Practical Limits of Function Prediction

Damien Devos and Alfonso Valencia*
Protein Design Group, CNB-CSIC, Madrid, Spain

ABSTRACT **The widening gap between known protein sequences and their functions has led to the practice of assigning a potential function to a protein on the basis of sequence similarity to proteins whose function has been experimentally investigated. We present here a critical view of the theoretical and practical bases for this approach. The results obtained by analyzing a significant number of true sequence similarities, derived directly from structural alignments, point to the complexity of function prediction. Different aspects of protein function, including (i) enzymatic function classification, (ii) functional annotations in the form of key words, (iii) classes of cellular function, and (iv) conservation of binding sites can only be reliably transferred between similar sequences to a modest degree. The reason for this difficulty is a combination of the unavoidable database inaccuracies and the plasticity of protein function. In addition, analysis of the relationship between sequence and functional descriptions defines an empirical limit for pairwise-based functional annotations, namely, the three first digits of the six numbers used as descriptors of protein folds in the FSSP database can be predicted at an average level as low as 7.5% sequence identity, two of the four EC digits at 15% identity, half of the SWISS-PROT key words related to protein function would require 20% identity, and the prediction of half of the residues in the binding site can be made at the 30% sequence identity level. Proteins 2000;41:98–107.** © 2000 Wiley-Liss, Inc.

Key words: protein function; sequence conservation; binding site; functional key words; classes of cellular function; genomics

INTRODUCTION

Sequence Similarity and the Prediction of Protein Structure and Function

The spectacular increase in the number of sequenced genomes is widening the gap between protein sequences and functions. For example, in the cases of the *E. coli* and yeast genomes, functional information was available for less than half of the ORFs at the time of their publication.[1,2] Because experimental assessment of the function of every protein of each newly sequenced genome (23 finished genomes at the time of writing) is beyond foreseeable resources, our knowledge of most of the new proteins will be from predictions based on similarity searches. The needs of consistent functional assignments goes beyond the whole genome analysis projects, toward the tremendous amount of data generated by the new techniques in functional and structural genomics.

The comparison of protein sequences is widely used to infer protein structure and function. The underlying hypothesis is that function and structure can be transferred between similar sequences because they have been conserved over long periods of time. This assumption can be confirmed in the case of protein structures, for which a direct relationship between sequence similarity and conservation of protein structure was observed.[3,4] In the case of protein function a similar relationship is commonly assumed, but it is far less justified.[5] Indeed the relation between fold and function was investigated by different authors[6–8] who used the EC classification as a reference. However, any direct relationship between sequence and function has only been partially assessed. For example, the direct prediction of EC classes based on sequence analysis was previously investigated.[9,10]

A Common Function Prediction Exercise

The description of a typical exercise in sequence analysis could be helpful in understanding computational function assignment. The quest for the function of a protein usually starts by searching for related sequences in public databases, using tools such us BLAST[11] or PSI-BLAST.[12] Potentially similar sequences are inspected to avoid obvious pitfalls.[13] The next step is to access the annotations of similar sequences to determine the adequate level of function that can be transferred.[14] The heterogeneity of the database annotations and the fact that most sequence annotations have probably been derived by similarity[5] can be misleading at this step but are difficult to avoid. At this point, function will be transferred directly from those database entries judged to be most similar. This process includes combining pieces of information such as function description (e.g., DE field in SWISS-PROT), biochemical function (EC number), or cell function. Moreover, in many cases function will be explored by site-directed mutagenesis of those residues identical to the binding site residues of the reference protein (for a review see Bork et al.[15]).

Abbreviations: EC, enzyme classification; BCC, binding site comparison class; ECC, enzyme comparison class; SCC, structure comparison class; KCC, keyword comparison class; IC50, identity level required for a conservation of **50%** of the functional characteristics.

Grant sponsor: GeneQuiz, European Union, TMR project; Grant number: FMRX-CT96-0019.

*Correspondence to: Alfonso Valencia, Protein Design Group, National Centre for Biotechnology, CNB-CSIC, Cantoblanco, Madrid E-28049, Spain. E-mail: valencia@cnb.uam.es

Received 24 February 2000; Accepted 2 June 2000

© 2000 WILEY-LISS, INC.

Practical and Theoretical Problems in Function Prediction

The difficulties of this process are related in part to the theoretical definition of function and also to practical problems. Some of them associated with the current analysis tools and the databases are (i) the identification of similar sequences in large databases, a field in which substantial progress has been made (PSI-BLAST,[12] Hidden Markov Models[16]); (ii) the persistence of systematic errors in homology detection due to compositionally biased regions of different nature[13,17]; (iii) the wrongly annotated sequences in different databases as has been amply described[17–20]; and (iv) the propagation of errors by repeated copying of annotations between similar sequences.[5,21]

Besides the practical difficulties mentioned above, function itself is an elusive concept. As a consequence, a general definition of function, supported by a well-defined ontology valid across domains and organisms, is still lacking; this precludes the construction of systematic annotations. It can be argued that the construction of a complete description of function requires extensive knowledge of the evolution of protein function that is not yet available.

We present here a systematic survey of the relationship between sequence and functional similarity. The analysis was performed on a large collection of structurally aligned proteins, which allowed us to cover the full range of similarities, including truly related protein pairs that are independent of any sequence similarity searching method. Indeed, protein structure similarity is accepted as the gold standard in the evaluation of similarity searching tools.[12,16,22] The analysis of protein function was performed at four different levels: (*i*) *enzymatic function classification*, representing a standard definition of the chemical model of the protein enzymatic function; (*ii*) *functional annotations in the form of keywords*, describing the biochemical function commonly defined by interactions with compounds, cofactors, substrates, regulators, and other cellular components; (*iii*) *cell function class*, capturing the main types of activity to which each protein contributes, e.g., "carbon compound metabolism" or "DNA biosynthesis"; and (*iv*) *conservation of the type of amino acid in the binding site*, related to the binding activity of the protein, and in many cases, the functional discrimination between different substrates and cofactors. The analysis presented here poses interesting questions about the reliability of current function prediction exercises and the intrinsic limitations of protein function prediction.

MATERIALS AND METHODS

Sequence similarity and protein structure or function were compared by using the FSSP database[23] as a reference. The FSSP database comprises an extensive set of protein structure alignments, generated by starting with a non-redundant set of representative proteins, with <25% pairwise similarity, to which the rest of the PDB files are structurally aligned. For the current analysis, we imposed some additional restrictions to (i) guarantee the reliability of the structural alignments, (ii) to facilitate comparison with related sequences in other databases, and (iii) to avoid problems with protein domains. Only those alignments were selected having lengths between 75 and 100% of the length of both sequences and containing >50 aligned residues. Consequently, this study does not directly address the problem of multi-domain proteins. In addition, self-comparisons were excluded, rejecting sequence identities of >95%. At this stage, of the 116,750 structural alignments in the FSSP database (23/03/99), 7,162 alignments were selected.

Quality of the Structural Comparisons

The analysis described here was repeated with a second set of alignments restricted to those protein pairs with a higher structural similarity score (FSSP, Z score of 3.5), considered to provide more reliable structural alignments.[23] The results remained similar, but the number of observations was drastically reduced in all categories analyzed. The full analysis was also repeated with a different structural database that does not include explicit pairwise alignments (SCOP database[24]). The results (not shown) were essentially indistinguishable from those presented here.

Structural Comparison

In some cases, two representative FSSP proteins, sharing <25% identity, could have detectable structural similarity, resulting in the generation of two FSSP files, with each one of the two proteins as file headers. In one of them, the representative protein A is aligned to B and in the other, the representative protein B is aligned to A. Both structural alignments are obviously identical and so redundant for the structural comparison step. In such cases, we excluded one of the two redundant alignments. For the comparison of protein functions, we maintained both pairs, because the comparison of any functional description of the reference structure A to B is different from the same functional comparison of the reference structure B to protein A.

The final number of alignments considered for the structural comparison was 5,876, representing 5% of the total FSSP database. Each of the structural alignments was classified into a Structure Comparison Class (SCC), defined as the number of FSSP family index digits hierarchically shared by the aligned proteins. For example, StrX, with FSSP code 345.1.1.1.1.1, shares two FSSP family index digits with StrY, code 345.1.2.1.1.1; thus, their SCC = 2.

Comparison of Enzyme Classification Number (EC)

EC numbers for each PDB chain were obtained form the Enzyme Structure database,[25] an extension of the ENZYME database.[26] For those protein pairs for which EC numbers were available (2,338, 2% of the FSSP pairs), the percentage of shared code digits was computed as the Enzyme Comparison Class (ECC). For example, PrtX, with EC code 1.4.1.1, shares two EC code digits with PrtY (EC code 1.4.2.1); therefore, the ECC value = 2.

Comparison of SWISS-PROT Keyword Functional Annotations

The SWISS-PROT codes corresponding to each PDB structure were retrieved from the pdbtosp.txt file (release

Corteva Exhibit 2026 - Page 2 of 10
Bayer v. Corteva - IPR2023-01038

37.0) of the SWISS-PROT database[27] or from the PDBsum database[28] for PDB entries corresponding to more than one SWISS-PROT file, i.e., multi-chain structures, mutants of natural proteins, different ligands or crystallization conditions. To avoid redundancies, the alignments of PDB files with the same SWISS-PROT code were considered only once. The keywords associated with each SWISS-PROT entry were retrieved from the sprot37.dat file. Only functionally informative keywords (as derived from our previous analysis[29]) were considered. At this level, we did not implement any other discrimination of the keywords by their significance. To avoid possible artifacts created by the comparison of proteins annotated with too few keywords, only alignments in which the FSSP representative structure has more than one informative keyword were considered, leading to a final number of 2,161 valid alignments corresponding to 1.8% of FSSP. Keyword conservation was evaluated by counting the percentage of keywords of the representative structure present in the aligned protein. The percentages are grouped in four bins of Keyword Comparison Class (KCC); KCC 25, from 0 to 25%, KCC 50, up to 50%, KCC 75, up to 75%, and KCC 100, up to 100%. An example could be Structure 1, with 5 keywords in the corresponding file (kw1, kw2, kw3, kw4, and kw5), aligned with Structure 2, annotated with three keywords (kw2, kw3, and kw6). They share two of the five keywords of Structure 1 (40%), corresponding to the KCC 50 class.

It should be noted that the SWISS-PROT keywords and functional class (which are keyword based) comparison curves are directly related to the process adopted by SWISS-PROT for the annotation of the sequences.

Comparison of the Annotations of Cell Function Class

The classes of cellular function were defined after the classification originally proposed by Riley for the *E. coli* genome[30] and later extended by the TIGR group, during the initial analyses of different genomes. The 14 functional classes used were Transcription; Central intermediary metabolism; Purines, pyrimidines, nucleosides, and nucleotide biosynthesis; Energy metabolism; Regulatory functions; Biosynthesis of cofactors, prosthetic groups, and carriers; Amino acid biosynthesis; Replication; Cellular processes; Fatty acid and phospholipid metabolism; Cell envelope; Translation; Transport and binding proteins; and Unclassified proteins.

For the automatic assignment of sequences to classes, we used a recent extension of our previous system based on the SWISS-PROT keywords[29] and (Tamames and Valencia, unpublished data). We selected 2,226 alignments for the analysis (1.9% of FSSP). The protein pairs were classified according to whether both proteins belong to the same functional class.

Comparison of the Amino Acids at the Binding Sites

It is difficult to obtain a well-annotated set of active or binding sites. Different authors[31–33] have simplified the problem, defining the binding sites as those residues in protein structures that are in physical contact with different compounds, including cofactors and reaction products (for a detailed description of the selection process for these heteroatoms, see Ouzounis et al.[31]). Residues are labeled as binding if any of their atoms lie within 4 Å of any atom of the bound heteroatom, after exclusion of the solvent molecule. This practical definition has the advantage of providing a large number of binding site residues derived from real protein three-dimensional structures. Even this is hampered by a number of uncontrolled factors, including variability in the ligands crystallized with different proteins, the replacement of the natural ligands with other molecules, the lack of some cofactors in the protein crystals, or cases in which ligand three-dimensional coordinates were not deposited in the corresponding PDB files. Also, important residues binding to different cofactors via a water molecule are not selected by this method. Similarly, residues binding to peptides or to other proteins were not selected because we did not want to mix the study of binding sites with the analysis of protein-protein interactions, even if, to some extent, it can be seen as an additional aspect of protein function.

Of 9,424 structures selected in the PDB release of 03/99, 1,725 were not defined by X-ray diffraction, and an additional 2,486 structures did not contain heteroatoms in contact with the structures. For the remaining 5,213 proteins, it was possible to detect binding residues. To avoid considering the heavy atoms included for the crystallization and other modifications, we excluded those heteroatoms bound by less than three residues. The final number of cases analyzed was 5,068, corresponding to 4.3% of the FSSP database.

The similarity of the binding sites was evaluated by comparing the chemical type of the binding site amino acid of the FSSP reference structure with the chemical type of the structurally equivalent residue in the corresponding FSSP alignment. The level of similarity between the binding sites of the two proteins is given by the percentage of identical residues over the total number of binding site residues in the reference structure. Four classes of binding site comparison class (BCC) were defined, covering the range of 0 to 100% identical binding site residues. This comparison scheme mimics a simple approach to the prediction of binding sites, in which the position of the binding site residues is transferred on the basis of the pairwise alignment information.

Homogeneous Sampling

One important and difficult issue in this kind of analysis is to obtain a homogeneous set of alignments meeting the selection criteria for each kind of analysis. Nevertheless, the selection of the FSSP files allowed us to obtain a balanced sampling, covering between 2 and 5% of the FSSP entries for the analysis of the different functional characteristics.

RESULTS

Transference of Protein Structure From Sequence Information

More than a decade has passed since the relationship between sequence and structure was first investigated.[3]

Corteva Exhibit 2026 - Page 3 of 10
Bayer v. Corteva - IPR2023-01038

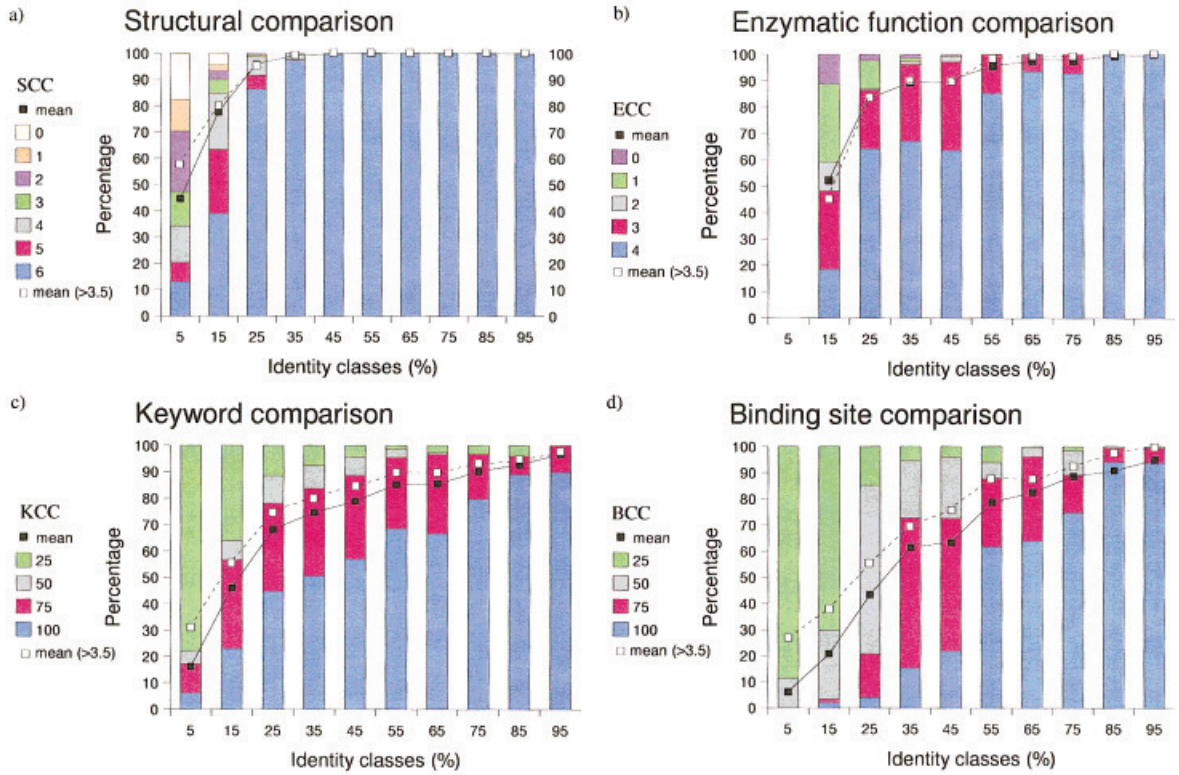

Fig. 1. Distribution of four protein characteristics at different levels of sequence identity. A large number of protein pairs, structurally aligned in the FSSP database, are classified by the level of sequence identity. For each of them, the different protein characteristics are analyzed. The number of pairs compared and the origin of the information is given in Materials and Methods. **a:** Structural similarity (SCC), representing the level of agreement of the FSSP six-digit codes classification. **b:** Enzymatic function (ECC), containing the level of agreement of the EC four-digit classification. In both cases, the level of matching of the two protein classifications is given, with 0 indicating no coincidence and the higher number (6 for FSSP and 4 for EC) indicating full matching of the classification of the two aligned proteins. **c:** SWISS-PROT keywords (KCC), representing the percentage of keywords of the reference sequence present in the paired protein. The bins of KCC 25 between 55 and 85 % identity, and the bin of KCC 50 at 65% identity, correspond to artifactual database annotations, i.e., a SWISS-PROT code erroneously assigned to a PDB file, or the same SWISS-PROT code assigned to a multichain protein. **d:** Coincidence of the amino acids in the binding site (BCC); the percentages are shown of amino acids of the representative sequence that are of the same chemical type as the corresponding residues in the aligned protein. Notes: The mean values are represented by lines and squares for the alignments with Dali Z score >2, black squares, and for Z score >3.5, open squares. Both sets (see Materials and Methods) show no substantial differences, except for the case of (d) where the number of observations at Z score >3.5 is too small to sustain further analysis.

The same relationship clearly appears in our analysis (Fig. 1a), in which proteins are automatically classified in the same classes (SCC = 6), until a region of low sequence similarity is reached. These results clearly show that the fold can be transferred reliably from a protein whose structure is known to an uncharacterized sequence, when the identity between both is >20% (in the conditions used in our analysis: >50 residues aligned; >75% of the protein lengths aligned). This result serves as reference for the functional comparisons presented below.

Transferring the EC Classification

Enzyme to Non-Enzyme Comparisons

We found a surprisingly large proportion of structural similarities between structures containing an EC number (putative enzymes) and structures without EC number label (putative non-enzymes). Approximately half of the alignments of our set corresponded to putative enzymes labeled with an EC number; of these 3,632 pairs, approximately one-third (1,294, 36%) involved a putative enzyme aligned with a putative non-enzyme. Even if the proportion of enzyme to non-enzyme alignments is smaller for the more similar pairs of proteins (Fig. 2a) still they represent a considerable number of cases in which a possible enzyme has not been properly labeled with the corresponding EC number. In contrast, in the low-similarity region, we found a large number of pairs in which the protein without EC number really is not an enzyme, including cases in which the active sites had been modified to retain binding activity but not catalytic capacity. Two examples are 5ptp, a bovine β-trypsin (EC code 3.4.21.4) 31% identical over 211 residues to 1ae5, a human heparin binding protein, without apparent enzymatic activity,[34] or 1auiB, a human calcineurin (EC code 3.1.3.16), aligned with 26% identity over 139 residues to 1rec, a bovine calcium-binding protein involved in vision, with no described enzymatic activity.

Comparing Enzymes With Assigned EC Number

For the 2,338 alignments in our set corresponding to pairs of proteins annotated with EC numbers, the level of sequence similarity was compared with the level of conservation of their corresponding EC classification (Fig. 1b).

Corteva Exhibit 2026 - Page 4 of 10
Bayer v. Corteva - IPR2023-01038

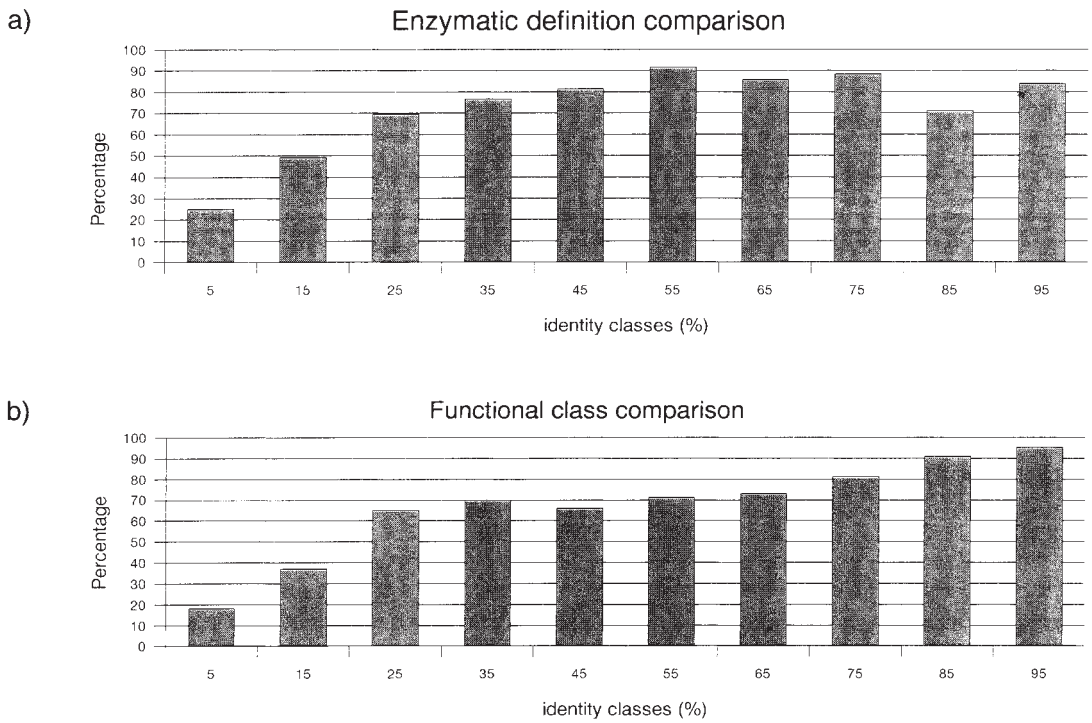

Fig. 2. EC code assignment and functional class comparison. **a:** Percentage of pairs in which both proteins have assigned EC codes. **b:** Percentage of pairs in which both proteins are assigned to the same functional class. The results are given independently for the 10 different identity classes. In (a) the proportion of pairs with both proteins labeled as enzymes becomes high (approximately 80%) as soon as a reasonable sequence identity level is reached (approximately 35% identity). In (b) the proportion of pairs with both proteins assigned to the same functional class only reaches 70% when sequence identity is between 30 and 70%.

The limit above which the EC codes were completely identical for both proteins was surprisingly high at 80% identity. Between 50 and 80% identity, only the first three EC digits were identical. For example, 1ppn and 1ppo, both Papaya hydrolases, share 70% identity. The former was labeled as a sulfhydryl proteinase and the latter as a thiol protease, with EC codes 3.4.22.2 and 3.4.22.30, respectively (even if sulfhydryl proteinase and thiol protease are functionally equivalent, there is an inconsistency in their description where the former is annotated in the EC database as "specific for peptide bonds with large hydrophobic side chain at P2 position," a specificity not stated for the second protein). Below 50% identity, the shift in enzymatic specificity begins to blur the signal, and the EC code is less conserved. Down to 30% identity, the dominant trend is to conserve at least the first three EC digits, an observation corroborated by analysis of the CATH database in which for 75% of the structural families, the first three EC numbers were conserved.[7]

In practice, below the 50% identity limit, it is difficult to select the completely correct annotations and below 30% identity, assignments of the EC code based on a pairwise alignment were found to be problematic. For example, with 41% sequence identity, 3lzt, a chicken hydrolase, and 1hfyA, a goat alpha lactalbumin, have different EC numbers (3.2.1.17 and 2.4.1.22, respectively, ECC = 0). But 1pgtA, the human glutathione S-transferase chain A, was 16% identical over 186 residues to 1gnwA, an Arabidopsis thaliana homologue, both with EC code 2.5.1.18.

Transference of Functional Annotation: the Case of the SWISS-PROT Keywords

SWISS-PROT contains a large set of manually curated functional annotations in the form of a restricted set of keywords. These include descriptions that are a mixture of aspects of protein function and other characteristics such as post-translational modifications. The keywords were less conserved between related proteins than the other features analyzed (compare Fig. 1c, with the structure distributions; Fig. 1a, and EC numbers; Fig. 1b). The probability of transferring correct functional annotations related to keywords is therefore quite low; for example, at 40% identity, an average protein pair will only share 70% of their keywords. Data analysis showed a general trend for keywords that contained less information about protein function, such as "disease mutation," "phosphorylation," or "lipoprotein," to be less conserved. In the case of two ferredoxin proteins that share 75% identity, 1awd from *Chlorella fusca* (FER_CHLFU) and 4fxc, from *Spirulina platensis* (FER_SPIPL), two informative keywords ("transport" and "iron-sulfur") were conserved, whereas a less general keyword was only included in the first sequence ("phosphorylation"), classifying the alignment in the KCC 75 bin. The heterogeneous nature of the keywords thus seems difficult to reconcile with their use during the evaluation of functional relations as recently proposed.[35–37]

Corteva Exhibit 2026 - Page 5 of 10
Bayer v. Corteva - IPR2023-01038

Functional Class Assignment

Functional classes are an attractive way of describing protein function at the general level of cellular activities. At levels of sequence identity as high as 70%, a significant number of cases were found in which both proteins were classified in different functional classes (Fig. 2b). Between 30 and 70% sequence identity, the average probability that two related proteins belong to the same functional class was only 70%. In the low sequence similarity range (approximately 20%), transfer of functional class was found to be essentially random. In general, the conservation of the functional classes was somewhat less conserved than the SWISS-PROT keywords.

Conservation of the Binding Site

The binding site is defined here as those residues in contact with different ligands in protein structures (see Materials and Methods). Binding sites are the least conserved feature between related proteins (Fig. 1d). We found extreme cases of binding sites that were very different at reasonably high levels of sequence similarity and cases of distantly related proteins that retained remarkably conserved binding sites. An interesting example is 1bcfA, the *E. coli* bacterioferritin, aligned to 1ryt, an electron transporter of *Desulfovibrio vulgari*, with 16% identity (Fig. 3a). The two manganese cations in the 1bcfA structure were bound by seven residues, of which six were identical in the corresponding positions of the 1ryt structure. The prediction of the cation-binding residues would have been completely correct, even for such a distant sequence relationship. It is interesting that 1bcfA also contained a bound heme ligand, whereas 1ryt is described as non-heme-binding protein. As expected, the structurally aligned residues around the heme binding site were not conserved between the two proteins, and as a consequence, the heme binding site would not have been predicted for 1ryt.

Binding site conservation follows a continuous drift directly parallel to the overall level of sequence identity; for example, at 75% identity, most of the proteins conserve 75% of the binding site residues. This point can be illustrated by following the conservation of the binding sites within a structural family. 1phnB, a phycocyanin from *Cyanidium caldarium*, is structurally similar to phycocyanin 1cpcL (75% identity), phycoerythrin 1liaL (54%), allophycocyanin 1b33B (39%), allophycocyanin 1allA (34%), phycocyanin 1cpcA (25%), and a sea cucumber hemoglobin 1hlb (14%) (Fig. 4). For this set, these overall levels of sequence similarity correspond directly to the conservation of their binding sites, with Binding site Comparison Class (BCC) values of 100, 75, 50, 50, 25, and 25 respectively (Fig. 4 and Table I).

Global Transference of Structure and Function

It is interesting to follow real cases of simultaneous transference of the different aspects of protein function (Table II). A first example of functional conservation is a pair of triosephosphate isomerases; 1amk, from *Leishmania mexicana* is 42% identical to 7timA, a yeast enzyme (Fig. 3b). Both proteins are TIM barrels (SCC = 6), with

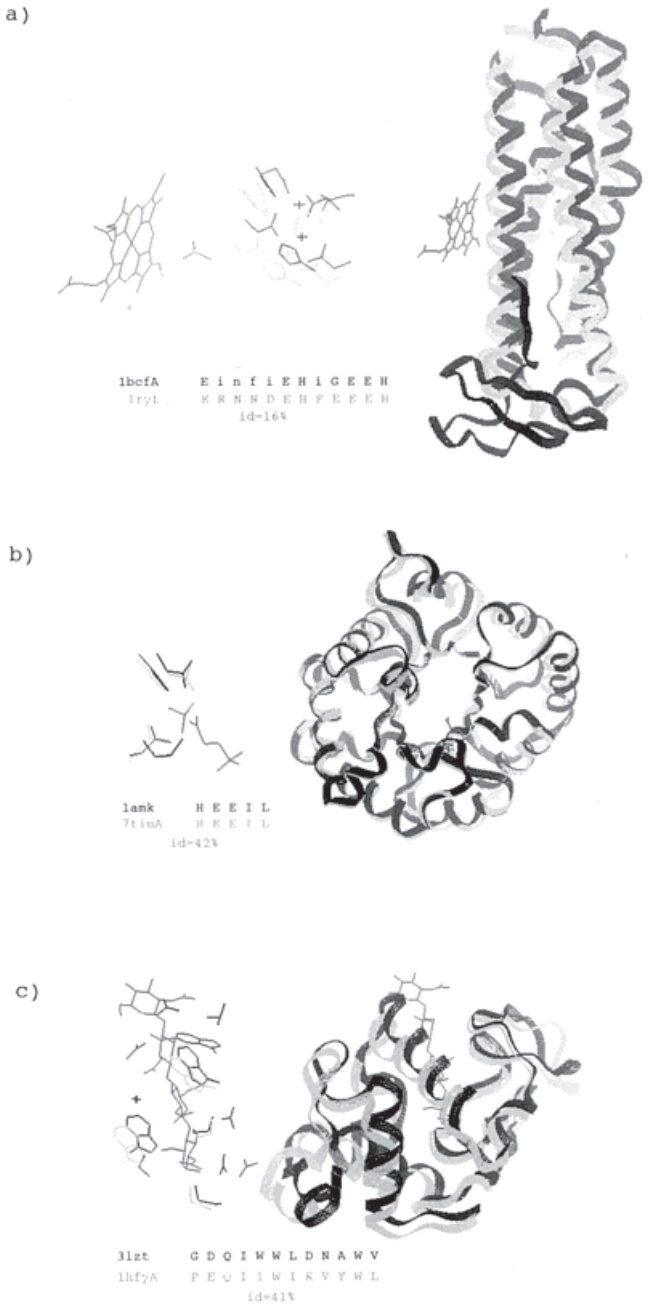

Fig. 3. Several examples of binding site conservation. Ribbon representations of different structural superpositions, with their corresponding binding sites showing the bound heteroatoms and side chains directly contacting the heteroatom. The sequence alignments corresponding to the structurally aligned binding sites residues are given below each structure. The reference structure is colored in black and the corresponding aligned structure in light gray. **a:** Conservation of the binding site in a case of low sequence similarity. The representative structure is compared with 1ryt, including the heme heteroatom of 1bcfA and the two ions bound to each protein. In the alignment, the heme binding residues are lower-cased. **b:** Conservation of the binding site of two related proteins. The selected pair is composed by 1amk and 7timA. The 2-phosphoglycolic acid heteroatom bound by 1amk is represented. **c:** No conservation of the binding site of two related proteins. The structure of 3lzt is compared with 1hfyA; in this case, four N-acetyl-D-glucosamine-bound heteroatoms are represented. The binding site residues and heteroatoms were derived from the related structure, 1lsz, 99% identical, because 3lzt does not contain any of the relevant heteroatoms.

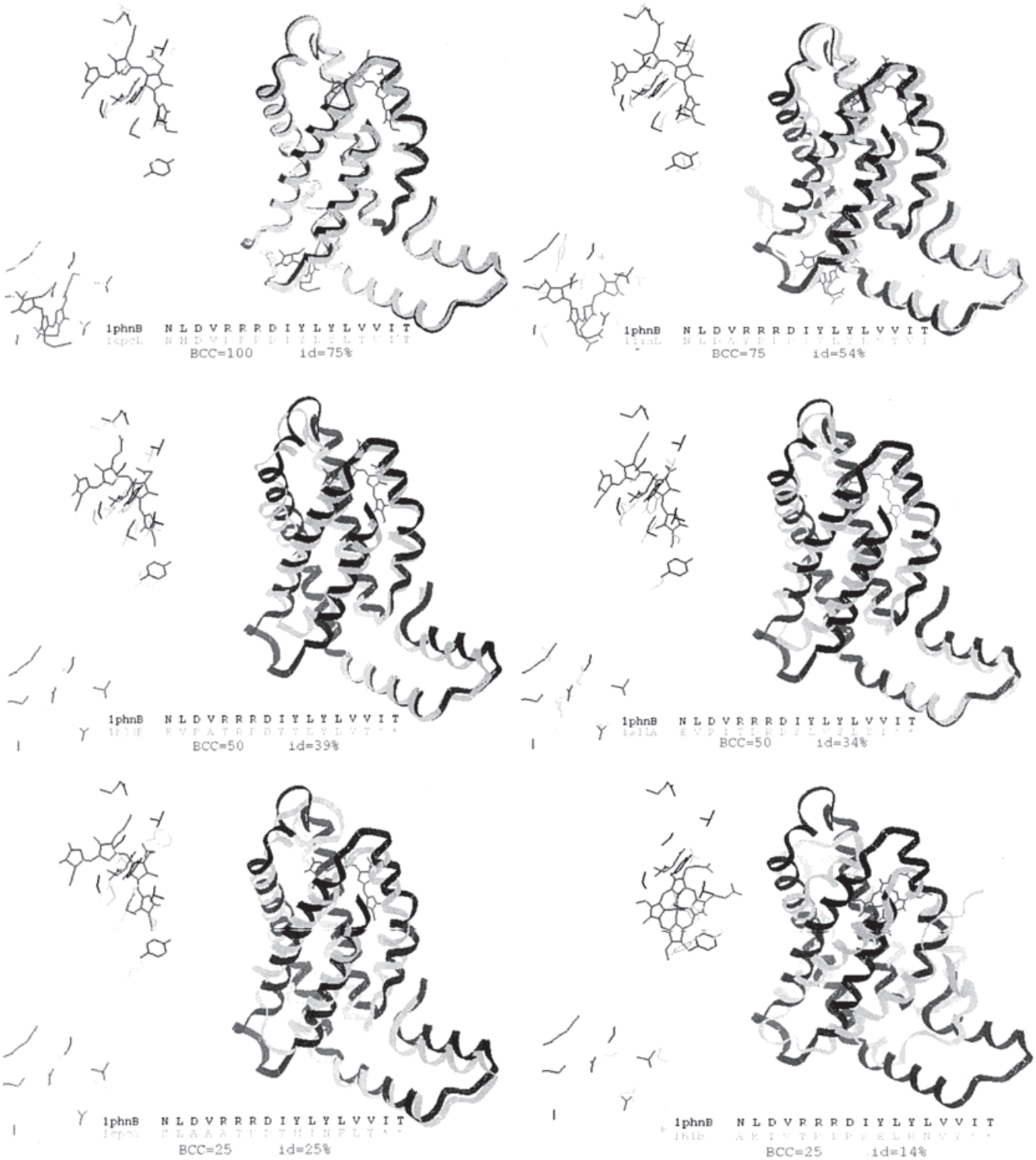

Fig. 4. Divergence of the binding site in different proteins aligned with phycocyanin, 1phnB. Ribbon representation of the structural superposition, together with the corresponding binding site structures and alignments (lower binding site in lowercase). The binding site comparison class (BCC) and the identity percentage are given for each case. The 1phnB structure and binding site are represented by black traces; the aligned proteins are in light gray. Only the heteroatoms of the aligned structure are represented, whereas the corresponding 1phnB heteroatoms are omitted for clarity, because they are in a position identical to that occupied in 1cpcL. For clarity, only the side chains contacting heteroatoms are represented. The proteins compared are phycocyanin (1cpcL), phycoerythrin (1liaL), allophycocyanins (1b33B and 1allA), phycocyanin (1cpcA), sea cucumber hemoglobin (1hlb).

TABLE I. Sequence, Structural, and Binding Site Divergence in Proteins Aligned With 1phnB

Reference protein	Aligned protein	Binding site match				Class		Identity (%)
		Upper (24)[a]	Lower (12)[b]	Total (36)	Total (%)	BCC[c]	SCC[d]	
1phnB	1cpcL	23	10	33	92	100	6	75
	1liaL	19	6	25	70	75	6	54
	1b33B	16	0	16	44	50	6	39
	1allA	11	0	11	31	50	6	34
	1cpcA	5	1	6	16	25	6	25
	1hlb	4	1	5	14	25	4	14

[a]Upper and [b]lower binding sites correspond to the two different binding sites at the opposite poles of the protein as represented in Figure 4.
[c]Binding site comparison class.
[d]Structural comparison class.

TABLE II. Examples of Global Conservation[†]

Structures	Identity (%)	SCC[a]	ECC[b]	KCC[c]	Fct class[d]	BCC[e]
1amk/7timA	42	6	4	100	1	100
3lzt/1hfyA	41	6	0	0	0	50
1pamA/1a47	69	6	? 4 or 0 ?	0	1	100

[†]For each structure pair, identity percentage and structural or functional conservation classes are indicated.
[a]Structural comparison class.
[b]Enzyme classification comparison class.
[c]Keywords comparison class.
[d]Functional classes comparison (1 = same, 0 = different).
[e]Binding site comparison class.

the same enzymatic function (EC code 5.3.1.1., ECC = 4), and identical keywords in their respective SWISS-PROT files (TPIS_LEIME and TPIS_YEAST, with keywords "glycolysis," "isomerase," "pentose shunt," "fatty acid biosynthesis," and "gluconeogenesis," KCC = 100), both of them were classified in the "Energy metabolism" class, and their binding site residues were also conserved (BCC = 100). A counterexample of low functional conservation at a similar level of sequence similarity could be the one of 3lzt, a chicken lysozyme, and 1hfyA, a goat alpha-lactalbumin (41% identity) (Fig. 3c). Both proteins have the same overall fold (SCC = 6), but they were classified in completely different EC classes (EC codes 3.2.1.17 and 2.4.1.17, ECC = 0). Their keywords are completely different (KCC = 0): LYC_CHICK is annotated with keywords "hydrolase," "glycosidase," and "bacteriolytic enzyme," whereas LCA-_CAPHI is described with the poorly informative keywords "milk," "lactose," and "glycoprotein." Correspondingly, their functional classes were also different, with the first sequence assigned to "Central intermediary metabolism" and the second to "Transport and binding proteins." Reflecting their possible common origin,[38] 48% of the binding site residues in equivalent structural positions are identical (BCC = 50). The functional transfer would have been misleading in this case.

One final example can be used to illustrate the difficulties due to incomplete database annotations. The structure of 1pamA, a *Bacillus sp*. cyclodextrin glucanotransferase, was aligned to 1a47, a *Thermoanaerobacterium thermosulfurigenes* cyclodextrin glycosylatransferase, with 69% identity. They share the same fold (SCC = 6) and the same binding site (BCC = 100). Both PDB sequences have the same EC code (ECC = 4, for EC code 2.4.1.19, cyclomaltodextrin glucanotransferase) and were classified in the "Central intermediary metabolism" functional class. Up to this point the information for the functional transfer appears clear, but an annotation discrepancy complicates the situation. The SWISS-PROT entry associated with 1a47, (AMY_THETU, complete sequence identity) is described as an alpha-amylase with different EC code (3.2.1.1, ECC = 0) and keywords (KCC = 0; CDGT_BACS0; "transferase" and "glycosyltransferase" and AMY_THETU; "hydrolase," "glycosidase." and "carbohydrate metabolism"). Therefore, the difference in annotation between PDB and SWISS-PROT for the same sequence would have led to different function predictions.

DISCUSSION

In this study, we address the crucial problem of function prediction. Despite the widespread use of database searching techniques followed by function inference as standard procedures in Bioinformatics, the results presented here illustrate that the transfer of function between similar sequences involves more difficulties than commonly believed. Our data show that even true pairwise sequence relations, identified by their structural similarity, correspond in many cases to different functions.

Calibrating the Sequence-to-Function Relationship

Structural similarity is conserved even at very low levels of sequence similarity.[3,4] We have explored the extent to which this relation can be extrapolated to protein function. Protein function is far less well defined than protein structure, because not only different experimental tech-

niques bring partial aspects of function but functional information is also captured only partially in different databases. These concurring factors make the systematic transference of protein function a difficult exercise. We have focused here on four partial aspects of protein function: (i) enzymatic function defined by the EC number, (ii) function description as illustrated in SWISS-PROT functional keywords, (iii) classes of cellular function as deduced from the database annotations, and (iv) amino acid composition of the binding site. Each of these definitions has a limited scope, but they are commonly used in molecular biology and genome analysis and can also serve as an example of the general limitations of function prediction based on sequence similarity. The EC schema has been used previously to assess functional conservation, at the sequence[9,10] or structure[7,8] level. The small degree of conservation of the EC classes between similar structures that we observed is in good agreement with these previous studies, but we have applied it in a more general analysis of the conservation of protein function.

The shape of the obtained curves mirror the ones obtained in a recent and complementary study[39] using different definitions of function and EC codes. A previous study by the same group[6] also complement nicely our results. Those studies, together with those of other groups,[7,8] contribute to the current view of the relationship between protein sequence-structure and function, as a complex evolutionary phenomenon that does not allow simplistic interpretations.

Limits of Function Prediction

Binding site, keywords, and functional class annotations were less conserved than EC numbers, and all of them in turn were less conserved than protein structure. Their different degrees of conservation can be described by comparing the percentage of identity at which the mean conservation of a given parameter (structural or functional) is 50%. This Identity level required for a Conservation of 50% (IC50) of the functional characteristics threshold was 7.5% for the structural classes, 15% for the EC codes, 20% for the keywords and functional classes, and 30% for the conservation of the type of amino acid in the binding sites. The observed margins of functional transfer may thus be useful in the future as reference points associated with function prediction exercises.

Practical Consequences for Genome Annotation

The new search methods based on family information, for example, PSI-BLAST[12] or Hidden Markov Models,[16] have produced a considerable improvement in the identification of distant sequence relationships. They are now able to delve into the twilight zone of sequence identity.[40] These improvements do not, however, imply per se a solution to the problem of predicting function, because many of the distant sequence relationships can only be translated partially into functional similarity (see for example Bork et al.[41]). Indeed, the analysis presented here shows that there are serious difficulties in transferring protein function based on sequence similarity, even if the analysis is based on structural alignments that provide better evidence of a relationship than any sequence database search.

In many cases, the problems highlighted here are related to erroneous or scarce functional annotations deposited in different databases. These problems appeared clearly during the analysis of the EC annotations that were available for only a fraction of the PDB structures, and for the analysis of SWISS-PROT keywords, which in some cases were insufficient for adequate description of protein function. It is important to stress that without further experimental or bibliographic investigation, it is impossible to differentiate between the cases in which functional transfer is erroneous because of inaccuracies in the underlying annotations from cases in which function is actually different.

Going one step beyond sequence searching and errors in databases, the analysis of complete protein families to identify sequences carrying equivalent functions might increase the reliability of the function transfer exercises.[42] The identification of orthologous sequences (sequences directly related by evolution with equivalent functions in different organisms) requires detailed phylogenetic analysis, best performed by human experts. Still this analysis can only be performed properly for protein families with few duplication events and only if they are represented in complete genomes for which all the information is available. Remarkably, even though this type of expert analysis should lead to better functional annotations, it has been observed repeatedly that there are no large differences from the results derived by simple analysis tools. This may be surprising because the automatic systems[43–45] are based on the direct transfer of function from similar sequences and do not include complex family analysis. The differences between the Genequiz annotations[43] and those provided by the best expert groups can be estimated at around 10% (8% in Brenner;[46] at least 4.5%—21 errors of 285 annotations Tables 2–4 in Galperin and Koonin;[47] and 4% in Ouzounis et al.[48]). Therefore, it seems that for small genomes, family analysis does not add crucial information, perhaps because there are a limited number of alternatives for the function of each protein, which in general tends to perform the same function as similar sequences in closely related genomes. The uncertainties that we found here in function prediction for large databases suggest that similar problems will arise in the analysis of complex genomes, e.g., the human genome.

Possible Theoretical Implications

In addition to the practical implications for function annotation, the results presented here can be interpreted in terms of protein evolution. Our analyses indicate that the chemical function of a protein related to its catalytic activity tends to be conserved (EC code), whereas the biochemical and cellular functions that are related to interactions and dynamic processes change more rapidly (keywords and functional classes). Finally, the specific composition of the cofactor binding sites, which in many cases is directly related to functional specificity, evolves even faster.

The emerging picture of the relationship between sequence, structure, and functional space shows that a large portion of sequence space is covered by a much smaller number of protein folds. At the same time, the structure of the functional space seems more complex. It includes regions in which many related sequences correspond to a single function, regions in which small changes in sequence correspond to important functional differences, and regions in which even unrelated sequences converged to the same function. This reveals that much remains to be done before we have a comprehensive picture of the relation between sequences and functions.

ACKNOWLEDGMENTS

We thank Christos Ouzounis (EBI-EMBL, Hinxton, UK), Burkhard Rost (Columbia University, NY), and Chris Sander (Whitehead Inst., Cambridge, MA) for interesting discussions about the role of conserved residues and the extrapolation of function, Nigel P. Brown and Catherine Mark for surveying the English version, and the members of Protein Design Group, CNB-CSIC for continuous support and stimulating discussions.

REFERENCES

1. Blattner FR, Plunkett III G, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K–12. Science 1997;277:1453–1462.
2. Winzeler EA, Davis RW. Functional analysis of the yeast genome. Curr Opin Genet Dev 1997;7:771–776.
3. Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
4. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
5. Karp PD. What we do not know about sequence analysis and sequence databases. Bioinformatics 1998;14:753–754.
6. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999;288:147–164.
7. Orengo CA, Todd AE, Thornton JM. From protein structure to function. J Mol Biol 1999;9:374–382.
8. Thornton JM, Orengo CA, Todd AE, Pearl FMG. Protein folds, functions and evolution. J Mol Biol 1999;293:333–342.
9. desJardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA. Prediction of enzyme classification from protein sequence without the use of sequence similarity. ISMB 1997;5:92–99.
10. Shah I, Hunterm L. Predicting enzyme function from sequence: a systematic appraisal. ISMB 1997;5:276–283.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
12. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;24:3389–3402.
13. Rost B, Valencia A. Pitfalls of protein sequence analysis. Curr Opin Biotechnol 1996;7:457–461.
14. Bork P, Koonin EV. Predicting functions from protein sequences—where are the bottlenecks? Nat Genet 1998;18:313–318.
15. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. J Mol Biol 1998;283:707–725.
16. Karplus K, Barrett C, Hughey R. Hidden Markov Models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
17. Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. Nat Genet 1994;6:119–129.
18. Smith TF, Zhang X. The challenges of genome sequence annotation or "The devil is in the details." Nat Biotechnol 1997;15:1222–1223.
19. Kyrpides NC, Ouzounis CA. Errors in genome reviews. Science 1998;281:1457.
20. Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: "going wrong with confidence." Mol Microbiol 1999;32:886–887.
21. Pallen M, Wren B, Parkhill J. "Going wrong with confidence": misleading sequence analyses of CiaB anc clpX. Mol Microbiol 1999;34:195.
22. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:349–354.
23. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
24. Murzin A, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
25. Laskowski RA, Wallace A. http://www.biochem.ucl.ac.uk/bsm/enzymes/.
26. Bairoch A. The ENZYME data bank. Nucleic Acids Res 1993;21:3155–3156.
27. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. Nucleic Acids Res 1997;25:31–36.
28. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 1997;22:488–490.
29. Tamames J, Casari G, Ouzounis C, Sander C, Valencia A. EUCLID: automatic classification of proteins in functional classes by their database annotations. Bioinformatics 1998;14:542–543.
30. Riley M. Functions of the gene products of *Escherichia coli*. Microbiol Rev 1993;57:862–952.
31. Ouzounis C, Perez-Irratxeta C, Sander C, Valencia A. Are binding residues conserved? Pac Symp Biocomput 1998;3:399–410.
32. Villar HO, Kauvar LM. Amino acid preferences at protein binding sites. FEBS Lett 1994;349:125–130.
33. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 1995;8:127–134.
34. Iversen LF, Kastrup JS, Bjorn SE, et al. Structure of Hbp, a multifunctional protein with a serine proteinase fold. Nat Struct Biol 1997;4:265–268.
35. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.
36. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
37. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.
38. McKenzie HA, White FH. Lysozyme and α-lactalbumin: structure, function, and interrelationships. J Adv Protein Chem 1991;41:173–315.
39. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
40. Doolittle RF. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. Mill Valley California: University Science Books; 1986. 103 p.
41. Bork P, Sander C, Valencia A. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin and hsp70 heat shock-proteins. Proc Natl Acad Sci USA 1992;89:7290–7294.
42. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.
43. Andrade MA, Brown NP, Leroy C, et al. Automated genome sequence analysis and annotation. Bioinformatics 1999;15:391–412.
44. Frishman D, Mewes HW. PEDANTic genome analysis. Trends Genet 1997;13:145–146.
45. Gaasterland T, Sensen CW. MAGPIE: automated genome interpretation. Trends Genet 1996;12:76–78.
46. Brenner SE. Errors in genome annotation. Trends Genet 1999;15:132–133.
47. Galperin MY, Koonin EV. Source of systematic error in functional annotation of genomes: domain rearrangement, non-hortologous gene displacement, and operon disruption. In Silico Biol 1998;1:0007.
48. Ouzounis C, Casari G, Valencia A, Sander C. Novelties from the complete genome of *Mycoplasma genitalium*. Mol Microbiol 1996;20:895–900.