doi: 10.1016/S0022-2836(02)00016-5 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 318, 595–608

JMB

Enzyme Function Less Conserved than Anticipated

Burkhard Rost[1,2]*

[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York NY 10032, USA*

[2]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York NY 10032, USA*

**Corresponding author*

The level of sequence similarity that implies similarity in protein structure is well established. Recently, many groups proposed thresholds for similarity in sequence implying similarity in enzymatic function. All previous results suggest the strong conservation of enzymatic function above levels of 50% pairwise sequence identity. Here, I argue that all groups substantially overestimated the conservation of enzyme function because their data sets were either too biased, or too small. An unbiased analysis suggested that less than 30% of the pair fragments above 50% sequence identity have entirely identical EC numbers. Another surprising finding was that even BLAST *E*-values below 10^{-50} did not suffice to automatically transfer enzyme function without errors. As expected, most misclassifications originated from similarities in relatively short regions and/or from transferring annotations for different domains. Both problems cannot be corrected easily by adjusting the thresholds for automatic transfer of genome annotations. A score relating sequence identity to alignment length (distance from HSSP-threshold) outperformed statistical BLAST scores for high sequence similarity. In particular, the distance score allowed error-free transfer of enzyme function for the 10% most similar enzyme pairs. The results illustrated how difficult it is to assess the conservation of protein function and to guarantee error-free genome annotations, in general: sets with millions of pair comparisons might not suffice to arrive at statistically significant conclusions. In practice, the revised detailed estimates for the sequence conservation of enzyme function may provide important benchmarks for everyday sequence analysis and for more cautious automatic genome annotations.

© 2002 Elsevier Science Ltd. All rights reserved

Keywords: genome annotation; conservation of protein function; enzyme classification; evolution, statistical significance; bootstrap, bioinformatics

Introduction

Missing analyses of annotation accuracy

The explosion of known sequences through large-scale sequencing projects unravelled the strength and weakness of today's bioinformatics. Raw sequences are of limited use: what we need are annotations. The most common way to mine entire proteomes is to transfer annotations from homologues.[1–14] The strength of bioinformatics lies in powerful search algorithms, such as the BLAST best-sellers[15] and PSI-BLAST[16] that are at the base for homology transfers. The weakness lies in the lack of large-scale evaluations of how sequence similarity translates to functional similarity. When experts make sense of database searches, they can easily separate the chaff from the wheat. Machines are less potent. However, large-scale annotation is based on algorithms, i.e. on automatically analysing putative similarities. Thus, we have many annotations and only vague ideas for when such annotations are adequate. Consequently, many annotations may be wrong. In fact, 10–30% of all genome annotations may be wrong.[17,18]

Abbreviations used: BLAST, fast sequence alignment method[15]; EC, Enzyme Commission number describing enzymatic function at increasing level of detail (there are four numbers, the first number distinguishes oxireductases, transferases, hydrolases, lyases, isomerases, and ligases); HSSP, data base of protein structure–sequence alignments[27]; HSSP threshold, threshold relating percentage of pairwise identical or similar residues to length of the alignment (equation (2))[19,24]; PSI-BLAST, position specific iterated database search[16]; SWISS-PROT, data base of protein sequences[46].

E-mail address of the corresponding author: rost@columbia.edu

0022-2836/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved

Relation between sequence and structural similarity is well established

The similarity between two protein structures is easy to measure. Many groups have explored how similarity in sequence translates to similarity in structure.[19,20–26] All groups agree that statistical scores are better indicators of structural similarity than are levels of pairwise sequence identity. Two of the three groups that evaluated raw alignment scores agreed that these are even better than BLAST *E*-values.[19,25] Surprisingly, very few groups implemented the relation between pairwise sequence identity and alignment length introduced by Sander & Schneider in their HSSP-threshold[27] to reflect the observation that levels of sequence identity do not mean the same for short and for long alignments.[24,28] In particular, two proteins of 50 residues with 25 of these identical may have different structures,[29] while we have no example of two different structures for proteins that have 50 in 100 residues identical.[19] In fact, an updated version of the original HSSP-threshold appears, at least, on par with raw alignment scores.[19] In general, transfer of structure is almost guaranteed to be correct above the twilight zone,[30] e.g. all known pairs of proteins with more than 33 residues in 100 identical have similar structure.[19] The transition into the twilight zone is characterised by an explosion of pairs, most of which have different structures.[19] Because profile-based comparisons are more sensitive than pairwise comparisons, profiles enable safe comparisons in the twilight zone.[26,31,32] However, most protein pairs with similar structure populate the midnight zone of random levels of sequence similarity.[23,33] Only threading methods reach into that region, sometimes.[34–41]

Sequence conservation of protein function studied extensively

Can we generalise the threshold for sequence similarity implying conservation of structure to function? The first large-scale study of functional conservation was based on the enzyme commission (EC) numbers.[42] The authors concluded that many enzyme classes are conserved in sequence and some are not. They published accuracy (enzymes with similar EC number found/all enzymes found) *versus* coverage (enzymes with similar EC number found/all enzymes with similar EC number) curves similar to those found for the conservation of structure. More recently, four groups based their analysis of functional conservation on proteins of known structure.[11,43–45] Devos & Valencia[43] established the levels of pairwise sequence identity indicative of conserved EC numbers, active sites, and keywords found in SWISS-PROT[46] and PDB.[47] They reported levels of pairwise identity taken from structural alignments. Two groups investigated the conservation of enzyme function between proteins of similar structure in detail.[44,45] In particular, the outstanding in-depth analysis by Todd *et al.* reviews the structural characteristics of enzymatic function;[44] sequence conservation is described by pairwise sequence identity obtained from PSI-BLAST alignments. Wilson *et al.* compare the performance of pairwise sequence identity to that of raw alignment scores and statistical scores derived from these raw alignment scores. They use dynamic programming alignments.[45] Pawlowski *et al.* compare EC numbers between proteins from *Escherichia coli* using their own profile-based dynamic programming algorithm;[11] they investigate pairwise sequence identity and statistical scores derived from the raw alignment scores. Despite the variety of methods and scores explored, all groups agreed that variations of function largely occurred below levels of 50% pairwise sequence identity, and that EC classifications are almost as conserved as structure. The two groups exploring statistical scores agreed that these are superior to thresholds in pairwise sequence identity.[11,45] None of the groups explored the performance of the HSSP-threshold that successfully distinguishes between proteins of similar and non-similar structure.[19,52]

Databases contain misleading bias

Seemingly, all groups investigating the sequence conservation of enzyme function addressed the question: given an alignment between protein A and B, and given the experimental annotation for the enzyme A, what is the probability that B has the same enzymatic activity as A? However, all groups used particular subsets of proteins to capture the universe of all proteins. Three groups based their selection on subsets of proteins with similar structures;[43–45] one group disregarded structural information, but restricted their analysis to *E. coli* proteins.[11] Hence, all four groups use constraints that may not be entirely representative for everyday sequence searches. Only one group used all proteins in an old version of SWISS-PROT.[42] A principal problem of any database analysis is that of bias related to the experimental focus of the groups that deposit their information. Assume we have 12 enzymes of experimentally known function from two families: {A: A1, . . ., A10} and {B: B1, B2}. When we compare all ten proteins in {A} and {B} with each other, we count 66 pair comparisons ($N(N-1)/2 = 12(11)/2$). Assume that the pairs within each family ($A1 = A2 = \cdots = A10$, $B1 = B2$) have similar function, and all across the families have not ($A1 \neq B1$). Now assume that we have a very naïve method that predicts all pairs to be similar in function ($A1 = A2 = \cdots = B1 = B2$). Obviously, a bad prediction method, nevertheless, we estimate its accuracy to be 70% (45 correct in {A}, one correct in {B} $\Rightarrow 46/66$). If we reduce the bias, we compare one representative of each family against all others. This gives 2×10 pairs. However, based on the unbiased set, we estimate an accuracy of $10/20 = 50\%$. Obviously, the unbiased estimate

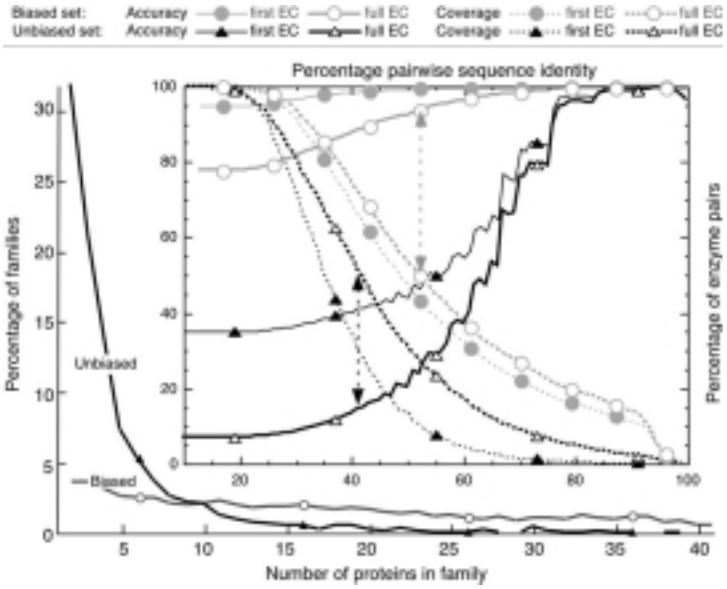

Figure 1. Biased *versus* unbiased results. The outer graph shows the distribution of family sizes. Families were defined by PSI-BLAST searches[16] including all similarities above an HSSP distance of 0 (equation (2)). The black line (filled triangles) gives the percentages of all families for the unbiased data set (1973 enzyme families), the grey line (open circles) marks the percentages for the biased set (26,342 enzyme families). The simplest way to rationalise the necessity of unbiased analyses for those who distrust statistics is that the black line mirrors the results of most entirely sequenced organisms.[51] The inset gives the percentage of enzyme pairs found at a given level of pairwise sequence identity. Enzyme pairs with fully identical EC numbers were considered as similar, others as non-similar. The grey lines mark the biased sets, the black ones the unbiased sets. Broken lines describe the coverage, i.e. how many of the pairs with identical EC numbers are found; continuous lines the accuracy, i.e. how many of the pairs found have identical EC numbers. The arrows point out the significant difference between the estimates of accuracy based on both sets: When finding 50% of all true pairs, the biased set estimates the accuracy above 90%, while the unbiased set corrects this estimate to below 15%.

(50%) reflects our intuitive judgement better than the biased one (70%). Currently, it is widely assumed that certain protein folds are more energetically favourable than others,[48–50] and that these folds are over-represented in the universe of protein structures rather than only in today's databases. If so, we may argue that we obtain the most representative estimates by accepting the bias. In order for this argument to be valid, we need to put up an additional assumption: the particular bias that we find in today's databases is representative in detail for the bias in the protein universe. We have no strong reasons to assume that today's bias is representative. Quite the contrary: for any simple feature of proteins, we see differences between entire proteomes and SWISS-PROT/PDB, in particular, in the composition of enzymes.[51] Another difference becomes obvious in the redundancy degree contained in various data sets. For instance, the biased set of 26,342 enzymes corresponded to 1973 families, in other words, less than 8% of the enzymes in SWISS-PROT were unique. In contrast, we found that the fraction of unique proteins varied between 25 and 50% for 32 of the entirely sequenced proteomes.[52] Furthermore, enzymes were not statistically over-represented in the set of proteins common to yeast, worm and human (J. Liu & B.R., unpublished results). Thus, the bias we see today is not representative of the bias we will see after another ten years of large-scale sequencing. Thus, we best begin with what physicists refer to as the minimal, or null assumption; namely, a set without bias.

Here, I analysed different thresholds to distinguish between enzymes of similar and non-similar function in everyday database searches. The Enzyme Commission numbers (EC) provided the standard-of-truth to measure functional similarity. In particular, I compared BLAST with PSI-BLAST searches, and the performance of BLAST *E*-values with those of pairwise sequence identity and distances from the structure-derived HSSP-threshold (equation (2)). The results suggested that the conservation of enzyme function has previously been overestimated considerably. Consequently, the percentage of errors in annotations may have been underestimated.

Results

Biased and unbiased results differed substantially

The 26,342 enzymes of annotated function clustered into 1973 families (see Materials and Methods). Thus, the bias in the data set was more than 13-fold. The distribution of family sizes illustrated the bias in a simple way: while the unbiased count revealed that most known enzyme families have fewer than five members, the few large families dominated the biased count (Figure 1, outer graph). The biased data set yielded results very similar to those obtained by others:[42–45] the entire EC number was conserved well above levels of about 50% identity (Figure 1 inset, grey curve with +). The picture changed dramatically when

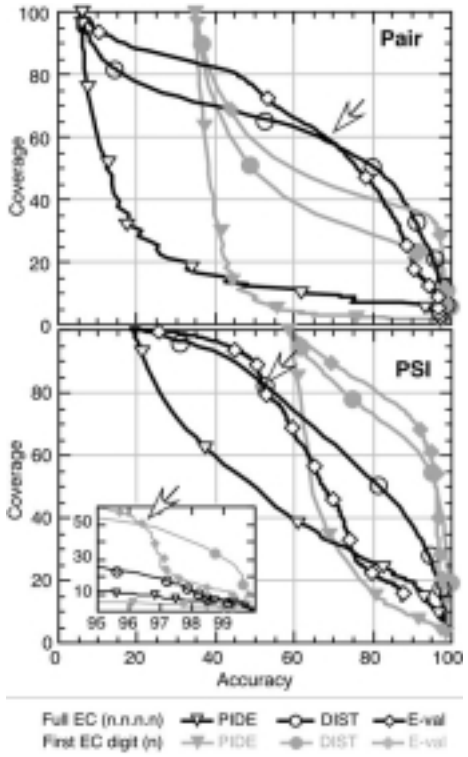

Figure 2. Accuracy *versus* coverage for pairwise BLAST and PSI-BLAST. For any given threshold applied to distinguish between similar and non-similar enzymes, there is a trade-off between accuracy (similar proteins found/all proteins found) and coverage (similar proteins found/all similar proteins): a particular threshold chosen corresponds to a particular coverage. The upper graph maps out the dependency of coverage on accuracy for all possible thresholds in pairwise BLAST searches, the lower graph gives the same data for PSI-BLAST searches. The better a threshold, the more the respective curves approach the upper right corner. The graphs separate between two ways of considering two enzymes to be of similar function. (1) Both have the same first EC digit, e.g. are isomerases (EC 5.*). (2) Both have identical EC numbers, e.g. are Mandelate racemases (5.1.2.2). The following scores are given: (i) *PIDE*, pairwise sequence identity, (ii) *DIST*, distance from HSSP-threshold (equation (2)), and (iii) *E-val*, BLAST and PSI-BLAST *E*-values. Pairwise sequence identity was obviously inferior to the other scores. *DIST* performed best for high levels of accuracy. Arrows mark the points at which the statistical scores start outperforming the HSSP-threshold. The transition corresponds to an HSSP-distance around 0 (arrows in Figure 3).

using the unbiased data set: less than 30% of all the pairs found at 50% sequence identity had identical EC numbers (Figure 1 inset, black curve with dark circles). Another way to illustrate the significance of bias was to look up the level of accuracy at a cut-off for which 50% of all enzymes of identical function were found (coverage). The biased estimate suggested a level above 90% accuracy (Figure 1 inset, grey broken-line arrow), while the unbiased estimate suggested a level below 15% accuracy (Figure 1 inset, black broken-line arrow).

PSI-BLAST powerful, pairwise sequence identity bad

Plotting the accuracy *versus* coverage (equation (3)) revealed a number of expected results (Figure 2). Firstly, PSI-BLAST searches found more similar enzymes at most levels of accuracy (Figure 2, upper *versus* lower graphs). However, the full power of PSI-BLAST searches did not occur at high levels of accuracy: pairwise BLAST and PSI-BLAST found 60% of all pairs with identical EC numbers at levels above 75% accuracy. Secondly, pairwise sequence identity proved once again to be a bad measure for functional similarity (Figure 2, triangles below all other lines). In particular, if an automatic annotation method based on pairwise sequence identity ignored the length of an alignment, it would generate a large number of incorrect assignments at very high levels of pairwise sequence identity. However, even for alignments of reasonable length, there are many pairs of different enzymatic activity above levels of 50% pairwise sequence identity (Table 1). Obviously, most of these similarities are spurious, in that they result from comparing different domains or are based on rather short regions that are aligned. Another result was that the distance from the HSSP-threshold (equation (2)) performed better than BLAST *E*-values for high levels of accuracy (transition points labelled by arrows in Figure 2). Another indication for the power of PSI-BLAST was that it considerably improved the performance, even when using the naïve measure of pairwise sequence identity to infer functional similarity: 50% coverage was reached at 50% accuracy by PSI-BLAST and only at 15% accuracy by pairwise BLAST (Figure 2, open triangles).

BLAST *E*-values bad for high level of accuracy

Figure 2 did hide the possibly most surprising result: even very low *E*-values did not yield 100% accuracy in inferring all four EC numbers (Figure 3, rightmost graphs). For example, "only" 86% of the pairwise BLAST hits with $E < 10^{-50}$ had identical EC annotations; at the same value of $E < 10^{-50}$ only 65% of all hits found by PSI-BLAST had identical annotations. In contrast, distances of more than 30 from the HSSP-threshold yielded levels close to 100% accuracy. Furthermore, using the HSSP-threshold to transfer EC annotations, PSI-BLAST performed better than pairwise BLAST even for high levels of accuracy. The HSSP-thresholds became worse than the BLAST *E*-values at levels that corresponded to HSSP distances around 0 and to *E* values around 10^{-3} (arrows in Figure 3). Another surprise was that very low *E* values ($<10^{-50}$) of the pairwise BLAST searches distinguished more accurately than low

Table 1. Examples for enzymes of similar sequence and different function

	Protein 1		Protein 2	
	Sequence identity (%)	Alignment length	HSSP-distance	BLAST *E*-value
Pair 1	94	29	15	2.3×10^{-6}
Name	chib_poptr (EC 3.2.1.14) [303] Acidic endochitinase WIN6.2b precursor		tp2b_chick (EC 5.99.1.3) [1627] DNA topoisomerase II, β-isozyme.	
Des	This protein functions as a defense against chitin-containing fungal pathogens		Control of topological states of DNA by transient breakage and subsequent rejoining of DNA strands. Topoisomerase II makes double-strand breaks	
Act	Hydrolysis of the 1,4-β-linkages of *N*-acetyl-D-glucosamine polymers of chitin		ATP-dependent breakage, passage and rejoining of double-stranded DNA	
Pair 2	88	21	6	2.6
Name	cyaa_neucr (EC 4.6.1.1) [2300] Adenylate cyclase		p3k3_dicdi (EC 2.7.1.137) [1585] Phosphatidylinositol 3-kinase 3	
Des	Essential in regulation of cellular metabolism by catalysing the synthesis of a second messenger cAMP			
Act	ATP ⇒ 3′,5′-cyclic AMP + pyrophosphate		ATP + 1-phosphatidyl-1d-myo-inositol ⇒ ADP + 1-phosphatidyl-1d-myo-inositol 3-phosphate	
Pair 3	83	28	1	0.00032
Name	chib_poptr (EC 3.2.1.14) [303] Acidic endochitinase WIN6.2b precursor		kdge_drome (EC 2.7.1.107) [1454] Eye-specific diacylglycerol kinase	
Des	This protein functions as a defense against chitin-containing fungal pathogens		Required for the maintenance of the photoreceptor; its absence leads to rhabdomere degeneration due to defective phospholipid turnover	
Act	Hydrolysis of the 1,4-β-linkages of *N*-acetyl-D-glucosamine polymers of chitin		ATP + 1,2-diacylglycerol ⇒ ADP + 1,2-diacylglycerol 3-phosphate	
Pair 4	78	19	0	24
Name	guna_psefl (EC 3.2.1.4) [962] Endoglucanase A precursor (endo-1,4-β-glucanase)		mdhp_flabi (EC 1.1.1.82) [453] Malate dehydrogenase [NADP]	
Des			The chloroplastic NADP-dependent form is essential for the photosynthesis C4 cycle, allowing plants to circumvent the problem of photorespiration in C4 plants; NADP-MDH activity acts to convert oxaloacetate to malate in chloroplasts of mesophyll cells for transport to the bundle sheath cells	
Act	Endohydrolysis of 1,4-β-D-glucosidic linkages in cellulose		L-Malate + NADP(+) ⇒ oxaloacetate + NADPH	
Pair 5	76	29	9	3.7×10^{-5}
Name	dhax_yeast (EC 1.2.1.3) [533] Aldehyde dehydrogenase		ppck_canal (EC 4.1.1.49) [553] Phosphoenolpyruvate carboxykinase [ATP]	
Des				
Act	Aaldehyde + NAD(+) + H_2O ⇒ acid + NADH		ATP + oxaloacetate ⇒ ADP + phosphoenolpyruvate + CO_2	
Pair 6	75	22	−1	0.00056
Name	maai_pseae (EC 5.2.1.2) [212] Maleylacetoacetate isomerase		gth_braol (EC 2.5.1.18) [76] Glutathione *S*-transferase	
Des			Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles	
Act	4-Maleylacetoacetate ⇒ 4-fumarylacetoacetate		RX + glutathione ⇒ HX + R-S-glutathione	
Pair 7	71	21	−3	0.048
Name	rpb1_plafd (EC 2.7.7.6) [2452] DNA-directed RNA polymerase II largest subunit		ubc2_yeast (EC 6.3.2.19) [172] Ubiquitin-conjugating enzyme e2−20 kDa	
Des	DNA-dependent RNA polymerase, catalyses the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates		Catalyses the covalent attachment of ubiquitin to other proteins. ubc2 is active on histones; it is required for postreplication repair of UV-damaged DNA and sporulation. ubc2 mediates e3-dependent ubc activity	
Act	*N*-Nucleoside triphosphate = *N* pyrophosphate + RNA(*n*)		ATP + ubiquitin + protein lysine ⇒ AMP + pyrophosphate + protein *N*-ubiquityllysine	
Pair 8	69	26	3	0.0025
Name	dpog_human (EC 2.7.7.7) [1239] DNA polymerase gamma		cya1_drome (EC 4.6.1.1) [2248] CA2 + /calmodulin-responsive adenylate cyclase	

(*continued*)

Table 1 Continued

	Protein 1		Protein 2	
	Sequence identity (%)	Alignment length	HSSP-distance	BLAST *E*-value
Des	Involved in the replication of mitochondrial DNA		Membrane-bound, calmodulin-sensitive adenylyl cyclase; inactivation of this cyclase leads to a learning and memory defect. Regulation: activated by CA(2 +)/ calmodulin and G protein	
Act	*N*-Deoxynucleoside triphosphate ⇒ *N*-pyrophosphate + DNA(*n*)		ATP ⇒ 3′,5′-cyclic AMP + pyrophosphate enzyme	
Pair 9	66	27	2	5.3×10^{-5}
Name	ig1r_human (EC 2.7.1.112) [1367] Insulin-like growth factor I receptor precursor		ptpu_human (EC 3.1.3.48) [1430] Protein-tyrosine phosphatase u precursor	
Des	This receptor binds insulin-like growth factor I (IgF I) with a high affinity and IgF II with a lower affinity; it has a tyrosine-protein kinase activity		Regulation of processes involving cell contact and adhesion such as growth control, tumor invasion, and metastasis	
Act	ATP + a protein tyrosine ⇒ ADP + protein tyrosine phosphate		Protein-tyrosine phosphate + H_2O ⇒ protein tyrosine + orthophosphate	
Pair 10	65	22	10	1×10^{-51}
Name	maai_pseae (EC 5.2.1.2) [212] Maleylacetoacetate isomerase (maai)		gstz_eupes (EC 2.5.1.18) [225] Glutathione *S*-transferase Zeta class	
Des	Pathway: catabolism of tyrosine; fourth step, catabolism of phenylalanine; fifth step			
Act	4-Maleylacetoacetate ⇒ 4-fumarylacetoacetate		RX + glutathione ⇒ HX + R-S-glutathione	
Pair 11	60	40	0	0.33
Name	syi_haein (EC 6.1.1.5) [941] Isoleucyl-tRNA synthetase (isoleucine–tRNA ligase)		pena_burce (EC 3.5.2.6) [313] β-Lactamase precursor (penicillinase)	
Des			Enables the organism to utilise penicillin as a carbon source	
Act	ATP + L-isoleucine + tRNA(Ile) ⇒ AMP + pyrophosphate + L-isoleucyl-tRNA(Ile)		A β-lactam + H2 ⇒ a substituted β-amino acid	
Pair 12	61	30	−4	0.0023
Name	fas2_yeast (EC 2.3.1.86) [1894] Fatty acid synthase (subunit alpha)		cyg4_human (EC 4.6.1.2) [732] Guanylate cyclase soluble, alpha-2 chain (gcs-alpha-2)	
Des	Catalyses the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADP-H. The α subunit contains domains for: acyl carrier protein, 3-oxoacyl-[acyl-carrier-protein] reductase, and 3-oxoacyl-[acyl-carrier-protein] synthase. This subunit coordinates the binding of the six β subunits to the enzyme complex		Has guanylyl cyclase on binding to the β-1 subunit	
Act	Acetyl-CoA + *n* malonyl-CoA + 2*n* NADPH ⇒ long-chain fatty acid + (*n* + 1)CoA + nCO_2 + 2 *n*NADP(+)		GTP ⇒ 3,′5′-cyclic GMP + pyrophosphate	
Pair 13	60	30	0	0.18
Name	plb1_yeast (EC 3.1.1.5) [664] Lysophospholipase 1 precursor (phospholipase B1)		metb_arath (EC 4.2.99.9) [563] Cystathionine γ-synthase	
Des	Catalyses the release of fatty acids from lysophospholipids			
Act	2-Lysophosphatidylcholine + H_2 ⇒ glycerophosphocholine + a fatty acid anion		*O*-Succinyl-l-homoserine + l-cysteine ⇒ cystathionine + succinate. Cofactor: pyridoxal phosphate	
Pair 14	59	27	−4	0.015
Name	odo2_fugru (EC 2.3.1.61) [409] Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex		p2bb_human (EC 3.1.3.16) [524] Serine/threonine protein phosphatase 2B catalytic subunit, β isoform	
Des	The 2-oxoglutarate dehydrogenase complex catalyses the overall conversion of 2-oxoglutarate to succinyl-CoA + CO_2		Calcium-dependent, calmodulin-stimulated protein phosphatase	
Act	Succinyl-CoA + dihydrolipoamide ⇒ CoA + *S*-succinyldihydrolipoamide		A phosphoprotein + H_2O ⇒ a protein + orthophosphate (this enzyme is serine/ threonine specific)	
Pair 15	58	29	−3	2.6×10^{-12}
Name	frk_human (EC 2.7.1.112) [505] Tyrosine-protein kinase FRK (nuclear tyrosine protein kinase RAK)		hser_cavpo (EC 4.6.1.2) [1076] Heat-stable enterotoxin receptor precursor (GC-c) (intestinal guanylate cyclase)	

(*continued*)

Table 1 Continued

	Protein 1		Protein 2	
	Sequence identity (%)	Alignment length	HSSP-distance	BLAST *E*-value
Des			Receptor for the *E. coli* heat-stable enterotoxin markedly stimulates the accumulation of cGMP in mammalian cells expressing GC-c; also activated by the endogenous peptide guanylin	
Act	ATP + a protein tyrosine ⇒ ADP + protein-tyrosine phosphate		GTP ⇒ 3,′5′-cyclic GMP + pyrophosphate	
Pair 16	58	36	4	5.7×10^{-32}
Name	gsa_profr (EC 5.4.3.8) [441] Glutamate-1-semialdehyde 2,1-aminomutase		gabt_ecoli (EC 2.6.1.19) [426] 4-Aminobutyrate aminotransferase	
Des	Catalytic activity: (s)-4-amino-5-oxopentanoate ⇒ 5-aminolevulinate		Catalytic activity: 4-aminobutanoate + 2-oxoglutarate ⇒ succinate semialdehyde + l-glutamate	
Act				
Pair 17	57	122	28	1.6×10^{-96}
Name	exoa_bacsu (EC 3.1.11.2) [252] Exodeoxyribonuclease		ape1_rat (EC 4.2.99.18) [316] DNA-(apurinic or apyrimidinic site) lyase.	
Des			Repairs oxidative DNA damages *in vitro*; may have a role in protection against cell lethality and suppression of mutations; removes the blocking groups from the 3′ termini of the DNA strand breaks generated by ionising radiations and bleomycin	
Act	Degradation of double-stranded DNA. It acts progressively in a 3′- to 5′-direction, releasing 5′-phosphomononucleotides		Endonucleolytic cleavage near apurinic or apyrimidinic sites to products with 5′-phosphate	
Pair 18	57	35	2	0.0025
Name	rpb1_plafd (EC 2.7.7.6) [2452] DNA-directed RNA polymerase II largest subunit		cn3b_rat (EC 3.1.4.17) [1108] cGMP-inhibited 3′,5′-cyclic phosphodiesterase b	
Des	DNA-dependent RNA polymerase, catalyses the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates		May play a role in fat metabolism	
Act	*N*-Nucleoside triphosphate ⇒ *N*-pyrophosphate + RNA(*n*)		Guanosine 3′,5′-cyclic phosphate + H_2O ⇒ guanosine 5′-phosphate	
Pair 19	57	28	−5	0.33
Name	gtfb_strmu (EC 2.4.1.5) [1476] Glucosyltransferase-I precursor (gtf-i)		am3b_orysa (EC 3.2.1.1) [438] Alpha-amylase isozyme 3B precursor	
Des	Production of extracellular glucans, that are thought to play a key role in the development of the dental plaque because of their ability to adhere to smooth surfaces and mediate the aggregation of bacterial cells and food debris		Important for breakdown of endosperm starch during germination	
Act	Sucrose + (1,6-alpha-D-glucosyl) (*n*) ⇒ D-fructose + (1,6-alpha-D-glucosyl) (*n* + 1)		Endohydrolysis of 1,4-alpha-glucosidic linkages in oligosaccharides and polysaccharides	
Pair 20	57	28	−5	0.062
Name	lys2_yeast (EC 1.2.1.31) [1392] Aminoadipate-semialdehyde dehydrogenase large subunit (alpha-aminoadipate reductase)		lcf2_yeast (EC 6.2.1.3) [744] Long-chain-fatty-acid-CoA ligase 2	
Des	Catalyses the activation of alpha-aminoadipate by ATP-dependent adenylation and the reduction of activated alpha-aminoadipate by NADPH		Esterification, concomitant with transport, of endogenous long-chain fatty acids into metabolically active CoA thioesters for subsequent degradation or incorporation into phospholipids. Preferentially acts on C9:0–C13:0 fatty acids although C7:0–C17:0 fatty acids are tolerated. The optimum activity is found at 25 degrees Celsius	
Act	L-2-Aminoadipate 6-semialdehyde + NADP(+) + H_2) ⇒ l-2-aminoadipate + NADPH		ATP + a long-chain carboxylic acid + CoA ⇒ AMP + pyrophosphate + an acyl-CoA	

All information, in particular the rows describing aspects about the biological function (Des) and the catalytic activity (Act), are taken directly from SWISS-PROT.[46] The values in the rows labelled Pair *N* refer to percentage pairwise sequence identity (second column), alignment length (third column), distance from HSSP-threshold (equation (2), fourth column), and the pairwise BLAST *E*-value (fifth column). The rows labelled Name give the SWISS-PROT identifier followed by the EC number (in parentheses), the length of the protein (in square brackets), and by the protein name.

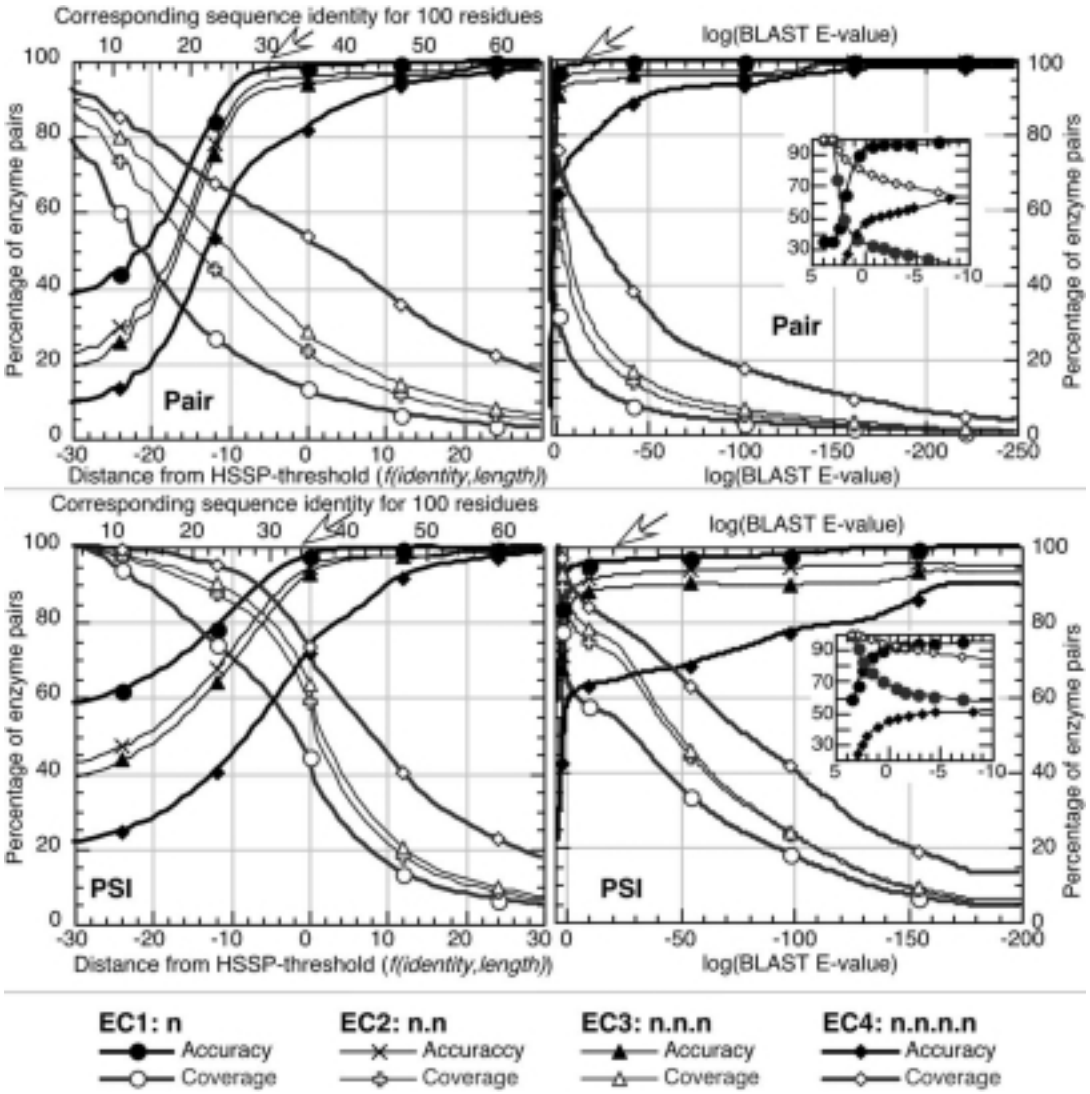

Figure 3. Estimates for accuracy and coverage at particular thresholds. The upper graphs describe the performance of pairwise BLAST searches, the lower graphs that of PSI-BLAST searches. Filled symbols and black lines describe accuracy (percentage of similar enzymes at given threshold), open symbols and grey lines describe coverage (percentage of enzymes found at given threshold). Similarity of enzymatic function is measured by the identity of the first EC number (thick lines with circles), the first two EC numbers (thin lines with crosses and +), the first three (thin lines with triangles), and all four EC numbers (thick lines with diamonds). The graphs on the left summarise the distances from the HSSP-threshold (equation (2)); corresponding levels of pairwise sequence identity (for alignment lengths of 100 residues) are shown on the top axis for orientation. The graphs on the right plot the logarithm of the BLAST *E*-values (note: log to the base 10). The arrows mark the points below which the HSSP-threshold become inferior to the statistical BLAST scores (Figure 2). The detailed view of the *E*-values given in the insets shows the data only for similarity in the first and all EC numbers. For example, 50% of all enzyme pairs found at a commonly used BLAST threshold 10^{-3} have four identical EC numbers. The difference between pairwise BLAST and PSI-BLAST is that the previous finds less than 80% of all pairs of identical EC number at that value, while the latter finds more than 90% of the identical pairs. Most surprisingly, *E*-values below 10^{-50} did not suffice to safely transfer enzyme function. In contrast, very high distances from the HSSP-threshold implied 100% accuracy in transferring function from homologues.

PSI-BLAST *E*-values. However, this was compensated by a considerably higher coverage. For example, at *E*-value = 1, PSI-BLAST found 68% of all enzymes with similar first EC number at 91% accuracy, while the pairwise BLAST found only 34% at 94% accuracy. The different main types of enzymes (first EC digit) had similar accuracy *versus* coverage curves (data not shown). However, for hydrolases (EC = 3.*) transfer of the full EC number was significantly worse than for all other classes; for lyases (EC = 4.*) marginally better. In contrast, the conservation of the first EC digit hydrolase was average, while transferases (EC = 2.*) had the best conservation of the class (first digit).

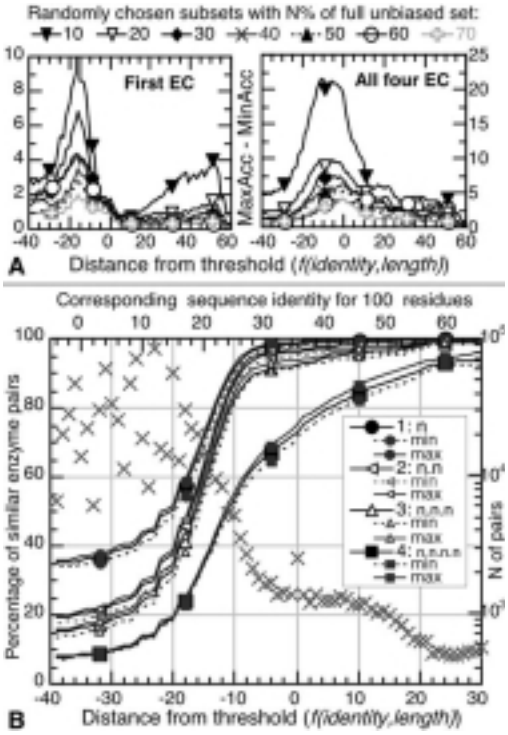

Figure 4. Bootstrap experiment to estimate the influence of the selection procedure. The set of representative proteins was divided randomly into smaller subsets, containing 10–70% of the representative proteins. The division was repeated 100 times and the minimal and maximal values found for each set were recorded. (a) The graphs on top show the difference between the sets with highest and lowest (min/max) accuracy at each threshold for a number of sizes of the random sets. Top left panel: enzymes were considered similar if they shared the first EC number; top right panel: all four EC digits were required to be identical. While restricting the analysis to 10–40% of the original representative proteins resulted in significant differences, the splits of 50–70% gave relatively similar results. (b) The lower graph compares the respective min/max values to the data shown in Figures 1–3; the min/max values were compiled on the basis of subsets comprising 60% of the original representative proteins; crosses give the number of pairs observed. Overall, the results appeared to be independent of the particular data set chosen.

Results are not sensitive to selection of data sets

When using only 10% of the data, the estimates differed considerably between different subsets of the representative proteins (Figure 4(a), 10% number marked by filled triangles). However, the results became rather stable when choosing more than half of the representative proteins: results differed by less than four percentage points (Figure 4(a) and (b)). Expectedly, different randomly chosen subsets varied most in the region in which the accuracy undergoes a sharp transition from accurate distinction to inaccurate distinction (between HSSP-thresholds of −20 and 0, equation (2)). This relative instability of the estimates for the accuracy were not caused by the fact that the respective region was not sufficiently populated by enzyme pairs, rather the "phase transition" occurred in the region of highest counts (crosses in Figure 4(b)). Note that such a sharp transition is a desirable feature of a threshold, since it facilitates the use of a stable and successful cut-off for sequence analysis. The transition was sharper for the HSSP-distances than for the BLAST *E*-values (Figure 2, lines with diamonds and circles). This was surprising, since the BLAST *E*-values explicitly account for the background statistics, while the HSSP-distance accounts for the background in a very empirical, *ad hoc* way.

Discussion and Conclusion

Transferring enzyme activity by sequence homology is more difficult than anticipated

Three groups related the conservation of enzyme function to pairwise sequence identity restricting their data to families of known structures.[43–45] Two groups[43,45] restricted both aligned proteins to those of known structure and with a single domain. Only one group implicitly uses a non-biased data set;[43] however, their results report levels of pairwise sequence identity from structural rather than from sequence alignments; these levels are typically lower for a particular protein, thus, the results overestimate conservation with respect to sequence alignments. Only one group reported results for PSI-BLAST alignments. No group compared accuracy *versus* coverage. All three groups agreed on levels above 90% accuracy for more than 50% pairwise identical residues.[43–45] However, due to the fairly different procedures, the data from the three groups differed considerably in detail (Figure 5). Two groups report that the full EC numbers start to diverge below 70% pairwise sequence identity,[43,44] one group finds a total conservation down to levels of 30% sequence

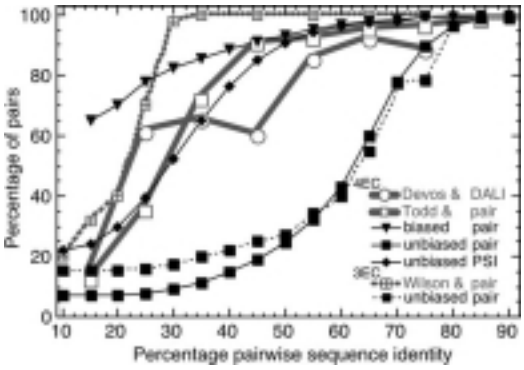

Figure 5. Comparison between results for different sets. The results given for the other groups were compiled from the respective publications. Devos & Valencia[43] used the structural alignments generated by DALI;[58] Todd *et al.*[44] used Needleman–Wunsch alignments to compile sequence identity, and Wilson *et al.*[45] used Smith–Waterman alignments. The thickness of the lines representing previously published results reflects the estimated error margin in reading the values off the publications. The values from Wilson *et al.* (broken grey line) differ from the other results in two ways: (i) the values are cumulative percentages; and (ii) similarity in function is defined as the identity of the first three EC digits (for comparison, the unbiased data set used here is shown for the same conditions: broken black line). All other results are averages over bins of ten percentage points (Devos & Valencia,[43] Todd *et al.*[44]) or over bins of five percentage points (all black lines), and the criterion for correct classification was the identity of all EC numbers.

identity.[45] Two groups find that the first digit begins to diverge only below 25%,[43,45] one group finds divergence already below 50%.[44] All previous groups used fairly small data sets (from about 6K[43] to about 30K[45] compared to the about 50,000 K used here for the unbiased set), none estimated the expected standard deviation for the estimates given. Thus, it is fairly difficult to compare the different results in detail. Nevertheless, all groups (including the results compiled here) appear to agree within the likely error margin above levels of 80% pairwise sequence identity. Between 50 and 80%, the results from the biased set (Figure 1) appear still relatively similar to those obtained by two of the three groups that agree in that region with each other.[43,44] In the same interval, the unbiased results deviate most from the biased ones. All sets differ amongst each other for levels between 30 and 50%; below levels of 20% pairwise sequence identity two of the structure-based results[43,44] approach the level suggested by the unbiased set. Overall, the two estimates for the performance of PSI-BLAST alignments given on the unbiased data set here, and by Todd *et al.* agree the most (Figure 5, filled diamonds and open squares). The unbiased data set revealed that for pair comparisons, less than 30% of all pairs above 50% sequence identity had fully identical EC numbers, rather than 90% as suggested by the biased set (Figure 1 inset, filled circles).

Statistical scores are better than sequence identity, but not necessarily best

Two groups established that statistical scores are better indicators for the similarity in enzyme function than are levels of pairwise sequence identity[11,45] using different versions of dynamic programming alignment methods and different statistical scoring schemes. Both the biased and the unbiased data confirmed these findings based on both pairwise BLAST and PSI-BLAST *E*-values. However, none of these groups noticed that a threshold relating sequence identity and alignment length (equation (2)) performed even better than the statistical BLAST scores for high levels of accuracy. A particular point appeared to be of interest to the community of BLAST users: even *E*-values below 10^{-50} did not always imply correct transfer of the full EC annotation (Figure 3, right panels). Basing the transfer of enzyme function on the HSSP distance proved much more reliable in this regime of greatest accuracy. Statistical BLAST scores outperformed the HSSP-threshold only in the regime for which it is no longer valid (below 0). Previously, pairwise BLAST searches had not been compared to the popular profile-based PSI-BLAST searches. Both data sets (biased and unbiased) suggested that at any given threshold, PSI-BLAST reaches a similar coverage as pairwise BLAST but at higher levels of accuracy (Figures 2 and 3).

Are thresholds for conservation of structure and function similar?

When two sequences are very similar we can safely infer that they have similar structures. The twilight zone characterises the region of sequence similarity in which we can no longer reliably infer structural similarity from sequence. The HSSP-threshold (equation (2)) optimally separates these safe and twilight zones for structural similarity. Surprisingly, this threshold optimised to capture structural similarity also proved more sensitive in identifying enzymes of similar function than the statistical BLAST scores for distances above 0 (Figure 3, arrows). The sequence conservation of structure and enzyme function differed at the point at which the transition from safe to twilight zone occurred. When considering two enzymes similar if they had identical EC numbers, the sequence signal for this similarity got blurred at much greater distances than the signal for similarity in structure. Consequently, two proteins of similar structure often differ in the details of their enzymatic function, as pointed out before.[43–45] Nevertheless, at an HSSP distance above 0 (similar to a BLAST score $<10^{-3}$) all pairs have similar structure,[19] and about 70–80% of these pairs have identical EC numbers (Figure 3, left graphs).

When considering two enzymes similar if they belonged to the same class (first EC digit), the transition from safe to twilight occurred much later than that for structure. In fact, more than 90% of all enzymes belonged to similar classes at distances at which less than 10% of all pairs have similar structures.[19] However, the most striking difference was the absence of a midnight zone: 90% of all pairs of similar enzymes could be detected at levels above 50% accuracy, i.e. experts could hope to identify the absolute majority of enzymes that belong to one of the six classes. In contrast, even experts fail to identify about 90% of all structural similarities from sequence.[23,33] Another similarity between the conservation of enzymatic function and structure was that both were characterised by relatively sharp transitions from safe to twilight. The general functional shape of the HSSP-threshold is explained by simple statistics,[28] although the precise equation (equation (2)) appears somehow arbitrary. In fact, the theoretical transition function marking the line between signal and noise depends on the particular details of the alignment algorithm, in particular on the substitution matrix chosen.[28] Is there a biological reason behind the similarities in the transitions from safe to twilight for enzyme function and protein structure? There may be. Different folds can perform the same function.[6,44,45,53,54] However, the surprising correlation between conservation of enzyme function and structure may indicate that conservation of structure drives function to a larger extent than we may have expected.[11]

Other aspects of protein function may be conserved differently

EC numbers capture only one aspect of protein function. Firstly, the assignment is somehow arbitrary: two experimentalists may assign different numbers to the same protein. Secondly, the full EC number does not describe all the details about a particular function: there may be considerable variation between two proteins of identical EC numbers.[44] Thirdly, enzymes differ from other functional classes; for example, the specificity of the binding of antibodies may be conserved differently from that of enzymatic activity. Why then focus on EC numbers? We may argue that the largest single class of existing proteins are enzymes, because about 30–50% of all proteins with experimentally characterised function have enzymatic activity. However, the real reason for bioinformatics to pick EC numbers is that they are well annotated and are numbers, i.e. can be compared easily on large data sets. For example, even the unbiased results were based on over 60,000 pairs of identical EC numbers. Hence, even if thousands of experimental annotations were wrong, this may still not impact the results considerably. Unfortunately, we cannot strictly generalise the estimates for when we can transfer EC numbers from homologues to other aspects of function. For example, Devos & Valencia[43] showed that the transfer of active sites information requires significantly higher levels of sequence similarity than the transfer of EC numbers. Wilson *et al.*[45] found that non-enzymatic functions are, on average, less conserved than enzymatic functions. More work remains to be done to explore the conservation of protein function further.

Impact on estimate of annotation errors

The differences between the biased and unbiased estimates impact estimates for errors in genome annotations.[18] Say we find N similarities at PSI-BLAST E-values $<10^{-3}$. Then we find, on average, $0.8N$ with E-values between 10^{-59} and 10^{-3}.[52] How did the estimates of accuracy differ between the biased and the unbiased sets in the corresponding interval of E-values? If genome annotations transfer only the first EC digit (e.g. hydrolase), then the levels of accuracy differed by a factor of 10! If annotations transfer all four EC digits, then the estimate provided by the unbiased set implied an error more than 20 times higher than anticipated previously. The good news was that more than 95% of the annotations transferring only the first EC digit are likely to be correct.

Materials and Methods

Data set of enzymes with experimentally known function

The data set of enzymes of known function were taken from the current merger of SWISS-PROT + SWISS-PROT_new:[46] 29,795 proteins had annotated EC numbers. In the next step, I excluded all those proteins for which either of the following applied: (1) EC number contained undecided digit (–); (2) more than one EC number was given (domains); (3) the keywords contained one of the "words" PROBABLE, PUTATIVE, BY SIMILARITY, or BY HOMOLOGY. This reduction left 26,342 enzymes of experimentally assigned function. The EC classification uses four digits to classify function. The first EC digit distinguishes the type of enzymatic activity: 1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; 6, ligases. The second EC digit specifies the substrate (oxireductases), the group transferred (transferases), the type of bond (hydrolases, lyases, ligases), or the type of reorganisation (isomerases). The third and fourth digits provide more detail. (For an excellent survey of structural aspects of enzymatic function, see Todd *et al.*[44])

Reduction of bias in data set: idea

Any database of today contains two different sorts of bias. The first is due to the past and current experimental techniques and the focus of experimental research. The second is due to evolutionary divergence that favoured neutral mutations, thus generating highly similar proteins with slightly different features. Typically, we are used to accounting for this bias by grouping proteins into families of homologues. If we want to estimate the accuracy in transferring function in the context of

genome annotation, we have to base our analysis on data sets that have a redundancy similar to that of entire proteomes. In order to reduce the bias from the set of enzymes of known function {B}, first we have to generate all-against-all alignments {B} × {B}. Then, we have to choose the maximal subset of the known-function set that fulfils the constraint that no pair in that subset is sequence similar {U}. The least biased comparison would now base statistics on all pairs {U} × {U}. Although this would still yield about four million comparisons for the set of known enzymes, the problem is that by definition of set {U}, none of the four million pairs has significant sequence similarity, i.e. we cannot derive a threshold for transfer of function between very similar proteins. Consequently, we have to accept some of the existing bias in {B}, i.e. we have to include some pair comparisons within the representative families {U}. One way to do this is by accepting one additional member from each family, yielding {U} × 2{U} pairs. However, I found that the resulting set of about eight million comparisons appeared far too small a data set to derive reliable estimates (Figure 4). Thus, I based the estimates on two alternatives: (1) biased set = {B} × {B}, corresponding to 700 million pairs, and (2) unbiased set = {U} × {B}, corresponding to 52 million pairs.

Reduction of bias in data set: procedure

(1) Align all 26,342 enzymes against each other by pairwise BLAST.[15] (2) Compile HSSP-threshold (equation (2)) for each pair. (3) Find all pairs likely to have similar structures (HSSP-threshold $\vartheta = 0$), and record the size of each such structural family. (4) Sort all pairs by the number of members in the respective structural family and by length (for families of similar size). (5) Cluster pairs by a simple greedy algorithm that starts from the top of the sorted pair list. For instance, for the first protein P1 in the list cluster 1 will contain all proteins that are more similar to P1 than a certain threshold ϑ. Note that the sorting of the list of pairs implies which protein is chosen as the representative (see below). The number of clusters we obtain depends on the choice of the threshold ($N_{\text{cluster}}(\vartheta)$). In general, ($N_{\text{cluster}}(\vartheta)$) is not a linear function but it is constantly growing, i.e. the smallest ϑ gives one single cluster (all pairs similar), and the highest ϑ results in 26,342 clusters (no pair similar). *A priori*, it is not clear which threshold to apply for this redundancy reduction. If we ignore the reality of protein structures and conceive sequences as simple strings of 20 letters, we can compile the probability that two particular sequences are similar by chance. Obviously, this probability is very low for a pair with 99% pairwise identical residues and very high for a pair with 5% pairwise identical residues. Interestingly, we observe a transition between proteins of similar structure that is almost as sharp as a state-transition in physics.[19,28] This observation suggests using the point of the transition to define the threshold used to cluster. In particular, I chose the threshold fulfilling the following condition:

$$N_{\text{cluster}}(\vartheta - 1) - 3 \leq N_{\text{cluster}}(\vartheta) \leq N_{\text{cluster}}(\vartheta + 1) + 3 \quad (1)$$

where $N_{\text{cluster}}(\vartheta)$ is the number of clusters resulting from clustering at a threshold of ϑ. The rationale for this particular way of reducing bias was to find a threshold that yields somehow stable cluster separations. In fact, I investigated all thresholds $\vartheta = -10, \ldots, 30$ and found that the most stable situation was that for $\vartheta = -2$, which fulfilled equation (1). After collecting all data, I verified that the major results were indeed insensitive to the particular choice of the threshold used to cluster. Furthermore, the results suggested that this initial choice indeed yielded the most stable results. Finally, I tested alternative ways of sorting the list of pairs; in particular, instead of sorting by family size, I sorted by longest proteins, shortest protein, and alphabet. The number of clusters at a given threshold differed indeed from the sorting by family size. However, the results displayed were not sensitive to the sorting.

Aligning the enzymes

Two different alignment schemes were explored: (1) pairwise BLAST[15] and (2) PSI-BLAST.[16] The particular protocol for finding similarities with PSI-BLAST applied the usual precautions to avoid drift and pollution.[55,56] Searches were restricted to three iterations, and the iteration parameter (*H*-value) to 10^{-10} was set. The search database was SWISS-PROT[29] + TrEMBL[29] + PDB.[47] Finally, enzyme pairs of known function were extracted from the resulting PSI-BLAST alignments.

Scores for measuring sequence similarity

The simplest way to measure sequence similarity is pairwise sequence identity, i.e. the percentage of residues identical between two proteins divided by residues aligned (not counting gaps). Since all results were based on BLAST and PSI-BLAST alignments, only statistical scores given by these programs (*E*-values) were used. Finally, the distance from the HSSP-threshold (*DIST*) was displayed; it is given by:[19]

$$DIST = PIDE - HSSP_PIDE(\vartheta)$$

$$HSSP_PIDE(\vartheta) = \vartheta + \begin{cases} 100, & \text{for } L \leq 11 \\ 480L^{-0.32/1+e^{-L/1000}}, & \text{for } 11 < L \leq 450 \\ 19.5, & \text{for } L > 450 \end{cases} \quad (2)$$

here: $\vartheta = 0$, where L is the length of the alignment between two proteins, *PIDE* is the percentage of pairwise identical residues, and $HSSP_PIDE(\vartheta)$ is the revised HSSP-threshold for the level ϑ. As described above, I chose $\vartheta = -1$ to reduce the bias. However, to compile distances, I chose the threshold of $\vartheta = 0$.

Defining accuracy and coverage

I used the following definitions:

$$\text{Accuracy} = 100\left(\frac{\text{number of similar pairs found above threshold}}{\text{number of all pairs found above threshold}}\right) \quad (3)$$

$$\text{Coverage} = 100\left(\frac{\text{number of similar pairs found above threshold}}{\text{number of all similar pairs}}\right)$$

with the thresholds given by (1) percentage pairwise sequence identity, (2) BLAST/PSI-BLAST *E*-values, or (3) distance from HSSP-threshold (equation (2)). For simplicity, I distinguished only between two alternatives for "similar pair". (A) The EC numbers are fully identical, i.e. all digits are the same. (B) The first digits of the EC numbers are identical, e.g. both proteins are hydrolases.

Estimating the error of the results

The results for the biased and unbiased data sets differed substantially (Figure 1). This fact raised the question about how stable the results were with respect to the particular choice of the unbiased set and to the particular set of enzymes of known function contained in SWISS-PROT. One way in which I addressed this question is by testing alternative ways of choosing the representatives for each cluster and the thresholds for the clustering (see above). However, I addressed the degree to which the sets were representative more explicitly by the following test, which is inspired by the bootstrap concept.[57] (1) Randomly choose a subset $\{SubU_f\}$ of f $(1 > f > 0)$ of all proteins in the unbiased set {U}. (2) Evaluate the dependency of the accuracy on the threshold for this subset. (3) Repeat the first two steps 100 times, and monitor the minimum and maximum values found for each threshold ϑ. I tested values of f between 0.1 and 0.7 in steps of 0.1. The more similar the results for any subset $i\{SubU_{if}\} \times \{B\}$ to that for all $\{U\} \times \{B\}$ pairs, the more representative the set {U}. Thus, large differences in the min/max tests indicate that the sets of the size chosen are not representative. On the other hand, if we find no difference for a fraction f close to 1, then this does not necessarily imply that the original set {U} is sufficiently representative for the universe of proteins. Instead, {U} may have some intrinsic consistency caused by the particular bias of today's experimental techniques. Nevertheless, the bootstrap experiment will reveal artefacts caused by the particular protocol of clustering and selecting representative subsets.

Acknowledgments

Thanks to Jinfeng Liu (Columbia) for computer assistance and the collection of genome data sets; to Dariusz Przybylski (Columbia) for software; to Rajesh Nair, Henry Bigelow (both Columbia) and Damien Devos (CNB Madrid) for important discussions. Also thanks to Henry Bigelow (Columbia) and the undisclosed reviewers for helping to improve the manuscript. The work was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

1. Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675–683.
2. Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393–403.
3. Casari, G., Andrade, M. A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C. *et al.* (1995). Challenging times for bioinformatics. *Nature*, **376**, 647–648.
4. Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Blenkowska, J., Adams, R. M., Smith, T. F. & Lindelien, J. (1997). Biology's new Rosetta stone. *Nature*, **385**, 29–30.
5. Doolittle, R. F. (1997). A bug with excess gastric avidity. *Nature*, **388**, 515–516.
6. Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711.
7. Frishman, D. & Mewes (1997). PEDANTic genome analysis. *Trends Genet.* **13**, 415–416.
8. Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences—a detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**, 302–310.
9. Koonin, E. v., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol.* **266**, 295–322.
10. Pawlowski, K., Zhang, B., Rychlewski, L. & Godzik, A. (1999). The *Helicobacter pylori* genome: from sequence analysis to structural and functional predictions. *Proteins: Struct. Funct. Genet.* **36**, 20–30.
11. Pawlowski, K., Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). Sensitive sequence comparison as protein function predictor. *Pac. Symp. Biocomput.* **8**, 42–53.
12. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
13. Walker, D. R. & Koonin, K. V. (1997). SEALS: a system for easy analysis of lots of sequences. In *Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 333–339, AAAI Press, Halkidiki, Greece.
14. Tamames, J., Ouzounis, C., Sander, C. & Valencia, A. (1996). Genomes with distinct function composition. *FEBS Letters*, **389**, 96–101.
15. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.
16. Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
17. Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132–133.
18. Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
19. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
20. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
21. Chung, S. Y. & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123–1127.
22. Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **249**, 816–831.
23. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679–689.
24. Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural

meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
25. Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355–368.
26. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210.
27. Schneider, R., de Daruvar, A. & Sander, C. (1997). The HSSP database of protein structure–sequence alignments. *Nucl. Acids Res.* **25**, 226–230.
28. Alexandrov, N. N. & Soloveyev, V. V. (1998). Statistical significance of ungapped sequence alignments. In *HICCS'98: Pacific Symposium on Biocomputing'98* (Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T., eds), pp. 463–472, World Scientific, Maui, HI.
29. Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing β-sheet into α-helix. *Nature Struct. Biol.* **4**, 548–552.
30. Doolittle, R. F. (1986). *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books, Mill Valley, CA.
31. Li, W., Pio, F., Pawlowski, K. & Godzik, A. (2000). Saturated BLAST an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics*, **16**, 1105–1110.
32. Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci.* **7**, 1431–1440.
33. Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**, S19–S24.
34. Lindahl, E. & Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
35. Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence–structure threading. *Curr. Opin. Struct. Biol.* **5**, 236–244.
36. Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
37. Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7**, 200–205.
38. Finkelstein, A. V. (1997). Protein structure: what is it possible to predict now? *Curr. Opin. Struct. Biol.* **7**, 60–71.
39. Fischer, D. & Eisenberg, D. (1999). Predicting structures for genome proteins. *Curr. Opin. Struct. Biol.* **9**, 208–211.
40. Rost, B. & Sander, C. (1996). Bridging the protein sequence–structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 113–136.
41. Rost, B. & O'Donoghue, S. I. (1997). Sisyphus and prediction of protein structure. *CABIOS*, **13**, 345–356.
42. Shah, I. & Hunter, L. (1997). Predicting enzyme function from sequence: a systematic appraisal. In *Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 276–283, AAAI Press, Halkidiki, Greece.
43. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
44. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
45. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
46. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
47. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
48. Finkelstein, A. V. & Reva, B. A. (1991). A search for the most stable folds of protein chains. *Nature*, **351**, 497–499.
49. Finkelstein, A. V., Badretdinov, A. Y. & Gutin, A. M. (1995). Why do protein architectures have Boltzmann-like statistics? *Proteins: Struct. Funct. Genet.* **23**, 142–150.
50. Thornton, J. M., Orengo, C. A., Todd, A. E. & Pearl, F. M. (1999). Protein folds, functions and evolution. *J. Mol. Biol.* **293**, 333–342.
51. Liu, J. & Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979.
52. Liu, J. & Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics*, in the press.
53. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
54. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **283**, 595–602.
55. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
56. Przybylski, D. & Rost, B. (2002). Alignments grow, secondary structure prediction improves. *Proteins: Struct. Funct. Genet.* **46**, 195–205.
57. Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Sci. Am.* **248**, 96–108.
58. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Edited by J. Thornton

(Received 10 September 2001; received in revised form 31 December 2001; accepted 23 January 2002)