The genome sequencing projects and knowledge of the entire protein repertoires of many organisms have prompted new procedures and techniques for the large-scale determination of protein structure, function and interactions. Recently, new work has been carried out on the determination of the function and evolutionary relationships of proteins by experimental structural genomics, and the discovery of protein–protein interactions by computational structural genomics.

Addresses

*Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK; e-mail: sat@biochem.ucl.ac.uk
†MRC Centre for Protein Engineering, Hills Road, Cambridge CB2 2QH, UK; e-mail: agm@mrc-lmb.cam.ac.uk
‡MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; e-mail: chc1@mrc-lmb.cam.ac.uk

Current Opinion in Structural Biology 2001. 11:354–363
0959-440X/01/$ — see front matter © 2001 Elsevier Science Ltd. All rights reserved.

Abbreviations

FBPase fructose-1,6-bisphosphatase
IMPase inositol monophosphatase
PDB Protein Data Bank

Introduction

The genome sequencing projects are the beginning of a change in perspective in molecular biology towards understanding how the various molecular attributes of a cell combine to produce its particular anatomical and physiological properties. Initially, this requires the determination of the structure, function, interactions and evolutionary relationships of all, or almost all, proteins that form different organisms. One of the methods that contributes to the achievement of this aim is to use computational procedures to detect homology between sequences from genome projects and those proteins of known structure and function. Jones [1] has provided a general review of these procedures. Teichmann et al. [2] have discussed in some detail their application to the analysis of sequences from genome projects. Contributions are also made by, for example, the experimental determination of the three-dimensional structures of the proteins that form model organisms and the use of this data, combined with those from experiments, to determine functions and evolutionary relationships, and by the computational analysis of genome sequences and of large-scale experiments to determine interactions between protein domains. Here, we review recent work in these two areas.

Protein structures from structural genomics projects

In a general sense, almost all new protein structures can make a contribution to structural genomics. In practice, the term is usually restricted to work on proteins belonging to model organisms for which the sequence of the whole genome has been determined. For this review, we examined recently determined structures that fall within the terms of this restricted definition. Amongst these, we selected proteins that fall into three categories:

1. Hypothetical proteins.
2. Proteins with a known function, but no apparent relationship to proteins of known structure.
3. Proteins with a known function and a distant relationship to proteins of known structure.

Proteins that are closely related to those of previously solved structure are not considered here. In Table 1, we list the 42 proteins discussed here and the category to which they belong. Most of the selected proteins had their structures determined as part of a structural genomics program. Some were determined because of their special interest.

Determination of function

Given the determination of the structure of a hypothetical protein, information on its function has been obtained from several sources:

1. The structure of the hypothetical protein shows that, in spite of insignificant sequence similarities, it is, in fact, homologous to a protein of known function and structure. The extent of functional similarity of the two proteins may be indicated by the extent to which active site residues are conserved.
2. The use of the structure, together with the sequences of homologous proteins, to find surface cavities or grooves in which conserved residues indicate an active site.
3. The discovery of bound cofactors in the structure.

For the hypothetical proteins listed in Table 1, the use of this kind of information, combined with experimental work, has given detailed functional information on about a quarter of the listed proteins. For another half, some functional information was obtained and for the other quarter, relatively little was discovered.
Two examples of structural genomics leading to detailed functional information for hypothetical proteins are the analysis of the product of gene 226 from *Methanococcus jan-naschii* (MJ-0226) [3•] and of the HdeA protein from *Escherichia coli* [4•]. The structural analysis of inositol monophosphatase (IMPase) from *M. jannaschii* [5•] provides an example of structural genomics giving an additional ‘new’ function to a protein of ‘known’ function.

Comparison of the structure of the protein product of gene MJ-0226 with those of other proteins [6] showed that it is structurally similar to the anticodon-binding domains of histidyl- and glycyl-tRNA synthetases and to nucleotide-binding domains of other proteins [3•]. Experiments to assess nucleotide binding by the protein found that it converted nucleotide triphosphates to monophosphates, and that xanthine triphosphate is the best substrate. This indicated strongly that the protein is a pyrophosphate-releasing XTPase [3•].

HdeA occurs in high concentration in the periplasmic space of certain enteric pathogens that pass through the acidic environment of the stomach before infection. Deletion of the *hdeA* gene abolishes *E. coli*’s resistance to acid conditions. A combination of structural and experimental studies showed that HdeA is a dimer that is transformed to monomers in acidic pH, and that this monomeric form acts as a chaperone to prevent the aggregation of proteins denatured at low pH [4•].

Fructose-1,6-bisphophatase (FBPase) is part of the gluconeogenesis pathway. Although FBPase activity and gluconeogenesis were detected in crude extracts from *M. jannaschii*, this enzyme could not be found in the genome sequence. However, IMPase, known as a distant relative of FBPase and involved in secondary messenger signalling in higher eukaryotes, is present. Structural analysis of this enzyme showed that, although its active site is very similar to that of human IMPase, there are structural differences in five adjacent loop regions [5•]. One of these loops has the same conformation as the loop in FBPase that binds a catalytic metal ion. The implication of this structure, that the enzyme possibly has both IMPase and FBPase activity, was shown to be the case by experimental measurements of its substrate specificity.

It has been suggested that comparisons of the geometry of the active sites of homologous and nonhomologous proteins may indicate functional similarities [7,8]. The procedures proposed for the application of such an approach include finding local functional motifs by comparison to libraries of three-dimensional amino acid or atomic templates [8–10], and the analysis of surface geometry and character [11].

### Determination of structural and evolutionary relationships

Determining the three-dimensional structure of a protein gives information that can be used to determine structural and evolutionary relationships. Structural similarities arise because the physics and chemistry of proteins favour, sometimes strongly, particular packing arrangements for secondary structures and certain chain topologies (see [12] for a recent review). This means that proteins that have no evolutionary relationships can have the same general arrangement for their major secondary structures and the same chain topology, that is, the same fold. Evidence for an evolutionary relationship between two proteins with the same fold requires the presence of detailed structural similarities and/or functional similarities.

Almost all the groups that determined the structures of hypothetical proteins compared their structures with those of other proteins; however, this was done by only a few of the groups that worked on proteins of known function. The comparisons were very largely carried out using the DALI structural superposition program [6]. In some cases, the databases that classify protein structures according to their structural and probable evolutionary relationships, SCOP [13] and CATH [14], were also used.

It is difficult to determine, from structural comparisons that use automatic procedures, whether or not two proteins that have the same fold are also evolutionary relatives. The automatic programs cannot, at present, extract the information needed to answer this question. In Table 1, we list the structural and evolutionary relationships of the 42 proteins, as determined by visual inspection and listed in the SCOP database [13]. To give an example of how these relationships are determined, we describe the classification of the ribosome anti-association factor [15].

The structures of ribosome anti-association factor from archaebial (*M. jannaschii*) and eukaryotic (*Saccharomyces cerevisiae*) organisms revealed five repeats an αββαβ structural unit arranged with approximate fivefold symmetry (Figure 1a). Because of this symmetry, this fold was named the pentein fold [15]. The close structural similarity between individual repeats strongly suggests that the probable origin of the pentein fold is gene duplication, despite the absence of internal sequence repeats. No significant similarity to a known structure was detected by automated searches, therefore the pentein fold was declared to be novel [15]; however, a very similar fold had been discovered previously in the structure of t-arginine-glycine amidinotransferase [16]. The amidinotransferase fold also consists of five αββαβ repeats arranged about an approximate fivefold axis (Figure 1b). Unlike the pentein fold, it is decorated with additional regions inserted into the turns and loops at one end of the constituent motifs. These form the active site of the amidinotransferase and result in its sequence being almost twice the size of that of ribosome anti-association factor. Relative to the pentein structure, small shifts in the amidinotransferase structure make its fivefold symmetry less perfect and do not allow an exact simultaneous superposition of all five motifs in the two.
Sequences and topology

This makes it difficult for their relationship to be detected by a high score from any of the existing automated search methods.

### Table 1

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB file(s)</th>
<th>Genome-gene</th>
<th>Fold</th>
<th>Superfamily</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purine regulatory protein</strong></td>
<td>1qu9</td>
<td>BS-yajf</td>
<td>Old(a)</td>
<td>Old(a)</td>
<td>[50]</td>
</tr>
<tr>
<td><strong>Conserved protein YgfF</strong></td>
<td>1qd9</td>
<td>EC-yjgf</td>
<td>Old</td>
<td>Old</td>
<td>[51]</td>
</tr>
<tr>
<td><strong>Heat-shock protein Hsp15</strong></td>
<td>1dm9</td>
<td>EC-yurH</td>
<td>Old</td>
<td>Old</td>
<td>[52]</td>
</tr>
<tr>
<td><strong>Hydrolase</strong></td>
<td>1jac</td>
<td>EC-ycaC</td>
<td>Old</td>
<td>Old</td>
<td>[53]</td>
</tr>
<tr>
<td><strong>Periplasmic protein</strong></td>
<td>1dbj, 1bg8</td>
<td>EC-hd6A</td>
<td>New</td>
<td>New</td>
<td>[4 at 54]</td>
</tr>
<tr>
<td><strong>eIF2 homologue</strong></td>
<td>1df</td>
<td>EC-ych</td>
<td>New(c)</td>
<td>New(c)</td>
<td>[55]</td>
</tr>
<tr>
<td><strong>RNA binding</strong></td>
<td>1hru</td>
<td>EC-yr6C</td>
<td>New</td>
<td>New</td>
<td>[56]</td>
</tr>
<tr>
<td><strong>YbaK homologue</strong></td>
<td>1dbu, 1dbx</td>
<td>EC-yrfH</td>
<td>Old</td>
<td>Old</td>
<td>[57]</td>
</tr>
<tr>
<td><strong>XTP pyrophosphatase</strong></td>
<td>1b78, 2mjp</td>
<td>Mj-0226</td>
<td>Old</td>
<td>New</td>
<td>[3 at 58]</td>
</tr>
<tr>
<td><strong>Maf</strong></td>
<td>1ex2, 1exc</td>
<td>EC-maf</td>
<td>Old(b)</td>
<td>New(b)</td>
<td>[59]</td>
</tr>
<tr>
<td><strong>ATP-binding protein</strong></td>
<td>1mjh</td>
<td>Mj-0577</td>
<td>Old</td>
<td>Old</td>
<td>[59]</td>
</tr>
<tr>
<td><strong>FMN-binding protein</strong></td>
<td>1eje</td>
<td>Mj-0152</td>
<td>Old</td>
<td>Old</td>
<td>[60 at 56]</td>
</tr>
<tr>
<td><strong>MT0538</strong></td>
<td>1eie</td>
<td>Mj-0538</td>
<td>Old</td>
<td>New</td>
<td>[61]</td>
</tr>
<tr>
<td><strong>Metal-binding protein</strong></td>
<td>1gh9</td>
<td>Mj-1184</td>
<td>Old</td>
<td>New</td>
<td>[60 at 56]</td>
</tr>
<tr>
<td><strong>Aptosis-related protein</strong></td>
<td>1elj</td>
<td>Mj-1615</td>
<td>Old</td>
<td>New</td>
<td>[60 at 56]</td>
</tr>
<tr>
<td><strong>PLP-binding protein</strong></td>
<td>1ct5, 1b54</td>
<td>SC-yb1063c</td>
<td>Old</td>
<td>Old</td>
<td>[62 at 63]</td>
</tr>
</tbody>
</table>

### Proteins of known function, but with no apparent relation to known structures

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB file(s)</th>
<th>Genome-gene</th>
<th>Fold</th>
<th>Superfamily</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pterin pyrophosphokinase</td>
<td>1hka</td>
<td>EC-folK</td>
<td>Old</td>
<td>New</td>
<td>[64]</td>
</tr>
<tr>
<td>Chaperone SecB</td>
<td>1ck</td>
<td>Hl-0064</td>
<td>Old(d)</td>
<td>New(d)</td>
<td>[65]</td>
</tr>
<tr>
<td>Diaminopimelate epimerase</td>
<td>1bwz</td>
<td>Hi-0750</td>
<td>New</td>
<td>New</td>
<td>[66]</td>
</tr>
<tr>
<td>Ferredoxin: α chain</td>
<td>1dj7</td>
<td>SS-Rov</td>
<td>Old</td>
<td>Old</td>
<td>[67]</td>
</tr>
<tr>
<td>Thioredoxin reductase β chain</td>
<td>1dh7</td>
<td>SS-RoC</td>
<td>New</td>
<td>New</td>
<td>[64]</td>
</tr>
<tr>
<td>Ribosome anti-association factor</td>
<td>1g61</td>
<td>Mj-0048</td>
<td>Old</td>
<td>Old</td>
<td>[65]</td>
</tr>
<tr>
<td></td>
<td>1g62</td>
<td>SC-ypr016c</td>
<td>Old(e)</td>
<td>Old(e)</td>
<td>[66]</td>
</tr>
<tr>
<td>Heat-shock protein Hsp20</td>
<td>1shs</td>
<td>Mj-0285</td>
<td>New</td>
<td>New</td>
<td>[67]</td>
</tr>
<tr>
<td>Cell-division protein FtsZ</td>
<td>1fs</td>
<td>Mj-0370</td>
<td>New/Old</td>
<td>New/Old</td>
<td>[68]</td>
</tr>
<tr>
<td>Eukaryote tip5a homologue</td>
<td>1eif, 2eif</td>
<td>Mj-1228</td>
<td>Old(d)</td>
<td>New(d)</td>
<td>[70]</td>
</tr>
<tr>
<td>RNA endonuclease EndA</td>
<td>1a7s, 1a7a</td>
<td>Mj-1424</td>
<td>Old/new</td>
<td>Old/new</td>
<td>[71]</td>
</tr>
<tr>
<td>Homoserine kinase</td>
<td>1fwk</td>
<td>Mj-1903</td>
<td>Old/Old</td>
<td>Old/new</td>
<td>[72]</td>
</tr>
<tr>
<td>RPB 10</td>
<td>1ef4</td>
<td>Mj-0040</td>
<td>Old</td>
<td>New</td>
<td>[73]</td>
</tr>
<tr>
<td>RPB 5</td>
<td>1eik</td>
<td>Mj-1048</td>
<td>New</td>
<td>New</td>
<td>[74]</td>
</tr>
<tr>
<td>Prefoldin: α subunit</td>
<td>1fs</td>
<td>Mj-0678</td>
<td>New</td>
<td>New</td>
<td>[75]</td>
</tr>
<tr>
<td>β subunit</td>
<td>1d5</td>
<td>Mj-1609</td>
<td>New(b)</td>
<td>New(b)</td>
<td>[76]</td>
</tr>
<tr>
<td>dTDP epimerase (RmIC)</td>
<td>1ep0</td>
<td>Mj-1790</td>
<td>Old</td>
<td>Old</td>
<td>[77]</td>
</tr>
</tbody>
</table>

### Proteins of known function and distantly related to known structures

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB file(s)</th>
<th>Genome-gene</th>
<th>Fold</th>
<th>Superfamily</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPase</td>
<td>1dk4</td>
<td>Mj-0109</td>
<td>Old</td>
<td>Old</td>
<td>[69]</td>
</tr>
<tr>
<td>Topoisomerase VI subunit</td>
<td>1d3y</td>
<td>Mj-0369</td>
<td>Old</td>
<td>Old</td>
<td>[70]</td>
</tr>
<tr>
<td>Fibrillarin</td>
<td>1fn</td>
<td>Mj-0697</td>
<td>Old</td>
<td>Old</td>
<td>[71]</td>
</tr>
<tr>
<td>Flap endonuclease-1</td>
<td>1at6, 1at7</td>
<td>Mj-1444</td>
<td>Old</td>
<td>Old</td>
<td>[72]</td>
</tr>
<tr>
<td>Phosphoserine phosphatase</td>
<td>1fs</td>
<td>Mj-1549</td>
<td>Old</td>
<td>Old</td>
<td>[73]</td>
</tr>
<tr>
<td>NMT adenyllytransferase</td>
<td>1f9a</td>
<td>Mj-0541</td>
<td>Old</td>
<td>Old</td>
<td>[74]</td>
</tr>
<tr>
<td></td>
<td>1e52</td>
<td>Mj-0150</td>
<td>Old</td>
<td>Old</td>
<td>[75]</td>
</tr>
<tr>
<td>OMP decarboxylase</td>
<td>1dij, 1dij</td>
<td>Mj-0219</td>
<td>Old</td>
<td>Old</td>
<td>[76]</td>
</tr>
<tr>
<td>Initiation factor IF2/eIF5B</td>
<td>1g7c, 1g7s, 1g7t</td>
<td>Mj-0259</td>
<td>Old</td>
<td>Old</td>
<td>[77]</td>
</tr>
<tr>
<td>Intracellular protease</td>
<td>1gi2</td>
<td>Mj-0704</td>
<td>Old</td>
<td>Old</td>
<td>[78]</td>
</tr>
</tbody>
</table>

*Pairs of proteins in bold are homologous. The criteria for the absence or presence of relationships to known structures (mentioned in the category headers) are from sequence comparison scores. The relations described in the fold and superfamily columns are determined by structure comparisons (see text). The genomics are BS; Bacillus subtilis; EC; E. coli; Hl, Haemophilus influenzae; Mj, M. jannaschii; MT, Methanobacterium thermoautotrophicum; PH, Pyrococcus horikoshii; SC; S. cerevisiae; SS, Synechocystis sp. PCC6803. The fold and superfamily information is taken from the SCOP database. ‘Old’ indicates the fold or superfamily is the same as that of a previously known structure; ‘new’ indicates that it has not been seen before. Two entries are given for those proteins that have two domains. Some of the proteins listed here are homologues of other entries in the table or of other recently determined structures: (a) same old fold and old superfamily as E. coli protein yjgF; (b) same old fold and new superfamily as XTP pyrophosphatase Mj-0226; (c) same new fold and new superfamily as human eIF1; (d) same fold and superfamily as the E. coli homologue; (e) same fold and superfamily as the M. jannaschii homologue; (f) two domains with the same folds and superfamilies as tubulin; (g) same fold and superfamily as human eIF5a; and (h) the α and β subunits of prefoldin are homologous.

---

All the individual αββαβ motifs in the two structures can be superimposed well (Figure 1c,d). The consensus structure of the five amidinotransferase motifs is essentially the
same as the consensus structure of the pentein motifs, and this includes not only the two helices and three strands but also, in each motif, three turns connecting the secondary structures. Therefore, it is most likely that the folds of the amidinotransferase and ribosome anti-association factor have originated by duplication of a common ancestral motif and that these proteins can be classified not only as having the same fold, but also as belonging to the same protein superfamily. In the SCOP database, there are many examples of superfamilies whose common ancestral folds are differently decorated in different members. Also, it is not uncommon to have enzymes and protein factors in the same superfamily [17].

There are 32 proteins in Table 1 that were not expected to be related to proteins of known structure before their structure determination. A few of the 32 proteins are homologous to each other or contain two domains or two subunits. Allowing for this, the 32 proteins contain 32 unique domains (see Table 1 for details). Of these 32 domains, 20 have a fold described previously for known protein structures and 11 of these also belong to a previously described superfamily.

Protein–protein interactions
Some of the structures discussed above are multidomain or multisubunit proteins. Analysis of the domain structure of the proteins in *Mycoplasma genitalium* showed that multidomain proteins are very common in its genome: about one-third contain a single domain and about two-thirds contain two or more domains as a result of recombination or gene duplication [18]. The size distribution of proteins in other organisms [19] indicates that the same proportions of single-domain and multidomain proteins occur in other bacteria, whereas in eukaryotes, there is a somewhat higher proportion of multidomain proteins. Therefore, interactions between protein domains within polypeptide chains, as well as between polypeptide chains, are very common.

This means that protein–protein interactions can be classified into three types: first, those that occur between domains within the same polypeptide chain (i.e. those that are the result of duplication or recombination); second, the interactions between domains on different chains in multimeric proteins; and third, the interactions in the transient complexes formed between independent proteins carrying out their function. Here, we review the recent advances that computational analyses of structures in the Protein Data Bank (PDB) and other data have made in our understanding of these types of protein interactions.

If the mode of interaction of two proteins or two protein domains is known, we might expect that any interactions that occur between their homologues would usually involve contacts of the same kind. However, there are a few cases in which homologous proteins have different modes of interaction; for example, nucleoside diphosphate kinases are known to occur as hexamers, as well as trimers, depending on the species [20]. Therefore, a complete three-dimensional library of protein–protein interactions would have to include, for the homologous pairs that interact with each other in different ways, this information and the distribution of such differences.

Interactions between domains within multidomain polypeptide chains
The interactions between domains within polypeptide chains in the PDB have been analysed by Park et al. [21••]. They used a threshold of five contacts between atoms no more than 5 Å apart to determine interactions between SCOP domains in PDB entries. In version 1.48 of the SCOP database, the domains that form the proteins of known structure clustered into 771 superfamilies. The domains that combine with other domains within the same chain in the PDB come from 334 different superfamilies and make 278 different interactions. Seventy of these interactions are contacts between domains from the same family within a polypeptide chain, that is, they are most probably the result of internal duplication. The majority of superfamilies (91%) are linked to members of only one or two other superfamilies, although there are a few superfamilies that are very versatile, making links with members from up to 14 different families. The
Protein family interaction map. Each family is represented by a shape according to its class in SCOP (ellipses: all-α; rectangles: all-β and small proteins; triangles: α/β; diamonds: α+β; pentagons: multidomain proteins; hexagons: membrane proteins and cell-surface proteins) and its identification number in version 1.48 of SCOP. The size of the shape of a family is proportional to the number of family interactions it undergoes. A black outline to a shape means some members of the family interact with each other. The interactions are colour-coded as follows: blue, interactions in PDB only; red, interactions in PDB and yeast; grey, interactions in yeast only. Intermolecular and intramolecular interactions can be distinguished by the type of line connecting two families: continuous line, intramolecular; thin-dashed line, intramolecular and intermolecular. More than one-third of the families are part of the big connected cluster shown here, centring on the immunoglobulin superfamily (2.1.1) and the P-loop NTP hydrolase superfamily (3.30.1). Other highly connected nodes are 1.110.1 (armadillo [ARM] repeat), 4.117.1 (protein kinases, catalytic core) and 2.41.1 (trypsin-like serine proteases). The names of all other families, shown here only by their SCOP identification number, can be obtained from http://www.biochem.ucl.ac.uk/bsdb/pdb_sc_int.html. This map shows 229 out of at least 572 protein families that interact in the PDB and yeast. The remaining families are part of smaller isolated clusters. More central nodes are shaded darker green. Adapted from [21*].
The use of gene order to predict protein–protein interactions

Like domain interfaces within chains, interactions between domains on separate polypeptide chains will be better understood when multichain structures become available on a large scale. In addition to the domain interactions between chains in the PDB, the structural assignment of pairs of proteins that are known to interact can be used to build a library of interactions of domains belonging to pairs of SCOP superfamilies. In the past two or three years, several computational methods for the prediction of protein interactions have been published. Below we review these methods and their limitations for the prediction of physical interactions between proteins.

If two polypeptide chains are part of the same protein or complex, it is often favourable for the two interaction partners to be expressed and regulated together at the DNA level. By relating data on protein interactions and gene expression, for instance with the PartsList database [26*], it can be seen that, although interacting yeast protein pairs, in general, do not have correlated gene expression, polypeptides that are part of the same protein complex do have significantly correlated gene expression.

For polypeptides that are subunits of the same protein, the genes are also often adjacent on the genome. Ouzounis and Karp [27] carried out a survey of the proteins and genes in the E. coli metabolic pathways, and found that different subunits of the same protein are almost always adjacent to each other on the genome (over 90% of cases). This means that it might be possible to predict the interactions between proteins computationally by simply looking for conservation of gene order across different genomes or by looking for genomes in which the genes for the two proteins are neighbours in one genome and are part of a single gene in another.

Conservation of gene order across different genomes among pairs of orthologues was studied by Dandekar et al. [28*]. They analysed gene order in six prokaryote and three archaean genomes, and identified about 100 proteins in each genome for which interactions can be predicted on the basis of gene order. In an assessment of prediction methods based on genomic context, Huynen et al. [29**] estimate that 63% of conserved adjacent genes to interact directly in a structural complex (as opposed to a functional relationship, such as being in the same metabolic pathway or some other functional process).

Overbeek et al. [30] and Lathe et al. [31] looked for conservation of gene neighbourhood for sets of genes across distantly related genomes. In both papers, this process is partially automated, but requires manual input at the level of decision on the merging of clusters of gene neighbours. Thus, there are hundreds or even thousands of candidate clusters in the 20–30 genomes discussed in these papers, but Overbeek et al. [30] concentrate on clusters of genes in individual metabolic pathways that are adjacent to one another on the chromosomes, and Lathe et al. [31] discuss gene neighbourhood in translation, flagellar proteins and ABC transporters. In agreement with the particular examples that they discuss, Huynen et al. [29**] estimate that only 3% of genes whose positions are conserved in same general neighbourhood physically interact. Of the genes in this category, 20% are in the same (metabolic) pathway and a further 26% are part of the same functional process.

A variation on this approach was used by Enright et al. [32*] and Marcotte et al. [33], who looked for examples in a genome where two or more orthologues are part of the same polypeptide chain, and then predicted that the instances of the orthologues in other genomes interact with each other. The methods used by the two groups are slightly different. Enright et al. [32*] used BLAST [34] to determine orthology between proteins and found, in four prokaryote genomes, 64 cases in which genes are fused in some cases, but not in others. Huynen et al. [29*], in their assessment of prediction methods based on genomic context, found that at least 63% of these predictions concern genes that physically interact and another 15% are genes in the same metabolic pathway.

Marcotte et al. [33] did not use the strict BLAST criterion for determining homology between individual protein domains, but rather used the much more distant homologies listed in the Pfam [35] and ProDom [36] protein family databases. This means they find many more putative protein interactions in their survey of interactions in the E. coli and yeast genomes. Essentially, this method assumes that, if two domains are seen in one protein, then all domains from those two families are expected to interact with each other. That this assumption will lead to a large number of false positives is obvious, in so far as the members of two interacting families are likely to have particular specificities for each other, rather than all interact with all members of the other family. Also, some families have different members that interact with partners from different superfamilies (see above).

---

**Table 2**

<table>
<thead>
<tr>
<th>Protein sets</th>
<th>Number of different types of interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intramolecular*</td>
</tr>
<tr>
<td>Known structures (SCOP)</td>
<td>356</td>
</tr>
<tr>
<td>Yeast</td>
<td>-</td>
</tr>
<tr>
<td>40 genomes</td>
<td>298 as in SCOP + 1009 novel</td>
</tr>
</tbody>
</table>

*The data in this column are from Apic et al. [23*] using version 1.53 of SCOP. †The data in this column are from Park et al. [21**] using version 1.48 of SCOP.
The use of phylogeny to predict protein–protein interactions

There are a few other methods of predicting protein–protein interactions computationally that are independent of the assumptions discussed above. A method that appears to reliably predict a loose functional correlation among proteins is the phylogenetic profile method of Pellegrini et al. [38]. The phylogenetic profile method relies on the detection of orthologues (or homologues, in a variation of the method) in a set of genomes. If the pattern of orthologue presence or absence is the same in a group of proteins, then these proteins are clustered together as belonging to the same functional class.

The phylogenetic patterns of clusters of orthologous groups of proteins, previously described by Tatusov et al. [39] and deposited in the COG database [40], could, in principle, be used for prediction in the same way. Huynen et al. [29**] found 34% of pairs in *M. genitalium* predicted to interact with this method and an additional 29% to belong to the same metabolic pathway or functional process.

The use of large-scale experiments to determine protein–protein interactions

Interactions between proteins in cells have been studied experimentally on a case-by-case basis for a long time, using a wide variety of biochemical, physical, genetic and other methods. For yeast, the results of the individual experiments have been collected in the MIPS database [41•]: http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/index.html.

Now, some of these methods are being automated and new methods are being developed so that protein–protein interactions can be detected on a large scale. For instance, the yeast-two-hybrid method has been automated and applied to yeast proteins on a large scale by two groups [42*,43*] and to *Helicobacter pylori* proteins by Rain et al. [44*]. The two large-scale experiments on yeast, together with the MIPS collection of individual experiments, mean that there are over 2500 experimentally determined pairwise interactions between yeast proteins. This is, of course, not the complete repertoire of interactions among the roughly 6200 yeast proteins. Also, the large-scale experimental data will contain false positives. It is, however, a major step toward the determination of the complete repertoire of protein interactions of one genome.

Schwikowski et al. [45•] have attempted to use the whole set of yeast interactions known from experiment to predict functions of unknown proteins and calculate a 72% success rate for functional annotation on the basis of protein interaction partners. Knowledge of the structure of a hypothetical protein with one or more interaction partners therefore promises to be very helpful for functional annotation, as both the structure and the interaction context will provide functional clues.

Several other methods are being applied on a smaller scale, with the prospect of being automated and scaled-up in the future. Mass spectroscopy [46] and 2D gels [47] have been used in proteomics for several years. Potentially, the presence of proteins in cells could be detected using antibody arrays such as those described in [48,49]. Protein–protein interactions could be more specifically detected individually if protein–protein arrays are developed for the large-scale detection of protein interactions in cells.

The experimental data described in the previous paragraphs can be compared to the interactions between protein families.
that are seen in the known structures for purposes of validation of the predictions and experiments, or for picking targets for structural genomics of protein complexes. Structural assignments to genome sequences allow classification of domain combinations or protein interactions.

To analyse the complete set of experimentally derived pairwise protein interactions in the yeast genome from the large-scale yeast-two-hybrid experiments [42•,43•] and from the MIPS database, Park et al. [21••] made structural assignments to yeast proteins. There are 99 protein pairs that are each completely matched by a single SCOP domain using PSI-BLAST-based methods. There are 55 SCOP superfamilies to yeast proteins. There are 99 protein pairs that are combinations, as shown in Table 2. The protein pairs with new combinations are clear targets for structure elucidation of domain–domain interfaces.

Conclusions

The hypothetical proteins in Table 1, together with those believed to be not related to proteins of known structure, contain, in all, 32 domains. The determination of their three-dimensional structures showed that, in fact, 20 of these domains have folds that are the same as previously known structures and 11 belonged to the same superfamilies as a previously known structure. In a number of cases, these relationships played an important role in helping to determine the functions of these proteins. The extensive novel structural data impending from structural genomics projects will greatly increase this ability to classify protein domains in terms of their evolutionary relationships. This knowledge, together with better computational methods for detecting functional features, should greatly increase our ability to suggest functions for proteins on the basis of their structure and, hence, to design experiments that can confirm, refine or refute these putative assignments.

The interactions made by a protein, or a protein domain, are an essential part of its function. The prediction of protein interactions on the basis of the position of genes in the genome provides some useful results, but have inherent limitations in coverage and accuracy. Although the current large-scale protein interaction experiments also have some limitations in coverage and accuracy, there is much scope for improvements and new developments. Structural assignments to genome sequences allow these data to be compared to interactions seen in the protein complexes whose structures are currently known and to domain interactions within polypeptide chains of known structure. For those domain complexes not observed previously, a structural elucidation is desirable in order to gain a complete picture of the evolution and mechanics of domain interactions.

Acknowledgements

We thank Michael Lappé for making Figure 2 available to us. SAT has a Beit Memorial Fellowship for Medical Research.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest


See annotation to [59•].


See annotation to [59•].


This paper describes an example of structural genomics giving an additional 'new' function to a protein of 'known' function.


The interactions between the domains in SCOP superfamilies are determined for the PDB and the yeast genome. Most superfamilies only interact with one or two other superfamilies, whereas some are very versatile and interact with 14 or more superfamilies. The interaction repertoire of intramolecular and intermolecular interactions overlaps for interactions between domains from the same family and enzyme domains.


In this paper, the patterns of domain combinations in the multidomain proteins of 40 genomes are analysed. Most domain families only combine with one or two other domain families, whereas a few families are very versatile and have domain neighbours from many different families. Many of the versatile families are common families, but they frequently combine in different ways in different genomes.


In this paper, all pairs of genes that are neighbours in at least three out of nine eukaryotic genomes were predicted for each genome. The coevolution of interacting proteins and applications for function prediction are discussed.


This dynamic database is centred on protein folds and structural protein domains, but contains information on protein interactions, gene expression and distribution of genes across genomes.


This evaluation of four protein function and interaction prediction methods based on genomic context uses the small M. genitalium genome as a benchmark. This analysis clarifies the precise types of functional or physical interactions that are predicted by the various methods, and compares the information gained with that from simple sequence homology.


One of two papers describing large-scale yeast-two-hybrid experiments on yeast proteins. About 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


In this paper, all pairs of genes that are neighbours in at least three out of nine genomes were predicted for each genome. The coevolution of interacting proteins and applications for function prediction are discussed.


The match of two or more separate proteins in one genome to a single protein in another genome is taken as an indication that the separate proteins physically interact. 64 fusion events were detected in three prokaryote genomes, many of which are known interactions between metabolic enzymes.


This paper quantifies the predominance of metabolic enzymes in gene-fusion events using E. coli as a test case and points out that 76% of genes that are fused in other genomes are part of complexes. Applications of this imbalance to the functional annotation of hypothetical proteins and to the reconstruction of metabolic pathways from genome sequences are suggested.


This evaluation of four protein function and interaction prediction methods based on genomic context uses the small M. genitalium genome as a benchmark. This analysis clarifies the precise types of functional or physical interactions that are predicted by the various methods, and compares the information gained with that from simple sequence homology.


This paper quantifies the predominance of metabolic enzymes in gene-fusion events using E. coli as a test case and points out that 76% of genes that are fused in other genomes are part of complexes. Applications of this imbalance to the functional annotation of hypothetical proteins and to the reconstruction of metabolic pathways from genome sequences are suggested.


This evaluation of four protein function and interaction prediction methods based on genomic context uses the small M. genitalium genome as a benchmark. This analysis clarifies the precise types of functional or physical interactions that are predicted by the various methods, and compares the information gained with that from simple sequence homology.


In this large-scale system of yeast-two-hybrid experiments on yeast proteins, about 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


One of two papers describing large-scale yeast-two-hybrid experiments on yeast proteins. About 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


One of two papers describing large-scale yeast-two-hybrid experiments on yeast proteins. About 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


One of two papers describing large-scale yeast-two-hybrid experiments on yeast proteins. About 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


In this large-scale system of yeast-two-hybrid experiments on yeast proteins, about 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.


One of two papers describing large-scale yeast-two-hybrid experiments on yeast proteins. About 10% of all possible combinations of proteins were tested, resulting in 183 interactions. Over half of these interactions were previously unknown.
Structural genomics Teichmann, Murzin and Chothia 363

member of the highly conserved YigF family. Proc Natl Acad Sci USA 2000, 97:13074-13079.


62. Crystal structure of yeast hypothetical protein Yb1036C. Protein Data Bank file 1b54 on World Wide Web URL: http://www.rcsb.org/pdb/cgi/explore.cgi?pid=11969986283632&pdbId=1CT5

63. Crystal structure of yeast hypothetical protein. Protein Data Bank file 1b54 on World Wide Web URL: http://www.rcsb.org/pdb/cgi/explore.cgi?pid=12047988237322&pdbId=1B54


