Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins

Hedi Hegyi and Mark Gerstein[1]

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

Annotation transfer is a principal process in genome annotation. It involves "transferring" structural and functional annotation to uncharacterized open reading frames (ORFs) in a newly completed genome from experimentally characterized proteins similar in sequence. To prevent errors in genome annotation, it is important that this process be robust and statistically well-characterized, especially with regard to how it depends on the degree of sequence similarity. Previously, we and others have analyzed annotation transfer in single-domain proteins. Multi-domain proteins, which make up the bulk of the ORFs in eukaryotic genomes, present more complex issues in functional conservation. Here we present a large-scale survey of annotation transfer in these proteins, using scop superfamilies to define domain folds and a thesaurus based on SWISS-PROT keywords to define functional categories. Our survey reveals that multi-domain proteins have significantly less functional conservation than single-domain ones, except when they share the exact same combination of domain folds. In particular, we find that for multi-domain proteins, approximate function can be accurately transferred with only 35% certainty for pairs of proteins sharing one structural superfamily. In contrast, this value is 67% for pairs of single-domain proteins sharing the same structural superfamily. On the other hand, if two multi-domain proteins contain the same combination of two structural superfamilies the probability of their sharing the same function increases to 80% in the case of complete coverage along the full length of both proteins, this value increases further to > 90%. Moreover, we found that only 70 of the current total of 455 structural superfamilies are found in both single and multi-domain proteins and only 14 of these were associated with the same function in both categories of proteins. We also investigated the degree to which function could be transferred between pairs of multi-domain proteins with respect to the degree of sequence similarity between them, finding that functional divergence at a given amount of sequence similarity is always about two-fold greater for pairs of multi-domain proteins (sharing similarity over a single domain) in comparison to pairs of single-domain ones, though the overall shape of the relationship is quite similar. Further information is available at http://partslist.org/func or http://bioinfo.mbb.yale.edu/partslist/func.

The ultimate goal of the genome projects is to determine the structure and function of all the newly identified gene products. Fundamentally, this will be carried out via annotation transfer, transferring the structural and functional annotation from an experimentally characterized protein (as in a model organism such as *Escherichia coli*) to a predicted protein in a newly sequenced genome that shares similarity in sequence. The degree of annotation transferred will depend on the degree of sequence similarity. This process is shown schematically in Figure 1. In this paper, we aim to address this major question in bioinformatics, specifically focusing on multi-domain proteins, as they make up the bulk of the proteome in eukaryotic organisms (Gerstein 1998).

Our work is a direct outgrowth of two previous analyses of ours that concentrated on single-domain proteins. In an earlier paper, we found that the different structural classes of the scop classification system have different propensities to carry out certain types of function (Hegyi and Gerstein 1999). In particular, while the alpha/beta folds were disproportionately associated with enzymes and all-alpha and small folds with non-enzymes, the alpha + beta structures had an equal tendency for both enzymatic and non-enzymatic functions.

[1]Corresponding author.
E-MAIL Mark.Gerstein@yale.edu
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr. 183801.

Wilson et al. (2000) compared a large number of protein domains to one another in a pair-wise fashion with respect to similarities in sequence, structure, and function. Using a hybrid functional classification scheme merging the ENZYME and FlyBase systems (Gelbart et al. 1997; Bairoch 2000), they found that precise function is not conserved below 30–40% identity, although the broad functional class is usually preserved for sequence identities as low as 20–25%, given that the sequences have the same fold. Their survey also reinforced the previously established general exponential relationship between structural and sequence similarity (Chothia and Lesk 1986).

Other Work on Establishing Relationships between Sequence, Structure, and Function

Several other groups have studied the relationship between sequence, structure, and function in detail, attempting to determine the extent to which functional transference between matching proteins is feasible (Shah and Hunger 1997; Martin et al. 1998; Thornton et al. 1999, 2000; Zhang et al. 1999; Shapiro and Harris 2000; Todd et al. 2001). Orengo et al. (1999) analyzed protein families in the CATH database and concluded that > 96% of the folds in the PDB are associated with a single homologous family. By investigating enzymatic folds they also found that more than 95% of homologous families show either single or closely related functions.

11:1632–1640 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

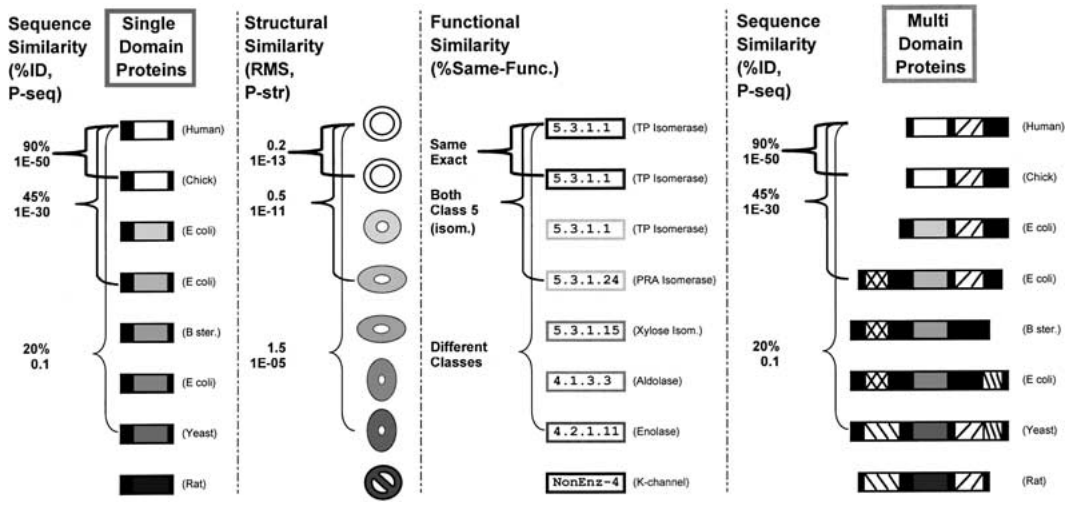

Figure 1 Schematic illustrating annotation transfer. This figure illustrates the process of annotation transfer for a group of hypothetical TIM barrel proteins. The *leftmost* panel represents sequence comparisons between idealized barrel domains from a number of organisms. The next panel shows analogous results for structural comparison, and the panel after that, functional comparison. The *rightmost* panel represents sequence comparisons between idealized multi-domain proteins that match over a single domain, the subject of much of this paper.

Pawlowski et al. (2000) studied the relationship between sequence and functional similarity in the twilight zone of 10%–15% sequence similarity and found a clear correlation between the two, with functional similarity based on the E.C. classification of enzymes.

Russell et al. (1997) analyzed binding sites in proteins with similar 3D structures and estimated that 90% of new remote homolog have common binding sites and similar functions. Eisenstein et al. (2000) evaluated the first results from the structural genomics projects and found that in many instances the protein structure itself offers an important clue to its biological function. Stawiski et al. (2000) found that function could be predicted rather successfully for just the proteases. Devos and Valencia (2000) presented a critical view of function transference between similar sequences, highlighting the limitations of this process due to errors in databases and the inherent complexity of the relationship between protein sequence-structure and function that does not allow "simplistic interpretations." They also found that binding sites are the least conserved features between related proteins while the catalytic activity of enzymes is the most conserved one.

Multi-Domain Proteins with Divergent Functions: How Common?

Most of these previous investigations focused on single-domain proteins or did not distinguish between single- and multi-domain ones. It is not clear how the multi-domain proteins with various functions behave with respect to functional conservation; namely, whether they are more or less conserved than their single-domain counterparts. In particular, as shown in Figure 1, if one multi-domain protein shares a single domain fold with another one, it is not clear the degree to which the functional conservation of these proteins is constrained by the shared part, and to what degree it is influenced by other domains that are not shared.

Specific groups of proteins that have the same combination of structural domains but dramatically different functions illustrate this situation. One example is the combination of the SH3-domain (scop superfamily identifier 2.24.2) and the P-loop containing NTP hydrolase (3.29.1). While in higher organisms this combination is associated with presynaptic and tumor suppressor functions (SWISS-PROT names SP02_HUMAN and DLGI_DROME, respectively), in the lower Dictyostelium it was found in myosin (MYSP_DICDI). Another example is the combination of the FAD/NAD(P)-binding superfamily and FAD-linked reductases C-terminal superfamily (3.4.1 and 4.12.1 superfamilies, respectively). In one group of proteins they appear in enzymes of the oxidoreductase group (e.g. OXDA_CAEEL or PHHY_PSEAE), while in another they are found in a dissociation inhibitor (e.g. GDIA_HUMAN). It should be noted that the proteins are not covered completely by the structural matches, so it is quite possible that the rest of them contain totally different domains that are responsible for the dramatically different functions. However, do these two examples show a rather rare or a more frequent phenomenon? How often do multi-domain proteins, sharing the same structural domain composition, differ in their functions?

In this paper, we attempt to provide a comprehensive answer to this question. This is particularly timely given that most of the unknown proteins in eukaryotic genomes are multi-domain. We use the same approach as in our previous analyses, comparing the sequences of the structural domains in scop to those of SWISS-PROT using BLASTP. We focus on the functional divergence of single and multi-domain proteins, extending previous investigations of single-domain proteins. Also, in comparison to previous work, we focus more on non-enzymatic functions and scop structural superfamilies, instead of folds.

RESULTS

Our Approach to Functional and Structural Assignment

We used the BLASTP program (version 2.0) (Altschul et al. 1997) to identify the scop 1.39 (Murzin et al. 1995) structural domains in SWISS-PROT (version 37) (Bairoch and Apweiler

2000) with $e = 10^{-4}$. We removed the hypothetical and fragment proteins. This resulted in two sets of proteins.

Single–Domain

Of the single-domain matches, only those that were almost completely covered with a match to a single structural domain were selected. (The maximum number of uncovered residues was set at 70 with an additional condition that a maximum of 40 residues on the N-terminal end and 30 residues on the C-terminus were allowed to be uncovered.) These criteria resulted in 1818 single-domain proteins being selected from SWISS-PROT.

Multi–Domain

We selected 4763 multi-domain proteins from SWISS-PROT. All of these matched (in different locations) at least two domains of known structure belonging to different scop superfamilies (see schematic in Figure 1). We also selected a subset of these proteins that have almost their entire length covered by matches with structural domains (allowing again a maximum of 70 uncovered residues). This selection resulted in 2829 proteins being selected from SWISS-PROT. (In all cases, duplicate matches were removed, i.e., a protein at a certain location matches only one structural domain.)

We set out to compare these two sets of proteins for functional divergence. As previously, we divided functions into enzyme and non-enzyme (Hegyi and Gerstein 1999). Enzymatic functions were classified by the EC system (Bairoch 2000). Comparisons of enzymatic functions were treated the same way as in our earlier analyses, that is, if they differ in the first three components of their respective EC numbers, they were considered different. This implied that our analysis dealt with a total of 112 enzymatic functions. Non-enzymatic functions were classified into 508 different categories based on a simple thesaurus we assembled of synonymous keywords drawn from SWISS-PROT description lines. In addition, we created 49 categories for functions that have an enzymatic component but which are not part of the EC system. This gave us a total of 669 functions (112 + 508 + 49). (The list of all the functional categories is described further in Table 2 below, and also can be found on the Web at http://bioinfo.mbb.yale.edu/partslist/func or http://partslist.org/func.)

fold 1	fold 2	FOLD PAIRS																			
		aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro
3.29	2.32	4	3	4	3	12	14	6	7	8	4	6	7	5	3	3	4	5	3	4	4
2.29	4.61	1	1	1	2	6	3	2	4	5	4	4	3	3	3	4	1	2	3	3	2
4.1	4.34	1	1	1	1	5	3	1	3	1	1	2	1	1	1	1	2	2	1	1	1
1.28	3.29	1	1	2	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
3.4	4.48	4	1	1	2	3	4	1	2	4	3	5	2		2	2	1	1	1	1	2
3.22	4.42	1	1	1	0	4	5	3	4	5	4	4	3	4	1	1	2	2	3	3	1
2.32	4.1	1	1	1	1	4	2	1	3	1	1	2	1	1	1	1	2	2		1	1
2.32	2.33	2	1	1	1	1	2	2	1	2	1		2	1	1	1	1	1	1	1	1
4.32	3.1	1	1	1	2	5	1	1	1	4	5	1	1	1	1	1	1	1	1	1	0
3.23	4.89	3	3	3	0	9	10	6	5	6	8	7	2	4	0	0	1	1	2	2	2
3.47	5.17	0	0	1	0	12	10	1	3	3	1	1	2	1	1	1	2	1	1	1	2
4.72	5.13	1	0	0	0	1	3	1	1	2	2	1	2	1	2	2	2	1	2	2	2
3.22	4.1	1	1	1	0	3	3	2	1	1	2	1	1	0	1	1	1	0	1	1	1
3.5	3.1	1	1	2	1			1	1	5	1		1	1	1	1	1	1	1	1	1
4.61	3.42	2	2	2	2	2	1	1	1	1	1	1	1	0	2	2	1	1	1	0	0
1.76	3.3	1	0	1	1	2	1	1	1	1	2	1	1	1	0	0	0	1	1	1	1
4.29	4.1	1		1	1	2	1	1	1	1	1		1		0	1	1	1	1	1	1
2.32	4.34				2	1	1	2	1	2	2	1	2	1	1	1	1	1	1	1	1
3.22	1.79	1	1	1	0	3	1	2	2	2	4	3	2	1	0	0	1	1	1	1	0
3.52	2.34	0	0	0	0	0	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1

Figure 2 Distribution of multi-domain combinations amongst the genomes. The figure shows the occurrence of multi-domain fold combinations in a number of genomes, indicating its great variability. Each row indicates a particular combination of scop fold pairs (using scop 1.39), where a fold pair is defined as two distinct folds occurring in tandem in a protein. Each column represents a different genome, using the four-letter codes in the PartsList system (Qian et al. 2001): Aaeo, *Aquifex aeolicus*; Aful, *Archaeoglobus fulgidus*; Bbur, *Borrelia burgdorferi*; Bsub, *Bacillus subtilis*; Cele, *Caenorhabditis elegans*; Cpne, *Chlamydia pneumoniae*; Ctra, *Chlamydia trachomatis*; Ecol, *Echerischia coli*; Hinf, *Haemophilus influenzae* Rd; Hpyl, *Helicobacter pylori*; Mthe, *Methanobacterium thermoautotrophicum*; Mjan, *Methanococcus jannaschii*; Mtub, *Mycobacterium tuberculosis*; Mgen, *Mycoplasma genitalium*; Mpne, *Mycoplasma pneumoniae*; Phor, *Pyrococcus horikoshii*; Rpro, *Rickettsia prowazekii*; Scer, *Saccharomyces cerevisiae*; Syne, *Synechocystis* sp.; Tpal, *Treponema pallidum*. The numbers in each intersection cell indicate the number of times the fold pairs occur in a genome. Only the 20 most common fold pair combinations are shown here; the remainder are shown on the Web site (http://partslist.org/func). If a cell is greater than 6, it is shaded black; between 3 and 6, gray; and below 3, white. The blank spaces show instances in which one of the pairs does not occur in the organism at all (indicated by a value of -1 in the data table on the Web site). The fold assignments are done in a fashion consistent with those in PartsList and associated systems (Gerstein 1997; Lin et al. 2000; Drawid et al. 2001; Harrison et al. 2001; Qian et al. 2001).

Overall Distribution of the Matches

Figure 2 shows the most commonly observed multi-domain combinations in a set of recently sequenced genomes. The occurrences of further combinations are available from the Web site. Clearly, the distribution is very skewed, with certain combinations, such as 3.29–2.32, and 2.29–4.61 tending to predominate.

Figure 3 shows the overall distribution of the single-domain and multi-domain matches in the different structural classes. The distribution of matches between enzymes and non-enzymes in multi-domain proteins largely agrees with that in the single-domain proteins. The multi-domain matches follow the overall tendency of the alpha/beta folds to be associated with enzymes to a larger extent and the all-alpha and small folds with non-enzymes. However, the values for the multi-domain matches are generally less extreme than for single-domains; for example, the 10-fold difference between single-domain alpha/beta enzymes and non-enzymes decreases to about twofold in multi-domain proteins. Another significant difference is the reduction in the number of multi-domain non-enzymes in the all-beta and alpha + beta structural classes compared to the single-domain matches. Altogether, there are more enzymes than non-enzymes among the multi-domain proteins (2805 enzymes vs. 1958 non-enzymes) whereas for single-domain proteins, the opposite is true (850 enzymes vs. 968 non-enzymes).

Table 1 summarizes the distribution of superfamilies and superfamily combinations among the major functional classes, i.e. whether they have only enzymatic, only non-enzymatic or both enzymatic and non-enzymatic functionality. Altogether, 215 superfamilies were found in single-domain proteins and 310 in multi-domain ones. As 70 superfamilies were found in both, altogether 455 distinct structural superfamilies matched a SWISS-PROT protein with our required coverage criteria (described above). Similarly, we apportioned the 281 superfamily combinations observed in multi-domain proteins amongst different broad functional categories.

In single-domain proteins there are about as many superfamilies with exclusively enzymatic functionality as there are those with exclusively non-enzymatic functions (82 vs. 78). In contrast, in multi-domain proteins this ratio increases

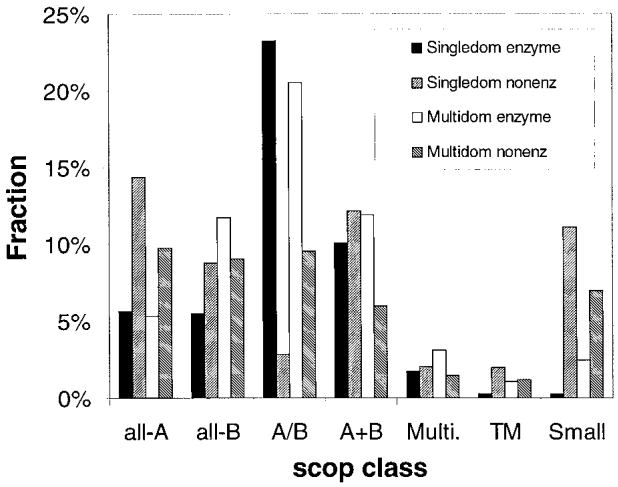

Figure 3 Distribution of proteins amongst broad structural and functional classes; the distribution of the matches among the seven structural and two functional classes in single- and multi-domain proteins. The single-domain and multi-domain matches each total 100%, independently of each other. The horizontal axis indicates the seven scop classes, which are (from 1 to 7): all-alpha, all-beta, alpha/beta, alpha + beta, multi-domain, membrane, and small protein.

to almost threefold (135 vs. 56). This agrees with the notion that most enzymes are multi-domain. Another difference between single and multi-domain proteins appears in the ratio of superfamilies with a single function compared to multifunctional ones. As it is apparent from Table 1, about a quarter of the superfamilies matched single-domain proteins with different functions (55 of 215), whereas in the multi-domain proteins, this ratio increased to more than a third (119 of 310).

Single-Domain Proteins

Table 2 lists the two functionally most diverse structural superfamilies in single-domain proteins with some representative functions. The most diverse superfamily, the 3.38.1 Thioredoxin-like, has 11 different functions associated with it, most of them with an oxidoreductase mechanism. For instance, THIO_BPT4 is a small disulphide-containing thioredoxin that serves as a general disulphide oxidoreductase, while TDX2_BRUMA is almost twice as long (199 aa) and serves as a thiol-specific antioxidant that acts against sulfur-containing radicals. Another interesting example of functional diversity is provided by the Scorpion toxin-like superfamily (7.3.6). While BRAZ_PENBA is a small protein that is known to be 2000 times sweeter than sucrose, the other members of the superfamily are associated with different host-defense mechanisms. In insects the superfamily possesses antifungal activity (DMYC_DROME) or acts as a toxin (SCX5_BUTEU). Interestingly, in plants it can also act as an antifungal (AF2B_SINAL) or as an inhibitor of insect alpha-amylases (SIA1_SORBI). It appears that many single-domain proteins are toxins or allergens, or are related in other ways to a host-defense response.

Based on the data we can also determine the probability of two single-domain proteins that match domains in the same superfamily category also carrying out the same function. Using Bayes' theorem:

$$P(F|S) = P(F)P(S|F)/((P(F)P(S|F) + P(\bar{}F)P(S|\bar{}F)) \quad (1)$$

where S is the probability that two proteins share the same superfamily, F is the probability that two proteins have the same function, and $\bar{}F$ is the probability that two proteins do not have the same function. Rearranging and simplifying the equation we get:

$$P(F|S) = 1/(1 + N(S,\bar{}F)/(N(S,F)) \quad (2)$$

where N is the number of times that the two events in the parentheses occur together in our database of 1818 single-domain proteins. This results in

$$P(F|S) = 1/(1 + 8501/12516) = 68\%.$$

That is, the probability that two single-domain proteins that have the same superfamily structure have the same function (whether enzymatic or not) is about 2/3.

Multi-Domain Proteins

Table 3 lists the combinations of superfamilies that have been associated with the greatest number of different functions in multi-domain proteins, with representative entries in SWISS-PROT. The combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Although it has twice as many different functions as the most diverse superfamily in

Table 1. Functional Distribution of Single-domain, Multi-domain Superfamilies, and Multi-domain Combinations

	Single-domain superfamilies		Multi-domain superfamilies		Multi-domain sfam combinations	
	Single function	Multiple function	Single function	Multiple function	Single function	Multiple function
Enzymatic	82	11	135	42	151	16
Nonenzymatic	78	23	56	30	70	27
Both functions	—	15	—	47	—	17
Total	160	55	191	119	221	60

The basic functional distribution of the superfamilies in single- and multi-domain proteins and the functional distribution of multi-domain combinations are shown. The first row lists the number of scop superfamilies that were associated only with enzymatic function in each category. The second row lists the number associated with only nonenzymatic functions, and the third row indicates the number of superfamilies that were associated with both types of function. Altogether, we characterized 160 + 55 = 215 single-domain and 191 + 119 = 310 multi-domain superfamilies, 70 of which overlapped in the two categories.

Table 2. Most Versatile Single-Domain Superfamilies

No. func	No. prot	Sfam comb	Function	SWISS-PROT ID	SWISS-PROT function
11	69	3.38.1	E1.11.1	GSHP_RAT	Plasma Glutathione Peroxidase (1.11.1.9)
			263#	DYL5_CHLRE	Dynein, Flagellar Outer Arm–*C. reinhardtii*
			D260#	BSAA_BACSU	Glutathione Peroxidase Homolog Bsaa
			268#	REHY_TORRU	Rehydrin–*Tortula ruralis* (Moss)
			266#	PHOS_HUMAN	Phosducin (33 Kd Phototransducing Protein)
			269#	REHY_ORYSA	Rad24 Protein–*Oryza sativa* (Rice)
			272#	THIO_BPT4	Thioredoxin (Bacteriophage T4)
			D271#272#	TDX2_BRUMA	Thioredoxin Peroxidase 2
			261#	BTUE_ECOLI	Vitamin B12 Transport Periplasmic Protein Btue
10	28	7.3.6	342#	BRAZ_PENBA	Brazzein–*Pentadiplandra brazzeana*
			376#336#	SCKK_TITSE	Neurotoxin Ts-Kapa (Tsk)–(Brazilian scorpion)
			341#356#	AF2B_SINAL	Cysteine-Rich Antifungal Protein 2b (Afp2b)
			343#	DEFA_ZOPAT	Defensin, Isoforms B And C–*Zophobas atratus*
			361#	DMYC_DROME	Drosomycin Precursor (Cysteine-Rich Peptide)
			361#376#	SCX5_BUTEU	Insectotoxin I5a–(Lesser Asian scorpion)
			336#	SCX3_LEIQH	Leiuropeptide Iii–(Scorpion)
			203#	SIA1_SORBI	Small-Pr Inhibitor Of Insect Alpha-Amylases
7	34	4.79.3	310#	AB18_PEA	Aba-Responsive Protein Abr18–Garden Pea
			311#	DRR3_PEA	Disease Resistance Response Protein Pi49
			231#	MPAA_CORAV	Major Pollen Allergen Cor A 1,–Eu. Hazel
			312#	L18B_LUPLU	Protein L1r18b (Llpr10.1b)
			E3.1.–	RNS2_PANGI	Ribonuclease 2 (3.1.–/–)–Panax Ginseng
			314#	SAM2_SOYBN	Stress-Induced Protein Sam22
7	43	1.26.1	184#	CSF2_SHEEP	Colony-Stimulating Factor
			381#564#184#	IL4_RAT	Interleukin-4 (B-Cell Igg Diff. Factor)
			185#	LIF_HUMAN	Leukemia Inhibitory Factor (Lif)
			187#	PRL_ANGAN	Prolactin Precursor (Prl)–
			186#	PLF3_MOUSE	Proliferin 3 Mitogen-Regulated
			188#	SOMA_PAROL	Somatotropin (Growth Hormone)

The most versatile superfamilies in single-domain proteins as determined from their functional description in SWISS-PROT, with some representatives. The keyword combinations in the fourth column were based either on the first three components of their EC numbers (for enzymes) or derived automatically by comparing the DE description line of SWISS-PROT entries to a list of synonymous keywords at http://bioinfo.mbb.yale.edu/partslist/func. A keyword number starting with a D indicates an enzyme that does not have an assigned EC number in its description in SWISS-PROT.

the single-domain proteins (22 vs. 11, respectively), careful examination reveals that all the proteins in this category are DNA-binding and most of them act as hormone receptors.

The second entry listed in the table is the combination of the 3.4.1 and 4.48.1 superfamilies associated with the FAD/NAD(P)-linked reductases. It is an all-enzymatic combination and always carries out an oxido-reductase function. All the proteins in this category are completely covered by matches with these two superfamilies. The 1.78.1–2.1.1 hemocyanin-immunoglobulin combination seems also to be fairly conserved; although the proteins in this category are called by eight different names, most of them turn out to be extracellular larval storage proteins, except for the copper-containing oxygen carrier hemocyanin itself (HCY_PALVU).

Following the same logic, we can also determine the probability that two proteins that have the same superfamily combination share the same function, viz:

$$P(F|S) = 1/(1 + 32242/134230) = 81\%$$

This means that we have significantly greater certainty in determining the function of a multi-domain protein with a particular superfamily combination than that of a single-domain protein containing a particular superfamily. We also determined a similar probability for those proteins that have an almost complete coverage with exactly the same type and number of superfamilies, following each other in the same order. The probability that the functions are the same in this case was 91%, a considerably higher value than above. However, if two multi-domain proteins share only a single superfamily, the probability that they share the same function drops to only 35%! This greater functional certainty from sharing a combination of superfamilies rather than just one is also reflected in Table 1. While one-fourth of the single-domain proteins and one-third of singularly matching superfamilies in multi-domain proteins have multiple functions, only about one-fifth of the multi-domain *combinations* possess multiple functions (60 of 281). It is also clear from the data that domains in larger proteins often lose their original function and no longer have an autonomous function.

Seventy Common Superfamilies and Their Functions Compared in Single-Domain and Multi-Domain Proteins

As mentioned above, of the 455 superfamilies in our analysis, only 70 occur in both single- and multi-domain proteins. Even more surprising is the small number of structural superfamilies (14) that have the same function in both single- and

Table 3. Most Versatile Superfamily Combinations in Multi-Domain Proteins

No. func	No. prot	Sfam comb.	Function	SWISS-PROT ID	SWISS-PROT function
22	176	1.95.1/7.33.1	29#	THB_RANCA	Thyroid Hormone Receptor Beta
			10#	HNF4_DROME	Transcription Factor HNF-4 Homolog
			31#32#	EAR2_MOUSE	V-Erba Related Protein Ear-2
			29#30#	ECR_MANSE	Ecdysone Receptor (Ecdysteroid Receptor)
			32#	ERBA_AVIER	Erba Oncogene Protein
			556#564#35#	NGFI_XENLA	Nerve Growth Factor Induced Protein I-B
			576#	NR42_HUMAN	Immediate-Early Response Protein Not
			36#	PPAT_HUMAN	Peroxisome Proliferator Activated Receptor
			37#	RXTG_CHICK	Retinoic Acid Receptor RXR-Gamma
			38#	TLL_DROVI	Tailless Protein
8	54	3.4.1/4.48.1	E1.8.2	DHSU_CHRVI	Sulfide Dehydrogenase (1.8.2.–)
			E1.8.1	DLDH_ZYMMO	Dihydrolipoamide Dehydrogenase (1.8.1.4)
			E1.6.4	TYTR_TRYCR	Trypanothione Reductase (1.6.4.8) (Tr)
			E1.16.1	MERA_STRLI	Mercuric Reductase (1.16.1.1)
			E1.6.99	NAOX_MYCPN	Probable NADH Oxidase (1.6.99.3) (Noxase)
8	23	1.78.1/2.1.1	19#	ARYB_MANSE	Arylphorin Beta Subunit–(Tobacco Hornworm)
			20#	CRPI_PERAM	Allergen Cr-Pi Precursor–(American Cockroach)
			21#427#	HCY_PALVU	Hemocyanin–(European Spiny Lobster)
			22#	HEXA_BLADI	Hexamerin Precursor–(Tropical Cockroach)
			23#	JSP1_TRINI	Acidic Juvenile Hormonne–Suppressible Protein
			24#	LSP2_DROME	Larval Serum Protein 2 Precursor (LSP-2)
			546#25#	SSP1_BOMMO	Sex-Specific Storage–Protein 1

Note that the combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Careful examination reveals that all the proteins with this combination are DNA-binding and most of them act as various hormone receptors. In particular, HNF4_DROME and NR42_HUMAN also have transcription activator functions. Note that these two proteins are considerably longer than the others in this group and are not covered completely by structural matches: A large C-terminal and a large N-terminal portion are left uncovered, respectively.

multi-domain proteins. These are listed in Table 4; 12 of them have enzymatic function, supporting the notion that enzymes are more conserved during evolution than non-enzymes. The two non-enzymatic superfamilies are the 4.29.1 ribosomal superfamily and the 5.4.1 superfamily in penicillin-binding proteins.

Table 5 presents several examples of the converse situation, shared superfamilies that have different functions in single and multi-domain proteins. Comparing parts A and B of the table highlights the fact that although both superfamilies in a multi-domain protein are often present in single-domain form as well, the functions in the different settings are only vaguely related. One example is the combination of the lipocalin superfamily (2.45.1) with that of the BPTI-like or Kunitz inhibitor (7.7.1), which in higher organisms forms a complex protein called alpha-1-microglobulin (AMBP_RAT). Another interesting example is the combination of the 2.5.1 Cupredoxin (occurring in the single-domain blue-copper protein, SOXE_SULAC) and the 6.5.1 Membrane all-alpha (single-domain representative: BACT_HALVA, a sensory rho-

Table 4. Superfamilies With the Same Function in Single- and Multi-Domain Proteins as Determined from Their Keyword Combination or First Three Components of Their EC Numbers

		Single-domain proteins		Multi-domain proteins	
Sfam	Function	SWISS-PROT ID	SWISS-PROT function	SWISS-PROT ID	SWISS-PROT function
1.81.1	E3.2.1	GUNY_ERWCH	Endoglucanase (3.2.1.4)	AMYG_NEUCR	Glucoamylase Precursor (3.2.1.3)
2.66.2	E3.5.1	URE2_YERPS	Urease Beta (3.5.1.5)	URE1_HELPY	Urease Alpha Subunit (3.5.1.5)
3.17.2	E6.3.5	NADE_MYCPN	NAD(+) Synthetase (6.3.5.1)	GUAA_YEAST	GMP Synthase (6.3.5.2)
3.37.1	E3.1.3	PTP2_NPVOP	Protein-Tyrosine Phosphatase 2 (3.1.3.48)	PTNB_RAT	Protein-Tyrosine Phosphatase (3.1.3.48)
3.67.1	E4.2.1	TRPB_VIBPA	Tryptophan Synthase (4.2.1.20)	TRP_YEAST	Tryptophan Synthase (4.2.1.20)
4.19.1	E5.2.1	FKB1_METJA	Peptidylprolyl *Cis-Trans* Isomerase (5.2.1.8)	FKB7_WHEAT	70 Kd Peptidylprolyl Isomerase (5.2.1.8)
4.2.1	E3.2.1	LYCV_BPP2	Lysozyme (3.2.1.17)	CHIX_PEA	Endochitinase Precursor (3.2.1.14)
4.29.1	85#	RS5_ACYKS	30s Ribosomal Protein S5	RS5_TREPA	30s Ribosomal Protein S5
4.52.1	E3.4.24	SNPA_STRCS	Extracellular Neutral Protease (3.4.24.–)	BMPH_STRPU	Collagenase 3 Precursor (3.4.24.–)
4.6.1	E3.5.1	URE3_YERPS	Urease Gamma (3.5.1.5)	URE1_HELPY	Urease Alpha Subunit (3.5.1.5)
5.10.1	E2.7.7	KANU_STAAU	Kanamycin Nucleotidyltransferase (2.7.7.–)	DPOB_XENLA	Dna Polymerase Beta (2.7.7.7)
5.4.1	161#	AMPH_ECOLI	Penicillin-binding Protein Amph	PBPX_STRPN	Penicillin-binding Protein 3x Pbp2x

Table 5. Examples of Superfamilies Present in Both Single- and Multi-Domain Proteins, Carrying out Different Functions

Table 5A. Single-Domain Proteins

Sfam	Funct #	SWISS-PROT ID	SWISS-PROT function
1.25.1	352# 183# E1.17.4 192#	FTN2_HAEIN NIGY_DESVH RIR4_YEAST NLP_HAEIN	Ferritin-like Protein 2 Nigerythrin (Ribonucleotide Reductase) (1.17.4.1) Ner-like Protein Homolog
1.4.3	196#	H1A_PLADU	Histone H1A, Sperm
1.81.2	E2.5.1	PFTB_PEA	Farnesyltransferase Beta Su (2.5.1.–)
2.45.1	226# 227# 228#412# 229# E5.3.99 230#421#	ERBP_RAT FAB3_CAEEL NGAL_MOUSE NP4_RHOPR PGHD_HUMAN VNS1_MOUSE	Epididymal-Tetinoic Acid Binding Protein Fatty Acid-Binding Protein Homolog 3 Neutrophil Gelatinase-Assoc. Lipocalin Nitrophorin 4 Precursor Prostaglandin-H2 D-Isomerase (5.3.99.2) Vesomeral Secretory Protein I
2.5.1	231# 232#427#	MPA3_AMBEL SOXE_SULAC	Pollen Allergen AMB A 3 (AMB A Iii) Sulfocyanin (Blue Copper Protein)
3.14.2	373#	RRF1_DESVH	Rrf1 Protein
3.29.1	E6.3.4 E2.7.4 D259# E2.7.1	PURA_CAEEL KTHY_YEAST VA57_VACCV KITH_VZVW	Adenylosuccinate Synthetase (6.3.4.4) Thymidylate Kinase (2.7.4.9) Guanylate Kinase Homolog Thymidine Kinase (2.7.1.21)
3.47.1	275# 276#	MBL_BACSU MREB_BACSU	MBL Protein Rod Shape-determining Protein Mreb
3.48.1	E3.1.3	PPA5_YEAST	Repressible Acid Phosphatase (3.1.3.2)
3.81.1	D281# 282#	AMIC_PSEAE LUXP_VIBHA	Aliphatic Amidase Expression-Regulator LUXP Protein Precursor
4.103.1	E2/4/2	TOX1_BORPE	Pertussis Toxin Su 1 (2.4.2.–)
4.105.1	291#	LECC_POLMI	Lectin–Polyandrocarpa Misakiensis
4.11.5	295#	TERP_PSESP	Terpredoxin
4.19.1	E5.2.1	FKB1_METJA	Pept-Prolyl *Cis-Trans* Isomerase (5.2.1.8)
6.5.1	E3.6.1 540#325#	ATPL_VIBAL BACT_HALVA	ATP Synthase (3.6.1.34) (Lipid-binding) Sensory Rhodopsin II (Sr-li)
7.35.4	E1.9.3 345#	COXB_RAT DESR_DESBI	Cytochrome C Oxidase (1.9.3.1) (Via*) Desulforedoxin (Dx)
7.7.1	349#	TAP_ORNMO	Tick Anticoagulant Peptide

(*Table continues on following page.*)

dopsin) superfamilies into a component of the respiratory chain, cytochrome C oxidase II (COOX_ZOOAN). All these examples demonstrate the evolutionary advantage of a domain fusion event, which creates a function that is more complex than either of the components.

Multifunctionality vs. Sequence Similarity

Previously, we presented a variety of graphs that show how the probability that two domains would share the same function varied with respect to sequence similarity (Hegyi and Gerstein 1999; Wilson et al. 2000). Figure 4 shows a similar graph with the calculations extended to multi-domain proteins. The figure shows that the functional divergence of a single domain in multi-domain proteins dramatically increases, more than twofold, compared to the single-domain ones. This reinforces our findings above, based only on superfamily content, that the certainty with which we can predict the function of a protein based on its sequence similarity with a domain in another multi-domain protein, is considerably less than for a comparable single-domain situation.

Table 5B. Multi-Domain Proteins

Sfam Comb.	Funct#	SWISS-PROT ID	SWISS-PROT function
1.25.1/7.35.4	104#	RUBY_METJA	Putative Rubrerythrin
1.32.1/3.81.1	11#	PURR_HAEIN	Purine Nucleotide Synthesis Repressor
	12#	DEGA_BACSU	Degradation Activator
	581#11#	SCRR_STRMU	Sucrose Operon Repressor
	582#11#	REGA_CLOAB	Transcription Regulatory Protein Rega
1.4.3/3.14.2	10#	SKN7_YEAST	Transcription Factor Skn7 (Pos9 Protein)
	11#	VIRG_AGRT5	Virg Regulatory Protein
	13#	RGX3_MYCTU	Sensory Transduction Protein REGX3
	190#	PFER_PSEAE	Transcriptional Activator Protein Pfer
	366#	PETR_RHOCA	Petr Protein
2.45.1/7.7.1	203#153#	HC_RAT	Alpha-1-Microglobulin/Trypsin Inhibitor
2.5.1/6.5.1	E1.9.3	COX2_ZOOAN	Cytochrome C Oxidase Ii (1.9.3.1)
3.29.1/3.48.1	E2.7.1	F26_RANCA	6-Phosphofructo-2-Kinase (2.7.1.105)
3.47.1/5.17.1	1#	YED0_YEAST	Heat Shock Protein 70 Homolog YEL030w
	1#83#	GR73_MAIZE	Ig-Binding Protein

DISCUSSION

Here we built on our previous studies on the relationship between protein structure and function to develop new results related to multi-domain proteins. Throughout the paper, we focused on superfamilies instead of folds, as the members of a superfamily are presumably of common evolutionary origin (Murzin et al. 1995).

We found that the 4763 multi-domain and 1818 single-domain proteins that met our selection criteria have about the same distribution of structural classes, with more enzymatic functions associated with the alpha/beta structural classes and more non-enzymatic ones with the all-alpha and small classes. We identified more than three times as many multi-domain proteins that were enzymes than single-domain ones (2805 and 850, respectively) and, conversely, about twice as many multi-domain proteins as single-domain ones that were non-enzymes (1958 vs. 968).

We focused on the functional divergence of the two groups and found that about a quarter of the superfamilies in single-domain proteins are associated with multiple functions, whereas only about a fifth of the multi-domain superfamily combinations are. Therefore, we can conclude that a combination of specific superfamilies results in a more specific functional assignment for a particular protein. However, about one-third of the superfamilies in the multi-domain proteins were associated with multiple functions, underlining the lesser autonomy of a domain function in multi-domain protein.

This latter finding was also supported by the difference in functional divergences between the two groups of proteins based on particular sequence similarities between the domains and SWISS-PROT proteins. As is shown in Figure 4, the average functional divergence of a single domain is much larger (more than twofold) in multi-domain proteins than in single-domain ones.

We also found that only 70 of a total of 455 superfamilies are shared between the multi-domain and single-domain proteins and only a small fraction (14) share their functions. This was rather surprising to us, and should be taken into consideration in functional characterization and annotation of new gene products. When the functions were related in single- and multi-domain proteins, we could observe an increasing functional complexity with the appearance of large multi-domain proteins.

Altogether, with the recent sequencing of the human genome and the genomes of other model organisms, we hope

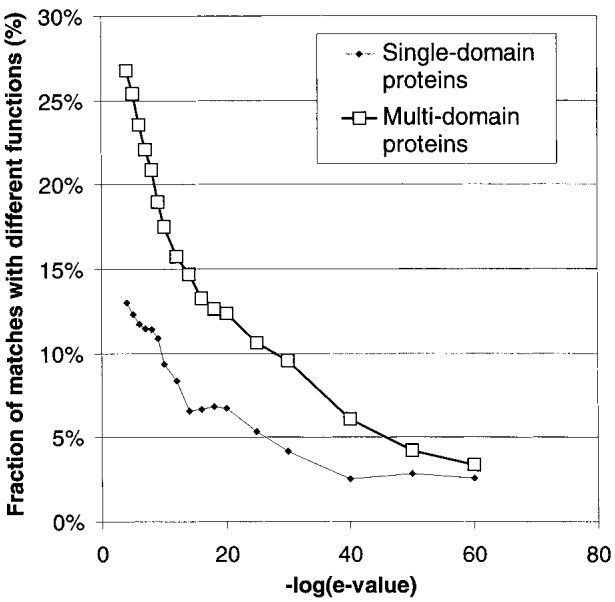

Figure 4 Divergence in function with respect to sequence similarity. Relative number of matching domains with multiple functions, as the function of *e*-value threshold. Diamonds represent single-domain proteins, squares multi-domain ones (matching just for a single domain), respectively. The first value on the *X*-axis starts at 4 (corresponding to an *e*-value=10^{-4}).

that this work can contribute to the successful annotation of the individual gene products, and will help to avoid some pitfalls associated with the functional characterization of large, complex proteins.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* **28:** 304–5.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–8.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41:** 98–107.

Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301:** 1059–1075.

Eisenstein, E., Gilliland, G. L., Herzberg, O., Moult, J., Orban, J., Poljak, R. J., Banerjei, L., Richardson, D. and Howard, A. J. 2000. Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. *Curr. Opin. Biotechnol.* **11:** 25–30.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Russo Twombly, S., Emmert, D., Ashburner, M., Drysdale, R. A., et al. 1997. FlyBase: A *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res.* **25:** 63–6.

Gerstein, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274:** 562–76.

———. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* **3:** 497–512.

Harrison, P. Echols, N. and Gerstein, M. 2001. Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *C. elegans* genome. *Nucleic Acids Res.* **29:** 818–830.

Hegyi, H. and 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.

Lin, J. and Gerstein, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Res.* **10:** 808–818.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. 1998. Protein folds and functions. *Structure* **6:** 875–884.

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

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L. and Thornton, J. M. 1999. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27:** 275–279.

Pawlowski, K., Jaroszewski, L., Rychlewski, L. and Godzik, A. 2000. Sensitive sequence comparison as protein function predictor. *Pac. Symp. Biocomput.* 42–53.

Pearson, W. R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **25:** 365–389.

Qian, J., Stenger, B., Wilson, C., Lin, J., Jansen, R., Krebs, W., Alexandrov, V., Echols, N., Teichmann, S., Park, J. et al.. 2001. PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.* **29:** 1750–1764.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. and Sternberg, M. J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269:** 423–439.

Shah, I. and Hunter, L. 1997. Predicting enzyme function from sequence: A systematic appraisal. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 276–283.

Shapiro, L. and Harris, T. 2000. Finding function through structural genomics. *Curr. Opin. Biotechnol.* **11:** 31–5.

Stawiski, E.W., Baucom A.E., Lohr S.C., and Gregoret L.M. 2000. Predicting protein function from structure: Unique structural features of proteases. *Proc. Natl. Acad. Sci.* **97:** 3954–8.

Thornton, J. M., Orengo, C. A., Todd, A. E. and Pearl, F. M. 1999. Protein folds, functions and evolution. *J. Mol. Biol.* **293:** 333–342.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. and Orengo, C. A. 2000. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7 Suppl**: 991–994.

Todd A.E., Orengo C.A., and Thornton J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307:** 1113–1143.

Wilson, C. A., Kreychman, J. and 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.

Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. and Godzik, A. 1999. From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8:** 1104–1115.

Received February 7, 2001; accepted in revised form June 19, 2001.