of 0.001, the FASTA program could detect nearly all the relationships that would be detected by a full Smith-Waterman comparison (Smith & Waterman, 1981). Specifically, they found that FASTA with a 0.001 threshold would find 16 % more of the structural relationships in SCOP than would be found by standard sequence comparison with a 40 % identity threshold. This rigorous benchmarking approach has been extended to assess transitive sequence comparison, through a third intermediate sequence and multiple-sequence matching programs such as PSI-blast (Park *et al*., 1997, 1998; Gerstein, 1998a; Salamov *et al*., 1999). In a related study Rost (1999) worked on characterizing the region after the twilight zone, which he called the ''midnight zone''. In a sense these benchmarking studies have culminated in the CASP fold recognition experiments (Moult *et al*., 1997; Sternberg *et al*., 1999).

Sequence-function

Although the exact dependence of functional similarity on sequence and structural similarity is not completely clear, initial indications of a gene product's function are most often based on simple sequence similarity (Bork *et al*. 1994, 1998). Often these are merely based on the best hit in database comparisons; see, for example, the annotation of some of the early genomes (Fraser *et al*., 1995, 1998). However, possibilities for more robust annotation transfer are increasingly available. One looks at the pattern of hits amongst different phylogenetic groups (Tatusov *et al*., 1997). Often these focus on the existence of key motifs and patterns associated with function (Zhang *et al*., 1998; Bork & Koonin, 1996; Attwood *et al*., 1999).

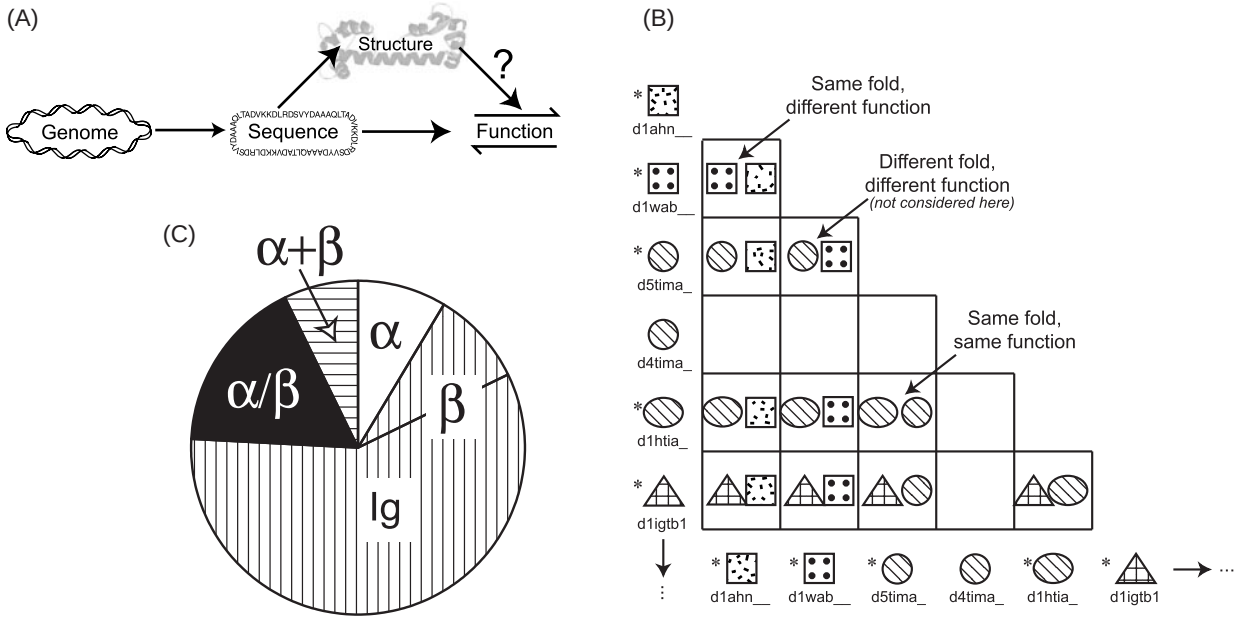


Figure 1. This Figure schematically depicts certain aspects of our comparison methodology. (a) The paradigm relating sequence to structure to function. There has not been as much assessment of functional annotation transfer based on structure as there has been with sequence-based structural and functional annotation transfer. (b) How we conceptualized our analysis in terms of pairs. A few examples of SCOP domains (identified on the left and bottom) are included from our comparison. In the Figure the shape represents fold, and the pattern represents function. We have highlighted some example categories of pairs: a pair that shares fold and function, a pair that shares fold but not function and a pair that shares neither fold nor function. The latter category of pairs is not considered in our investigation; we looked only at paired domains with the same fold. In constructing our pairs, we used only a representative set of SCOP domains. This is illustrated in the Figure by the domains flagged with asterisks. Note, in particular, that the SCOP domain d4tima_ is not paired with anything because it is represented by d5tima_, which is the same species and protein. For each level of pairs (fold, superfamily, family), cluster representatives were chosen for the level below: (i) for family pairs, one representative was selected from each species/protein, the level below, and then paired with all the other representatives within its family; (ii) for superfamily pairs, one representative was chosen from each family, unless there were domains in the family that shared less than 40 % sequence identity, in which case additional representatives were included, each not more than 40 % identical with the other representatives from the family (this occurs, for instance, for the globins); and (iii) likewise for fold pairs, one representative was chosen from each superfamily, more if there were domains with less than 40 % sequence identity. (c) Subdivides the pairs into the four SCOP classes from which they were composed: (i) all-α, domains consisting of α-helices; (ii) all-β, domains consisting of β-sheets; (iii) α/β, domains with integrated α-helices and β-strands; and (iv) α + β, domains with segregated α-helices and β-strands. We initially set apart the immunoglobulins from the rest of the all-β pairs because we realized that their large number biases our data. However, we compared the results for the immunoglobulin pairs to all other pairs and found that they generally exhibit the same behavior as the other pairs. Therefore we decided to leave them in the comparison.

Sequence-structure-function

One way that the better-defined sequence-structure relationship can assist in function prediction is initially to predict the structure of an uncharacterized sequence and then predict the function based on the limited repertoire of functions known to occur with that structure. To some degree this was achieved by Fetrow and co-workers (Fetrow *et al*., 1998; Fetrow & Skolnick, 1998). They predicted structural profiles based on threading and *ab initio* methods, and then searched with these against profiles of known structures in order to predict function.

In related work, Russell *et al*. (1998) discussed using identification of structural binding sites in predicting protein function. In a comprehensive study, Hegyi & Gerstein (1999) investigated to what degree folds were associated with functions. They found that most folds were associated with one or two functions with the exception of a few special folds, such as the TIM barrel, that could carry out numerous functions. Furthermore, they found that particular folds were often confined to distinct phylogenetic groups, an additional fact that can feed into an integrated sequence-structure-function analysis (Gerstein & Hegyi, 1998; Gerstein, 1997, 1998b,c).

Here, we look at pairwise comparisons of protein sequence, structure and function among proteins that share the same fold. We assess the

trends relating sequence, structure and function and consider the implications for structural and functional annotation transfer.

New developments: probabilistic scoring and growth of the databank

The past studies regarding sequence, structure and function relationships often used RMS separation and percent sequence identity (or a linear variant of it, such as the fraction of mutated residues) to express similarities in structure and in sequence, respectively. However, it has become increasingly common to use probabilistic scoring schemes (*P*-values) to express the quality of a match in terms of statistical significance rather than an arbitrary raw score such as percent identity (Pearson, 1998; Karlin & Altschul, 1990, 1993; Karlin *et al*. 1991; Altschul *et al*. 1994; Bryant & Altschul, 1995; Abagyan & Batalyov, 1997). With *P*-values, scores from different investigations can be compared in a common framework. Recently, it was found that sequence and structure similarity significance can be expressed as *P*-values in the same unified statistical framework (Levitt & Gerstein, 1998). Here, we use such probabilistic scoring methods to overcome the limitations of the more traditional scores.

Another recent development is the tremendous growth in the number of solved structures. The RCSB Protein Data Bank (Bernstein *et al*. 1977) now contains more than 10,000 protein structures. These structures are broken into more than 18,000 domains, and then domains that share a fold are paired up with each other for comparison (Figure 1(b)). Here, we survey ~30,000 pairs of protein domains that are known to have the same fold, approximately 1000 times the number compared by Chothia & Lesk (1986). The large scale of this comparison affords greater statistical weight to the results.

Alignment of 30,000 pairs from SCOP

The basic unit of comparison: a pair of protein domains

The protein domains that we studied were classified by SCOP, a Structural Classification of Proteins (Murzin *et al*. 1995; Brenner *et al*. 1996; Hubbard *et al*. 1997), a hierarchy of five levels: (i) class, domains that have the same secondary structural content (all-α, all-β, α/β, or $\alpha+\beta$); (ii) fold, domains that geometrically share the same tertiary fold; (iii) superfamily, domains descended from the same ancestor (but which lack measurable sequence similarity); (iv) family, domains in the same protein sequence family (which have appreciable sequence similarity); and (v) species and protein.

Pairs of protein domains that are grouped together at the fold, superfamily or family level form the basic unit of our comparisons.

Selection of pairs

There is potentially a huge number of pairs of domains that can be constructed out of the relationships in SCOP. For instance, in the current version of SCOP there are ~3.9 million potential pairs between domains sharing the same fold. Most of these are between nearly identical structures. In order to keep the number of pairs manageable, we used a straightforward clustering scheme, described in the legend to Figure 1. We selected 29,454 representative pairs from the total in SCOP. To achieve a wide range of similarities, we constructed the pairs on three levels of the SCOP hierarchy: (i) family pairs, 19,542 pairs of domains in the same family; (ii) superfamily pairs, 4220 pairs of domains in the same superfamily but different families; and (iii) fold pairs, 5692 pairs of domains in the same fold but different superfamilies.

All the selected domains were at least 50 residues in length and were drawn from the four major SCOP secondary-structural classes: all-α, all-β, α/β, and $\alpha+\beta$ (Figure 1(c)).

We automatically aligned each of our selected domain pairs twice, once by global Needleman-Wunsch sequence comparison (Needleman & Wunsch, 1971; Myers & Miller, 1998) and then by structure (Gerstein & Levitt, 1996, 1998), calculating scores for sequence and structural similarity.

Web-accessible database

The results of all the pairwise comparisons are available *via* a searchable database on the web at http://bioinfo.mbb.yale.edu/align The query engine allows searches of individual SCOP pairs, all pairs that include a given SCOP domain, or all pairs containing any SCOP domain contained in a given PDB entry.

Traditional scores: RMS and percent identity

The sequence-structure relation, as expressed by the root-mean-square (RMS) of the aligned C^{α} distances and percent sequence identity, has been previously characterized as an exponential function by Chothia & Lesk (1986) and others (Flores *et al*. 1993; Russell & Barton, 1994; Russell *et al*. 1997). As Figure 2 illustrates, our data display a similar trend. (Exact equations are given in the legend to Figure 2.) However, we have one thousand times as many data points as in Chothia and Lesk's original study (30,000 as opposed to 30).

The main difference between our results and the previous studies is due to differences in RMS ''trimming'' methods. By trimming we refer to the process of removing the worst-fitting aligned atoms from the RMS calculation, to arrive at a structural ''core.'' This was first developed in Lesk's sieve-fit procedure (Lesk & Chothia, 1984) and has been refined in numer-

ous studies (e.g. Gerstein & Altman (1995)). This is done because the small distances between well-matched alpha carbon atoms have much less of an effect on the RMS than do the very large distances between poorly matched atoms. The untrimmed score of divergent protein domains is then concerned primarily with the poorly matched residues instead of the conserved core. Trimming alleviates this effect by restricting the RMS calculation to include only those residues believed to be in the conserved core. However, the degree of trimming is to some extent arbitrary, and this choice affects the baseline of the reported RMS scores. Here we considered only the better half (50%) of matched residues in a given pair of protein domains. Chothia & Lesk (1986) chose a somewhat different threshold. Figure 2(c) and (d) demonstrate the effect of trimming.

Analogous alignment similarity scores: Smith-Waterman score and structural comparison score

The dependence of the RMS separation on trimming method restricts its usefulness in comparing data. Likewise, there are many problems with using percent identity as a measure of sequence similarity. For instance, a match of non-identical but still similar residues (e.g. Arg *versus* Lys) scores the same as one between completely different residues (e.g. Arg *versus* Val), and gaps do not enter in the score calculation. Consequently, we now turn to alignment similarity scores, which eliminate some of the problems with traditional scores.

For sequence alignments, an alignment score is defined as the sum of the similarity matrix values for the alignment, minus the total gap penalty. This is sometimes called the Smith-Waterman score (Smith & Waterman, 1981). An analogous alignment score for structure is the structural comparison score, described by Levitt & Gerstein (1998). We will refer to these two similarity scores as S_{seq} and S_{str}, respectively. Note that they both increase for more similar pairs, whereas RMS increases for more divergent pairs. Specifically, S_{str} is the score maximized by the structural alignment program we used (Gerstein & Levitt, 1998). It can be calculated from any pair of aligned structures according to the function:

$$S_{str} = M \sum \left(\frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} - \frac{N_{gap}}{2} \right) \qquad (1)$$

M and d_0 are constants, usually set to 10 and 5 Å, N_{gap} is the number of gaps in the alignment, d_i is the distance between each aligned pair of C^α atoms, and the sum is carried over all aligned pairs, i.

The main advantage of S_{str} over RMS in describing structural similarity is that the C^α to C^α distance, d_i, appears in the denominator of the calculation. This means that the smallest distances, corresponding to the best matches in the conserved core, are most significant in determining the score. Hence, the need for trimming is eliminated. S_{str} is also advantageous because it takes gaps into account and because of the fundamental analogy between this score and S_{seq}.

Figure 3(a) displays the relationship between structural and sequence similarity as expressed by S_{str} and S_{seq}. Figure 3(c) and (d) show calibration curves relating each of these scores back to approximate RMS separation and percent identity, respectively. Calibration curves help one get an intuitive feel for the degree of relationship in terms of the more traditional scores. Figure 3(b) adds a third axis, alignment length, and demonstrates that S_{str} depends greatly on this quantity. Although S_{str} and S_{seq} are ''better'' scores than RMS and percent sequence identity, the heavy dependence of both of these on length limits their usefulness in many situations. In other words, two pairs of similar domains with equal percent sequence identities but different lengths can have drastically different S_{seq} scores.

Probabilistic scores: *P*-values expressing the significance of sequence and structure similarity

Probabilistic scores can, to a great degree, overcome the length-dependence problems associated with the alignment scores. Probabilistic measures are advantageous because they express similarity not by an arbitrary ''score'' but by a statistical significance: the likelihood that such a similarity could be achieved by chance. This likelihood is also called the ''*P*-value.'' We used calculations (described in detail in the legend to Figure 4) based on those given by Levitt & Gerstein (1998) to obtain *P*-values based directly on S_{str} and S_{seq}; we refer to these calculated *P*-values as P_{str} and P_{seq}, respectively. For P_{seq} we could equally well have used the numbers from one of the popular sequence search programs (i.e. BLAST or FASTA) as all these values have been shown to be perfectly proportional to each other (Levitt & Gerstein, 1998; Brenner *et al*. 1998).

P_{seq} and P_{str} can be used to express the relationship between structure and sequence similarity on a more fundamental level. Figure 4(a) shows a log-log (base 10) plot of P_{str} against P_{seq}. Because it is log-log, trends can be visualized as straight lines. Two straight lines are necessary to fit the points well, with the discontinuous boundary between the lines located at the beginning of the twilight zone. The different slope of the line at low sequence similarity reveals that in the twilight zone there is a different relationship between the significance of structural similarity and that of sequence similarity. In particular, for domain pairs

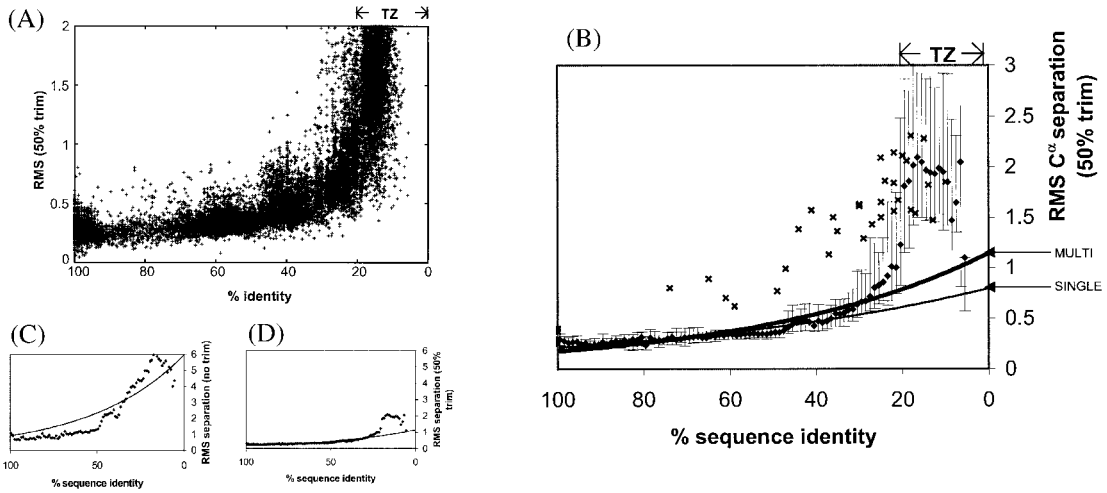


Figure 2. RMS as a function of percent identity. (a) A simple scatter plot of our pairs, relating RMS separation to percent sequence identity. This is similar to the presentation given by Chothia & Lesk (1986), but in this survey we looked at 30,000 pairs, 1000 times the number they compared. Outliers (pairs with RMS scores further than two standard deviations from the mean for their percent identity) are excluded from this graph; they represent domains that are very closely related with the exception of a conformational change. (b) A simplified graph with a number of fits to the data. For each percent identity bin we show the median RMS value, indicated by (◆) and the top and bottom quartile RMS values, indicated by the bars. Two fits are drawn through the median RMS values. The thin line, labeled SINGLE, is a simple exponential fit through the medians. It has the form:

$$R = 0.21\mathrm{e}^{0.0132H}$$

where R is the RMS deviation after least-square fitting, H is the percent difference between the sequences (H for Hamming distance), and $H = 100\,\% - I$, where I is the percent sequence identity. The thick line, labeled MULTI, is a multigraph fit, which is described in the legend to Figure 4. The relation between RMS and percent identity according to this fit is expressed by the equation:

$$R = 0.18\mathrm{e}^{0.0187H}$$

The twilight zone of sequence identity and below is labeled TZ. In this region, sequence similarity is not significant and not reliable for predicting structural similarity. This is why the median values in this area of the graph deviate significantly from the fits, which consider only data above 20 % sequence identity. For reference we include the original data points from Chothia and Lesk's, 1986 paper (A.M. Lesk, personal communication), indicated by X. Their data follow the form:

$$R = 0.40\mathrm{e}^{0.0187H}$$

The difference between the Chothia & Lesk trend and our relationship is due to the different trimming methods used in calculating the RMS score. Chothia and Lesk imposed a 3 Å cut-off in determining the conserved core residues; we defined the core as the better matching (in terms of C^{α} distances) half (50 %) of the residue pairs. (c) and (d) The effect our trimming has on median RMS values. The RMS values in (c) are calculated from all the matched residues in each pair; the values in (d) are calculated from the better matching 50 % of the residues.

in the twilight zone (according to the percent identity to P_{seq} calibration in Figure 4(b)), structural similarity is more significant than sequence similarity (having a smaller P-value or more negative log P-value). In contrast, for pairs with more than ~30 % identity, the situation is reversed, with a given pair having more significant sequence similarity than structural similarity. One possible interpretation of this reversal is as follows. Structure is always more highly conserved than sequence, so usually a given amount of structural similarity is not as significant as a corresponding amount of sequence similarity. However, this is true only when meaningful sequence similarity actually exists; thus, it does not apply in the twilight zone, where sequence similarity is by definition not significant. Note that all pairs in our comparison share at least the same fold, implying that they always have a significant amount of structural similarity.

In other words, for closely related sequences, differences in sequence similarity are more meaningful, whereas for highly diverged sequences that share the same fold, the differences in structural similarity are more significant.

Fitting two lines to the P_{str} *versus* P_{seq} graph suggests that the same might be done for other scoring schemes. It is possible to some degree to fit

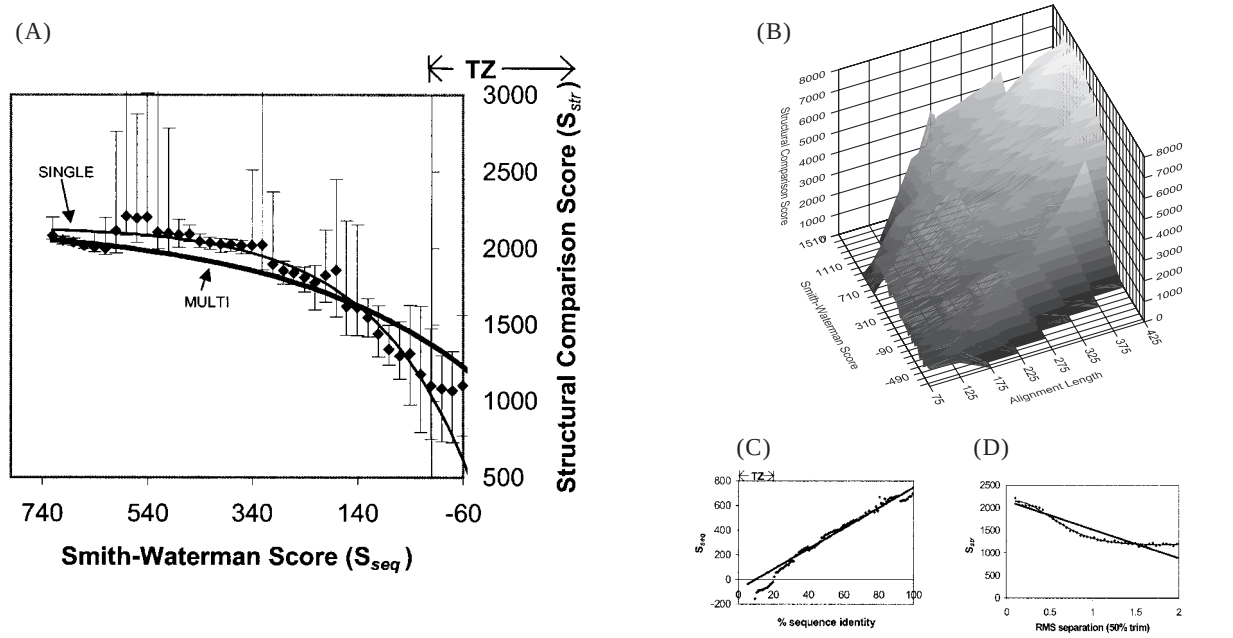


Figure 3. Similarity scores: structural comparison score as a function of Smith-Waterman score. Alignment similarity scores S_{str} and S_{seq} have certain advantages over RMS and percent identity scores for expressing the sequence-structure relation. S_{str} is calculated according to equation (1) in the text (Gerstein & Levitt, 1998; Levitt & Gerstein, 1998). S_{seq} is calculated using the BLOSUM50 matrix (Henikoff & Henikoff, 1992) with gap opening and extension penalties of −12 and −2, respectively. (a) This is analogous to (b) in Figure 2. From the original 30,000 pairs we show the median S_{str} value for each S_{seq} bin, along with quartile bars above and below. Again the twilight zone and below is labeled TZ. The thin line, marked SINGLE, is a simple fit to the median S_{str} values in this graph; it has the form:

$$S_{str} = 2144 - 1106\exp(-0.00544S_{seq})$$

The thick fit, marked MULTI, is the multigraph fit, explained below. It follows the equation:

$$S_{str} = 2157 - 787\exp(-0.0028S_{seq})$$

The equations presented here provide an approximation of the observed trends; as (b) illustrates, they are nothing more than simple approximations. The main disadvantage of S_{str} as a measure of structural similarity is its heavy length dependency for pairs of structurally similar protein domains. (b) Surface plot of the median S_{str} as a function of S_{seq} and alignment length (the number of matched residue pairs). It is clear that the size of the aligned domains plays a major role in the resulting S_{str}, even though our fits do not take length into account. (c) and (d) Relate S_{seq} and S_{str} to the more familiar percent identity and RMS measures. The fits were used to convert between scoring schemes in constructing the multigraph fit. We derived the multigraph fit in order to create one set of equations and parameters that would relate sequence and structural similarity using either the percent identity and RMS scheme or the S_{seq} and S_{str} scheme, and allow translation between them. We simultaneously performed least-squares fits to the median values in four graphs: Figures 2(b) and 3(a) and the calibrations of S_{seq} to percent identity and S_{str} to RMS, (c) and (d), respectively. In all cases, we ignored data in and below the sequence identity twilight zone (labeled TZ). The parameters in (a) are dependent on the parameters in Figure 2(b) *via* the mentioned calibrations.

the traditional RMS *versus* percent identity graph (Figure 2) with two straight lines instead of an exponential cruve. However, in this case, we opted for the more conventional presentation.

Class differences

The division of SCOP into classes based on secondary-structural composition allows easy investigation as to whether there are any deviations from the common similarity relationships on account of secondary-structure characteristics. Figure 5(a) reveals that secondary structural composition does not markedly affect the trends in sequence and structure similarities. This is consistent with the data given by Wood & Pearson (1999). However, the larger average length of α/β domains compared with domains in the other classes results in a deviation in the length-dependent S_{str} (Figure 5(b)). The consistency among length-independent scores applies for certain individual folds as well. The immunoglobulin fold makes up an appreciable fraction of all the β-pairs (Figure 1(c)), yet the results are not affected if these pairs are left out.

Linking sequence and structure to function

Difficulties of functional comparison

There is a clear, well-characterized relationship between sequence and structure similarity, which

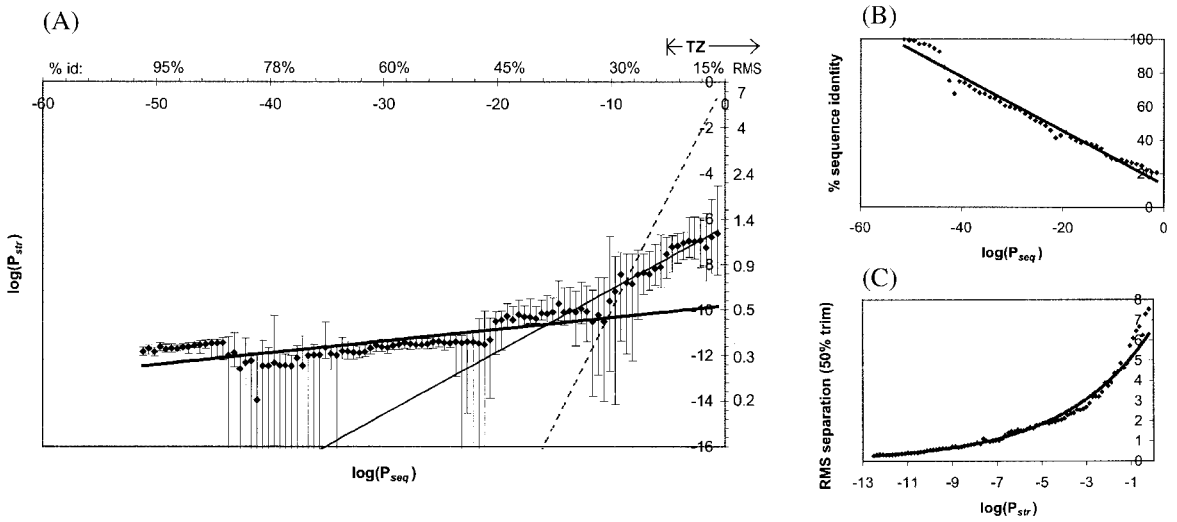


Figure 4. Probabilistic scores: *P*-values. P_{seq} and P_{str} are *P*-values calculated from S_{seq} and S_{str} according to the formalism given by Levitt & Gerstein (1998). Both quantities have the same overall functional form in terms of an extreme value distribution:

$$P = 1 - \exp(-\exp(-Z))$$

where P is either P_{seq} or P_{str}. For P_{seq}, $Z = S_{\text{seq}}/a - 2\ \ln M - b/a$, where $a = 5.84$, $b = -26.3$, and M is the geometric mean of the lengths of the two sequences (i.e. $M^2 = nm$, where n and m are the two sequence lengths). For P_{str}, Z is a function of S_{str} and N, the number of matched residues: For $N < 120$:

$$Z = (S_{\text{str}} - c \ln^2 N - d \ln N - e)/(f \ln N + g)$$

For $N \geqslant 120$:

$$Z = (S_{\text{str}} - a \ln N - b)/(f \ln 120 + g)$$

At $N = 120$, continuity implies that:

$$a \ln 120 + b = c \ln^2 120 + d \ln 120 + e \quad \text{and} \quad a = 2c \ln 120 + d$$

This, in turn, allows the calculation of the constants:

$$a = 171.8,\ b = -419.4,\ c = 18.4,\ d = -4.50,\ e = 2.64,\ f = 21.4,\ g = -37.5$$

(a) of this Figure is analogous to Figures 3(a) and 2(b), with the exception of the fits. It is a log-log (base 10) plot relating P_{seq} and P_{str}. We show the median $\log(P_{\text{str}})$ value for each $\log(P_{\text{seq}})$ bin, along with quartile bars above and below. We have added approximate percent identity and RMS values to the x and y axes to aid interpretation of the graph in terms of more familiar scores. The values were calculated using the calibration curves in (b) and (c). The straight-line nature of the log-log plot reveals distinct relations inside and outside the twilight zone, labeled TZ. (The area of percent identity below the twilight zone does not appear in P_{seq} graphs, there is no significance for such low sequence similarity; thus all data points in that zone appear at $P_{\text{seq}} = 1$ or $\log[P_{\text{seq}}] = 0$.) The thick line in the figure is fit to the median P_{str} values for P_{seq} values outside the twilight zone; its equation is:

$$P_{\text{str}} = 10^{-10} P_{\text{seq}}^{0.05}$$

The thin line is fit to the data inside the twilight zone; it follows the relation:

$$P_{\text{str}} = 10^{-6} P_{\text{seq}}^{0.274}$$

For reference we include the dotted line, representing the function $P_{\text{str}} = P_{\text{seq}}$, where sequence and structural similarity are equally significant. See the text for a discussion of how the two trends might be interpreted with respect to this line.

can be used to transfer precisely structural annotation based on the degree of sequence homology. In genome analysis, however, one is usually more interested in finding a functional annotation for an open reading frame based on similarity to well-known proteins; yet the sequence-function and structure-function relationships have not been as explicitly characterized. The fundamental obstacle to extending this and similar investigations to deal with function is the absence of a clear measure of

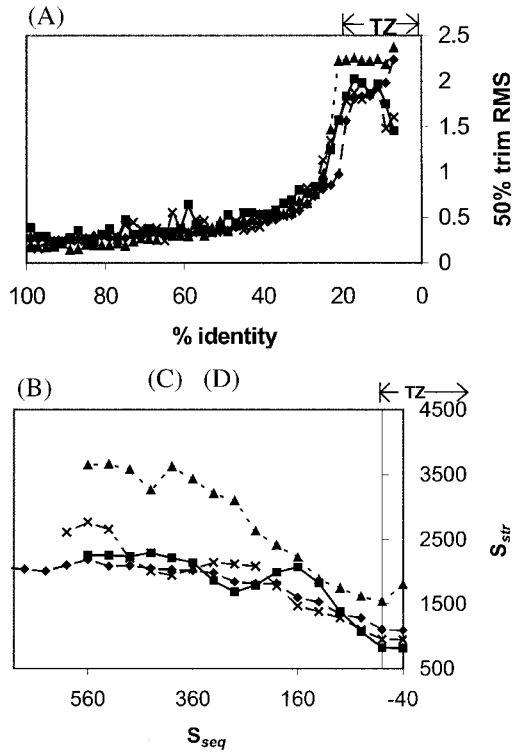


Figure 5. SCOP class differences. Previously it has been observed that secondary structural composition does not cause deviations from the trends in structure and sequence similarity (Flores *et al.* 1993). To test this observation we looked at the scores divided by SCOP class. The following legend applies to the graphs: (— ■ —), all alpha; (– ◆ –), all beta; (- - ▲ - -), alpha/beta; (– - × – -), alpha + beta. (a) Median RMS values for each percent identity bin. The traditional scores reveal no dependency on class. However, in (b) α/β pairs consistently score higher S_{str} scores than pairs in other classes. This is a consequence of the dependence of S_{str} on length; domains in the α/β class are longer, on average, than in the other classes.

functional similarity. Although we were able to present three different quantitative measures of structural relatedness, an analogous situation for function does not exist. How can one express quantitatively the degree of similarity between a triosephosphate isomerase and a glucose-6-phosphate isomerase? How do they compare to trp repressor?

The absence of a clear measure of functional similarity is not the only obstacle in transferring the functional annotations between proteins with different degrees of homology. The definition of function itself is often vague. More specifically, at present there is an absence of such important information as a standardized vocabulary for protein functional annotations with an associated numbering scheme, descriptions of monomer functions of subunits of multisubunit proteins and hierarchical functional assignments for proteins with multiple functions. As a consequence of these difficulties there is no functional equivalent to the hierarchical fold classification for domains in PDB.

As signs of progress in this direction, several functional classifications have been developed to date. One is the ENZYME system developed by the Enzyme Commission (EC) to classify enzymes by reaction type (Webb, 1992). This system has the advantage that it is ''universal,'' applicable to proteins in many different organisms, and is in wide use. However, it also has several drawbacks. First of all, it does not consider catalytic reaction mechanisms (Riley, 1998a), often ignoring obvious similarities. Second, it presumes a 1:1:1 relationship between gene, protein and reaction, although this is often not the case (an enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function). Perhaps the most significant drawback of the EC classification is that it applies to only enzymes.

A number of more comprehensive schemes have been developed, which classify non-enzymes as well as enzymes. Most of these focus on individual organisms. Several such schemes exist, for instance, GenProtEC/EcoCyc for *E. coli* (Karp *et al.*, 1998b; Riley & Labedan, 1996; Riley, 1998b), MIPS for yeast (Mewes *et al.*, 1998), Ashburner's functional classification for Drosophila, which is connected to FLYBASE (Ashburner & Drysdale, 1994), and EGAD for human ESTs (Adams *et al*, 1995). These classifications possess some advantages. They have additional levels of hierarchy that help present a more comprehensive picture of genotype-phenotype relationships. On the other hand, these classifications still leave much room for improvement. For example, there is no standardized vocabulary to allow for keyword searches among multiple databases and across organisms, and there are inconsistencies in category numbering style.

Finally, there has been some promising work going beyond the ENZYME and organism-focused classifications. There has been progress on completely automated functional classification (des Jardins *et al.*, 1997; Tamames *et al.*, 1997), which has the potential for putting function assignments on a more objective basis. There are a number of databases synthesizing the various enzyme functions into coherent pathways and systems (e.g. KEGG and WIT, Ogata *et al.*, 1999; Selkov *et al.*, 1998). There also have been some very recent attempts to develop cross-species classifications of non-enzyme functions in the framework of the Gene Ontology Project (GO, geneontology.org). GO is a joint project between FlyBase, the Saccharomyces Genome Database and Mouse Genome Informatics, attempting to merge the fly, yeast and mouse functional classification schemes. However, a truly universal system for classifying all protein functions in all organisms within the same framework remains quite a challenge because of the

sheer diversity of organisms and distinct protein functions.

Our simple functional classification of SCOP domains: FLY+ENZYME

Given the discussed limitations, we constructed a simple functional classification for the SCOP domains included in our comparison; our classification is based on a merger of two of the existing functional annotations and a cross-referencing of subsets of this combination with some of the organism-specific schemes. First, we used pairwise comparison to cross-reference the PDB domains against the Swissprot database (Bairoch & Apweiler, 1998), as described by Hegyi & Gerstein (1999). We chose to assign protein functions according to Swissprot because it provides more comprehensive functional annotations than SCOP.

We were initially able to divide all entries into enzymes and non-enzymes, a division that represents the highest level of functional difference in our classification scheme (Figure 6). For the enzyme category, we transferred EC (Webb, 1992) numbers to those SCOP domains with a one-to-one match to a Swissprot enzyme. Only one-to-one matching entries could be considered because Swissprot assigns ENZYME numbers to entire proteins, whereas SCOP is a domain-based classification; therefore we could be confident about the classification of only those domains which map to an entire Swissprot entry.

In the absence of an EC-type classification for non-enzymes, we assigned functions to non-enzymatic SCOP domains according to Ashburner's original classification of Drosophila protein functions. This classification is derived from a controlled vocabulary of fly terms. It is available on the web and loosely connected with the FLYBASE database (Ashburner & Drysdale, 1994). For clarity, we precisely describe the specific files and version (1.55, 1997) of the classification that we used in the caption to Figure 6, and we will hereafter refer to these data files as constituting the original FLY classification.

The FLY classification is a dynamic object, changing as more is learned about the fly and other organisms. This is particularly true of late with the imminent completion of the Drosophila genome. In fact, since the completion of our analysis, the FLY classification has been superceded by the new GO classification (see above).

The hierarchical structure of the FLY classification makes it well suited for classifying non-enzymatic SCOP entries in a manner comparable to the ENZYME assignments for the enzymes. Another advantage of this classification is that it is more compatible with the makeup of the PDB than the *E. coli* and yeast classifications, as Drosophila is a multi-cellular organism, and many of the known structures come from animals. We were able to use the original FLY classification as a framework to which we added functional categories and individual proteins. For instance, we added ''Hemoglobin'' to the ''Physiological Processes - Respiration'' category. Another example is the ''Physiological processes - Immunity'' category (Figure 6(b)), to which we added immune system proteins. Many of the additions would not be necessary in the context of the new cross-species GO system. We also modified slightly the numbering scheme in the original FLY classification in order to assign a unique hierarchical number to each protein domain (Figure 6(b)). We will refer to our augmented FLY classification as the FLY+ scheme, and our merged scheme as the FLY+ ENZYME classification.

As discussed earlier, the universal functional classification of proteins is very challenging and may not be possible with the current level of knowledge about genes, proteins and genomes. Consequently, the FLY + ENZYME classification of SCOP proteins is somewhat incomplete and inconsistent and retains many of the limitations of its components (Hegyi & Gerstein, 1999; Riley, 1998a). It is not yet broad enough to include many plant, virus and bacterial proteins. Nevertheless, it was sufficient for our analysis, as we were able to classify a very large number of the total 30,000 pairs.

Determining functional similarity

Using our compound functional classification, we were able to assign a level of functional similarity to each domain pair. According to our scheme, a pair can have no functional similarity (an enzyme paired with a non-enzyme) or it can have one of three levels of similarity:

(i) General similarity. Both domains are enzymes or both are non-enzymes.

(ii) Same functional class. Both domains share the first component of their ENZYME or FLY + numbers, e.g. 1.1.1.1 alcohol dehydrogenase and 1.3.1.1 cortisone beta-reductase (for enzymes), or 3.3.2.1.2 calcicyclin and 3.6.3.2.1 calmodulin (for non-enzymes).

(iii) Same precise function. Both domains share three components of their ENZYME or FLY + number, e.g. 1.1.1.1 alcohol dehydrogenase and 1.1.1.3 homoserine dehydrogenase (for enzymes) or 1.2.9.1.1.1 Arc repressor and 1.2.9.1.1.1 C-jun (for non-enzymes; both are transcription factors). A pair that shares precise function must also, by definition, share functional class and general similarity.

Based on those assignments we calculated the percentage of total pairs at a given level of sequence or structural similarity possessing each level of functional similarity. The results appear in Figure 7.

Sequence and function

The relation between sequence similarity and functional similarity behaves as one might expect, with sigmoidal curves that drop off sharply at particular conservation thresholds, and with the three levels of functional similarity (precise function, functional class and general similarity) having progressively lower thresholds. Figure 7(a) shows that precise function is not conserved below 30-40 % sequence identity, whereas functional class is conserved for sequence identities as low as 20-25 %. Below 20 %, general similarity is no longer conserved; among pairs of approximately 7 % sequence identity, about 40 % are enzymes paired with non-enzymes. It is important to note that in all the pairs considered here, the domains share the same fold. Functional similarity at low percent identities (e.g. 7 %) would be much less for all possible pairs of domains rather than just for those with the same fold. It is also important to remember that our thresholds for functional conservation are statistical averages over many sequences; one will, of course, be able to find individual cases that diverge more or less rapidly.

There are differences between the functional conservation thresholds of enzymes and non-enzymes, with enzymes appearing to more highly conserve precise function than non-enzymes, but non-enzymes conserving functional class more highly than enzymes. This may reflect that in our classification, the non-enzyme functional classes are broader and hence easier to conserve than those of the enzymes, while the non-enzymatic precise functions are more specific.

When P_{seq} is used as the measure of sequence similarity (Figure 7(b)) the results look somewhat different, it appears that functional class is conserved for the entire range of sequence similarities. In this case, percent identity is actually more discriminating than P_{seq} because functional class diverges only at sequence similarities that are low enough that they have little or no statistical significance, i.e. for P_{seq} the divergence is compressed near the vertical axis of the graph.

Structure and function

The relation between similarity in structure and function is somewhat less straightforward than that between similarity in sequence and function. Figure 7(c) shows the relationship between RMS and functional similarity. Broadly, it appears similar to that for percent identity and functional similarity; however, the thresholds for conservation of the various types of functional similarity are less sharp.

RMS is more revealing with respect to functional similarity than the non-traditional structural scores, S_{str} and $P_{str.}$ (Data for S_{str} and P_{str} are not shown but are available from the website.) The reason is that, while very structurally similar pairs all have RMS scores clustered between 0 and 0.5 Å, S_{str} has a large range of scores for similar pairs due to the length dependency, and P_{str} does not have any limit for maximum similarity. The wide range of possible S_{str} and P_{str} scores for similar structures tends to blur the broad sigmoid curves so much so that they are no longer apparent.

Alternative functional classifications: MIPS and GenProtEC

To get some perspective on the degree to which our results reflected the particularities of our combined FLY + ENZYME classification, we decided to try the same comparisons based on the well-known functional classifications for yeast and *E. coli*, MIPS and GenProtEC (Mewes *et al.*, 1998; Riley & Labedan, 1996; Riley, 1998b). These classifications have the advantage that they integrate enzyme and non-enzyme functions from the start and are widely used. However, as they are only applicable to individual organisms, we could only use them to classify a considerably smaller subset of the known structures than the compound FLY + ENZYME system.

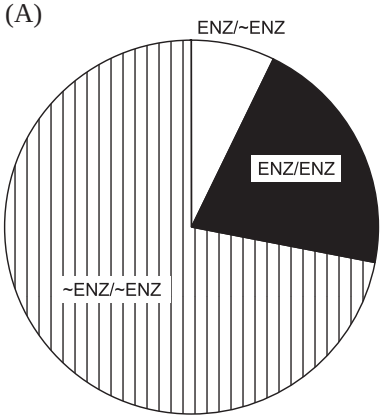
The specific way we used the MIPS and GenProtEC classifications to assign function to structures and to calculate functional similarities is described in the legend to Figure 7. Our results in terms of functional conservation (precise and class) at various levels of percent identity are shown in Figure 7(d). We observe the same general relationships as we did for our FLY + ENZYME scheme. That is, the functional conservation curves have a sigmoidal shape and have cut-offs for precise functional similarity after 40 % and for functional class similarity at lower values. However, because the MIPS and GenProtEC classifications are restricted to individual organisms, each curve represents considerably fewer data points than do the curves based on the FLY + ENZYME scheme; this required us to ''bin'' the MIPS and GenProtEC curves in a somewhat coarser fashion.

Discussion and Conclusion

Here, we assessed the transfer of functional and structural annotation by analyzing the relationships between similarity in sequence, structure and function. The ~30,000 protein domain pairs of varying levels of similarity (at least the same fold) that we constructed out of the SCOP classification show quantitative sequence-structure relationships consistent with previous research. The exponential relationship is consistent across the secondary-structural classes and holds for newer probabilistic scoring methods.

The sequence-function and structure-function relationships have not been studied as precisely due to the lack of a robust functional classification and measure of functional similarity. To overcome

(A)



(B)

Category type	Specific example	Original	Modified
Function primitive	Cell structure	0.	3
Process or Molecule	Nucleus	1.	3.1
Pathway or Macromolecule	Nuclear Membrane structure	2.	3.1.1
Individual instances	Nuclear envelope protein	3.	3.1.1.1
Individual instances	Lamin	4.	3.1.1.1.1
Individual instances	LaminA	5.	3.1.1.1.1.1
	...		
Process or Molecule	Nucleocyt. transport	2.	3.1.2
Individual instances	Nuclear pore structure	3.	3.1.2.1
	...		
Process or Molecule	Cell motility	1.	3.12
Pathway or Macromolecule	Motor protein	2.	3.12.1
Individual instances	Myosin	3.	3.12.1.1
Process or Molecule (new)	Immunity	1.	5.6
Pathway or Macromolecule (new)	Immunoglobulin	2.	5.6.1
Individual instances(new)	Immunoglobulin binding prot.	3.	5.6.1.1

(C)

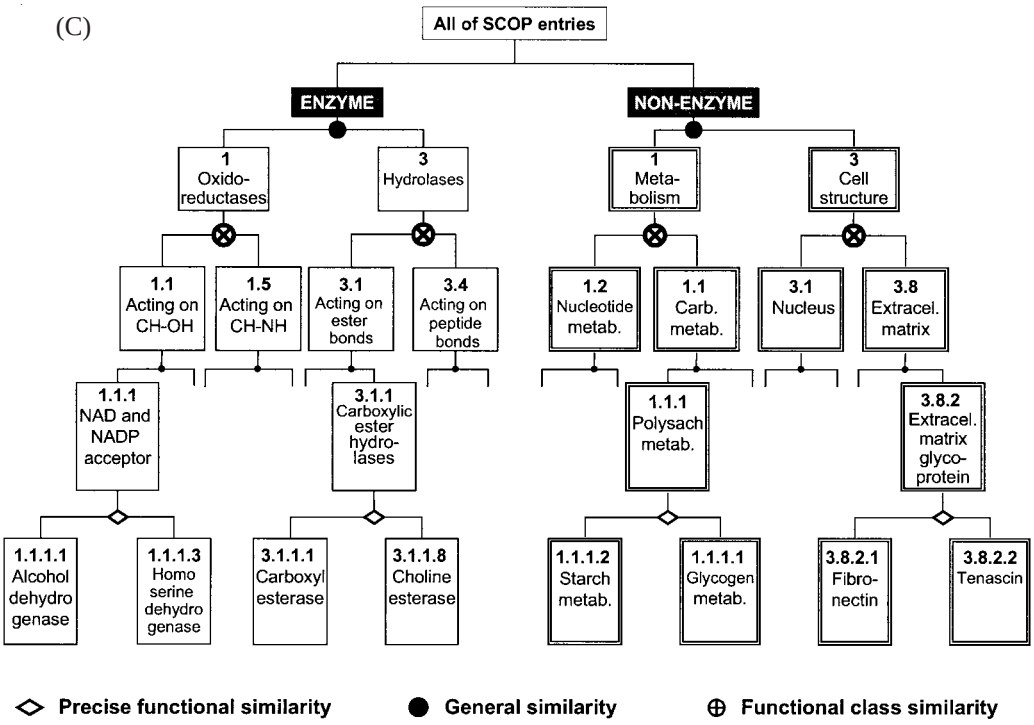


Figure 6. Functional classification of enzymes and non-enzymes. (a) Divides the pairs by general function. There are three categories of pairs: (i) enzymes paired with non-enzymes (no general functional similarity), labeled ENZ/ ~ ENZ; (ii) enzymes paired with enzymes (same general function), labeled ENZ/ENZ; and (iii) non-enzymes paired with non-enzymes (same general function). Pairs for which one or both domains could not be identified as enzyme or non-enzyme are not included in this chart. Enzymes are classified according to the EC system (Webb, 1992). The first component of the number represents the nature of reaction and is called class. There are six classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The next level is subclass. It refers to the chemical groups on which the enzyme acts. For example, the first class, oxidoreductases, has 19 subclasses that are arranged according to the donor group that undergoes oxidation (CH-OH, aldehyde or oxo group, CH-CH group, etc). For another group of enzymes (hydrolases) subclass is determined by the nature of the bond: ester bond, peptide bond, etc. The next level is sub-subclass. For oxidoreductases this indicates the acceptor group: NAD(+) and NADP(+), or cytochrome; for hydrolases the sub-subclass represents the nature of substrate (carboxylic ester hydrolases, thiolester hydrolases, etc.). The fourth level represents a unique number for each individual enzyme, for example, 1.1.1.1: alcohol dehydrogenase. (b) Shows how we adapted the functional classification of Drosophila gene products developed by M. Ashburner. This classification is loosely connected with FLYBASE (Ashburner & Drysdale, 1994). We used version 1.55 (4 August 1997) that was available from Ashburner's website:

http : //www.ebi.ac.uk/ ~ ashburn

The specific files that we used were taken from the ftp directory:

ftp.ebi.ac.uk/databases/edgp/misc/ashburner

this we constructed our own classification by merging and extending the ENZYME and FLY schemes and assigning levels of functional similarity. Our measures of functional similarity provide curves relating function to sequence and structure; when relating functional conservation to sequence divergence, we find distinct thresholds at ~40% for precise function and ~25% for functional class.

One of the interesting results that emerges from this is that percent identity is more useful for quantifying functional divergence than the newer probabilistic scores. In general, modern probabilistic scores, such as P_{seq}, are better at discriminating amongst highly diverged sequences (near the twilight zone) than percent identity, since they better take into account gaps and conservative substitutions (of similar amino acids). However, for very similar pairs of sequences, percent identity is a simpler and more direct measure of divergence (essentially a Hamming distance). Since divergence in precise function takes place before that in structure (well before the twilight zone), it is quite reasonable that percent identity is more successful at measuring the former than the latter and that the converse is true for the probabilistic scores. In other words, percent identity is better calibrated for discriminating amongst very close, significant relationships and P_{seq} for more distant ones.

Practical implications

The sequence-structure and sequence-function relationships described here provide practical information for genome annotation in terms of folds and functions. Table 1 summarizes the relative advantages of the different scoring methods we used. Using the trends in sequence and structure similarity, one can assess the degree to which structural annotation can be transferred between sequences at a given level of sequence similarity. The sequence and function similarity thresholds potentially establish minimum requirements of sequence similarity for reliable function prediction. Note that because the protein domain pairs considered here all share the same fold, the numbers for all possible pairs will differ in the region of very little sequence identity, in which the sequence similarity is not enough to indicate the same fold.

We refer to these as constituting the original FLY classification. Recently, the FLY classification has been superceded by the GO (Gene Ontology) Project classification, which merges fly, mouse and yeast annotation. Files related to the GO classification are available from www.geneontology.org In the original FLY classification all members of the highest level are labeled 0, representatives of the next level are labeled 1, and all lower levels are labeled 2 through to 9. We changed the numbering scheme so that it will reflect the hierarchical nature of the classification. This Figure illustrates sections of the original and modified classification. The top level in the FLY classification scheme is called ''Function primitive'' (level 0) and includes five classes: ''Metabolism,'' ''Intracellular protein traffic,'' ''Cell structure,'' ''Developmental process,'' ''Physiological process,'' and ''Behavior.'' The next level after ''Function primitive'' is ''Process'' or ''Molecule'' (level 1 in Ashburner's classification). For ''Function primitive - Metabolism'' the processes are ''Carbohydrate metabolism,'' ''Nucleotides and nucleic acids metabolism,'' etc. For ''Function primitive - Cell Structure'' the ''Process'' can be ''Nucleus,'' ''Mitochondrion,'' ''Membrane,'' etc. The next level is ''Pathway'' or ''Macromolecule'' (level 2 in the original classification). ''Pathway'' can include ''Metabolic pathway,'' ''Signaling pathway,'' or ''Developmental pathway.'' The ''Macromolecule'' category includes ''Protein'' and ''Nucleic Acid''. We added categories to the original classification in order to classify some mammalian proteins that are widely represented in SCOP but are absent from the original FLY scheme. These categories include immune system proteins (labeled ''new'' in (b) and respiratory proteins such as hemoglobin and myoglobin that we added to ''Function primitive - Physiological process - Respiration''. We call our adaptation of the original FLY scheme, FLY + . Further information on this adaptation is available at:

http : //bioinfo.mbb.yale.edu/align/func

(c) The overall hierarchy of our final scheme and identification of the different levels of similarity. If two proteins are both enzymes or both non-enzymes, then they possess general functional similarity. If they share the first component of their classification numbers, then they are in the same functional class. If they share the first three components of their enzyme numbers (or the equivalent for non-enzyme numbers, depending on category) then they have the same precise function. A significant difference between the two main branches of the hierarchy is that the levels of the ENZYME classification do not correspond exactly to those in the FLY+ system because the fly classification is more extensive than the enzyme classification. For instance, the FLY classification takes into account aspects of cellular (cytoskeleton, metabolic pathways, etc.) and phenotypic function (morphology, physiology, behavior) that are absent from the ENZYME scheme. This makes our classification of SCOP proteins somewhat unbalanced, as non-enzymes have much broader and more loosely defined functional classes. As a consequence, while each enzyme is assigned a four-component number, the length of a non-enzyme number varies, depending on the functional category to which it belongs. For example, myosin is assigned a number that happens to have the same length as EC numbers: 3.12.1.1. However, transcription factors are numbered 1.12.9.1.1.1. We took into account this varying hierarchy depth in deciding how many components are necessary to identify precise function in each category. Note that what we mean by domains having the same precise function is not the same as the domains coming from the same essential protein.

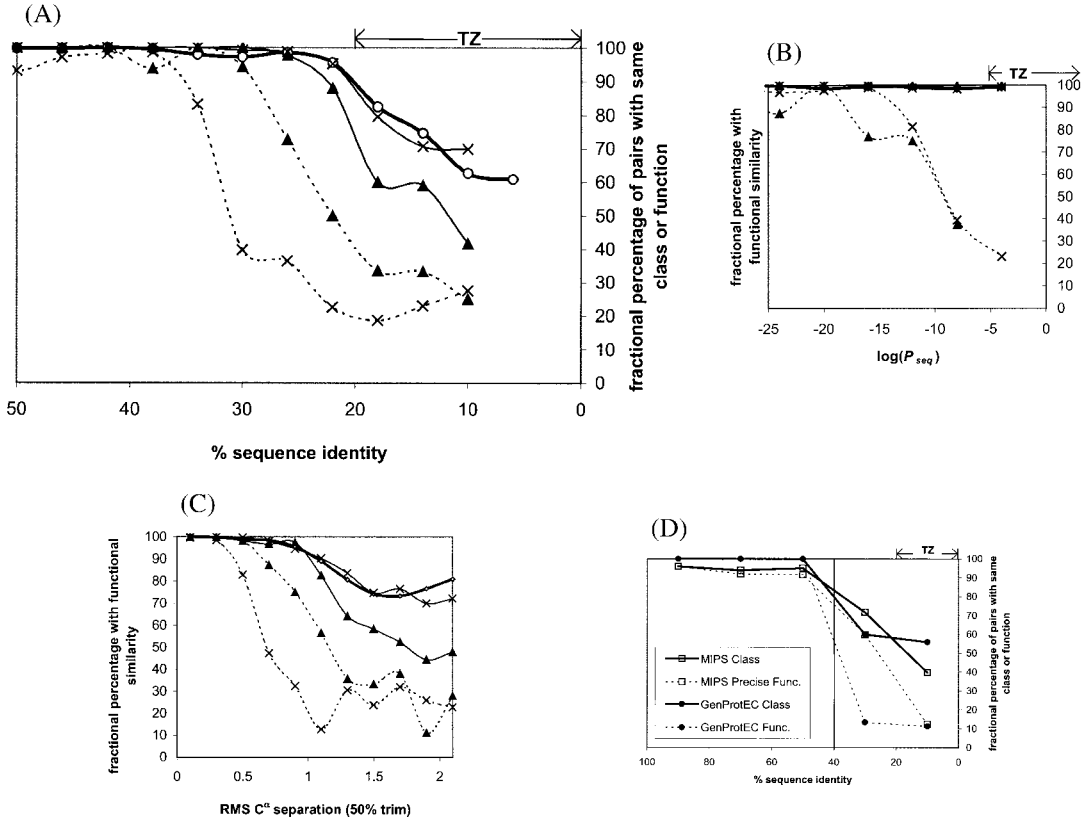


Figure 7. Linking sequence, structure and function. We express functional similarity as the fractional percentage of pairs at a given level of sequence/structural similarity for which the paired domains share a precise function, functional class, or general similarity (according to our classification, see Figure 6). The following legend applies to (a) through (c): (—○—), general similarity; (— × —), non-enzymes with same functional class; (—▲—), enzymes with same functional calss; (--- × ---), non-enzymes with same precise function; and (---▲---), enzymes with the same precise function. (a) Relates functional similarity to sequence similarity in terms of percent identity. The functional similarity appears as a sharp sigmoid, with distinct thresholds of divergence for precise function, functional class, and general similarity. Enzymes are paired with non-enzymes only at very low percent identity, in and below the twilight zone (labeled TZ). At slightly higher sequence identity, pairs diverge with respect to functional class, and beyond 40 % identity with respect to precise function. Note that 50-100 % identity is not shown because almost all domains that are that similar share function with their counterparts. (b) Shows the same data using P_{seq} as the measure of sequence similarity. Only the divergence in precise function is visible because there is such little significance for the low sequence similarity at which functional class and general similarity diverge, all data points in that region appear near $P_{\mathrm{seq}} = 1$ or $\log[P_{\mathrm{seq}}] = 0$ (the y-axis). (c) Illustrates that the structure-function relation is not as clearly defined as that for sequence and function. Functional similarity expressed in terms of RMS separation appears as a broad sigmoid curve; there are thresholds of divergence for precise function, but the divergences in functional class and general similarity are more gradual. The thresholds are apparent only because RMS clusters the most structurally similar pairs between scores of 0 and 0.5 Å. For this reason, RMS is better at discerning functional similarity than S_{str} and P_{str}, which do not cluster the most similar pairs around a set limit. (d) Shows the same relationships (functional conservation *versus* percent identity) as in (a), except that for this graph functional similarity is determined in terms of the MIPS (Mewes *et al.*, 1998) and GenProtEC (Riley, 1998b) classifications rather than the FLY + ENZYME scheme. The legend appears as the inset on the graph. We assigned MIPS and GenProtEC classifications to SCOP domains based on sequence comparisons to classified yeast and *E. coli* open reading frames (ORFs), respectively. The SCOP domain most closely matching each ORF classified in MIPS or GenProtEC was assigned the corresponding MIPS or GenProtEC function number. Only matches of 80 % sequence identity or greater were considered. We used this SCOP domain as a functional representative; when determining functional similarity, we assigned to SCOP domains with no MIPS or GenProtEC functional designation the function of the closest representative with at least 85 % sequence identity, if one existed. GenProtEC functional identifiers are three-component numbers. We consider a pair of domains sharing the first component of their functional designation to be in the same functional class. Domains that share all three components are said to have the same precise function. For MIPS the functional designation is not as straightforward, as one ORF can be assigned multiple functions. Therefore we consider domains which have at least one function in common to share functional class. Domains with all functions in common, the same combination of identifiers, share precise function. Because MIPS and GenProtEC each classify the proteins of a single organism, yeast and *E. coli*, respectively, these classifications can determine the functional similarities of only a small fraction of all our SCOP domain pairs. The data based on these classifications, appearing in (d), are therefore very sparse compared to the data in (a)-(c). Despite the coarseness of the data, functional similarity based on the MIPS and GenProtEC classifications follows the same general relation to sequence similarity as does functional similarity based on the more comprehensive FLY + ENZYME scheme. Vertical line indicates an approximate threshold of functional divergence at 40 % identity.

Table 1. Summary of scoring methods

	Sequence similarity	Structural similarity	Features	Limitations
Traditional scores	Per cent sequence identity	RMS C^{α} separation	Well understood, in use; percent identity better for looking at functional similarity	RMS depends most highly on worst matches, requiring arbitrary trimming; percent identity is insensitive to gaps and conservative substitutions
Alignment similarity scores	S_{seq}	S_{str}	Analogous similarity scores, S_{str} depends most highly on best matches	Dependence on alignment length
Modern probabilistic scores	P_{seq}	P_{str}	Statistical significance, unified framework for different comparisons	Not as familiar as RMS and percent identity

The Table lists the schemes presented here for characterizing the sequence-structure relationship, along with their relative advantages and disadvantages.

Practically, then, when one searches an uncharacterized open reading frame against known structures, if the open reading frame matches a structure with a good *e*-value or percent identity, then the curves presented here can be used to check how the functional and detailed structure annotation will transfer. For example, if an unknown open reading frame matches a PDB structure with an *e*-value of 0.001 and a percent identity of 30%, then one can be assured that it has the same fold (Brenner *et al*., 1998) and according to our analysis it has a two-thirds chance of having the same exact function. Furthermore, it has a ~99% chance of having the same functional class and its structure probably diverges from the known structure by a trimmed RMS of less than 0.7 Å.

Future directions

There are a number of directions in which we might extend this analysis. With respect to the sequence-structure relation, we can reduce the overrepresentation of the immunoglobulins and improve the calculation of P_{str} (by redoing the fit to the extreme value distribution reported by Levitt & Gerstein (1998) to eliminate residual length-dependency.

In the functional realm, we can investigate if and how the sequence-function and structure-function relationships vary for different categories of proteins. For example, although we found consistency of the sequence-structure relationship among secondary structural classes, Hegyi & Gerstein (1999) found that the distribution of enzymes and non-enzymes varies with secondary structural class. A related issue is that of conformational changes. It is conceivable that among domains with very similar sequences but structures that differ by a conformational change, function is less conserved than it is among similar sequences with more similar structures.

Perhaps the most important direction in which to further this work is the augmentation of the functional classification. With the growing amount of fully sequenced genomes there is a need for the development of a comprehensive system for functionally classifying proteins, a complete classification for the entire universe of protein functions. It will be a difficult process, as many existing organism-specific classifications will have to be merged, but the end result will have the advantage of not being biased towards any one organism. Such a universal classification will allow much more reliable transfer of functional annotation.

Acknowledgments

We thank A. Lesk for helpful conversations and supplying us with reference data for Figure 2, S. Brenner for providing carefully curated SCOP domain sequences, and H. Hegyi, W. Krebs and V. Alexandrov for assistance with the sequence comparisons, development of the FLY + ENZYME scheme, and design of the web database. M.G. thanks the Keck and Donaghue foundations for financial support.

References

Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., Venter, J. C., *et al*. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature,* **377**, 3-174.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tools. *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.

Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotech.* **8**, 675-683.

Ashburner, M. & Drysdale, R. (1994). Flybase: the *Drosophila* genetic database. *Development,* **120**, 2077-2079.

Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N. & Wright, W. (1999). PRINTS prepares for the new millennium. *Nucl. Acids Res.* **27**, 220-225.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366-376.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707-725.

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132-133.

Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.* **266**, 635-643.

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.

Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236-244.

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.

des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB,* **5**, 92-99.

Doolittle, R. F. (1987). *Of Urfs and Orfs*, University Science Books, Mill Valley, CA, USA.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402**, 86-90.

Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949-968.

Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703-711.

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

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Venter, J. C., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science,* **270**, 397-403.

Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren, E., Hardham, J. M., McLeod, M. P., Salzberg, S., *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science,* **281**, 375-388.

Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics,* **14**, 707-714.

Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.

Gerstein, M. (1998c). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des.* **3**, 497-512.

Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161-175.

Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277-304.

Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *ISMB,* **4**, 59-67.

Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* **7**, 445-456.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

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

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236-239.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA,* **87**, 2264-2268.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA,* **90**, 5873-5877.

Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175-203.

Karp, P. D. (1996). A protocol for maintaining multidatabase referential integrity. *Pac. Symp. Biocomput.* 438-445.

Karp, P. (1998a). What we do not know about sequence analysis and sequence databases. *Bioinformatics,* **14**, 753-754.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998b). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50-53.

Karp, P. D., Ouzounis, C. & Paley, S. M. (1996b). HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. *ISMB,* **4**, 116-124.

Lesk, A. M. & Chothia, C. (1984). Mechanisms of domain closure in proteins. *J. Mol. Biol.* **174**, 175-191.

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

Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucl. Acids Res.* **26**, 33-37.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins: Struct. Funct. Genet.* **1**, 2-6.

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

Myers, E. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11-17.

Needleman, S. B. & Wunsch, C. D. (1971). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29-34.

Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 349-354.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-259.

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.

Riley, M. (1998a). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* **8**, 388-392.

Riley, M. (1998b). Genes and proteins of *Escherichia coli* K-12. *Nucl. Acids Res.* **26**, 54.

Riley, M. & Labedan, B. (1996). *E. coli* gene products: physiological functions and common ancestries. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F., Curtiss, R., III, Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umbarger, H. E., eds), 2nd edit., pp. 2118-2202, ASM Press, Washington, DC.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85-94.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* **244**, 332-350.

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

Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds - binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.

Salamov, A. A., Suwa, M., Orengo, C. A. & Swindells, M. B. (1999). Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12**, 95-100.

Selkov, E., Jr, Grechkin, Y., Mikhailova, N. & Selkov, E. (1998). MPW: the metabolic pathways database. *Nucl. Acids Res.* **26**, 43-45.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-198.

Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **9**, 368-373.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science,* **278**, 631-637.

Webb, E. C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*, Academic Press, New York.

Wood, T. C. & Pearson, W. R. (1999). Evolution of protein sequences and structures. *J. Mol. Biol.* **291**, 977-995.

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res.* **26**, 3986-3990.

Edited by F. E. Cohen

(Received 2 September 1999; received in revised form 5 January 2000; accepted 6 January 2000)