The proteome: structure, function and evolution

Keiran Fleming¹, Lawrence A. Kelley^{1,2}, Suhail A. Islam^{1,2},
Robert M. MacCallum^{2,†}, Arne Muller^{1,2,‡}, Florencio Pazos^{1,¶}
and Michael J. E. Sternberg^{1,2,*}

¹Structural Bioinformatics Group, Centre for Bioinformatics, Division of Molecular Biosciences,
Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK

²Biomolecular Modelling Laboratory, Cancer Research UK, 44 Lincoln's Inn Fields, London
WC2A 3PX, UK

This paper reports two studies to model the inter-relationships between protein sequence, structure and function. First, an automated pipeline to provide a structural annotation of proteomes in the major genomes is described. The results are stored in a database at Imperial College, London (3D-GENOMICS) that can be accessed at www.sbg.bio.ic.ac.uk. Analysis of the assignments to structural superfamilies provides evolutionary insights. 3D-GENOMICS is being integrated with related proteome annotation data at University College London and the European Bioinformatics Institute in a project known as e-protein (<http://www.e-protein.org/>). The second topic is motivated by the developments in structural genomics projects in which the structure of a protein is determined prior to knowledge of its function. We have developed a new approach PHUNCTIONER that uses the gene ontology (GO) classification to supervise the extraction of the sequence signal responsible for protein function from a structure-based sequence alignment. Using GO we can obtain profiles for a range of specificities described in the ontology. In the region of low sequence similarity (around 15%), our method is more accurate than assignment from the closest structural homologue. The method is also able to identify the specific residues associated with the function of the protein family.

Keywords: bioinformatics; proteome annotation; protein function

1. INTRODUCTION

Knowledge of the structure and function of the proteome is central to the exploitation of the wealth of biological information available in the post-genome era. This knowledge provides fundamental understanding of biological processes and can inform the systematic development of novel pharmaceuticals (e.g. see in this volume see (Blundell *et al.* 2005)). Bioinformatics has a central role in providing this knowledge by modelling the inter-relationships between the sequences, structures and functions of proteins. Moreover, bioinformatics provides a tool to explore evolutionary relationships between component molecules within different species.

This paper reports two studies from our laboratory to model protein sequence–structure–function inter-relationships. First, we describe the development of an automated pipeline for proteome annotation (Muller *et al.* 2002; Fleming *et al.* 2004). Analysis reveals the

extent of structural and functional annotation of the proteomes. Comparative studies provide insights into the evolution of the proteomes in the different species. The second topic addresses the problem that, as a result of structural genomics initiatives, increasingly protein structures are being determined without prior knowledge of their function (e.g. Goldsmith-Fischman & Honig 2003; Kinoshita & Nakamura 2003; Laskowski *et al.* 2003). We report a new method, PHUNCTIONER (Pazos & Sternberg 2004), to assign protein function to a structure.

2. PROTEOME ANNOTATION

(a) Strategy for annotation

3D-GENOMICS (www.sbg.bio.ic.ac.uk/3dgenomics) is a database containing structural annotations for fully sequenced proteomes (Muller *et al.* 2002; Fleming *et al.* 2004). At the time of writing, the database contains data for 173 species, including 18 eukaryotes, 17 archaea and 138 bacteria. The strategy for annotation is as follows: as a first step, we identify sequence features such as signal peptides (IPSORT, Bannai *et al.* 2002), transmembrane regions (HMMTOP, Tusnady & Simon 2001), coiled coils (COILS, Lupas *et al.* 1991), repeats (PROSPERO, Mott 2000) and low complexity regions (SEG, Wootton & Federhen 1996). PSI-BLAST (Altschul *et al.* 1997) is then utilized to infer homology to sequences in a database of all available non-redundant sequences, including those for which a

* Author for correspondence (m.sternberg@imperial.ac.uk).

† Present address: Stockholm Bioinformatics Center, AlbaNova University Center, Stockholm University, 106 91 Stockholm, Sweden.

‡ Present address: Bioinformatics, Sanofi Aventis, 13 quai Jules Guesde, 94403 Vitry-sur-Seine Cedex, France.

¶ Present address: Protein Design Group, Centro Nacional de Biotecnología (CNB-CSIC), Campus Universidad Autónoma, Cantoblanco, 28049 Madrid, Spain.

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

Figure 1. Current state of annotation in the 3D-GENOMICS database for 14 representative proteomes from the three branches of life: eukaryotes, archaea and bacteria. Coverage is reported as the percentage of residues in each proteome that are annotated. Structural annotations represent homology to a known structure from the PDB. Functional annotations have no structural annotation but homology is detected to an entry from SwissProt with a description that does not contain the words: hypothetical, probable, predicted, putative. Any homology represents a sequence match to a SwissProt entry containing any of these words. Non-globular regions are those predicted as low complexity, coiled coil, low complexity or transmembrane. Finally, orphan regions do not have a match to any sequence in the NR database.

three-dimensional structure is available; i.e. structural classification of proteins (SCOP; Andreeva *et al.* 2004) and the protein data bank (PDB; Bourne *et al.* 2004). The way in which we use PSI-BLAST has been tailored to allow detection of remote homologues (Müller *et al.* 1999). Masking of sequence features, identified above, prior to running PSI-BLAST aids detection of true homologues and reduces the number of false positives introduced into the position-specific scoring matrix (PSSM). Additionally, the sequences in the PSSM at the end of each iteration are inspected for signs of drift (i.e. sequences erroneously added to the PSSM causing a gradual drift away from true homologues). If drift is found to occur, PSI-BLAST parameters are altered in an iterative manner until either a threshold is reached or no drift occurs. Benchmarking against SCOP sequences has enabled the selection of optimal PSI-BLAST parameters for proteome annotation. PSI-BLAST-detected relationships are not symmetric (i.e. sequence A might not have a significant hit to sequence B but sequence B could have a significant hit to sequence A). As such, we additionally use IMPALA to assign SCOP/PDB sequences to genome sequences. Functional annotation was achieved by using HMMer (Eddy 1998) to identify matches to domains in the rapidly expanding PFAM protein families database (Bateman *et al.* 2004). Links to cluster of orthologous groups domains provide additional information about function (Tatusov *et al.* 2003). Three-dimensional homology models have been recently introduced, and are currently available for sequences in 13 proteomes. Our plans are to use the next version of our fold recognition program (PHYRE; Bennett-Lovsey *et al.* work in progress; <http://www.sbg.bio.ic.ac.uk/~phyre>) to predict structures for remote homologues undetectable by sequence-based methods.

(b) Current status of annotations

Statistics for the current state of annotations of 14 proteomes, containing representatives from the three branches of life, can be seen in figure 1. The status of structural and functional annotations is expressed as a percentage of the total number of residues in a proteome. For the four eukaryotic proteomes shown, approximately 30–40% of the residues can be structurally annotated via homology to SCOP domains or PDB sequences. For the prokaryotic proteomes, the equivalent figure is higher at approximately 45–50% of the total number of residues in each proteome. However, the fraction of residues that can be ascribed a functional annotation is similar for both the eukaryotes (*ca* 25–35%) and the prokaryotes (*ca* 25–40%). These figures have undergone only small increases of a few percent compared to the study we undertook 3 years ago (Muller *et al.* 2002), despite increases in the number of PDB sequences available. There may be several reasons for this. The first is that, despite the best efforts of structural genomics initiatives, the majority of PDB structures published each month do not describe novel folds but are often homologous to other structures already in the database. Additionally, the accuracy of our analysis is based on the accuracy of gene prediction for each genome. Genes identified, particularly in the eukaryotes, over the past few years have been those more likely to fall into the ‘orphan’ bin in our figure due to the way in which genes are predicted: the ‘easy’ proteins with homology to known proteins are identified early on during sequencing.

(c) Reliability of annotation

Figure 2 provides a measure of the reliability of functional annotations. Assignment of PFAM domains, except those with unknown function,

Figure 2. Reliability of functional annotations for 14 representative proteomes. The bins are cumulative. The first three bins represent 'reliable' annotations, where sequence identity to a structural hit is above 30% or has a PFAM hit of known function. The last three bins represent 'fuzzy' annotations, where sequence identity to a non-structural hit is above 30% and some functional annotation is available.

represents a reliable functional annotation. Approximately 30% of the residues in eukaryotic genomes match a PFAM domain of known function, with approximately 20% of these also having homology to a known structure. This figure increases to 43–58% for prokaryotic species. For these residues, we can be reasonably confident in assigning a function, and clearly this represents a high proportion of each proteome. Below 30% sequence identity, the ability to infer function becomes more complicated since it is widely accepted that homologues often have diverged to perform completely different functions below this threshold.

Compared to equivalent figures from 3 years ago (Muller *et al.* 2002), it can be seen that the reliability of functional annotation has increased substantially. This is primarily due to the expansion of the high-quality PFAM database. Moreover, it reveals that despite only very small increases in the percentages of residues for which we have structural or functional annotations, the reliability of these annotations is increasing at a much greater rate. Indeed, the ability to map biochemical function to sequence is arguably of primary importance in sequence annotation, since it provides additional biological value beyond that afforded by traditional structure annotation. Our results are, therefore, encouraging and we intend to incorporate new databases, such as gene ontology (GO; Harris *et al.* 2004), that provide annotations of proteome sequences using a common set of terms to describe function. Presumably, this will lead to further increases in the coverage reported above, leading to an improved 3D-GENOMICS resource for users.

(d) Transmembrane proteins

Figure 3 provides the latest figures for the proportion of residues predicted to occur in transmembrane proteins. In addition to the metazoan proteomes of human, worm and fly, we provide average figures for all of the yeast, bacteria, and archaea in the 3D-GENOMICS database. In human and fly, approximately 21% of residues can be found in transmembrane proteins, with 5% of these in actual membrane spanning regions or the short loops between them, while the remainder of the residues form globular domains attached to these transmembrane regions. The worm proteome displays a higher proportion of transmembrane residues, and this has been linked to expansion of a few families of 7-transmembrane helical proteins that act as chemoreceptors (Liu *et al.* 2002). Due to its inability to see or hear, chemosensation is the main method by which *Caenorhabditis elegans* detects its environment. In addition, a greater proportion of residues are involved in membrane-spanning regions, but this is as expected considering the expanded families are 7-transmembrane.

The bacterial species have on average 20% of their residues in transmembrane proteins, but a large proportion of these residues (7%) are found in the membrane spanning or short loop regions with the remaining 13% being buried within the membrane. Similarly, 6% out of the 16% transmembrane protein residues in archaea are membrane spanning or in loops. This relative higher proportion of residues may be explained by a larger fraction of completely membrane integral proteins in these species (Muller *et al.* 2002).

Figure 3. Fraction of transmembrane proteins in proteomes. The first bin represents globular (non-membrane) proteins. The remaining three bins represent the percentage of residues involved in the globular region of transmembrane proteins, the small loops joining transmembrane regions, and the transmembrane helices themselves. The bar chart does not total 100% as certain non-globular regions such as signal peptides and coiled-coils are not included.

(e) SCOP superfamilies

Table 1 shows the number of domains found in each proteome for the 12 most commonly occurring SCOP superfamilies in the human proteome. For the metazoan species, i.e. eukaryotes not including yeast, the most common superfamilies are similar, with the top 12 in human mapping to those in at least the top 25 of worm and fly. In contrast, the prokaryotic species (i.e. bacteria and archaea) comprise a completely different mix of superfamilies apart from the ubiquitous P-loop containing nucleotide triphosphate hydrolases, and to a lesser extent, the membrane all-alpha superfamily. This implies that the most common superfamilies reflect functions common to all higher eukaryotes. Indeed, this is borne out when one considers that these superfamilies are involved in functions such as cell-cell interaction, signalling, and cell-surface receptors. The observation that C-type lectins are observed at high frequency in the worm genome with respect to human and fly was observed previously (Muller *et al.* 2002), but no explanation was available. Proteins of the C-type lectin superfamily are carbohydrate binding and involved in immune cell signalling. Reports (Loukas & Maizels 2000) have implicated proteins in this superfamily in evading the host immune system in parasitic nematodes and although *C. elegans* is a free-living nematode, it is closely related to these parasites. We may be observing a superfamily expansion that occurred early in the nematode phylogenetic branch, allowing a parasitic lifestyle. Comparison with other parasitic and free-living nematodes is required to investigate this hypothesis more thoroughly.

The most common superfamilies in bacteria and archaea are very different (see table 2). These superfamilies are predominantly involved in nucleic acid binding (e.g. the winged helix DNA binding domain, homeodomain-like, and nucleic-acid binding protein) or enzyme cofactor binding (e.g. NAD(P)-binding

Rossmann-fold domain and FAD/NAD(P)-binding domain). The absence of superfamilies involved in cell-cell interactions or cell-surface proteins is unremarkable and clearly reflects the absence of multicellular organization. Indeed, the unicellular yeast have more in common with the prokaryotic species than they do with the multicellular eukaryotes in terms of commonality of structural superfamilies. This has also been discussed previously and we observe no differences from our earlier conclusions (Muller *et al.* 2002).

(f) Expansion of SCOP superfamilies

Here, we discuss the relative expansion of SCOP superfamilies that provide evidence for particular evolutionary functional adaptations made, and allow an estimation of when these adaptations might have occurred.

Figure 4a gives an indication of the extent of domain duplication based on the expansion of SCOP superfamilies for the 10 most common superfamilies in humans. Taking the frequency in humans as a reference (i.e. at a value of 100%), it can be seen that for the majority of these SCOP superfamilies expansion has occurred between fly/mosquito and *Fugu*. This is as expected given the approximate doubling in proteome size between these species. For the *Fugu* species, several superfamilies are observed in slightly greater numbers than in the human proteome. However, the current estimated size of the *Fugu* proteome stands at approximately 33 500 proteins, whereas the human genome is at approximately 26 500 sequences. One of the quirks of whole-genome shotgun sequencing is that overestimation of true genome size occurs. It may be that the *Fugu* genome is actually smaller than this, and therefore we may be overestimating the number of SCOP superfamily members in this genome.

One clear trend is the sudden expansion of the C2H2 and C2HC zinc finger domains from *Fugu* to mouse and human, given the roughly equal genome

Table 1. Commonly occurring SCOP superfamilies relative to the top 12 most abundant superfamilies in the human proteome. (R, rank of a superfamily within a proteome; N, number of domains within a proteome.)

SCOP superfamily	human		mouse		worm		fly		yeast		archaea		bacteria	
	N	R	N	R	N	R	N	R	N	R	N	R	N	R
C2H2 and C2HC zinc fingers	4971	1	9767	1	216	11	865	1	50	11	1	359	—	—
immunoglobulin	2182	2	2099	2	755	1	576	3	19	44	3	102	4	76
EGF/laminin	1226	3	1196	4	488	4	381	4	—	—	—	—	—	—
P-loop containing NTP hydrolases	1091	4	1220	3	684	2	629	2	404	1	149	1	149	1
fibronectin type III	908	5	954	5	307	7	231	8	3	263	1	359	2	175
cadherin	900	6	880	6	152	22	212	9	—	—	—	—	—	—
membrane all-alpha	737	7	849	7	246	10	143	21	35	17	9	29	12	23
protein kinase-like (PK-like)	642	8	674	8	548	3	258	5	131	2	4	69	5	73
RNA-binding domain, RBD	541	9	554	9	264	9	255	6	123	4	—	—	—	—
PH domain-like	467	10	450	10	146	25	145	20	28	25	—	—	—	—
EF-hand	421	11	341	15	151	23	165	14	31	20	2	159	1	322
ankyrin repeat	393	12	388	11	182	16	147	19	21	35	—	434	1	251

sizes. This may be explained by the fact that the *Fugu* genome, although containing a similar amount of coding regions to the human genome, is actually only one-eighth of its size. As such, such an expansion in zinc fingers may point to an increased need for processes associated with transcriptional regulation and mRNA stability in genomes containing a lot of 'junk' DNA. The apparent doubling of zinc finger domains in the mouse genome, with respect to the human genome, is particularly puzzling, and we can offer no explanation as to why this should be.

P-loops and the RNA-binding domain display the smallest increase in domain expansion across the board from bacteria to higher eukaryotes, but this is unsurprising given their central roles in housekeeping.

The rice genome is particularly interesting, with the majority of the common superfamilies in human observed at low frequency here. However, the protein kinase-like superfamily, in particular, has undergone major expansion and indeed is the most common SCOP superfamily in rice overall. The protein kinase-like superfamily is involved in the majority of signalling and regulatory processes in the eukaryotic cell and our results presumably reflect the different requirements for a plant cell.

Figure 4b shows the relative domain frequencies expressed as the fraction of the total assigned SCOP domains for the top 10 most common superfamilies in the human genome. Zinc fingers account for greater than 25% of all domains in mouse, and over 15% in humans, so are obviously utilized to a huge degree. At an average length of just 27 residues, this corresponds to just over 1% of the residues in the human proteome and 1.75% in the mouse proteome. Their abundance may be due to their versatility coupled with the low costs in terms of cellular resources expended during their production. In contrast, the relative frequencies of P-loops decrease from the unicellular to multicellular species, exposing the increasing importance of other processes during adaptation to a metazoan lifestyle. In rice, an increased frequency of PK-like domains and reduced numbers of many of the other domains associated with multicellular eukaryotes highlights its status as the only plant in our study.

With the exception of the zinc fingers, the relative frequencies of the remaining superfamilies are fairly constant across all of the animal proteomes. These results point to the general trend in metazoans that for the most popular superfamilies in a proteome, the relative abundance of those superfamilies is not substantially different in other proteomes, i.e. the expansion of the most common superfamilies between proteomes is proportional to the increase in overall proteome size.

(g) *e-Protein: a distributed proteome annotation pipeline*

Proteome annotation methodology is not well established and accordingly several groups throughout the world are exploring different procedures and generating other protein annotation databases (e.g. PEP, Carter *et al.* 2003; SUPERFAMILY, Madera *et al.* 2004; WILMA, Prlic *et al.* 2004). Indeed, this volume reports the Gene3D database by the Orengo

Table 2. Top 12 most commonly occurring SCOP superfamilies for the groups of organisms under study. (The ranking of the superfamilies is based on the number of observations of a particular superfamily domain in a proteome. Rankings were based on average figures for the yeast, archaea and bacteria.)

rank	human	mouse	worm	fly	yeast	archaea	bacteria
1	C2H2 and C2HC zinc fingers	C2H2 and C2HC zinc fingers	immunoglobulin	C2H2 and C2HC zinc fingers	P-loop containing NTP hydrolases	P-loop containing NTP hydrolases	P-loop containing NTP hydrolases
2	immunoglobulin	immunoglobulin	P-loop containing NTP hydrolases	P-loop containing NTP hydrolases	protein kinase-like (PK-like)	winged helix DNA-binding domain	NAD(P)-binding Rossmann-fold domains
3	EGF/laminin	P-loop containing NTP hydrolases	protein kinase-like (PK-like)	immunoglobulin	Trp-Asp repeat (WD-repeat)	NAD(P)-binding Rossmann-fold domains	winged helix DNA-binding domain
4	P-loop containing NTP hydrolases	EGF/laminin	EGF/laminin	EGF/laminin	RNA-binding domain, RBD	S-adenosyl-L-methionine-dependent methyltransferases	periplasmic binding protein-like II
5	fibronectin type III	fibronectin type III	glucocorticoid receptor-like (DNA-binding domain)	protein kinase-like (PK-like)	FliG	4Fe-4S ferredoxins	S-adenosyl-L-methionine-dependent methyltransferases
6	cadherin	cadherin	C-type lectin-like	RNA-binding domain, RBD	NAD(P)-binding Rossmann-fold domains	CBS-domain	homeodomain-like
7	membrane all-alpha	membrane all-alpha	fibronectin type III	trypsin-like serine proteases	TPR-like	PYP-like sensor domain	CheY-like
8	protein kinase-like (PK-like)	protein kinase-like (PK-like)	nuclear receptor ligand-binding domain	fibronectin type III	actin-like ATPase domain	FAD/NAD(P)-binding domain	FAD/NAD(P)-binding domain
9	RNA-binding domain, RBD	RNA-binding domain, RBD	RNA-binding domain, RBD	cadherin	thioredoxin-like	nucleic acid-binding proteins	nucleic acid-binding proteins
10	PH domain-like	PH domain-like	membrane all-alpha	FliG	S-adenosyl-L-methionine-dependent methyltransferases	nucleotide-diphosphate sugar transferases	ATPase domain of HSP90 chaperone/DNA topoisomerase II
11	EF-hand	ankyrin repeat	C2H2 and C2HC zinc fingers	spectrin repeat	C2H2 and C2HC zinc fingers	PLP-dependent transferases	PLP-dependent transferases
12	ankyrin repeat	spectrin repeat	homeodomain-like	trp-asp repeat (WD-repeat)	nucleic acid-binding proteins	thiamin diphosphate-binding fold (THDP-binding)	alpha/beta-hydrolases

group (Buchan *et al.* 2003; Marsden *et al.* 2005). Prof. Jones at University College London is applying his protein sequence analysis tools to proteomes to generate the genomic threading database (McGuffin *et al.* 2004). Closely related to these annotations is the work from Prof. Thornton's group on using their catalytic site atlas to assigning protein function to protein structures (Porter *et al.* 2004). The aim of e-Protein is to provide a single front end to the 3D-GENOMICS, Gene3D and GenTHREADER databases and to augment this with functional annotations from the Thornton group. In addition to integrating the information in individual databases into a single resource, comparison of the results from the different annotations for the same protein will provide a method to estimate the confidence an end user should place on the results.

The e-Protein project makes the annotations from each of the contributing sites available through the distributed annotation system (DAS) server (Dowell *et al.* 2001; <http://www.biodas.org/>). This way, no central repository of feature annotations is necessary, and each site has full control over the updates and modification to their data. Additionally, no agreement on database schema is needed as a DAS annotation server is database and schema agnostic. The current status is that the DAS server is operational and provides the required single front end to distributed proteome annotation resources. The client is easily accessible to the community from <http://www.e-protein.org/e-proteinindastypr.html> (must have Flash installed) or via the main page of the e-protein web site (www.e-protein.org).

Bioinformatics research has benefited from the availability of low-cost desktop computers, which have been used to build large processor farms. However, costs and the physical demands for space, power and cooling limit the size of these clusters at a single location. The recent advances in GRID technology are now providing proper management and security tools that make it a practical solution to share the available resources across different sites (Foster & Kesselman 2004; Hey & Trefethen 2005). The other aspect of e-Protein is to develop and apply GRID technology to establish the protocol for distributed computing for proteome annotation.

3. PREDICTION OF FUNCTION FROM STRUCTURE

(a) *Structural genomics and the need for function prediction*

Current structural genomics projects are yielding structures for proteins whose functions are unknown. (Goldsmith-Fischman & Honig 2003; Kinoshita & Nakamura 2003; Laskowski *et al.* 2003). Today there are more than 500 proteins annotated as 'hypothetical' in PDB (roughly one per 50 entries; Bourne *et al.* 2004). This demonstrates that improvements are urgently required in methods to assign function to proteins both just from their sequence and after the structure has been determined.

Assignment of protein function is complicated, for reviews see (Laskowski *et al.* 2003; Whisstock & Lesk

2003; Jones & Thornton 2004). Ultimately, one is interested in function at the level of the phenotype but an important step towards this goal is the identification of molecular function. The complexity of protein function makes the establishment of any functional classification complex. One major advance in providing a bioinformatics resource to describe function is the classifications used in the GO project (Harris *et al.* 2004). GO provides a functional hierarchy that progresses from general functions to more specific functions. In GO protein function ranges from the very general (e.g. enzyme activity) through broad terms (e.g. hydrolase) down to more specific (e.g. hydrolysis of O-glycosyl compounds).

Functional assignment is commonly performed via transfer from the closest homologue of known function and/or via sequence motifs/profiles such as those in InterPro (Mulder *et al.* 2005). Other methods exploit co-location on the genome, domain co-location, phylogenetic profiles or inferences from the interactome (Huynen *et al.* 2003; Gabaldon & Huynen 2004). Recently (see Lichtarge & Sowa 2002; Jones & Thornton 2004), several groups have developed algorithms to identify functionally important residues often employing sequence conservation and/or structural information including the widely used strategy known as the evolutionary trace (Lichtarge *et al.* 1996; Madabushi *et al.* 2002). However, evolutionary trace and related approaches primarily focus on the identification of functional residues which is distinct from actually assigning a function to the protein.

We report here a new automatic method PHUNCTIONER (Pazos & Sternberg 2004) for structure-based function prediction using automatically extracted functional sequence profiles generated by using the GO classification to supervise extraction of sequence motifs.

(b) *Methods*

The concepts underpinning our approach PHUNCTIONER are:

- (i) Sequence alignments between proteins with low (less than 30%) sequence identity are more reliable when based on a structural alignment than from sequence alone.
- (ii) GO provides a coherent computational approach to represent protein function at a variety of levels.
- (iii) GO can be used to supervise the grouping of proteins into functional families and this presents an alternate approach to supervising classification via a phylogenetic approach.

The strategy used in PHUNCTIONER is described in figure 5. PHUNCTIONER starts with the FFSP structural alignment (Holm & Sander 1996) of a set of protein homologues and uses the GO classification to extract subfamilies with a common GO term. The multiple sequence alignment from a subfamily (derived from the structural alignment) is used to generate a PSSM. The structure of the protein whose function needs to be assigned (X) is then added to the FSSP

Figure 4. (a) Superfamily expansion for the 10 most abundant in the human proteome. Expansion of a superfamily relative to the human proteome is plotted as the number of domains in superfamily X in proteome Y divided by the number of domains in superfamily X in human (multiplied by 100). This gives a base figure of 100% for all superfamilies in human. (b) Relative superfamily expansion within each proteome. Number of domains in a superfamily normalized by the number of domains in all superfamilies for a proteome (multiplied by 100).

multiple alignment and hence the sequence equivalences generated. The sequence of X is then scored against each PSSM for the alignment and the highest scoring match is taken as the prediction of function. The observed score is expressed as the number of standard deviations the score is above the mean of a

distribution of scores from matching scrambled sequences (see Pazos & Sternberg 2004).

To benchmark the approach, we applied a leave-one-out strategy on the FSSP database. We obtained 4753 sub-alignments (profiles) comprising 121 different GO terms in different levels of the GO. For comparison, we

Figure 5. Schema of the PHUNCTIONER method. (a) Structural alignment from FSSP. Proteins with different GO functional annotations are depicted in different colours. (b) The initial structural alignment is split into function-specific sub-alignments according with the GO annotation. Conservation patterns due to function are identified in those sub-alignments. The conserved residues in those sub-alignments are used to construct position-specific scoring matrices (PSSM). (c) These PSSM are used to assign a new structure to the corresponding function by scoring the sequence against the profiles in the search for the one where it fits better (thick arrow).

contrasted the accuracy of function assignment against that from inheritance from the closest homologue (SEQID) within the FSSP multiple alignment.

(c) Results and discussion

The accuracy of each method is quantified as the percentage of cases in which the first hit predicted by the method is correct. The results show that that PHUNCTIONER is significantly better than the SEQID method in the region when the closest homologue has less than 20% identity to the query. The percentages of correct GO assignments for PHUNCTIONER and SEQID are 75 and 60%, respectively. For the 30% sequence identity cut off, the accuracies of both methods are comparable at 90%.

In the generation of the PSSM, PHUNCTIONER identifies the residues that are responsible for a specific function. Figure 6 shows the residues automatically extracted by this method for the GO term 'GTP binding activity' (GO:0005525) in the 1ctqA.fssp structural alignment, mapped on the structure of the *Ras* oncogene (PDB 1ctq_A). All the residues are close to the bound nucleotide. We have used several other contemporary approaches to identify functionally important residues and the residues identified by them are more widely distributed around the protein and can be distant from the bound nucleotide (see Pazos & Sternberg (2004) for details).

Figure 6. Functional residues associated with the 'GTP binding activity' GO annotation (GO:0005525). The nucleotide (GTP) is in yellow and the residues are represented in magenta.

Recently, other groups have exploited the power of GO to assist in assigning function. The Barton group developed GOtcha that uses a weighted pool of GO annotations to homologous sequences to suggest the

best GO assignment (Martin *et al.* 2004). Pal & Eisenberg (2005) developed a consensus method using sequence and structural information to assign a GO function.

4. CONCLUDING REMARKS

This paper described two topics: the first was our relational database 3D-GENOMICS which reports structural and functional annotation of proteome sequences. The second topic was a new approach, PHUNCTIONER, to predict protein function from structure function of a protein given its structure. Other groups have developed similar resources. A major development in bioinformatics over the last few years has been the exploitation of computational power to link these bioinformatics resources so the community can benefit from consensus methods of annotation and prediction. Linking across different modalities of biological information will be increasingly important as bioinformatics moves from molecules to systems.

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 data base search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. 2004 SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229. (doi:10.1093/nar/gkh039)
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. 2002 Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18**, 298–305. (doi:10.1093/bioinformatics/18.2.298)
- Bateman, A. *et al.* 2004 The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141. (doi:10.1093/nar/gkh121)
- Blundell, T. L., Sibanda, B. L., Montalvão, R. W., Brewerton, S., Chelliah, V., Worth, C., Harmer, N. J., Davies, O. & Burke, D. 2005 Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Phil. Trans. R. Soc. B* **361**, 413–423. (doi:10.1098/rstb.2005.1800)
- Bourne, P. E. *et al.* 2004 The distribution and query systems of the RCSB protein data bank. *Nucleic Acids Res.* **32**, D223–D225. (doi:10.1093/nar/gkh096)
- Buchan, D. W., Rison, S. C., Bray, J. E., Lee, D., Pearl, F., Thornton, J. M. & Orengo, C. A. 2003 Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.* **31**, 469–473. (doi:10.1093/nar/gkg051)
- Carter, P., Liu, J. & Rost, B. 2003 PEP: predictions for entire proteomes. *Nucleic Acids Res.* **31**, 410–413. (doi:10.1093/nar/gkg102)
- Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R. & Stein, L. 2001 The distributed annotation system. *BMC Bioinform.* **2**, 7. (doi:10.1186/1471-2105-2-7)
- Eddy, S. R. 1998 Profile hidden Markov models. *Bioinformatics* **14**, 755–763. (doi:10.1093/bioinformatics/14.9.755)
- Fleming, K., Muller, A., MacCallum, R. M. & Sternberg, M. J. E. 2004 3D-GENOMICS: a database to compare structural and functional annotations of proteins between genomes. *Nucleic Acids Res.* **32**, D245–D250. (doi:10.1093/nar/gkh064)
- Foster, I. & Kesselman, C. 2004 *The GRID 2—blueprint for a new computing infrastructure*. San Francisco: Morgan Kaufmann Publishers.
- Gabaldon, T. & Huynen, M. A. 2004 Prediction of protein function and pathways in the genome era. *Cell. Mol. Life Sci.* **61**, 930–944.
- Goldsmith-Fischman, S. & Honig, B. 2003 Structural genomics: computational methods for structure analysis. *Protein Sci.* **12**, 1813–1821. (doi:10.1110/ps.0242903)
- Harris, M. A. *et al.* 2004 The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261. (doi:10.1093/nar/gkh066) Database issue.
- Hey, T. & Trefethen, A. E. 2005 Cyberinfrastructure for e-science. *Science* **308**, 817–821. (doi:10.1126/science.1110410)
- Holm, L. & Sander, C. 1996 The fssp database—fold classification based on structure structure alignment of proteins. *Nucleic Acids Res.* **24**, 206–209. (doi:10.1093/nar/24.1.206)
- Huynen, M. A., Snel, B., von Mering, C. & Bork, P. 2003 Function prediction and protein networks. *Curr. Opin. Cell Biol.* **15**, 191–198. (doi:10.1016/S0955-0674(03)00009-7)
- Jones, S. & Thornton, J. M. 2004 Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* **8**, 3–7. (doi:10.1016/j.cbpa.2003.11.001)
- Kinoshita, K. & Nakamura, H. 2003 Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **13**, 396–400. (doi:10.1016/S0959-440X(03)00074-5)
- Laskowski, R. A., Watson, J. D. & Thornton, J. M. 2003 From protein structure to biochemical function? *J. Struct. Funct. Genomics* **4**, 167–177. (doi:10.1023/A:1026127927612)
- Lichtarge, O. & Sowa, M. E. 2002 Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27. (doi:10.1016/S0959-440X(02)00284-1)
- Lichtarge, O., Bourne, H. R. & Cohen, F. E. 1996 An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358. (doi:10.1006/jmbi.1996.0167)
- Liu, Y., Engelman, D. M. & Gerstein, M. 2002 Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol.* **3**, research0054.1 research0054.12.
- Loukas, A. & Maizels, R. M. 2000 Helminth C-type lectins and host-parasite interactions. *Parasitol. Today* **16**, 333–339. (doi:10.1016/S0169-4758(00)01704-X)
- Lupas, A., van Dyke, M. & Stock, J. 1991 Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164.
- Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. 2002 Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154. (doi:10.1006/jmbi.2001.5327)
- Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. & Gough, J. 2004 The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* **32**, D235–D239. (doi:10.1093/nar/gkh117)
- Marsden, R. L. *et al.* 2005 Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Phil. Trans. R. Soc. B* **361**, 425–440. (doi:10.1098/rstb.2005.1801)
- Martin, D. M., Berriman, M. & Barton, G. J. 2004 GOfcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinform.* **5**, 178. (doi:10.1186/1471-2105-5-178)
- McGuffin, L. J., Street, S. A., Bryson, K., Sorensen, S. A. & Jones, D. T. 2004 The genomic threading database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res.* **32**, D196–D199. (doi:10.1093/nar/gkh043)

- Mott, R. 2000 Accurate formula for *P*-Values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300**, 649–659. (doi:10.1006/jmbi.2000.3875)
- Mulder, N. J. *et al.* 2005 InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201–D205. (doi:10.1093/nar/gki106)
- Müller, A., MacCallum, R. M. & Sternberg, M. J. E. 1999 Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.* **293**, 1257–1271.
- Müller, A., MacCallum, R. M. & Sternberg, M. J. E. 2002 Structural characterization of the human proteome. *Genome Res.* **12**, 1624–1641. (doi:10.1101/gr.221202)
- Pal, D. & Eisenberg, D. 2005 Inference of protein function from protein structure. *Structure (Camb)* **13**, 121–130. (doi:10.1016/j.str.2004.10.015)
- Pazos, F. & Sternberg, M. J. E. 2004 Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA* **101**, 14 754–14 759. (doi:10.1073/pnas.0404569101)
- Porter, C. T., Bartlett, G. J. & Thornton, J. M. 2004 The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32**, D129–D133. (doi:10.1093/nar/gkh028)
- Prlic, A., Domingues, F. S., Lackner, P. & Sippl, M. J. 2004 WILMA-automated annotation of protein sequences. *Bioinformatics* **20**, 127–128. (doi:10.1093/bioinformatics/btg380)
- Tatusov, R. L. *et al.* 2003 The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4**, 41. (doi:10.1186/1471-2105-4-41)
- Tusnady, G. E. & Simon, I. 2001 The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–850. (doi:10.1093/bioinformatics/17.9.849)
- Whisstock, J. C. & Lesk, A. M. 2003 Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340. (doi:10.1017/S0033583503003901)
- Wootton, J. C. & Federhen, S. 1996 Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.