BIOINFORMATICS
Vol. 1 no. 1 2001
Pages 1–7

An Insight into Domain Combinations

Apic Gordana[1], Julian Gough[1] and Sarah A. Teichmann[2]

[1]MRC Laboratory of Molecular Biology, , Hills Road, Cambridge, CB2 2QH, UK and [2]Department of Biochemistry & Molecular Biology, University College London, Darwin Bldg., Gower Street, London, WC1E 6BT, UK

Received on February 6, 2001; revised and accepted on March 29, 2001

ABSTRACT

Domains are the building blocks of all globular proteins, and are units of compact three-dimensional structure as well as evolutionary units. There is a limited repertoire of domain families, so that these domain families are duplicated and combined in different ways to form the set of proteins in a genome. Proteins are gene products, and at the level of genes, duplication, recombination, fusion and fission are the processes that produce new genes. We attempt to gain an overview of these processes by studying the structural domains in the proteins of seven genomes from the three kingdoms of life: Eubacteria, Archaea and Eukaryota. The domain and superfamily definitions in the Structural Classification of Proteins Database are used, so that we can view all pairs of adjacent domains in genome sequences in terms of their superfamily combinations. We find 624 out of the 764 superfamilies in SCOP in these genomes, and the 624 families occur in 585 pairwise combinations. Most families are observed in combination with one or two other families, while a few families are very versatile in their combinatorial behaviour. This type of pattern can be described by a scale-free network. Finally, we study domain repeats and we compare the set of the domain combinations in the genomes to those in PDB, and discuss the implications for structural genomics.

Contact: apic@mrc-lmb.cam.ac.uk

INTRODUCTION

All proteins consist of one or more domains, with the exception of some disordered proteins. Domains are units of compact three-dimensional structure (Murzin *et al.*, 1995; Orengo *et al.*, 1997), and also units of evolution (Riley & Labedan, 1997). There is a limited repertoire of types of domains (Chothia, 1992; Wolf *et al*., 2000), such that domains from this set are duplicated and combined in different ways to form the set of proteins in a genome. Proteins are gene products, and at the level of genes, duplication, recombination, fusion and fission are the processes that produce new genes.

We attempt to gain an overview of these processes by studying the structural domains in the proteins of seven genomes. We use structural assignments to genome sequences, because proteins of known three-dimensional structure have clearer domain definitions and evolutionary family relationships than sequences of unknown structure. The domain and superfamily definitions in the Structural Classification of Proteins Database (Murzin *et al.*, 1995) are used, so that we can view all pairs of adjacent domains in genome sequences in terms of their superfamily combinations. This allows us to survey and compare the set of domain family combinations present in the archaeal, bacterial and eukaryote genomes. We study pairwise domain combinations and domain repeats. We also compare the set of the domain combinations in the genomes to those in PDB, and discuss the implications for structural genomics.

DOMAINS IN THE STRUCTURAL CLASSIFICATION OF PROTEINS (SCOP) DATABASE

As mentioned above, the domain definition used here is that of the Structural Classification of Proteins (SCOP) database developed by Murzin *et al.* (1995). SCOP contains domain definitions and describes evolutionary relationships between the proteins whose three-dimensional structures have been determined. The domain is the unit of classification in SCOP.

We used SCOP version 1.48 containing 21,828 structural domains clustered into 764 superfamilies from 9580 PDB entries. In this work, we studied the 764 superfamilies assigned to genomic sequences in order to elucidate the domain structure and combinations in seven genomes. We use the term “family” in the remainder of the text meaning a SCOP superfamily.

SEVEN GENOMES AND THREE PHYLOGENETIC GROUPS

SCOP domain assignments to genome sequences are the data analysed here. The set of completely sequenced genomes we chose are diverse, so that we hope to

cover much of the domain family and domain combinatorial space in nature. The genomes are two archaea (*Archaeoglobus fulgidus*, AF; *Methanobacterium thermoautotrophicum* , MT), two eubacteria (*Escherichia coli*, EC; *Bacillus subtilis*, BS), one unicellular eukaryote (*Saccharomyces cerevisiae*, SC) and two multicellular eukaryotes (*Caenorhabditis elegans*, CE; *Drosophila melanogaster*, DM). These organisms come from very different external environments: for instance their optimal temperatures range from room temperature (SC) to 85°C in deep marine subsurface oil areas (AF). These genomes cover multicellular and unicellular organisms with different modes of life, from autotrophs (MT) to optional parasites (EC, BS).

In some parts of this investigation, it is convenient to compare the three different phylogenetic groups. To create one phylogenetic group, we take two genomes from this group and make a non-redundant union of the features studied in the two genomes, for instance domain combinations or tandem domains. When we refer to the phylogenetic group in the text we mean these non-redundant unions. The ,,Archaea“ group is composed of AF and MT, ,,Eubacteria“ groups together the gram-positive BS and the gram-negative EC and ,,Eukarya“ merges the unicellular SC and the multicellular CE.

ASSIGNING DOMAINS TO GENOME SEQUENCES

The SCOP domains were assigned to the genome sequences using Hidden Markov Models (HMMs) (Eddy, 1996; Krogh *et al.*,1994). HMMs are one of the most sensitive sequence comparison methods currently available (Park *et al.*, 1998). For each non-identical SCOP domain, at the 95% sequence identity level, an HMM was generated by Gough and co-workers (Gough *et al.*, 2000) using the iterative SAM-T99 method (Karplus *et al.*, 1998). The genome sequences were scanned against this Hidden Markov Model library and this resulted in assignment of SCOP domains to the genomic sequences. Using this method, we obtained a precise description of the domain composition for 10-30% of the proteins in these genomes that are matched for the entire length of their sequences. We obtained an insight into about one-half of the proteins in these genomes, if the protein matches at least one SCOP domain but also contains an unassigned region of more than fifty residues in length.

Previously, Teichmann *et al.* (1998) and Gerstein (1998) observed that about two-thirds of the genomic sequences consist of more than one domain. In accordance with this, the majority of the genomic proteins that have structural assignments in our data set are multi-domain proteins, which have evolved through gene duplication, recombination, fusion and fission.

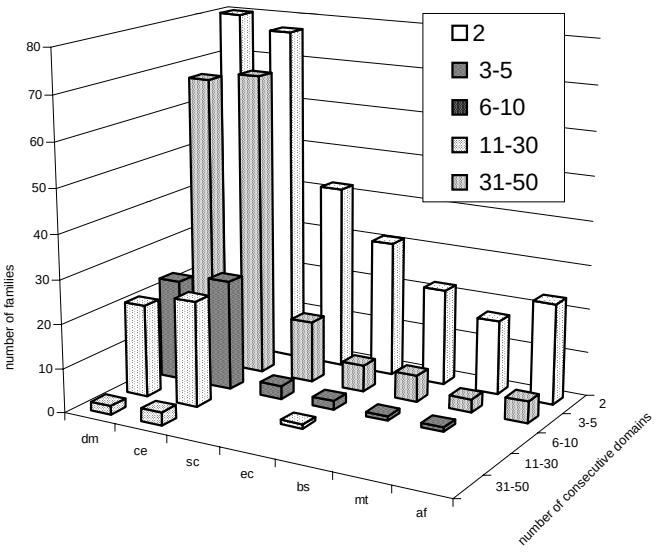

Fig. 1. Distribution of the sizes of repeated domains in seven genome. The distributions of tandem domains are plotted here for each individual genome. The number of families that form repeats of a certain length determines the height of the bar. We observe that a repeat length of two domains occurs most frequently in all genomes. The distribution is very similar for all unicellular organisms (AF, MT, BS, EC, SC) with very few repeats that are five to ten domains long. Multicellular organisms (DM and CE) have a significant fraction of much longer repeats, from 10 to 50 domains long.

With the extensive information we have on the domain structures and evolutionary relationships of the proteins in seven completely sequenced genomes, we want to investigate the patterns of domain combinations and thus reveal the driving

forces for the evolution of more complex proteins. First we will turn our attention to tandem domains in proteins, and then to combinations of different types of domains. Finally, we ask whether more complex proteins have evolved by the creation of the new protein families, or the recombination of existing families.

TANDEM DOMAINS FROM THE SAME FAMILY IN POLYPEPTIDE CHAINS

Tandem domains from the same family within one polypeptide chain, also called domain repeats, may have evolved by recombination or fusion, in the same way as adjacent domains from two different families. However, tandem domains may have also evolved by a different mechanism, internal duplication. Therefore, we consider this type of domain combination separately from combinations of different domain types. We define tandem domains as adjacent domains from the same family with

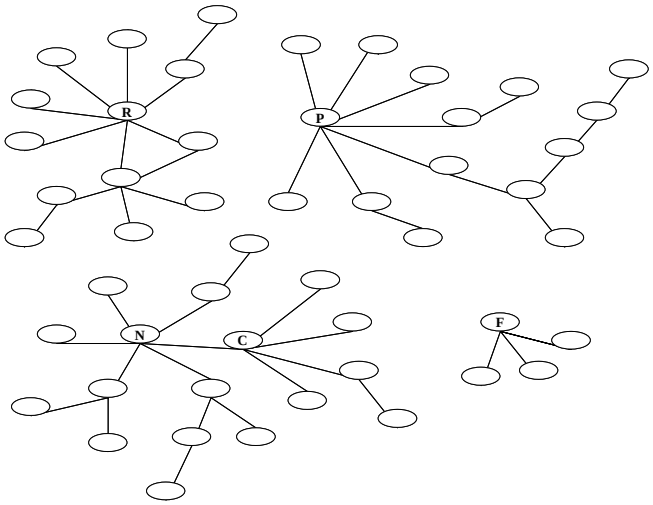

Fig. 2. Network of domain combination partners for the AF genome.The graph representing the five most versatile domain families and their combination partners is given here for the AF genome. Each ellipsoid represents a different protein family. Ellipsoids with a letter inside them are the most versatile families ("R" for Rossmann domain, "P" for P-loop hydrolases, "N" for nucleic acid binding proteins, "C" for Class II synthases, and "F" for FAD/NAD(P) binding domain). For example, Rossmann domain combines with seven different domains, and one of them combines with four different domains. The graph resembles a scale-free network. The hubs are the most versatile families, which have unique repertoires of combination partners connected to them. AF is the simplest genome studied here in terms of domain combinations. The four key groups of connections surrounding the hubs are not interconnected here. In Eubacteria and Eukarya the hubs have more combination partners and the key groups are more interconnected, still preserving the fairly unique repertoire of combination partner for each key family.

less than thirty residues between them. We are interested in how frequently evolutionary mechanisms have resulted in tandem domains in the different groups of genomes, and the particular protein families involved.

We observe that less than ten percent of genomic sequences contain tandem domains, and only a small fraction of genomic families, 10-20%, are seen as tandems. Domains of the same family can be internally duplicated just once, resulting in a tandem of two domains, or they can be duplicated more times, resulting in many consecutive domains. The number of consecutive domains from the same family against the number of occurrences in the individual genomes is plotted in Figure 1. We observe that multicellular organisms contain much longer repeats than the unicellular organism.

The question arises as to the nature of these long repeats of domains in the metazoa. Are the repeats of domains from families in the protozoa just longer in metazoa, or are they repeats of families that arose later in evolution, and are thus specific to metazoa? Almost all families that are seen in repeats, that are present in all three phylogenetic groups, are enzymatic families and their repeats are rarely longer than two domains in any organism.

The families involved in the longest repeats in the metazoan organisms, repeats of 30 to 50 domains, are specific to metazoa. These are extracellular domains involved in cell adhesion and signaling, or intracellular regulatory and signaling families. Cell adhesion and complex signaling, as well as regulatory mechanisms, became important as multicellular organisms evolved. We show here that some of the additional demands in these organisms were met by internal duplication of metazoa-specific protein families. For the families involved in cell adhesion or other functions related to the cellular or physiological structure of an organism, the domain repeats provide a structural role, such as immunoglobulin domains in muscle proteins or laminin domains in proteins in the extracellular matrix.

Many of the families specific to metazoa involved in long repeats are flexible in the number of domains adjacent to each other in the proteins. For instance in CE, the immunoglobulin repeats are present in thirteen different lengths ranging from two to fifty-two domains. It was shown for the cadherin family in CE and DM (Hill *et al.*, 2000) and immunoglobulins in CE (Teichmann *et al.*, 2000) that gene predictions for the long genes involving these families were often incomplete. Therefore, the eukaryotic domain repeats may be even longer than shown here.

Table 1. Fraction of families that combine with one, two or three or more families.

Genomes	% of families combining with 1 family	% of families combining with 2 family	% of families combining with 3 family
AF	38	7	5
MT	40	8	5
BS	45	8	4
EC	43	10	7
SC	38	5	4
CE	35	10	11
DM	36	9	11

COMBINATIONS OF DOMAIN FAMILIES

As for the tandem domains from the same family, we consider a pair of domains to be "neighbours" in general if they are not more than thirty residues apart. If domains are neighbours in one polypeptide chain, we say that they

combine with each other in the course of evolution, as the definition of a SCOP entry is an independent evolutionary unit.

In the seven genomes from the three kingdoms of life, we observe that only a small fraction of families combines with more than one other family. About half of the families are not observed combined with other domains, and about one third of them are seen as neighbours to only one other family. The details are given for each individual genome in Table 1. Thus for the majority of families that have domain neighbours, the adjacent domains are from one or two types families.

A few families are very versatile in their combination partners, however, as shown in Table 2. Most of these families are also the most abundant ones in genomes (Teichmann *et al.*, 1999). The reason for the abundance and versatility of these families is their function. For instance, the energy for motion and reactions in the cell is often provided by the P-loop nucleotide triphosphate hydrolases. Domains from this family hydrolyse ATP or GTP and can act as kinases and transferases on their own or combined with different families. Rossman domains are similar in that they provide oxidising or reducing energy through oxidation or reduction of the NAD(P)(H) cofactor. Transcription and translation are tightly regulated by proteins that consist of nucleic acid binding motifs, such as "winged helix" DNA binding domains or RNA binding domains, combined with other domains responsible for the specificity of the regulation.

The pattern of few families combining with many other domains, and most families having one or few partners, is that of a power law, as shown in Figure 2. This power law implies that the graph of domain combinations is a scale-free network. The part of the graph surrounding the five most versatile families in *Archaeoglobus fulgidus* is shown in Figure 3. A scale-free type of network, recently described for the World Wide Web (Albert *et al.*, 2000) and metabolic pathways (Jeong *et al.*, 2000), has a couple of key nodes that are connected with many other nodes. All the other nodes have only a few connections. By definition the key nodes, or hubs, in this type of network have a fairly unique repertoire of nodes connected to them. This means that the versatile families, such as P-loop nucleotide triphosphate hydrolases and Rossmann domains have a unique repertoire of combination partners that they do not share with other key families. With the future progress in homology detection and new protein structures, the domain combination network is going to expand, but there is no reason to believe that its form will change significantly.

From the results presented above, we conclude that the number of types of neighbours is small for most families. Only a few protein families are very versatile in their combination partners, and these versatile families are also the largest families in the genomes.

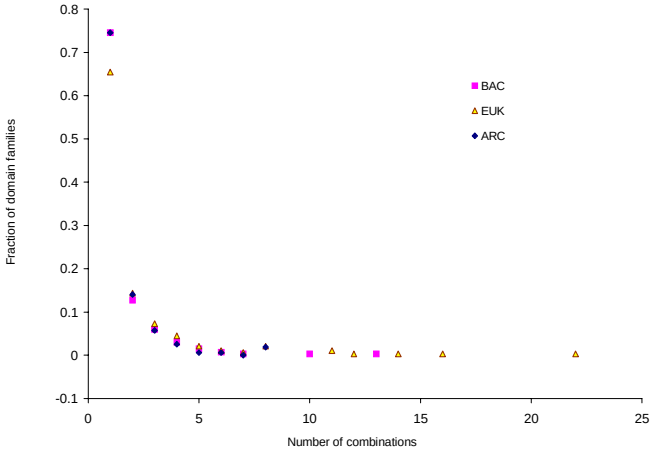

Fig. 3. Graph of fraction of domain families against number of combination partners. The number of partner families is given on the x-axis. On the y-axis, the fraction of domain families out of all families that combine with other domains is plotted. We observe that about 75 percent of the families combine with only one family in all phylogenetic groups. "ARC" is Archaea, "BAC" is Eubacteria, and "EUK" is Eukarya. Only a few families in the Eubacteria and Eukarya combine with more than ten families, representing less than 1 percent of families. A power law can be fitted to the data as follows: for ARC $y=0.47x^{-3.8}$ and $R^2=0.44$; for BAC $y=0.6x^{-2.2}$ and $R^2=0.95$; for EUK $y=0.47x^{-1.8}$ and $R^2=0.93$.

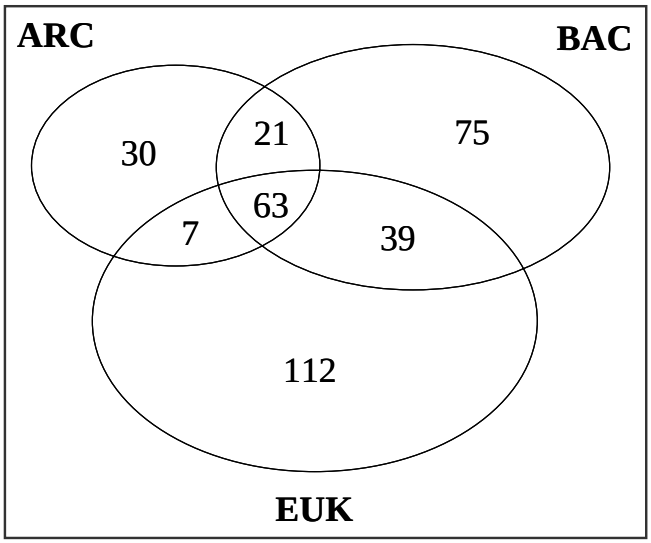

Fig. 4. Unique and Shared Domain Combinations. The diagram shows the pattern of domain combinations involving domain families common to the three phylogenetic groups: ARC for Archaea, BAC for Eubacteria and EUK for Eukarya. There are 63 family combinations that are shared among all three kingdoms.

Table 2. The most versatile families. The number of combination partners for each family is given in brackets next to the family name. The most versatile families in all genomes are Rossmann domains and P-loop hydrolases.

Genomes	AF	MT	BS	EC	SC	CE	DM
Rank	Domain family (number of combination partners)						
	Ploop hydrolase (8)	Ploop hydrolase (12)	Ploop hydrolase (12)	Ploop hydrolase (14)	Ploop hydrolase (14)	Ploop hydrolase (18)	Ploop hydrolase (17)
	Rossmann domain (8)	Nucleic acid binding domain (8)	Rossmann domain (7)	Rossmann domain (12)	Rossmann domain (8)	EGF (16)	Ig (13)
	Nucleic acid binding domain (7)	4Fe4S Ferredoxin (5)	Winged Helix DNA binding domain (6)	Winged Helix DNA binding domain (9)	Nucleotidyl transferase (5)	EFhand (12)	Protein kinase (11)
	Class II amino acyl tRNA synthases (5)	Rossmann domain (4)	Class II amino acyl tRNA synthases (6)	Homeodomain like (7)	Class II amino acyl tRNA synthases (5)	Protein kinase (11)	SH2 (11)
	FAD/NAD(P) binding domain (4)	Homeodomain like (3)	Nucleic acid binding domain (5)	Class II amino acyl tRNA synthases (7)	Glutamine amidotransferase class II (4)	SH3 domain (10)	EFhand (12)
	Gluthatione synthetase ATP domain (4)	Nucleotidyl transferase (3)	Nucletidyl transferase (4)	Che Y like (6)	Homing endonocleases (4)	Ctype lectin (10)	Rossmann domain (10)

CONSERVATION AND VARIATION OF DOMAIN COMBINATIONS AMONG THE THREE PHYLOGENETIC GROUPS

In order to assess the extent to which domain combinations are variable in the different phylogenetic groups, we take families common to all three phylogenetic groups, and observe how these families combine with other families in each group of genomes. We have compared the unions of domain combinations in AF and MT, EC and BS, and SC and CE. The three sets of domain combinations from the different kingdoms of life share 255 families, which we call “common families”. Most of the families in Archaea (80%), more than a half in Eubacteria (60%), and about half of the eukaryotic families are “common families”.

The domain combinations that involve “common families” represent most of the domain combinations present in each phylogenetic group, ranging from two-thirds in Eukarya to 90% in Archaea (Table3). The 255 “common families” occur as combinations among themselves, in other words “common family-common family combinations”, or in combination with families that are not common to all three phylogenetic groups. 63 of the “common family-common family combinations” are found in Archaea, Eubacteria and Eukarya, as shown in Figure 4. These 63 combinations represent families involved in macromolecule and small molecule metabolism. There are seven combinations that are shared only between Archaea and Eukarya. These combinations involve Rossmann domains, ADC domains, P-loop nucleotide triphosphate hydrolases and Ferredoxins, combined with the cell-cycle regulatory families, cyclins and Cdc48. The cyclins and Cdc48 homologues in Archaea are transcription initiation factors and AAA family ATPases respectively, that are distantly related in sequence and function to the eukaryotic proteins.

One third of the combinations involving “common families” in Eubacteria and Eukarya, and one-quarter in Archaea, are “common family” combinations specific to that kingdom. The question arises whether specific combinations for one kingdom are from combinations of “common families” among themselves, or whether the diversity comes from combining “common families” with other families that are specific to this kingdom. The results (Table 3) show that 60 to 90 percent of the combinations present in only one kingdom are made of “common families” combining with each other in a way not observed in the genomes of other kingdoms. Assuming that at least a large fraction of these combinations is genuinely not present in all three kingdoms, this shows that novel combinations of domains, even among ancient families, are an important part of the process of divergence of genomes that includes sequence divergence, and expansion and contraction of domain families.

Table 3. Domain Combinations Involving Common Families. The total number of combinations gives all the combinations of all families in a kingdom. 60-90% of these are combinations that involve common families, as given in the third column of the table. Some of these combinations involving common families are specific for one kingdom. Combinations that are specific for one kingdom are largely made of unique combinations of families shared between kingdoms, and to a small extent of common families combined with families that are specific for that particular kingdom.

Group	Total no. of combinations	No. of combinations of common families				
		with all families	specific to the kingdom	with families specific to the kingdom / with families common to all kingdoms	with families shared with one kingdom	with families shared with both other kingdoms
Archaea	132	121	30	3 / 27	28	63
Eubacteria	230	198	75	17 / 50	60	63
Eukarya	335	221	112	44 / 68	46	63

DOMAIN COMBINATIONS IN GENOMES ONLY: TARGETS FOR STRUCTURAL GENOMICS

The whole set of domain combinations that we observe in all seven genomes comprises 585 combinations among 624 families. 201 of these combinations are also observed in the PDB, which contains 272 domain combinations. Thus we detect three-quarters of all combinations in the PDB in the seven genomes. There are 71 combinations in the PDB from organisms different than those studied here that we do not see in the seven genomes. One-half of these combinations come from different phylogenetic groups such as mammals, viruses and plants, and the other half is from different types of bacteria and fungi.

This means that we detect 384 novel combinations of structural superfamilies, not seen in PDB. Structural genomics projects have the aim of extending the fold library and hence solving the structures of sequences without assigned PDB structures (Burley, 2000; Brenner, 2000), but another aim of the structural genomics efforts should be to elucidate the way domains interact with each other. This will provide important insights into the function of individual proteins and protein-protein interactions. The 384 novel combinations of structural superfamilies, not seen in PDB, are suitable to as targets for structural genomics.

CONCLUSIONS

Here we provide a first quantitative insight into the domain combinations in the three kingdoms of life. The majority of genomic proteins are multi-domain proteins created as a result of combination of domains. We observe a pattern of domain combinations where only few domain families combine with many types of domains and most families have one or few combination partners. This pattern can be described as a scale-free network.

The majority of domain combinations in all three kingdoms of life are those involving families that are shared among the three kingdoms, the “common families”. We show that there are many combinations between common families that are specific for one kingdom, and that the domain recombination of ancient families among each other contributes more to the process of divergence at the level of domain combinations than ancient families combining with kingdom-specific families. This implies that in creating new functions, nature more frequently combines old building blocks than inventing new ones. The analysis of domain repeats shows that special functional requirements that appeared with the evolution of metazoa, such as cell adhesion and cell-signalling, were fulfilled by the formation of repeated domains of metazoa-specific families that became important late in evolution.

ACKNOWLEDGEMENTS

We thank Cyrus Chothia for helpful discussions and encouragement. GA has an LMB Cambridge Overseas Fellowship and SAT has a Beit Memorial Fellowship for Medical Research.

REFERENCES

Albert,R., Jeong,H. and Barabasi,A.L.(2000) Error and attack tolerance of complex networks. *Nature,* **406,** 378-382.

Brenner,S.A.(2000) Target selection for structural genomics. *Nature Struct. Biol.,* **7 Suppl.,** 967-969.

Burley,K.S.(2000) An overview of structural genomics. *Nature Struct. Biol.,* **7 Suppl.,** 932-934.

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

Eddy,S.R.(1996) Hidden Markov Models. *Curr. Opin.Struct.Biol.,* **3,** 361-365.

Gerstein,M. (1998) How representative are the known structures of the proteins in complete genome? A comprehensive structural census. *Folding and Design,* **3,** 497-512.

Gough,J., Chothia,C., Karplus,K., Barrett,C. and Hughey,R.(2000) Optimal Hidden Markov Models for all sequences of known structure. In Miyano,S., Shamir,R., Takagi,T.(eds) *Currents in Computational Molecular Biology,* Universal Academic Press, Tokyo, Japan, p.124-125.

Hill,E., Broadbent,I., Chothia,C. and Pettitt,J. (2000) The cadherin superfamily of proteins in Caenorhabditis elegans and Drosophila melanogaster. *J. Mol. Biol.,* **in press.**

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabasi,A.L.(2000) The large-scale organisation of metabolic networks. *Nature,* **407,** 651-654.

Karplus,K., Barrett,C. and Hughey,R.(1998) Hidden Markov Models for detecting remote protein homologies. *Bioinformatics,* **14,** 846-56.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications in protein modelling. *J. Mol. Biol.,* **235,** 1501-1531.

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

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH a hierarchic classification of protein structure. *Structure,* **5,** 1093-1098.

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

Pearson,W.R. and Lipman,D.J.(1998) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* **85,** 2444-2448.

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

Teichmann,S.A., Park,J. and Chothia,C. (1998) Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc.Natl.Acad.Sci. USA,* **95,** 14658-14663.

Teichmann,S.A. and Chothia,C.(2000) Immunoglobulin superfamily proteins in Caenorhabditis elegans. *J. Mol. Biol.,* **296,** 1367-1383.

Wolf,Y.I., Grishin,N.V. and Koonin,E.V.(2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.,* **299,** 897-905.