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
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 C² 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).
**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 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-$\alpha$, all-$\beta$, $\alpha/\beta$, 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, 4,220 pairs of domains in the same superfAMILY but different families; and (iii) fold pairs, 5,692 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-$\alpha$, all-$\beta$, $\alpha/\beta$, 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 $C^\alpha$ 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 $S_{\text{seq}}$ and $S_{\text{str}}$, respectively. Note that they both increase for more similar pairs, whereas RMS increases for more divergent pairs. Specifically, $S_{\text{str}}$ 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:

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

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

The main advantage of $S_{\text{str}}$ over RMS in describing structural similarity is that the $C^\alpha$ to $C^\alpha$ 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. $S_{\text{str}}$ is also advantageous because it takes gaps into account and because of the fundamental analogy between this score and $S_{\text{seq}}$.

Figure 3(a) displays the relationship between structural and sequence similarity as expressed by $S_{\text{str}}$ and $S_{\text{seq}}$. 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 $S_{\text{str}}$ depends greatly on this quantity. Although $S_{\text{str}}$ and $S_{\text{seq}}$ 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 $S_{\text{seq}}$ 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 $S_{\text{seq}}$ and $S_{\text{str}}$; we refer to these calculated $P$-values as $P_{\text{seq}}$ and $P_{\text{str}}$, respectively. For $P_{\text{seq}}$ 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).

$P_{\text{seq}}$ and $P_{\text{str}}$ 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 $P_{\text{str}}$ against $P_{\text{seq}}$. 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
in the twilight zone (according to the percent identity to $P_{\text{seq}}$ 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 $P_{\text{str}}$ versus $P_{\text{seq}}$ graph suggests that the same might be done for other scoring schemes. It is possible to some degree to fit
the traditional RMS versus percent identity graph (Figure 2) with two straight lines instead of an exponential curve. 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 \( \alpha/\beta \) domains compared with domains in the other classes results in a deviation in the length-dependent \( S_{\text{str}} \) (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 \( \beta \)-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
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

\[
P = 1 - \exp(-\exp(-Z))
\]

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

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

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:

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

(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 \( P_{\text{seq}} \) and \( P_{\text{str}} \). 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 \( P_{\text{seq}} \) 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 \( P_{\text{str}} \) values for \( P_{\text{seq}} \) 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.
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 calcyclin 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 $P_{\text{seq}}$ 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 $P_{\text{seq}}$ because functional class diverges only at sequence similarities that are low enough that they have little or no statistical significance, i.e. for $P_{\text{seq}}$ 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, $S_{\text{str}}$ and $P_{\text{str}}$. (Data for $S_{\text{str}}$ and $P_{\text{str}}$ 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, $S_{\text{str}}$ has a large range of scores for similar pairs due to the length dependency, and $P_{\text{str}}$ does not have any limit for maximum similarity. The wide range of possible $S_{\text{str}}$ and $P_{\text{str}}$ 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
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(+) acceptor; 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 \( \sim 40\% \) for precise function and \( \sim 25\% \) 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 \( P_{\text{seq}} \), 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 \( P_{\text{seq}} \) 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 percent identity suffers from the same fold tiers, 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: 3.12.9.1.1. 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; (— x —), non-enzymes with same functional class; (— — —), enzymes with same functional class; (— — —), 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 $P_{\text{seq}}$ 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 $P_{\text{seq}} = 1$ or $\log P_{\text{seq}} = 0$ (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 $S_{\text{str}}$ and $P_{\text{str}}$ which do not cluster the most similar pairs around a set limit. (d) Shows the same relationships (functional conservation $\text{CONSIS}$ 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.
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


Edited by F. E. Cohen

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