Advances in structural genomics

Sarah A Teichmann*†, Cyrus Chothia*‡ and Mark Gerstein§

New computational techniques have allowed protein folds to be assigned to all or parts of between a quarter (*Caenorhabditis elegans*) and a half (*Mycoplasma genitalium*) of the individual protein sequences in different genomes. These assignments give a new perspective on domain structures, gene duplications, protein families and protein folds in genome sequences.

Addresses
*MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK
†e-mail: sat@mrc-lmb.cam.ac.uk
‡e-mail: chc1@mrc-lmb.cam.ac.uk
§Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520, USA;
e-mail: Mark.Gerstein@yale.edu

Current Opinion in Structural Biology 1999, **9**:390–399

http://biomednet.com/elecref/0959440X00900390

Abbreviations

LC	low-complexity
MG	*Mycoplasma genitalium*
ORF	open reading frame
PDB	Protein Data Bank
TIGR	The Institute for Genomic Research
TM	transmembrane

Introduction

The purpose of structural genomics can be defined as the assignment of three-dimensional structures to the protein products of genomes (proteomes) and the investigation of the biological implications of these assignments. If the structure assigned to a new protein is homologous to one already known, it provides an indication of its probable function and evolutionary relationships. If structures can be assigned to all or to a significant fraction of the products of a whole genome, it will provide a much better understanding of the evolution and physiology of an organism.

The assignment of structures to proteomes can be carried out on two levels — experimental and computational. The experimental level involves the directed, large-scale determination of the protein structures using NMR spectroscopy or X-ray crystallography [1••,2,3]. The computational level involves the assignment of structures to proteins using calculations that mostly involve demonstrating homology to proteins of known structure. Here, we review the recent advances made at the computational level. The first part of the review deals with methods that have been used to assign structures to genome products. The second part reviews the biological implications of this work. Throughout, we pay particular attention to work related to the genome of *Mycoplasma genitalium* (MG), the second genome to be sequenced [4]. With only 479 proteins, this genome has emerged as the initial focus and bench-mark for computational investigations in structural genomics.

Methods for assigning structures to genome sequences

Three classes of computational methods are used to assign structures to genome sequences: the detection of distant homologies (this usually involves pairwise or multiple sequence comparisons); fold recognition (which tries to determine whether the sequence of a new protein fits a fold that is close to that of a known structure); and predictions based on statistical rules derived from structures. (These are used to predict secondary structure, transmembrane [TM] helices and coiled-coils.)

Detection of distant homologies

Pairwise sequence comparison

Not long after the first few complete bacterial genome sequences were published, their sequences were matched to the sequences of proteins of known structure (PDB sequences) using pairwise sequence comparison methods, such as FASTA [5], Smith–Waterman [6] and BLAST [7]. Sometimes these comparisons took the whole sequence of the known structure as the ‘query’, but the sequences of individual structural domains are more useful, as these correspond to the functional and evolutionary units of proteins. The domains that have been used in the assignment of structures to genome sequences are those described in the SCOP [8] and CATH [9] databases.

Pairwise matches between genome sequences and known structures form part of genome analysis systems such as GeneQuiz [10], PEDANT [11,12•] and GeneCensus [13•–15•]. Early calculations matched between 8 and 12% of the proteins from different genomes to known structures [10,11,12•,13•]. The rapid increase in the number of known structures has, for more recent calculations, increased these proportions to between 11 and 20% [15•]. For MG, these matched sequences comprise about 16% of the residues in the proteome.

Pairwise sequence comparisons detect only about half of the evolutionary relationships between proteins with 20–30% sequence identity, however, and, for related proteins with less than 20% identity, the proportion is much smaller [16•,17•]. There are many protein families whose members diverge to the point at which they have sequence identities well below 30% and, consequently, many homologous relationships between known structures and genome sequences cannot be detected by pairwise comparison.

Assignment of structures to genome sequences using PSI-BLAST

In order to try to overcome the limitations of pairwise comparisons, search procedures based on the shared characteristics of sets of related sequences have been developed. One of the most widely used of these procedures is the PSI-BLAST (Position-Specific Iterated Basic Local Alignment and Search Tool) program [18••]. This program does an initial, gapped BLAST search to collect close homologues of the query in a sequence database and then builds a profile of the query sequence and its close homologues. The profile is then matched to the database and more homologues are collected. These new homologues are added to the profile and another search is carried out. This process can be iterated as many times as the user specifies or until no more homologues are found.

A quantitative assessment of PSI-BLAST [17•] showed that, for evolutionary relationships among proteins whose sequence identities are less than 30%, it can detect three times as many relationships as pairwise comparisons. Of course, PSI-BLAST is significantly more computationally demanding and complicated to use than pairwise comparison methods. On a current work station (500 MHz DEC alpha), building up a PSI-BLAST profile can take between 1 and 30 min and can consume a considerable amount of disk space.

This past year, PSI-BLAST has been used by three groups to assign structures to genome sequences. All these attempts included detailed assignments for the MG genome. We summarise the work below, quoting, where appropriate, updated results from web sites, rather than the original data from the papers. There were differences in both the parameters that were used in the three calculations and the manner in which they were carried out [19••–21••]. The performance of PSI-BLAST is affected both by such differences [17•] and by the particular database from which the homologues are collected. Together, these factors account for most of the variations in the number of matches to MG sequences that were made by the different calculations.

Huynen *et al.* [19••] were the first to use PSI-BLAST to match PDB sequences to MG proteins. Both sets of sequences were preprocessed to remove regions of low complexity (LC), TM helices, coiled-coils and cysteine-rich proteins, as these readily give false matches. They found that 184 regions in 172 MG sequences (37%) matched PDB sequences, with different regions of 12 of the MG sequences matched by two different PDB sequences. Overall, these matches cover 23% of all the MG residues.

Teichmann *et al.* [20••] also used PSI-BLAST to match PDB domain sequences to MG proteins. They did this comparison in a two-way fashion, first searching using PDB domain sequences as queries and MG sequences as targets embedded in a large, nonredundant sequence database (NRDB90 [22]) and then using MG sequences as queries and PDB sequences as targets embedded in NRDB90. In the most recent version of this calculation, PSI-BLAST matched 314 PDB domains to all or part of 223 MG proteins (47%) (http://www.mrc-lmb.cam.ac.uk/genomes/MG/). Sixty-four of the matched MG proteins had different regions matched by between two and five PDB domain sequences. Overall, the matched regions cover 33% of all the MG residues.

Wolf *et al.* [21••] used PSI-BLAST to assign structures to the genomes of MG, ten other prokaryotes, *Saccharomyces cerevisiae* and *C. elegans*. These calculations matched PDB domain sequences to all or part of 181 (39%) of the MG sequences, to 19–34% of sequences in the other prokaryotes, to 24% of sequences in *S. cerevisiae* and to 21% of sequences in *C. elegans*. On average, 11% of the matched proteome sequences had between two and five PDB domain sequences matching different regions.

Profile-profile matching

The BASIC procedure, developed by Rychlewski and coworkers [23], provides a further refinement to the PSI-BLAST approach. Homologues are collected by PSI-BLAST for query and target sequences and profiles created for both sets of sequences. Relationships are then detected by profile-profile matching. Rychlewski *et al.* [24••] used this procedure to match 1151 representative PDB sequences to the MG proteome. Using this method, 139 (29%) MG proteins were matched. (These are updated results subsequent to publication.)

Model building of three-dimensional structures

Sanchez and Sali [25••] used a pairwise comparison to find matches between sequences of the yeast *S. cerevisiae* [26] and 1151 representative PDB sequences. All or part of 2256 (36%) *S. cerevisiae* sequences matched a PDB sequence. Of these sequences, 1071 had a good enough match for a detailed three-dimensional model to be built. For the other matched sequences, the divergence of structure, which occurs commonly for more distantly related proteins, only allows the construction of outline models.

Threading

Threading procedures cover a variety of techniques that try to determine whether the sequence of a protein with an unknown structure is compatible with that of a known structure [27–29]. The first detailed assignment of structures to the MG proteome used one such technique — the fold prediction method of Fischer and Eisenberg [30••]. With this method, the compatibility of the query sequence with each of the folds in a library of known structures is determined by both its predicted secondary structure and its sequence characteristics, as given by a matrix of residue similarities. The query sequence can be used by itself or with homologues. The most recent use of this method matched PDB sequences to 160 MG sequences, of which 75 could also be found using pairwise comparisons.

Table 1

A comparison of different calculations of the number of MG proteins that are homologous, all or in part, to PDB sequences*.

Authors of the calculation	Number of MG proteins matched to a PDB sequence	Percentage of the MG proteins matched using other calculations that are the same as those found using the calculations in column 1						Percentage of matches made using the column 1 calculations that are common to at least one other calculation
		T	W	H	F	R	G	
Teichmann *et al.*	223	–	92	97	98	92	97	93
Wolf *et al.*	181	74	–	79	80	77	88	94
Huynen *et al.*	172	74	75	–	83	76	91	98
Fischer and Eisenberg	160	70	71	77	–	78	94	99
Rychlewski *et al.*	139	57	59	61	78	–	60	94
Gerstein (representative FASTA)	90	39	44	47	53	39	–	99

*This table shows a comparison of the updated assignments of structures to MG sequences made by Teichmann *et al.* (T) [20••], Wolf *et al.* (W) [21••], Huynen *et al.* (H) [19••], Fischer and Eisenberg (F) [30••] and Rychlewski *et al.* (R) [24••]. For comparison, a representative result of FASTA assignments is listed as well (Gerstein [G] [13•]). The comparisons are based just on common ORFs (open reading frames). All the comparisons are based on the original TIGR (The Institute for Genomic Research) MG ORF file [4], which contained 468 genes, rather than the most current one, which contains 479 genes. Note that the W matches are based on some alternative gene definitions and so have ORF matches that do not correspond to either of the TIGR ORF files. We provide a more detailed comparison table on the Web (via http://bioinfo.mbb.yale.edu/genome/MG or http://www.mrc-lmb.cam.ac.uk/genomes/MG). In addition, many of the matches are collected together in the PRESAGE database [41•].

Grandori [31] used the threading program ProFIT [28] to match PDB sequences to *M. pneumoniae* sequences that are shorter than 200 residues. Matches were found for 12 genome sequences that could not be found using pairwise comparisons.

Secondary structure prediction

Secondary structure predictions were carried out on five genomes [12•] using PREDATOR [32] and on eight genomes [13•,15•] using GOR [33]. One of the more interesting results to emerge from these calculations was that the genomes have a similar overall composition in terms of secondary structure, although they have very different amino acid compositions. This was unexpected in light of the well-known and markedly different secondary-structure propensities of individual amino acids.

Membrane proteins

Several groups have carried out calculations to determine the occurrence of membrane proteins in genome sequences [13•,14•,34–40]. The overall number of membrane proteins found depends somewhat on the prediction method and threshold used. Nevertheless, there seems to be a broad agreement that all or part of 20–30% of the proteins in microbial genomes are membrane proteins. Membrane protein structures can be classified by how many TM helices they have. In all the surveys, the occurrence of membrane proteins with a given number of TM helices falls off rapidly as the number of helices increases; thus, only a small fraction of membrane proteins have large numbers of TM helices.

The current state of structural annotation of the *M. genitalium* genome

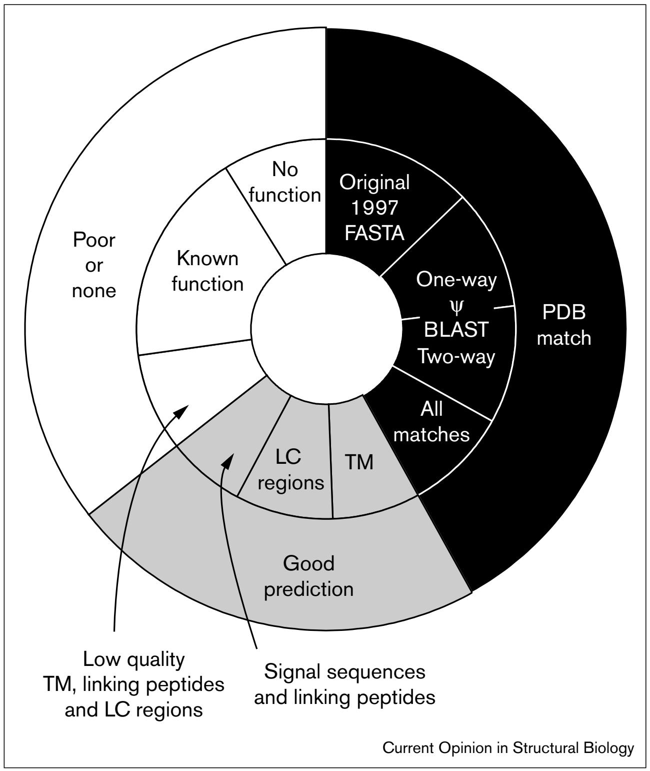
As described in the previous section, a number of groups have used pairwise sequence comparison, PSI-BLAST, profiles or threading to match sequences of proteins with known structure to sequences from the genome of *M. genitalium* [10,12•,14•,15•,19••–21••,24••,30••,41•] (see Table 1). Most of these groups have made their published results available on the web and their sites often carry 'updated results' that have been obtained subsequent to the original publications (see Figure 1 for details).

Overall, the results produced up to early 1999 by the different groups show a high degree of consistency, as shown in Table 1. When matches to an MG protein are made by more than one group, as is commonly the case, they are matched to the same PDB sequence or to a homologue in the large majority of cases. On a more detailed level, of the 352 SCOP domains assigned to MG proteins, 88 are assigned by one group only and only 12 (3.5%) are assigned different superfamilies by the different groups. Small differences in the lengths of the matched regions are also common. (The details of these assignments can be found at the web sites cited elsewhere in the text.) Examining the union of the matches made by many of the different groups, we find that more than half (242) of the MG sequences are matched, all or in part, by a PDB sequence and these matches cover more than 40% of all the MG residues (see Figure 1).

In addition to the regions matched to PDB sequences, about 79 MG sequences have the characteristics of integral membrane proteins and about 65 have long, nonglobular regions. This results in a total of about two-thirds of the MG sequences having some structural annotation (Figure 1) (some of the assignments are to different regions of the same protein).

The complete structural characterisation of the MG sequences will not be achieved in the near future if structures continue to be solved in an untargeted fashion. This

Figure 1

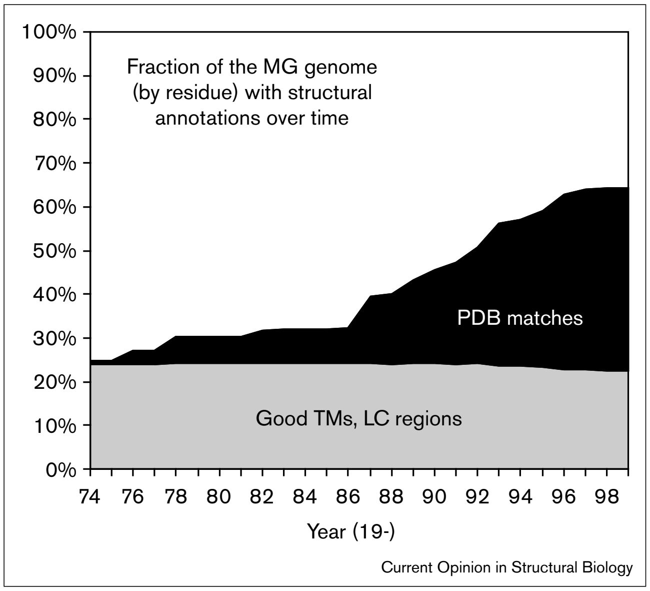


Current Opinion in Structural Biology

A pie chart showing the current status of the structural annotation of the MG genome (as of January 1999). The different parts of the pie chart are described in detail in Table 2. For each of the representative PSI-BLAST calculations, we used the results described as 'two-way PSI-BLAST'. These are updated versions of those results described previously [20••]. All of the calculations for the pie chart were based on the original TIGR MG ORF file [4], which contained 468 genes, rather than the most current one, which contains 479 genes. This was to enable us to merge other annotations, which were often based on the earlier ORF file. The current status of the level of annotation of MG is available from http://bioinfo.mbb.yale.edu/genome/MG and http://www.mrc-lmb.cam.ac.uk/genomes/MG. These web sites report data for both the current 479 gene ORF file and the original 468 gene file.

is shown by Figure 2, a graph of how the structures homologous to MG sequences have been determined over the past 25 years — the development of MG structural annotation over time. Experimental structural genomics projects, many of which have started recently, will target regions of the proteome that have neither a matching PDB sequence nor a different type of structural annotation (LC, TM and so on). Thus, they will increase the gradient of the graph in Figure 2 such that genomes should soon be almost completely structurally annotated. To be optimally efficient about the target choice for experimental structural genomics projects, the uncharacterised regions must be clustered at the sequence level. A list of the uncharacterised regions of MG and their sequence clusters will be made available on the web (http://www.mrc-lmb.cam.ac.uk/genomes/MG or http://bioinfo.mbb.yale.edu/genome/MG).

Figure 2



Current Opinion in Structural Biology

A time-line showing how MG structural annotation is changing each year. The panel shows how the fraction of residues in the MG genome that have been characterised increases each year with the addition of new structures to the PDB – imagining that the complete sequences of MG and the current sequence-matching techniques (e.g. PSI-BLAST) were known a quarter of a century ago. In particular, the time-line shows how the black 'PDB match' section changes over time. This time-line is based on exactly the same 'sequence masking' methodology discussed previously [15•,53]. In contrast to the previous analysis [53], however, we come to a somewhat more optimistic conclusion – that a large fraction of a genome can be structurally annotated. There are two reasons for this difference. Firstly, we focus on one small genome, rather than on an average of all the known genomes and, secondly, we use the union of all the known structural matches made from advanced methods (e.g. PSI-BLAST), rather than the matches generated by one rather conservative method (FASTA).

Biological implications of structural assignments of genome sequences

Many of the results for sequences of MG that are discussed in the following sections of the review are based on the updated, two-way PSI-BLAST assignments of 314 domains to 223 MG proteins [20••].

Domain structure of genome sequences

Small proteins and most medium-size proteins contain a single domain. Large proteins comprise two or more domains, of which the large majority are known to undergo independent duplications and recombinations [8,42,43•]. The average size of the domains in proteins of known structure is about 175 residues.

The distributions of the lengths of sequences in bacterial and archaeal genomes have been to found to follow very similar extreme-value distributions [15•]. The most common length, 190 residues, is roughly the length of a single PDB domain and the average length is approximately 280 residues in archaea and 330 residues in

Table 2

Description of methods used to determine annotation.

Type of structural annotation	Number of ORFs with annotation	Additional fraction of total residues with annotation	Description of methods used to determine annotation
Original 1997 FASTA	90	13%	Base-line structural annotation: MG regions matched to PDB domain structures in 1997 using FASTA [5] with a very conservative E-value threshold of 0.01 and an old database (SCOP 1.35) [8].
One-way PSI-BLAST	161	10%	Additional regions matched to PDB structures (beyond the above), based on running PSI-BLAST [18••]. A more recent 1998 version of the PDB domains (SCOP 1.38) (excluding coiled-coils and small leucine- and cysteine-rich proteins) was run against MG embedded in NRDB (which had been masked in the default fashion by SEG [54]). These comparisons used 20 iterations and an inclusion threshold into the matrix of 0.0005, an overall match cut-off of 0.0001 and matches were continuously parsed from output.
Two-way PSI-BLAST	223	10%	Additional matches from running PSI-BLAST in a two-way fashion, plus preclustering the ORFs in MG [20••]. By 'two way', we mean that the PDB was first run against MG embedded in NRDB and then unmatched regions of MG were cut out and run against the PDB embedded in NRDB. The preclustering was done with GEANFAMMER [55]
All matches	242	9%	Additional matches by considering all the MG matches discussed in Table 1 (i.e. various PSI-BLAST approaches, threading etc [19••,21••,24••,30••]).
TM	79	7%	Surest annotation for TM helices in integral membrane proteins. These were segments of at least 20 residues with an average GES hydrophobicity less than −1 kcal/mol [13•,37] in a protein that had at least one TM segment with an average hydrophobicity less than −2 kcal/mol (adapted from Boyd and Beckwith's MaxH criteria [38].) Only about 7% of the residues are flagged as sure TM segments, but these occur in ~17% of the sequences.
LC	65	8%	Very long, LC regions. These are thought not to fold into globular protein structures. They were identified using the SEG program, with a trigger complexity K(1) of 3.4, an extension complexity K(2) of 3.75 and a window of length 45 [54]. In addition, the whole LC region had to be longer than 150 residues.
Signal sequence and linking peptides	258	7%	Hydrophobic signal sequences and linking peptides. Signal sequences have the pattern of a charged residue within first seven residues, followed by a stretch of 14 hydrophobic residues. Segments of sequence already accounted for thus far, that is, PDB matches, LC or TM helices, are considered to be 'characterised' regions. Short sequences (<80 residues) between characterised segments are considered to be linkers.
Low quality TM, LC and linking peptides	42	9%	This category consists of much lower quality structural annotation of TM helices and LC regions. That is, LC regions according to the same criteria as discussed above, but shorter than 150 amino acids, and TM helices with an average hydrophobicity less than −1 kcal/mol, but that are in proteins that do not meet the MaxH criteria.
Known function	131	18%	These regions have no other structural annotation, but occur in proteins given functional annotation by Mushegian and Koonin [56] or TIGR [4] and, thus, probably fold into globular structures.
No function	70	9%	Region has no structural annotation and occurs in a protein that is given no functional annotation by TIGR or Mushegian and Koonin [56] (as of January 1999).

eubacteria. The distributions of sequence lengths in *S. cerevisiae* and *C. elegans* are similar to those in prokaryotes, but there is a greater preponderance of long sequences. This results in larger average lengths (465 for yeast and 425 for the worm). These results imply that a significant fraction of the proteins produced by genomes contain two or more domains.

Based on the various types of structural annotation shown in Figure 1 and Table 2, it is possible to roughly estimate the number of soluble protein domains in MG. The two-way PSI-BLAST calculation shows that 223 (47%) MG sequences are matched, all or in part, by a PDB sequence. Of these, 83 MG sequences were completely matched by a single, known structural domain and 39 by between two and five domains [20••]. Another 101 sequences were matched to between one and four domains and had unmatched regions that are long enough for at least one additional domain to be present. These figures show that, for the MG sequences matched by PDB sequences, close to one-third of the sequences contain one domain and two-thirds have two or more domains.

So, according to this calculation, 314 PDB domains match all or part of a total of 223 MG proteins. These matches cover 33% of the MG proteome. Excluding the well-characterised TM, LC and linker regions, as well as the 314 PDB domains, we are left with regions that, presumably, code for soluble proteins with globular structures, but

Table 3

Common superfamilies in genomes*.

Eubacteria						
	M. genitalium		*B. subtilis*		*E. coli*	
Rank	Superfamily	Number of domains	Superfamily	Number of domains	Superfamily	Number of domains
1	Δ P-loop hydrolase	60	Δ P-loop hydrolase	192	Δ P-loop hydrolase	216
2	= SAM methyl-transferase	16	⊗ Rossman domain	133	⊗ Rossman domain	113
3	⊗ Rossman domain	13	● Phosphate-binding barrel	51	§ Periplasmic-binding protein II-like	80
4	Class I synthetase	12	♦ PLP transferases	44	● Phosphate-binding barrel	42
5	Class II synthetase	11	§ Periplasmic-binding protein II-like	44	♦ PLP transferases	38
6	OB-fold nucleic acid binding domain	11	Acyl-carrier protein	40	* CheY-like domains	36
Total ORFs		479		4100		4265
ORFs with common superfamilies		105 (22%)		438 (11%)		489 (11%)

Archaea				
	M. thermoautotrophicum		*A. fulgidus*	
Rank	Superfamily	Number of domains	Superfamily	Number of domains
1	Δ P-loop hydrolase	104	Δ P-loop hydrolase	124
2	◊ Ferredoxins	60	⊗ Rossman domain	71
3	⊗ Rossman domains	45	◊ Ferredoxins	52
4	● Phosphate-binding barrel	40	● Phosphate-binding barrel	37
5	Thiamin-binding	18	Thiolase	29
6	= SAM methyl-tranferase	17	Firefly luciferase	26
Total ORFs		1869		2409
ORFs with common superfamilies		246 (13%)		300 (13%)

Eukaryotes				
	S. cerevisiae		*C. elegans*	
Rank	Superfamily	Number of domains	Superfamily	Number of domains
1	Δ P-loop hydrolase	249	× Protein kinase	429
2	× Protein kinase	123	Δ P-loop hydrolase	411
3	⊗ Rossman domain	90	Ligand-binding nuclear receptor domain	254
4	RNA-binding domain	75	C-type lectin	253
5	= SAM methyl-transferase	63	α/β hydrolase	180
6	Ribonuclease H-like	57	Immunoglobulin superfamily	149
Total ORFs		6218		19,099
ORFs with common superfamilies		560 (9%)		1676 (8%)

*The superfamily descriptions are from the SCOP database, with the exception of the Rossmann domains and the phosphate-binding α/β barrels (see [17•]). Domains in this table can occur in multiple copies within one gene. This means that the total number of genes in which they occur is smaller than the total number of domains. Some numbers are almost certainly underestimates because PSI-BLAST cannot find all distant homologous relationships [17•] and conservative E-values were used to select matches. For example, using a hidden Markov model, we find 453 immunoglobulin domains in 76 proteins in *C. elegans.*

that do not have a known fold. As indicated in Figure 1, these 'uncharacterised regions' currently comprise about 40% of the MG genome (by residue). They are formed from about 270 whole or partial sequences. Of these, about three-quarters contain less than 200 amino acids and most of these probably have a single domain. Of the remainder, most probably have multiple domains. After putting the results of the known PDB matches and the uncharacterised regions together, one comes up with a very rough estimate of 700 soluble, globular domains in MG, of which about 200 form single-domain proteins.

Protein domains produced by gene duplications

If the total number of families to which most proteins belong is small [44], high levels of domain duplication would be expected in the genome sequences. Pairwise comparison of genome sequences and the clustering of matched sequences into families indicated that the proportion of sequences that have arisen by gene duplication is between a quarter (in small bacterial genomes) and half (in large bacterial genomes) [45–48]. The sequences in individual bacterial genomes, however, have relatively few pairs with residue identities that are greater than 30% (see, for example, [20••]). This means that duplication rates based on pairwise sequence comparison must be underestimates. For distantly related proteins, the detection of evolutionary relationships requires structural, functional and sequence information, which is only available for proteins whose structures have been solved.

In the two-way PSI-BLAST assignment of structures to MG proteins, 223 were matched, all or in part, by 314 PDB domain sequences. The inspection of the superfamily assignment given to the PDB domains in SCOP shows that they belong to 124 different superfamilies. Eighty-two MG sequence regions are unique representatives of their superfamily and 232 sequences belong to one of 42 superfamilies, with each having between 2 and 60 homologues. Thus, the proportion of these MG sequences that has arisen by gene duplication is (314–82–42)/314, that is, 60%. This proportion is more than twice that found from pairwise sequence comparisons [20••].

Using PSI-BLAST, the sequences of proteins of known structure can be matched to 30 and 27% of the protein sequences in *S. cerevisiae* and *C. elegans*, respectively ([21••]; SA Teichmann, C Chothia, unpublished data). These matched regions form, respectively, 18% and 15% of the amino acids in the two genomes. Carrying out calculations similar to those described above shows that the proportion of domains produced by gene duplications in matched regions is 88% for *S. cerevisiae* and 95% for *C. elegans*.

Protein families and folds in genomes

Proteins have evolutionary and structural relationships. Proteins with evolutionary relationships are descended from a common ancestor. For more closely related proteins, this can be detected from sequence similarities, which allow proteins to be clustered into families. For distantly related proteins, the detection of evolutionary relationships requires structural, functional and sequence information. This information is used collectively in the SCOP database to cluster proteins of known structure into superfamilies.

Proteins can also have structural similarities that arise not from common descent, but as a result of physics and chemistry favouring certain secondary-structure packing and chain topologies ([see [49] for a recent review). Proteins that have the same major secondary structures with both the same arrangement and the same topology are clustered into folds that are described in the SCOP and CATH databases. It is important to note that two proteins having the same fold does not, by itself, indicate their descent from a common ancestor.

Bacterial genome sequences have been clustered into families using pairwise comparisons [13•,43•,46–48]. These calculations showed that the sizes of the families have an exponential character — many families with one or a few sequences and a few families with many sequences. Subsequently, using the SCOP classification, a number of groups have determined the superfamily and fold membership of the genome sequences that match known structures [13•,14•,20••]. Wolf *et al.* [21••] have described the most common folds in 13 genomes. Lists of the six largest superfamilies found in various genomes are given in Table 3 (SA Teichmann, C Chothia, unpublished data). The distributions of the superfamilies (Table 3) show systematic differences when small parasitic bacteria are compared with free-living bacteria and when both are compared with eukaryotes, a fact previously noted with regard to fold distributions [14•,21••,50].

In Table 3, the size of a superfamily is measured by the number of different homologues within the genome. Folds and superfamilies can also be ranked by their level of mRNA expression [14•] or even by the direct measurement of protein levels in the cell. These will give different rankings, in particular, elevating ribosomal folds, which are highly expressed, but not highly duplicated.

It should be noted that the current information on the superfamilies and folds in genome sequences is limited to the 15–35% of the genome that can be matched to sequences of known structure and that, in genomes, there are undetected homologues of known structure, as well as common folds that are not related to the structures known at present. Consequently, one should take the current numbers (Table 3) as only rough and somewhat biased approximations. Nevertheless, it is remarkable that the common folds identified in the early calculations [13•,14•] are largely similar to those identified using the newer PSI-BLAST methods [21••].

Conclusions

The work described here has shown that pairwise comparisons, PSI-BLAST, profiles and threading techniques can assign structures to all or part of between one-quarter and one-half of the sequences in different genomes and that these matches cover between 15 and 40% of all residues in the genome. We can expect these proportions to increase rapidly as a result of improvements in computational techniques and experimental structural genomics projects. The most powerful sequence matching technique, which uses hidden Markov models [17•,51,52], has not been used for large-scale matching so far. On the experimental side, we see from Figure 1 that most of the structures that match approximately 40% of the MG genome were determined over the past 12 years. The rapid increase in the number of both structure determinations and, particularly, programs for experimental structural genomics should mean that the time required to determine the globular structures that occur in the remaining approximately 35% of the genome should be much shorter.

Although the current results only cover parts of genomes, they are of great interest. The matched regions are often the product of gene duplications of domain sequences and their recombination. A few families have many members and play a major role. There is no reason, at present, to believe that results of the same kind will not be found for globular domains in the regions that, up to now, have not been assigned structures. The current results support the hypothesis that the domains that form the most proteins come from a small number of superfamilies. Also, the observation that many of the proteins involved in the most basic functions of simple cells are the product of duplications and recombinations implies that these processes initially occurred in cells that were much simpler than any now known [46].

Note added in proof

Jones [57••] recently published structural assignments to 218 MG ORFs with a high reliability, which is 46% of the proteins and 30% of the amino acids. This calculation found 17 assignments to MG ORFs not found by any of the groups in Table 1.

Acknowledgements

We thank Eugene Koonin for providing information prior to publication and for comments on the manuscript. We thank Steven Brenner and Leszek Rychlewski for data. SAT thanks the Boehringer Ingelheim Fonds for support and MG thanks the Donaghue Foundation.

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

1. •• Zarembinski TI, Hung L-W, Muller-Dieckmann H-J, Kim K-K, Yokota H, Kim R, Kim S-H: **Structure-based assignment of the biochemical function of a hypothetical protein: a test case for structural genomics.** *Proc Natl Acad Sci USA* 1999, **95**:15189-15193.

This paper reports the first structure determined by a structural genomics project, that of protein MJ0577 from the hyperthermophile *M. jannaschii*. The gene is a representative of a family of hypothetical proteins and the structure shows an ATP-binding pocket and, possibly, a dimerisation interface.

2. Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J: **Class-directed structure determination: foundation for a protein structure initiative.** *Protein Sci* 1998, **7**:1851-1856.

3. Shapiro L, Lima CD: **The Argonne structural genomics workshop: Lamaze class for the birth of a new science.** *Structure* 1998, **6**:265-267.

4. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270**:349-548.

5. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.

6. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-197.

7. Altschul SF, Gish W, Miller W, Myers EW, Lipmann DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.

8. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.

9. Orengo CA, Michie AD, Jones S, Jones DJ, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093-1108.

10. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C: In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, California: AAAI Press; 1994:348-353.

11. Frishman D, Mewes H-W: **PEDANTic genome analysis.** *Trends Genet* 1997, **13**:415-416.

12. • Frishman D, Mewes H-W: **Protein structural classes in five complete genomes.** *Nat Struct Biol* 1997, **4**:626-628.

After assigning structures to the sequences of five genomes using FASTA, secondary-structure prediction using the program PREDATOR was carried out on the remaining sequences. Similar proportions of protein structural classes are found in all genomes.

13. • Gerstein M: **A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure.** *J Mol Biol* 1997, **274**:562-576.

This paper compares the first genomes sequenced from each of the three kingdoms of life (yeast, *H. influenzae* and *M. jannaschii)* in terms of protein structure, focusing on supersecondary structure. It is shown that the most common folds shared among all three kingdoms have a remarkably similar supersecondary structure architecture, containing a central sheet with helices packed onto at least one face. In terms of predicted supersecondary structures, it is shown that yeast has a preponderance of consecutive strands and *H. influenzae*, consecutive helices.

14. • Gerstein M: **Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census.** *Proteins* 1998, **33**:518-534.

This paper compares eight genomes in terms of their patterns of fold usage. It is shown that the genomes can be clustered into a plausible tree according to fold usage (rather than sequence similarity), that common (highly duplicated) folds are often superfolds and that folds can be ranked by expression, as well as by duplication. Membrane protein folds were surveyed and were also found to be fairly similar among the genomes.

15. • Gerstein M: **How representative are the known structures of the proteins in a complete genome? A comprehensive structural census.** *Fold Des* 1998, **3**:497-512.

This paper describes how representative the known structures (from the PDB) are of the proteins encoded by a complete genome. It was found that proteins in the genomes differ in terms of length (being longer than PDB proteins) and composition (having more lysine, isoleucine, aspargine and glutamine and less cysteine and tryptophan). A procedure for segmenting genome sequences into various regions (known structure, TM helices etc) was introduced and secondary structure was predicted for 'uncharacterised regions'. These are found to be more helical than proteins in the PDB, but have a rather constant secondary-structure content, despite their differences in amino acid composition.

16. • Brenner SE, Chothia C, Hubbard TJP: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc Natl Acad Sci USA* 1998, **95**:6073-6078.

See annotation to [17•].

17. • Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia, C: **Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284**:1201-1210.

This paper, together with [16•], describes assessments of sequence comparison procedures. [16•] deals with pairwise comparisons and [17•] discusses comparisons that use multiple sequences. Their relative and absolute success was assessed, as well as the accuracy of their scoring schemes. To find the relationships among homologous sequences with identities of less than 30%, the hidden Markov model procedure SAM-T98 does somewhat better than PSI-BLAST and both do about three times better than pairwise comparisons. All the methods, however, fail to find a significant fraction of distant relationships.

18. •• Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

This paper describes the PSI-BLAST program, which has played a major role in the assignment of structures to genome sequences.

19. •• Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P: **Homology-based fold predictions for *Mycoplasma genitalium* proteins.** *J Mol Biol* 1998, **280**:323-326. [URL: http://dove.embl-heidelberg.de/3D/]

See annotation to [21••].

20. •• Teichmann SA, Park J, Chothia C: **Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplication and domain rearrangement.** *Proc Natl Acad Sci USA* 1998, **95**:14658-14663. [URL: http://www.mrc-lmb.cam.ac.uk/genomes/MG_strucs.html]

See annotation to [21••].

21. •• Wolf YI, Brenner SE, Bash PA, Koonin EV: **Distribution of protein folds in the three superkingdoms of life.** *Genome Res* 1999, **9**:17-26. [URL: ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/FOLDS/index.html]

Papers [19••–21••] describe the large-scale assignment of structures to sequences from genomes using the PSI-BLAST program. Huynen *et al.* [19••] and Teichmann *et al.* [20••] assigned structures to MG sequences. Wolf *et al.* [21••] assigned structures to sequences from 13 genomes. Huynen *et al.* [19••] also described a calibration of the parameters used in PSI-BLAST. Teichmann *et al.* [20••] discuss the implications of their matches on the domain structure, gene duplications and major families in MG. Wolf *et al.* [21••] list the major protein folds in the matched sequences of 13 genomes.

22. Holm L, Sander C: **Removing near-neighbour redundancy from large protein sequence collections.** *Bioinformatics* 1998, **14**:423-429.

23. Jaroszewski L, Rychlewski L, Zhang BH, Godzik A: **Fold prediction by a hierarchy of sequence, threading and modelling methods.** *Protein Sci* 1998, **7**:1431-1440.

24. •• Rychlewski L, Zhang B, Godzik A: **Fold and function predictions for *Mycoplasma genitalium* proteins.** *Fold Des* 1998, **3**:229-238. [URL: http://cape6.scripps.edu/leszek/genome/cgi-bin/genome.pl?mp]

The authors describe the assignment of folds to MG genome sequences using a profile-to-profile matching method. Function predictions are made for some MG proteins without functional annotations using their homology to *E. coli* proteins.

25. •• Sanchez R, Sali A: **Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome.** *Proc Natl Acad Sci USA* 1998, **95**:13597-13602.

This was the first attempt at making three-dimensional models of as many sequences as possible for an entire genome. This was possible for 17% of the yeast genome. Making homology models for all of the sequences in a genome is one of the aims of structural genomics projects.

26. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M: **Life with 6000 genes.** *Science* 1996, **274**:546.

27. Finkelstein AV, Reva BA: **A search for the most stable folds of protein chains.** *Nature* 1991, **351**:497-499.

28. Sippl MJ, Weitckus S: **Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations.** *Proteins* 1992, **13**:258-271.

29. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358**:86-89.

30. •• Fischer D, Eisenberg D: **Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*.** *Proc Natl Acad Sci USA* 1997, **94**:11929-11934. [URL: http://www.doe-mbi.ucla.edu/people/frsvr/preds/MG/MG.html]

This was the first paper to assign structures to a complete genome. A fold recognition method developed by the group was calibrated and applied to the MG genome. An estimate of the time and the number of new structures needed for the complete assignment of this genome was made.

31. Grandori R: **Systematic fold recognition analysis of the sequences encoded by the genome of *Mycoplasma pneumoniae*.** *Protein Eng* 1998, **11**:1129-1135.

32. Frishman D, Argos P: **Seventy-five percent accuracy in protein secondary structure prediction.** *Proteins* 1997, **27**:329-335.

33. Garnier J, Gibrat JF, Robson B: **GOR method for predicting protein secondary structure from an amino acid sequence.** *Methods Enzymol* 1996, **266**:540-553.

34. Goffeau A, Slonimski P, Nakai K, Risler JL: **How many yeast genes code for membrane-spanning proteins?** *Yeast* 1993, **9**:691-702.

35. Rost B, Fariselli P, Casadio R, Sander C: **EXTRA-REF: prediction of helical transmembrane segments at 95% accuracy.** *Protein Sci* 1995, **4**:521-533.

36. Rost B: **PHD: predicting one-dimensional protein secondary structure by profile-based neural networks.** *Methods Enzymol* 1996, **266**:525-539.

37. Arkin I, Brunger A, Engelman D: **Are there dominant membrane protein families with a given number of helices?** *Proteins* 1997, **28**:465-466.

38. Boyd D, Schierle C, Beckwith J: **How many membrane proteins are there?** *Protein Sci* 1998, **7**:201-205.

39. Jones DT: **Do transmembrane protein superfolds exist?** *FEBS Lett* 1998, **423**:281-285.

40. Wallin E, von Heijne G: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7**:1029-1038.

41. • Brenner SE, Barken D, Levitt M: **The PRESAGE database for structural genomics.** *Nucleic Acids Res* 1999, **27**:251-253.

The database PRESAGE is a repository of the structural assignments made for proteome sequences. It is likely to be a useful resource of information for those interested in structural genomics and is on the World Wide Web at http://presage.stanford.edu/.

42. Rossmann MG, Moras D, Olsen KW: **Chemical and biological evolution of a nucleotide-binding protein.** *Nature* 1974, **250**:194-199.

43. • Riley M, Labedan B: **Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module.** *J Mol Biol* 1997, **268**:857-868.

This paper was the first to show that the multidomain character of many protein structures is also common to many proteins in genomes.

44. Chothia C: **One thousand families for the molecular biologist.** *Nature* 1992, **357**:543-544.

45. Labedan B, Riley M: **Widespread protein sequence similarities: origins of *E. coli* genes.** *J Bacteriol* 1995, **177**:1585-1588.

46. Brenner SE, Hubbard T, Murzin A, Chothia C: **At least one third of the proteins in *Haemophilus influenzae* arose from gene duplications.** *Nature* 1995, **378**:140.

47. Koonin EV, Tatusov RL, Rudd KE: **Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications.** *Proc Natl Acad Sci USA* 1995 **92**:11921-11925.

48. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archeael and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for archaea.** *Mol Microbiol* 1997, **25**:619-637.

49. Chothia C, Hubbard T, Brenner S, Barns H, Murzin A: **Protein folds in the all-β and all-α classes.** *Annu Rev Biophys Biomol Struct* 1997, **26**:597-627.

50. Gerstein M, Levitt M: **A structural census of the current population of protein sequences.** *Proc Natl Acad Sci USA* 1997, **94**:11911-11916.

51. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology – applications to protein modelling.** *J Mol Biol* 1994, **235**:1501-1531.

52. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.

53. Gerstein M, Hegyi H: **Comparing microbial genomes in terms of protein structure: surveys of a finite parts list.** *FEMS Microbiol Rev* 1998, **22**:277-304.

54. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266**:554-571.

55. Park J, Teichmann SA: **DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins.** *Bioinformatics* 1998, **14**:144-150.

56. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

57. Jones DT: **GenTHREADER: an efficient and reliable protein fold
•• recognition method for genomic sequences.** *J Mol Biol* 1999, **287**:797-815.

This paper describes structural assignments to MG using a combined sequence alignment and threading method.