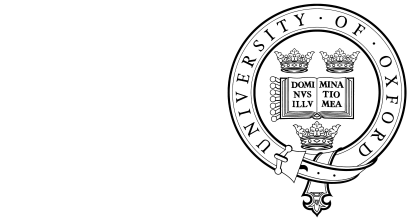
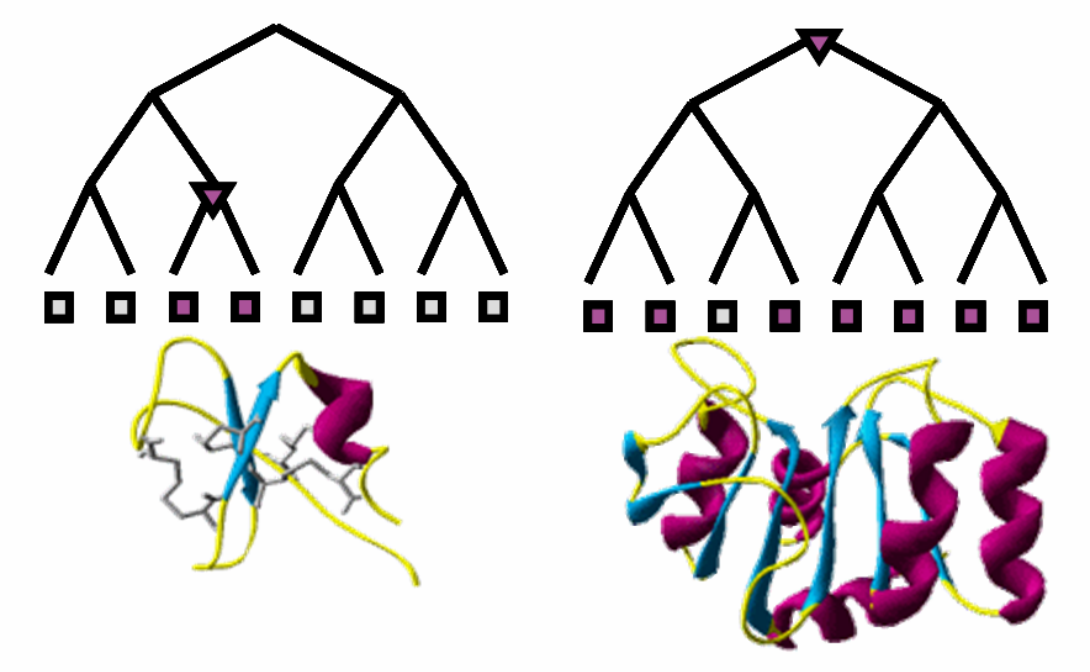


Protein fold evolution on completed genomes: distinguishing between young and old folds



Sanne Abeln

Jesus College
University of Oxford

Submitted for the degree of D.Phil.
Hilary Term 2007

Department of Statistics, 1 South Parks Road,
Oxford OX1 3TG

Abstract

We review fold usage on completed genomes in order to explore protein structure evolution and assess the evolutionary relevance of current structural classification systems (SCOP and CATH). We assign folds on a set of 150 completed genomes using fold recognition methods (PSI-BLAST, SUPERFAMILY and Gene3D). The patterns of presence or absence of folds on genomes gives us insights into the relationships between folds and how we have arrived at the set of folds we see today. In particular, we develop a technique to estimate the relative ages of a protein fold based on genomic occurrence patterns in a phylogeny. We find that SCOP's 'alpha/beta' class has relatively fewer distinct folds on large genomes, and that folds of this class tend to be older; folds of SCOP's 'small protein' class follow opposite trends. Usage patterns show that folds with many copies on a genome are generally old, but that old folds do not necessarily have many copies. In addition, longer domains tend to be older and hydrophobic amino acids have high propensities for older folds whereas, polar - but non-charged - amino acids are associated with younger folds. Generally domains with stabilising features tend to be older. We also show that the reliability of fold recognition methods may be assessed using occurrence patterns. We develop a method, that detects false positives by identifying isolated occurrences in a phylogeny of species, and is able to improve genome wide fold recognition assignment sets. We use a structural fragment library to investigate evolutionary links between protein folds. We show that 'older' folds have relatively more such links than 'younger' folds. This correlation becomes stronger for longer fragment lengths suggesting that such links may reflect evolutionary relatedness.

Acknowledgements

I would like to thank everyone who helped and supported me during the preparation of this thesis.

In particular, I would like to thank my supervisor, Charlotte Deane, for guiding me during the last three and half years with my research, and for improving the value of my work significantly.

Also thanks to Henry Winstanley and Carlo Teubner for very pleasant collaborations on research projects: estimating the relative ages of folds and improving fold recognition methods respectively.

I am grateful to Pao-Yang Chen and Ramazan Saeed for providing me with precalulated protein-protein interaction datasets and helpful comments.

Peter Crowe deserves a special thanks for proof-reading my thesis, including this section, and helpful advice.

I am also grateful to Taane Clark, Frank von Delft, Gesine Reinert, Peter Clifford, Jotun Hein, Gil McVean, Andrew Dalby and Rhodri Saunders for helpful comments and suggestions.

Finally, I would like to acknowledge the organisations who have supported my work on this thesis: the Department of Statistics - University of Oxford, the Engineering and Physical Sciences Research Council, the Jesus Old Members' Group - Years to 1969, Jesus College and the Oxford Supercomputer Centre.

Contents

List of Figures

List of Tables

List of Abbreviations

SCOP - Structural Classification of Proteins (database)
PDB - Protein Data Bank (database and structure format)
CATH - Class - Architecture - Topology - Homology (database)
PB - set of genomic assignments generated by PSI-BLAST, using SCOP domains
SF - set of genomic assignments taken from the SUPERFAMILY database, using SCOP domains
G3 - set of genomic assignments taken from the Gene3D database, using CATH domains
HGT - Horizontal Gene Transfer
LUCA - Last Universal Common Ancestor
RMSD - Root Mean Square Distance
URMSD- Unit vector Root Mean Square Distance
NMR - Nucleic Magnetic Resonance
HMM - Hidden Markov Model
PSSM - Position Specific Scoring Matrix (used in PSI-BLAST)
COG - Clusters of Orthologous Groups of proteins
NR - the non-redundant sequence database, containing all available non-identical protein sequences
CD - Contact Density

In Figures and Tables

fam - family
sfam- superfamily
A - Archaea
B - Bacteria
E - Eukaryotes
sse - secondary structure element

List of Publications

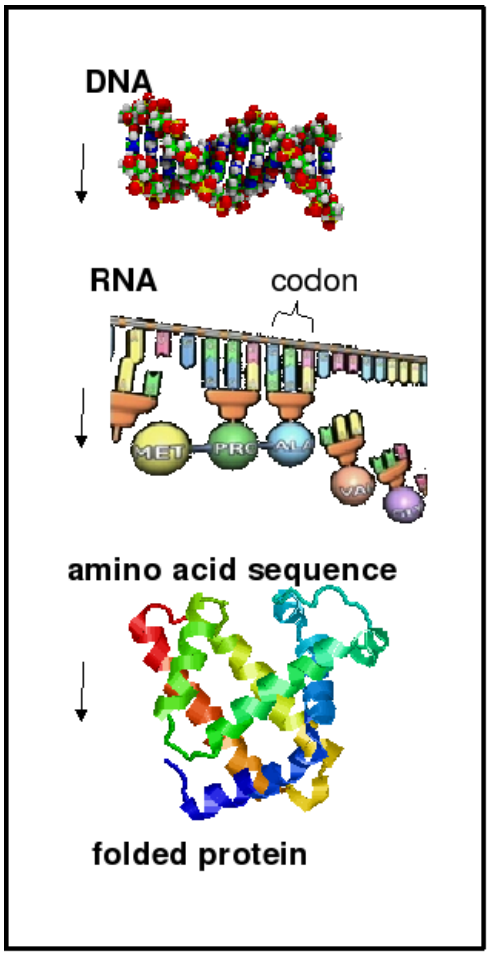
Abeln S. and Deane C.M., July 2005
Fold usage on genomes and protein fold evolution.
Proteins 60(4), 690-700

Winstanley H.F., Abeln S. and Deane C.M., June 2005
How old is your fold?
Bioinformatics Suppl 1, 449-458

Abeln S., Teubner C. and Deane C.M., January 2007
Using Phylogeny to Improve Genome-Wide Distant Homology Recognition
PLoS Comput Biol 3(1), e3

Chapter 1

Background



1.1 Proteins and protein structure

1.1.1 Proteins

Proteins are the most important actively functioning molecules within a cell. In each species a set of specific proteins are encoded by genes on its genome. The DNA code of a gene is transcribed into RNA, which is subsequently translated into a sequence of amino acids (a protein). Then the polypeptide chain folds into a three dimensional structure, determined by its amino acid sequence, to start performing its function within the cell. The structure of a protein is often vital for its function: for example the surface of a protein, can be recognised by a binding partner; other proteins might have a cleft in their surface to bind a particular substrate or ligand, and catalyse a chemical reaction.

The amino acid sequence of a protein is determined by the DNA sequence of the gene encoding it. A triple of nucleic acids (or codon) in the DNA sequence encodes a single amino acid. The translation table for this process has been determined (Khorana, 1966) and hence it is possible to work out the amino acid sequence from the gene sequence. Peptide bonds connect the amino acids together and form the 'backbone' of the protein; each of the 20 different amino acids has a unique side chain (see Table 1.1.1). The properties of the side chains determine its functional three-dimensional structure.

There is, however, no universal technique that can predict the three dimensional structure of a protein given its amino acid sequence. A better understanding of protein structure and the process of folding, is essential to the understanding of biochemical processes and, in practical terms, could aid the discovery of new drugs (Congreve et al., 2005).

1.1.2 Protein structure

Protein structures are available through experimental techniques such as X-ray diffraction and nuclear magnetic resonance (NMR). When analysing the structure of a protein one can consider four different levels (Branden and Tooze, 1999):

Please see refrence

Table 1.1: Amino acids and their properties
Source `http://folding.stanford.edu/education/AminAcid.html`

The primary structure: this is the amino acid sequence of a protein. Table 1.1.1 shows all 20 amino acids, their chemical structures and chemical properties.

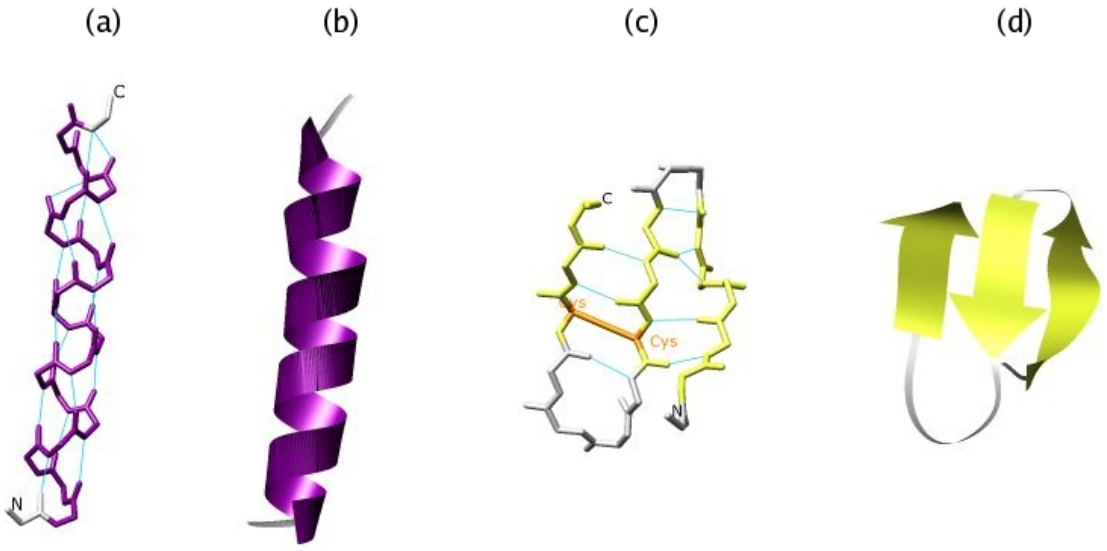


Figure 1.1: Secondary structure
(a) and (b) show an alpha helix from 1hus. (c) and (d) show an anti parallel beta-sheet from 1jk4. Alpha helix is shown in violet, beta strand in yellow and loop region in grey. H-bridges are shown in cyan in backbone models (a) and(c); (b) and (d) show ribbon representations of the same structures. A disulphide bridge is shown in orange in (c).

Secondary structure: these are local regular structures, which are energetically favourable confirmations of the backbone of the protein. H-bonds between the hydrogen of the (NH) group and oxygens of the (CO) group of residues in the backbone stabilise these structures. There are two different types of secondary structure elements: alpha helix and beta strand, with loop regions connecting these elements.

In alpha helical structure the backbone forms a right handed helix, with each residue (i) connecting with another residue (i+4) through a h-bond. Figure 1.1.2 (a) and (b) show an example of an alpha-helix.

Beta sheets are formed through strands of the backbone lying in parallel to each other. These sheets are again stabilised by h-bonds between the backbone residues. The direction of the chain can differ: two strands running in the same direction are called parallel strands, whereas those in opposite direction are referred to as anti-parallel strands. Sheets that combine both conformations (or topologies) also exist and are referred to as

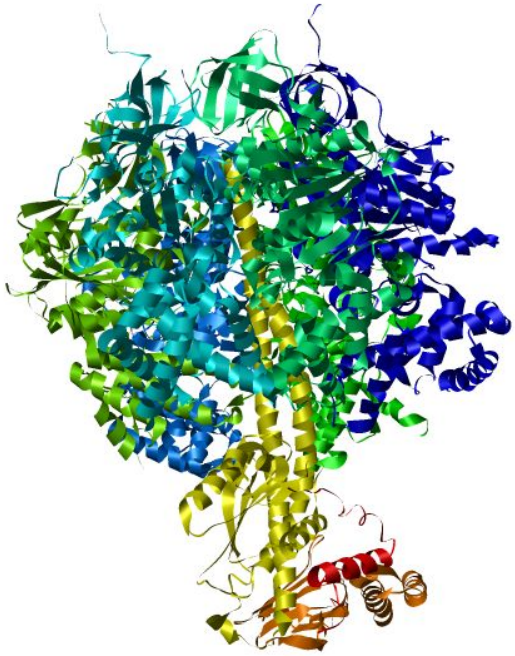


Figure 1.2: Molecular motor
Bovine f1-ATPase (1E79). This ATP synthase unit is responsible for the production of ATP within the cell. Each chain is displayed in a different colour. Together these chains make up the quaternary structure.

mixed beta sheets. Figure 1.1.2 (c) and (d) show different depictions of an anti-parallel beta sheet.

A loop can be described as the region of a protein connecting two other secondary structure elements. Loops are often more variable and less regular than alpha helices or beta-strands. However, loop regions are in general stable within the structure and should not be confused with 'disordered' regions of proteins, which do not form a stable three-dimensional structure in their natural environment.

Tertiary structure: this is the final three dimensional structure of the protein. It is determined by the amino acid sequence. It is thought that the major force behind protein folding is the assembling of hydrophobic residues in the core of the protein. Proteins in their natural environment are surrounded by water molecules, and hydrophobic residues favour positions within the protein where there is no contact with water. This self assembling process of hydrophobic parts may be compared to oils that spontaneously form

drops after being mixed with water (Table 1.1.1 indicates which of the amino acids are hydrophobic).

Quaternary structure: this describes the structure and interaction of multiple polypeptide chains. Many proteins can work together as a functional unit; see for example Figure 1.1.2, which displays the ATP synthase unit of a cow.

1.2 Introduction to structure evolution

After the first few protein structures were experimentally determined around 1960 (Kendrew et al., 1958; Perutz et al., 1960), it became clear that proteins with a similar sequence often have a very similar structure. It was already known by this time that different proteins have different amino acid compositions, different biological activities and different physical properties. The similarity in structure and amino acid content of these first few proteins was quickly linked to evolutionary conservation (Perutz et al., 1960, 1965), even though only low-resolution structures and amino acid composition, rather than sequences, had been determined.

When a sufficiently large set of proteins were sequenced within a family, i.e. with a similar sequence and function, it became possible to work out evolutionary relations between these proteins. It was shown that a 'plausible' evolutionary tree (or phylogeny) of the protein family could be achieved by modelling changes in amino acid sequence (Fitch and Margoliash, 1967).

During the next decade, more protein structures became available and as with sequence, an approach was made to compare structures systematically (Rossmann and Argos, 1976). It was noted that in fact structure seemed to be more conserved than sequence (Rossmann and Argos, 1976), particularly in the area around binding (active) sites of proteins; this is still the most important guideline to establish evolutionary relationship between proteins. The same research also suggested that the amount of variation in sequence correlates with the variation in structure for evolutionary related (or homologous)

proteins.

In 1971 the Protein Data-Bank (PDB) was established to collect the (slowly) growing set of protein structures (Bernstein et al., 1977). To organise this set of structures several classification schemes were proposed, initially for specific groups of folds (Richardson, 1977; Rossmann and Argos, 1976). Later classifiers applicable to all proteins were used, such as secondary structure elements and domains (Levitt and Chothia, 1976; Richardson, 1981). Many current classification schemes are still based on these rules.

The first comprehensive and maintained databases to classify protein structures were introduced more than a decade later (Holm et al., 1992; Murzin et al., 1995; Orengo et al., 1997; Holm and Sander, 1998). These schemes try to classify proteins to reflect evolutionary as well as structural relationships. Many of the classification rules are based on a mixture of concepts. Ideally the classifications would resemble evolutionary history up to the highest level. However, due to a lack of evolutionary-signal in protein sequences it is often very difficult to establish if proteins are evolutionarily related, and therefore difficult to study evolutionary changes in protein structure.

Understanding the evolution of protein structures is not only interesting from a theoretical point of view, but also has important implications for protein structure prediction. A simple example is provided by homology modelling, which is based on the assumption that structure is more conserved than sequence and the corollary that similar sequences often have a similar structure. Homology modelling uses homologs, identified by fold recognition techniques (see Section 1.4), to build a model for a given sequence based on a known structure of an evolutionary-related protein. It is therefore important to be able to recognise distantly related proteins and to get a better understanding of structural divergence during evolution.

There are still many unresolved issues regarding protein structure evolution (see Section 1.6.4). Here we investigate what kind of information can be obtained from completed genomes with respect to protein *structure* evolution, and in particular with respect to changes in topology. A description of resources and techniques used in this research follow.

1.3 Structural databases

1.3.1 PDB

The Protein Data Bank (Bernstein et al., 1977; Berman et al., 2000) facilitates electronic access to, and a standardised format for, biological macromolecules. Moreover it presents a near-complete resource for all experimentally determined protein structures. Each structure is assigned a four character code containing both letters and numbers (e.g. 1E79 in Figure 1.1.2).

The database predominantly contains protein structures accompanied by water, ligands, DNA or RNA fragments. It contains biological structures, ranging from 3 (1TN1) to 150000 (1VRI) atoms (source `http://www.ebi.ac.uk/thornton-srv/databases/pdbsum`). The PDB currently holds $> 30,000$ structures, but is highly redundant: when all PDB entries are filtered at 70% sequence identity, only around one third of the structures remain.

The set of available structures in the PDB is thought to be biased due to experimental methods and biological interest. For example, trans-membrane proteins are under-represented since low protein expression and bad solubility makes crystallisation of membrane proteins considerably more difficult than crystallisation of globular proteins (McCloskey and Poo, 1984; Schein, 1990). Over-representation of certain proteins due to biological interest is a clear problem; for example mutation studies (e.g. Herning et al. (1992)) and multiple PDB entries for bound and unbound structures of the same protein (e.g. 1HHO and 2HHB) could create such a bias .

Studies comparing the set of sequences of the PDB with sequences from more general protein sequence databases highlight such biases (Gerstein, 1998; Boeckmann et al., 2003; Peng et al., 2004). Gerstein (1998) shows that the length of protein sequences in the PDB are significantly shorter than in a set of genes from completed genomes, and that some amino acids are significantly over- (Lys, Ile, Asn, Gln) and under- (Cys, Trp) represented. Wallin and von Heijne (1998) shows that around 25% of open reading frames are predicted to be membrane proteins. On the other hand less than 1% of the structures within the

PDB are currently classified as membrane proteins (Nilsson et al., 2005). Peng et al. (2004) showed through predictive methods that covalent bonds between two cysteines (disulphide bridges) occur significantly more often, and signal and trans-membrane regions less often in the PDB than in SWISS-PROT.

In the last two decades many structures from difficult to crystallise proteins have become available through NMR techniques and recently more membrane structures have been determined at high resolution. Moreover, the appearance of high throughput methods in combination with an effort to specify more varied targets has already led to an increase in the diversity of newly determined structures (Todd et al., 2005).

Although PDB entries often contain a description of the function(s) of protein(s) and often gives links to related entries, the PDB does not provide a hierarchical classification for the structures. If one wishes to understand evolutionary or structural relations between the PDB entries, a structural classification database such as SCOP, CATH or the DALI Domain Database should be used.

1.3.2 SCOP

SCOP (a Structural Classification of Proteins) was originally built to facilitate research about the relationship between protein structure and sequence. It attempts to reflect evolutionary relationships as accurately as possible, based on current human knowledge. The database is therefore 'essentially' built through visual inspection and comparison of structures (Murzin et al., 1995). Some computational aid for structure and sequence comparison is used to find possible matches.

SCOP is based on a four level hierarchy (Figure 1.3). The top level in the hierarchy groups structure on the basis of their secondary structure elements. The second level is the fold level which describes the topology of a protein structure. The next level divides folds into superfamilies; proteins within a superfamily are thought to be evolutionarily related by experts in the field. The final level is the sequence family, in which the proteins have a high level of sequence identity, indicating clear evolutionary relations. All these

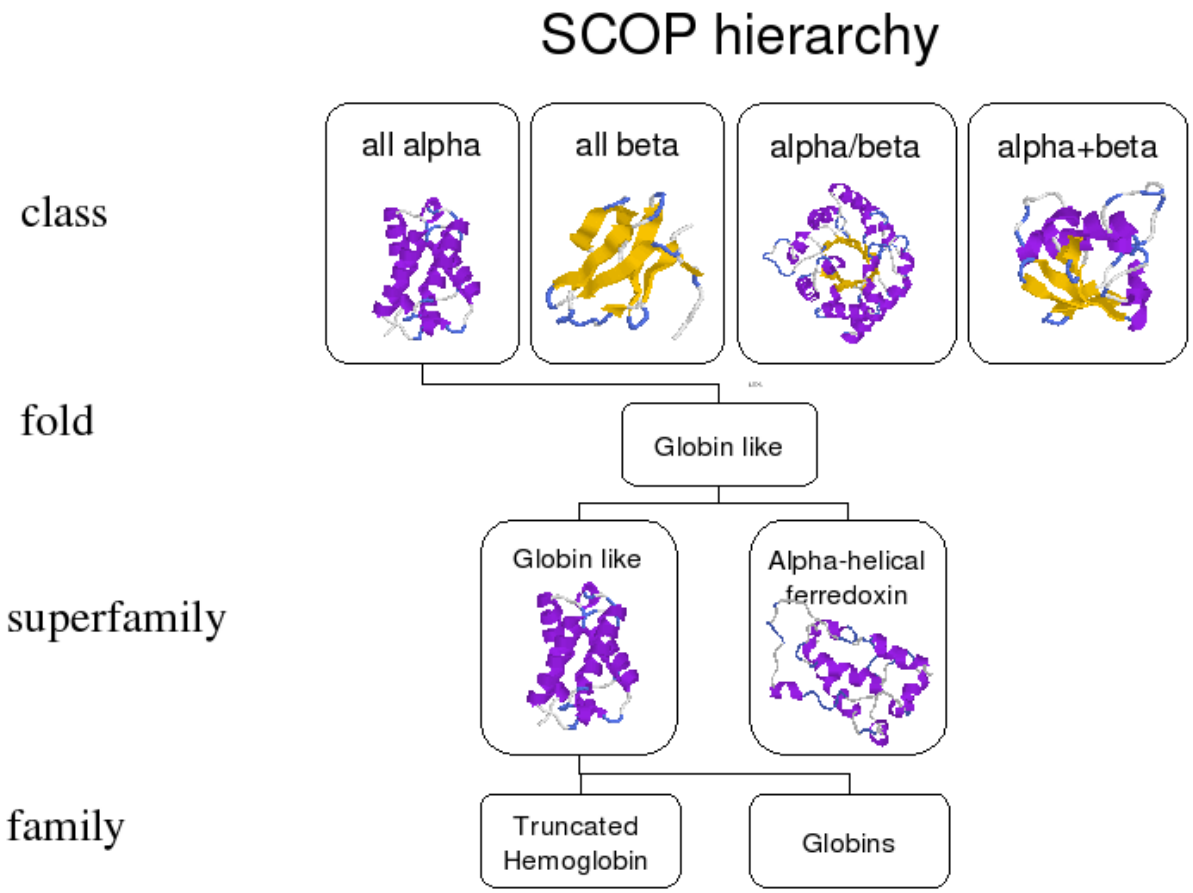


Figure 1.3: Structural Classification of Protein Structures
The hierarchical structure of the SCOP classification.

concepts are described in more detail below:

Domains

SCOP uses structural domains as the entities to be classified, rather than the structures of entire proteins. Many structural classification databases use the concept of a structural domain, but the defining rules are often different. This can make it difficult to compare the systems.

In SCOP a domain is defined as an independent evolutionary unit of protein structure which occurs at least once in isolation. This rule implies that a domain should be able to fold in isolation. SCOP domains are defined by its authors, and a protein can only be split up into several domains if its parts are homologous to known independent domains.

Class

Class provides the top level of SCOP's hierarchy. The first four classes (see Table 1.2) in SCOP are mainly based on secondary structure content, using a classification that was proposed nearly two decades before the first release of SCOP (Levitt and Chothia, 1976; Richardson, 1981).

Perhaps the most remarkable decisions at this level of the classification is the separation of secondary structure classes with alternating 'alpha/beta' elements and segregated 'alpha+beta' elements. The main difference between these two classes is the connectivity of the beta-strands. However, when the scheme was proposed it was also noted that alpha/beta folds consist mainly of parallel beta sheet (usually a single sheet), whereas alpha+beta folds mainly consist of multiple anti-parallel sheets. This rule still holds for the majority of folds within these classes, see Section 5.3.3.

SCOP defines other classes to accommodate all structures from the PDB. Some are artificial, but necessary to keep the classification biologically relevant (e.g. low-resolution and designed protein classes); other classifications do not attempt this, and therefore classify a smaller fraction of available PDB entries (see Figure 1.4). The multi-domain class consists of those folds which cannot be segregated by SCOP domain rules and appear to have sub-units belonging to different structural classes. It is possible that folds within this class will be split up into separate domains when evidence, in the form of homologs, becomes available.

The trans-membrane class contains mainly folds with trans-membrane helices and/or strands. It is likely that the number of folds, superfamilies and families within this class is highly under-represented due to the lack of experimental data (see Section 1.3.1).

Small proteins are grouped together due to their lack of a hydrophobic core. In larger globular proteins this core provides stability and is thought to be the driving force behind the folding process. Hence proteins without a hydrophobic core need to be stabilised by another physical force, such as disulphide bridges, heme or metal ligands. These features are typical for the small protein class.

class	code	description	true class
all alpha	a	structures essentially formed by alpha-helices	Yes
all beta	b	structures essentially formed by beta-sheets	Yes
alpha/beta	c	alpha helices and beta-strands largely intersparsed	Yes
alpha+beta	d	alpha helices and beta-strands largely segregated	Yes
multi-domain	e	containing domains of different classes	Yes
membrane	f	membrane and cell surface proteins and peptides	Yes
small proteins	g	usually dominated by metal ligand, heme, and/or disulfides	Yes
coiled-coil	h	coiled-coil regions	No
low resolution	i	low resolution protein structures	No
peptides	j	peptides and fragments	No
designed	k	structures with essentially non-natural sequences	No

Table 1.2: Structural classes of SCOP

Fold

A fold in SCOP is defined as a group of domains with the same major secondary structure elements in the same arrangement with the same topological connections. The authors of SCOP stress that two proteins with the same fold are not necessarily evolutionary related; two proteins could have evolved independently into a similar topology if it is favourable due to physical forces (convergent evolution).

Superfamily

Domains within the same superfamily are thought to be evolutionarily related by the authors of SCOP, even though there might be a very low sequence similarity. There are no clear cut rules to define a superfamily. Grouping of families into a superfamily is usually established through a combination of any of the following: structural similarity, functional similarity, conserved residues and/or similarity of the active site.

Family

Families are defined as a group of domains with a sequence similarity higher than 30%, based on single linkage. However, some domains, which have lower sequence identities but a highly conserved function and structure, can also be grouped together under a single family (e.g. globins).

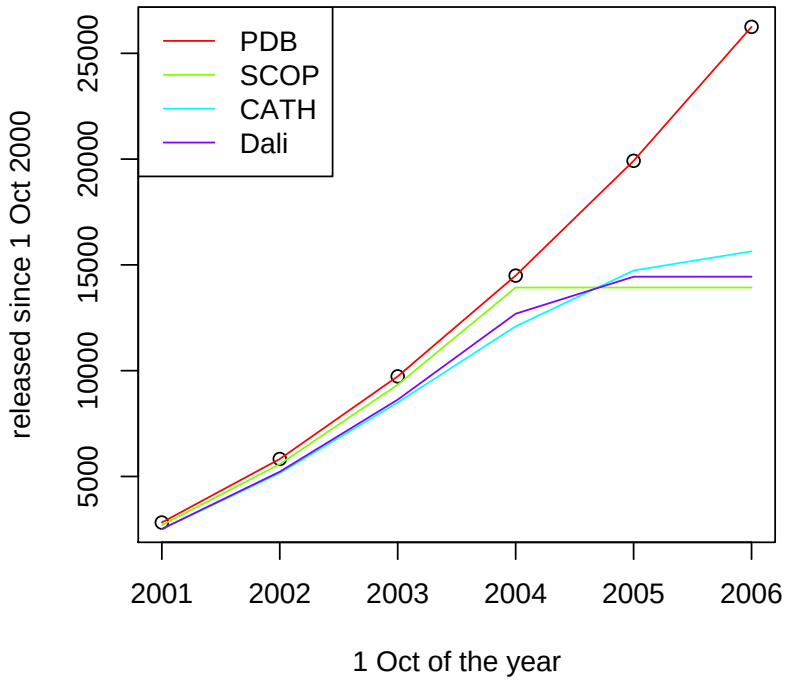


Figure 1.4: Delay in classification
Year versus number of released PDB entries since 1 October 2000. The figure shows the delay in classification for SCOP, CATH and DALI. Sources SCOP version 1.69, CATH version 3.0.0 and DALI last update in March 2005. Note that Dali does have an automatic service available for PDB entries released after this date.

SCOP classifications are given in a four level code, starting with the class as given in the second column of Table 1.2, followed by a numerical code for the fold, superfamily and family. For example the PDB entry 1AA7 (see Figure 1.5) is classified as a single domain with SCOP code: a.95.1.1.

One of the major difficulties for SCOP, and other structural classification systems, is

the rapidly growing number of solved structures, since human intervention is necessary for all new assignments. This can result in huge delays for classifications, with many proteins structure waiting to be classified for 2-3 years (see Figure 1.4).

1.3.3 CATH

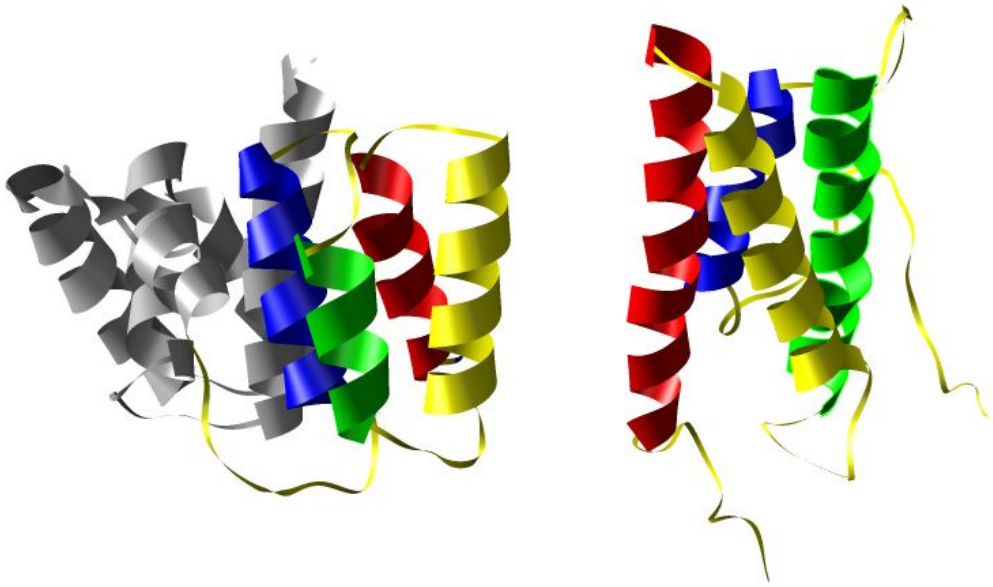


Figure 1.5: SCOP and CATH domain examples
Left: 'Influenza virus matrix protein', 1AA7. Right: 'Solution structure of four helical up-and-down bundle domain of the hypothetical protein 2610208M17Rik similar to the protein FLJ12806', 1UG7. CATH classifies the N-terminal domain of 1AA7 (left, coloured region) and 1UG7 (right) into the same architecture: 'up and down bundle'. Following the path of the secondary structure elements (coloured sequentially: red, yellow, green and blue) it is clear that the 4 helices are differently connected and have thus another topology. SCOP classifies both proteins under the same class: 'all alpha'. CATH defines two separate domains for 1AA7 (grey, coloured), whereas SCOP defines the entire protein as a single domain.

CATH (Orengo et al., 1997) is, like SCOP, a hierarchical structural classification database, based on similarity of structure, function and sequence. There are, however, some significant differences.

The authors of CATH attempt to automate their classification process as much as possible without losing biological relevance. In addition, CATH puts more weight on the *structural* aspects of the classification. To reflect this, CATH uses an extra level

in its hierarchy when compared to SCOP. CATH consists of four main levels: Class, Architecture, Topology and Homology (CATH). Here topology is similar to SCOP's fold level, and homology is similar to SCOP's superfamily level. The 'architecture' level is specific to CATH and represents the shape defined by the assembly of secondary structures *without* considering the connectivity (see Figure 1.5). Below these main levels, there are a number of family levels with boundaries of 35%, 60%, 95% and 100% sequence identity, all with a minimal alignment-overlap of 80%.

CATH classifications are given in a four level code, starting with the class as given as a numeric values (1 = alpha, 2 = beta, 3 = alpha-beta, 4 = few secondary structure elements), followed by a numerical code for the architecture, topology and homology. For example the PDB entry 1AA7 (see Figure 1.5) is classified as two domains, with CATH codes 1.20.91.10 and 1.10.10.180.

The top level of the CATH-hierarchy, class, uses different definitions than SCOP. Most importantly, a single alpha-beta class is designated for SCOP's alpha/beta and alpha+beta classes. The argument for this treatment, is that alpha/beta and alpha+beta structures can not be segregated with automatic procedures without considering secondary structure connections (Michie et al., 1996); such differences in topology are reflected at a lower level of CATH's hierarchy.

CATH uses several computational tools to facilitate classification of new PDB entries. The procedure is split into four major steps (Pearl et al., 2003). Firstly, low resolution structures are filtered out of the set of new PDB entries to be classified. Secondly, sequences are matched against domains which have already been classified; here an intermediate sequence search is used, whereby intermediate homologous sequences may be used to find matches. Thirdly, a structural comparison is made. If no sequence intermediate was found, this process is used to identify potential structures, otherwise to verify those found previously. For structure comparison a fast prefilter, GRATH (Harrison et al., 2002), is first used, and then a more rigorous but slower method SSAP (Orengo and Taylor, 1996) is applied to identify structural similarity. Finally, new PDB entries are split into sequence families, underneath the homology level.

Despite CATH using significantly more automated processes to produce its classification than SCOP, a significant amount of the assignments need human intervention: 26% of single domain proteins and 79% of multi-domain proteins (Orengo et al., 1997). It becomes clear from these statistics that automatic *domain* assignments are especially problematic. Although improvements have been made since the start of the database (Pearl et al., 2005), this problem remains one of the most difficult to overcome in automatic structure classification; domain boundaries can only be assigned without human intervention when a new PDB entry has an 80% sequence identity and a SSAP score > 80 (out of 100) with a previously classified chain. Due to the high amount of visual inspection, a large number the new PDB structures have not been included into the latest version of the classification (see Figure 1.4).

1.3.4 FSSP and Dali-domain server

FSSP (families of structurally similar proteins) is a fully automated structural classification database based on structure alone (Holm et al., 1992; Holm and Sander, 1998); protein sequence and function are not taken into account. The classification procedure is essentially an all-against-all comparison of structures with the structural alignment program DALI. This comparison is performed on a subset of the PDB, filtered for sequence identity. By average linkage clustering, fold groups can be formed (FSSP: Holm et al., 1992) and structural domains can be defined for the Dali Domain Database (Dali Domain Database: Holm and Sander, 1998). The classification is hierarchical: levels are based on structural similarity scores. The hierarchy is cut at Dali Z-score levels 2, 4, 8, 16, 32 and 64. The first level ($Z > 2$) can be used as an operational definition of folds.

1.3.5 Comparing structural classification databases

Studies comparing the three databases, SCOP, CATH and DALI, show that there is generally a reasonable agreement between the classifications (on most levels of comparison $> 50\%$). The largest discrepancy is between domain assignments, with DALI splitting

proteins into more domains than either SCOP or CATH (Day et al., 2003), and CATH splitting up proteins into slightly more domains than SCOP (Hadley and Jones, 1999); Figure 1.5 shows an example of different domain splitting in SCOP and CATH.

At fold level, DALI and SCOP have a higher level of agreement than either has with CATH. This is mainly due to CATH having a few very large 'super-topologies', which are split into multiple folds in both DALI and SCOP. However, CATH defines a larger number of folds than SCOP or DALI (Hadley and Jones, 1999; Day et al., 2003). At homology level SCOP and CATH have a better agreement than at fold level (Hadley and Jones, 1999).

1.3.6 Difficulties in structural classification

It is clear that classification of protein structures remains far from trivial. Moreover, different classifications use very different rules to separate sets of structures based on a combination of multiple concepts such as secondary structure content, topology, sequence and function.

Defining the entities to be classified also remains a difficult problem, with each scheme using different domain splitting rules. A common explanation for this problem is the so-called 'Russian doll' effect (Richardson, 1981; Orengo et al., 1997). This states that protein structures form compact subunits at multiple levels, allowing perhaps a continuous view of the space describing differences between protein structures. Whether or not such an effect actually exists is still under debate. Evidence has been found of structural similarities between different folds below the domain level (sub-domains) (Shindyalov and Bourne, 2000; Harrison et al., 2002). Most of these sub-structures occur in alpha/beta folds. In addition, this class contains more folds that are self-similar (Taylor et al., 2002). However, these studies also show that most folds have very little similarity with other known folds. Consequently the question remains open, whether hierarchical classifications are the most suitable form to subdivide protein structures, or if perhaps networks or continuous representations may be more useful.

Automatic classification remains another difficult problem: Table 1.3 shows a clear anti-correlation between the amount of automation and the number of references to the database (indicating usefulness for (biological) research). Figure 1.4 shows the delay in classification of newly released PDB entries.

database	automation	citations
SCOP	low	2188
CATH	medium	841
FSSP	high	305

Table 1.3: Citations of structural classification databases
Level of automation versus number of citations. Sources: scholar.google.com, 16 Sept. 2006, Murzin et al. (1995) , Orengo et al. (1997) and Holm and Sander (1998)

1.4 Fold recognition methods

1.4.1 Structure prediction

As described above, prediction of protein structure from sequence is an unsolved problem and is becoming increasingly important as the gap between the number of experimentally determined sequences ($> 6,000,000$) and structures ($< 35,000$) widens.

When predicting the structure for a target sequence, a major step is achieved when an evolutionarily related protein with a known structure is identified. Since structure is more conserved than sequence, it is presumed that the target sequence has a similar fold to the related protein. This process is called fold recognition and has been a major force behind improvement in structure prediction over recent years (Moult, 2005).

Currently there are several ways to recognise a fold for a given sequence. If there is close homology between the target sequence and a known structure in the PDB, a simple sequence search, such as BLAST (Altschul et al., 1990), will be sufficient to identify the fold. In order to detect more distant homologies we can use position specific scoring methods such as PSI-BLAST (Altschul et al., 1997) and Hidden Markov Model (HMM) based methods such as SAM-T98 (Karplus et al., 1998). These sequence based methods are the least expensive forms of fold recognition. There are more computationally expensive

methods that take structural information into account; an example of such a technique is THREADER (Jones et al., 1992).

A major challenge for all fold recognition techniques is to discriminate a true homologue from a false positive (specificity) using confidence scores such as e-values. E-values (expectation-values) indicate how likely it is that an alignment with the search sequence would occur by chance in a given database, i.e. they should reflect the chance of a false positive assignment.

Some structure prediction techniques do not depend directly on identifying an evolutionary related protein. These techniques are classified as 'ab initio' methods, even though most are still knowledge-based, i.e. dependent on libraries of known protein structures. An example of such a method is ROSETTA (Simons et al., 1999), which assembles models from fragments of existing protein structures.

Although structure prediction has made significant progress, there are still major difficulties to overcome, such as the refinement of models based on homology modelling, distant homology recognition and ab initio modelling (Moult, 2005). Such problem areas alongside progress are highlighted by CASP (Critical Assessment of Techniques for Protein Structure Prediction, (Moult et al., 1995)), a bi-annual assessment exercise, where structure predictions of different scientific groups and automatic servers are compared.

In this work we will use fold recognition methods to predict structural domains on the genes of completed genomes. The methods used are summarised below.

1.4.2 BLAST

BLAST (Basic Local Alignment Search Tool) was developed to find similar sequences to a search sequence in a large database (Altschul et al., 1990). It remains one of the most frequently used bioinformatics tools to perform database searches, and is commonly used as a pre-filter for more computationally expensive methods. BLAST approximates a solution to the exact dynamic programming algorithm for local alignment (Needleman and Wunsch, 1970), with a major decrease in computational time. The main increase

in performance regarding computation-time comes from a preprocessing step, where the database of sequences is encoded in 'words' of a specific length, with direct links to all other occurrences of similar words in the database. The search sequence is also encoded in words. Matches between words from the search sequence and the database sequences form starting points for an alignment. Dynamic programming is then used to extend the alignment of the search sequence and the sequence with the word match in the database .

1.4.3 PSI-BLAST

PSI-BLAST (Altschul et al., 1997) is an iterative version of BLAST, which uses a Position Specific Scoring Matrix (PSSM) to include information about homologous sequences. The PSSM is used to identify the amino acids that are most likely to occur at a given position in the sequence. During each run sequence information, from all hits with an e-value below a threshold, is added to the PSSM. Final e-values between the target and each database sequence are based on sequence similarity to the PSSM, which is subsequently normalised for amino acid composition with respect to the entire database. PSI-BLAST easily outperforms a simple BLAST search (e.g. Madera and Gough (2002)), with relatively little added computational cost.

1.4.4 Hidden Markov Model based methods

The sequences of a (homologous) family of proteins can be described statistically using a Hidden Markov Model (HMM). The model describes the family with a number of states (start, end, match, insert and delete) and a number of transition probabilities between these states. The model is labelled 'hidden' since the match and insert states emit an amino acid character with emission probabilities dependent on its state.

In order for such a model to be useful, a number of problems first need to be solved. To model a family, transition and emission probabilities need to be estimated. Some algorithms use manual alignments from a sequence family to estimate these parameters; others can create such an alignment automatically while building the model. In many

cases the model is iteratively improved by adding newly identified sequences.

To score a sequence with the given the model, a probability needs to be calculated for the likelihood that the sequence is produced by the model. In order to do this all possible paths through the HMM for the sequence should be considered. Simply trying all paths would computationally be very expensive $O(e^N)$. The forward-backward or Viterbi algorithm, which is essentially a dynamic programming approach ($O(N^2)$), estimates the negative logarithm of the probability of the single most likely path for the sequence (see Rabiner (1989) for details).

HMMer and SAM-T99 are currently the most widely used packages to build HMMs for protein families and to search protein sequence databases with the produced HMMs. PSI-BLAST and HMM based methods are all classified as sequence-profile methods; this indicates that a specific search profile is created against which individual search sequences are scored. HMM based methods such as SAM and HMMer generally outperform PSI-BLAST (e.g. Madera and Gough (2002)).

1.5 Databases of fold recognition assignments

1.5.1 Superfamily

The SUPERFAMILY database is a resource that contains predictions of SCOP domains on completed genomes (Gough and Chothia, 2002). These predictions are generated using an automated fold recognition procedure SAM-T99, which was was fine tuned with expert knowledge to recognise *superfamilies* as defined by SCOP.

For each sequence in a superfamily, filtered at 95% sequence identity, HMMs were built using homologues from a non-redundant sequence database. The HMMs for each seed were then scored against the set of genes from the completed genomes. Interestingly, this procedure gave better results than forming a single HMM from a structural alignment of all sequences in a SCOP superfamily (Madera and Gough, 2002). In the SAM-T99 procedure e-values are normalised by the reversed score of the search sequence.

1.5.2 Gene3D

Gene3D is a resource similar to SUPERFAMILY, but assigns CATH domains rather than SCOP domains to completed genomes. To save computational time, the genes on the genomes are first clustered into families based solely on sequence information. HMMs are built for CATH domains using the SAM-T99 technology. The HMMs are then scored against representative sequences from the family clusters, filtered at 35% sequence identity (Lee et al., 2005). The final domains are assigned through a 'Domain Finder' method, which checks for significance (e-values) and overlap.

1.6 Protein evolution

1.6.1 Structure vs. sequence

Family correlations

Several studies have investigated the correlation between sequence and structure. The results of these studies greatly depend on the subset and range of sequence identity between pairs used. Between 30 and 100% sequence identity, the vast majority of protein pairs have close evolutionary relations. If we compare pairs within such families, there is a correlation between sequence identity and root mean square distance (RMSD) of the backbone atoms (Chothia and Lesk, 1986). If the scores are normalised for alignment length or transformed to statistical z-scores these correlations become linear, with strong correlation factors (Wood and Pearson, 1999; Koehl and Levitt, 2002; Panchenko et al., 2005). No specific family property (e.g. family size or secondary structure content) has been found to strongly correlate with plasticity of protein families, where plasticity is defined by the slope of the correlation between structure and sequence deviation. However, families with a high number of conserved disulphide bridges show perturbed correlations, which might be explained by the additional structural restraints imposed by the disulphide bridges (Panchenko et al., 2005).

Structure only

Rost (1999) and Yang and Honig (2000) show that a decrease in sequence identity to a level between 20% - 30% (twilight-zone), gives a dramatic increase in the number of sequence pairs that are not either evolutionary related or structurally similar. In fact there is only a small shift in the distribution between random sequence pairs and pairs with a similar structure (Yang and Honig, 2000). This explains why sequence based fold recognition creates so many false negatives at a reasonable level of accuracy. However, sequence identity is not necessarily the best measure to predict structural relatedness: Rost (1999) shows that if the sequence similarity is higher than the sequence identity, a structural match becomes more likely .

1.6.2 Proteins and completed genomes

In the last decade significant progress has been made in determining sequences of entire genomes. There are now more than 200 completed genome sequences, including examples from all three kingdoms of life. In the field of protein bioinformatics, this newly available data has helped to investigate various existing problems. In particular several studies have been performed which use fold recognition methods, like those described above, to assign structural domains to genes on completed genomes (e.g. Pellegrini et al. (1999); Qian et al. (2001)).

In the area of protein function, it was discovered early on that protein families, with similar occurrence patterns on a set of completed genomes, often function together in a biological complex or pathway (Pellegrini et al., 1999; Wu et al., 2003; Marcotte et al., 1999).

Ranea et al. (2004) showed that structure predictions on genomes could also cluster proteins together in functional groups. To achieve this the correlation between genome size and the number of predicted copies of a superfamily on a genome were examined. For example, many of the superfamilies which are not dependent on genome size with respect to their number of copies are involved in protein translation and biosynthesis,

while superfamilies that have an increasing number of copies with increasing genome size are mostly involved with metabolism and gene regulation.

It has also been shown by several groups that evolutionarily plausible phylogenies of completed genomes can be achieved by clustering occurrence patterns of superfamilies or folds (Qian et al., 2001; Yang et al., 2005). Furthermore, fold recognition on completed genomes allows research into possible domain combinations for proteins (Apic et al., 2001; Vogel et al., 2005). Attempts have also been made to estimate the total number of folds (see Section 1.6.4) and to build a protein-structure phylogeny (Caetano-Anollés and Caetano-Anollés (2003), see Section 2.4).

Structural Genomics has been another important development for the area of protein bioinformatics (Baker and Sali, 2001). Many new protein structures are to be determined experimentally by a number of high-throughput projects. Selection of target proteins for these projects is regulated to prevent multiple projects working on the same proteins. Targets can be identified by clustering genes, from the fast growing set protein sequences, into families and subsequently appointing a single representative for each cluster (Brenner, 2000). Such family clusters have recently been used to determine the number of structures which need to be experimentally determined before an appropriate amount of fold space can be covered (Lee et al., 2003; Yan and Moult, 2005; Sadreyev and Grishin, 2006).

1.6.3 Detailed studies structure/topology evolution

In this study we will focus on global trends of fold and superfamily evolution. It is important to be aware of documented mechanisms through which such evolutionary changes can arise. Evidence for significant changes in secondary structure and small topological changes are described by Murzin (1998), who considers several extreme cases of divergent evolution.

Furthermore, Grishin (2001) found evidence for several topology changes between evolutionary-related proteins. For example insertions, deletions and substitutions of one or more secondary structure elements, for example an alpha helix substituted by an anti-

parallel beta sheet, can lead to small changes in topology (see Figure 1.6 (a),(c)). Circular mutations can lead to more drastic topology changes: a circular mutation could occur during evolution, after a duplication event of an entire domain, followed by partial deletion at the termini. The new termini would then be found at different places within the fold, changing the topology. Rearrangements of beta sheets can create topological changes through strand invasions in, or withdrawals from, a sheet, or through beta hairpin flips or swaps, by which the order of subsequent strands in the sheet is changed (see Figure 1.6 (b)). Later Kinch and Grishin (2002) also report evolution through domain multiplication and through beta strand slips of a few residues. Such slips can occur at very high sequence identities.

Figure 1.6: Fold evolution mechanisms
(a) Substitution of a beta strand for two beta/alpha/beta subunits. (b) Beta hairpin flip. (c) Some examples of secondary structure elements involved in insertions, deletions or substitutions. All figures from Grishin (2001)

1.6.4 Currently unanswered questions

Convergent versus divergent evolution of topologies

It remains unknown if most proteins classified within the same folding topology are evolutionarily related. There are known examples of super-secondary structure motifs that are thought to have developed convergently (Murzin, 1998; Krishna and Grishin, 2004).

Divergent evolution is suggested for most superfamilies under a fold (Koonin et al., 2002; Deeds et al., 2004; Panchenko et al., 2005). These studies suggest that divergent evolution is more likely by considering the observed distribution of protein sequences and structures, and comparing these to evolutionary models.

However, it remains very difficult to find direct evidence to support the scenario of divergent evolution as the major factor in the current universe of protein structures.

The number of folds

A related discussion is the total number of folds in nature. It has been suggested several times that there is a limited set of folds in nature (Koonin et al., 2002; Coulson and Moult, 2002; Liu et al., 2004; Wolf et al., 2000; Zhang and DeLisi, 1998; Govindarajan et al., 1999; Yan and Moult, 2005; Sadreyev and Grishin, 2006). However, more recent studies tend to estimate considerably higher numbers of folds than the earlier studies.

In addition it is not clear whether there are more fold topologies possible than those which currently exist in nature. Evolution might have reached a limit on the total number of possible fold topologies (Crippen and Maiorov, 1995) or there might be more topologies to be explored, e.g. for larger alpha/beta proteins not all possible secondary structure connections have yet been found (Taylor et al., 2002). Convergent evolution would become more likely if nature has already reached or is close to reaching all possible fold topologies.

1.7 Overview

Currently protein structures are classified using hierarchical classification schemes, such as SCOP and CATH, where different classification levels are based on a mixture of structural properties and evolutionary relations. These classification schemes are not only important for our understanding of protein evolution, but also act as benchmarks for structure prediction methods. However, it is at this point unclear how well such classifications reflect evolution and if there are biases in these databases. Furthermore, there is a classification gap above the fold level. It is not clear if and how domains may be linked above this level in an evolutionary relevant way.

Recently a wealth of evolutionary information has become available through sequencing projects on entire genomes. Here we will use this information to investigate protein fold evolution by predicting structural domains on a set of completed genomes. This provides us with information if (occurrence) and how many times (copies) a superfamily or a fold appears on a given genome.

In Chapter 2, we will compare different classification systems (SCOP and CATH), different classification levels (superfamily, fold), different kingdoms (archaea, bacteria, eukaryotes) and different structural classes (alpha, beta, alpha/beta, alpha+beta) with respect to occurrence and copy data, questioning if these follow the same general trends. In addition, we investigate the hierarchical aspect of such classification systems, by measuring the number of superfamilies per fold on a fixed set of genomes.

Power-law-like distributions are found for copies and the number of superfamilies per fold, in all datasets. More variable results are found for distributions of the number of genomes on which a fold occurs.

These occurrence patterns are further investigated in Chapter 3. We develop a parsimony based method to estimate the relative age of a fold, given an occurrence pattern and a species tree. Correlations between the fold age estimates and measures previously thought to be influenced by protein age (copies and protein-protein interactions) are assessed. We found that such measures follow the expected tendencies: folds with

many copies/interactions are usually old, but old folds do not necessarily have many copies/interactions. In addition, functional categories of superfamilies and structural classes of folds show variation in their age distributions.

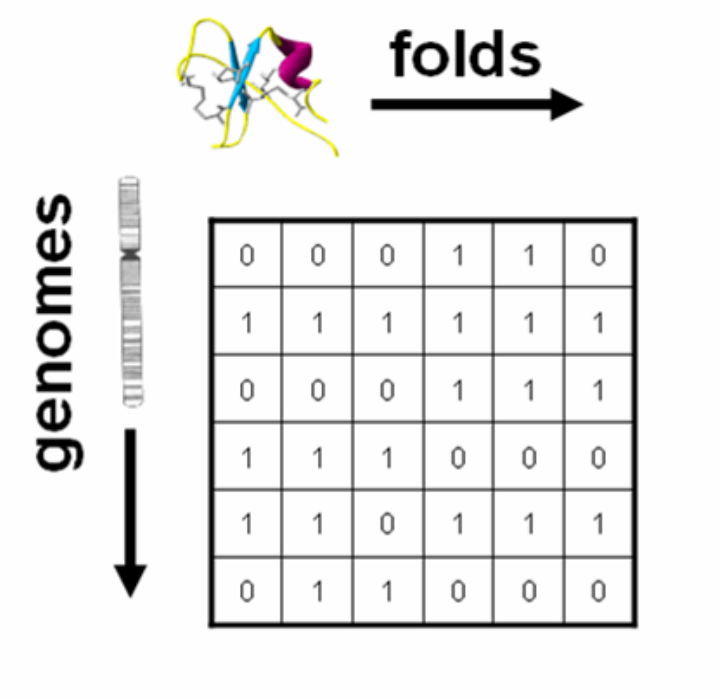
In Chapter 4 we explore occurrence patterns as a way to assess fold recognition techniques. In particular, we examined whether isolated occurrences within a species tree are less reliable than others. This led to a method that can to improve sets of genome wide fold assignments.

To further explore the observed variance in age distribution for structural classes, other structural domain properties are examined for correlations with fold age in Chapter 5, while questioning if such correlations may be due to biases in fold recognition methods. Fold ages are found to be dependent on domain length, the number of residue contacts, amino acids composition and secondary structure contents. Generally, domains with stabilising features tend to be older.

The relation between copies and superfamily age suggests that the size of homologous groups are influenced by age. In Chapter 6 we explore such influences imposed on measures linking domains above the fold level. We find that domains with many links based on shared structural fragments are likely to be old. The correlation between age and the number of links, becomes stronger for longer fragments or when sequence similarity is included, indicating that that such links may be evolutionary relevant.

Chapter 2

Fold usage on genomes



2.1 Introduction

Protein folds are very unevenly distributed over the currently known set of completed genomes; some folds are observed far more often than others (Coulson and Moult, 2002; Qian et al., 2001; Hegyi et al., 2002). Here we exploit this variation to investigate evolutionary relationships between protein structures.

Here we use fold recognition techniques to predict structural domains on a set of completed genomes. This information allows us to investigate the structural content of genomes at different classification levels (e.g. superfamily or fold). We also compare usage patterns for two different structural classification systems, SCOP and CATH.

2.1.1 Measures of fold usage

Fold usage can be explored with several different measures such as occurrence across the genomes, copies per genome and the number of sequence families under the same fold. Comparing these measures for different folds can give us an indication of the age of each fold and can give us insights into the mechanisms behind fold evolution. First we will describe common measures of fold usage and how these have been used previously. In the descriptions below SCOP's fold level is used, but other classification levels (family, superfamily) and the classification levels of CATH (homology, topology, architecture) have also been used in this study.

(1) The number of (super)families per fold

This is estimated as the number of subclassifications (e.g. families) of a classification level (e.g. fold) found on a fixed set of genomes. The distribution for the number of families per fold has been described to follow many different functions including an exponential (Zhang and DeLisi, 1998), a highly-stretched exponential (Govindarajan et al., 1999), a logarithmic function (Wolf et al., 2000) and a power-law (Koonin et al., 2002). Coulson and Moult (2002) modelled this distribution in a different way by separating folds into

three different groups: unifolds with only one sequence family per fold, super folds with very many sequence families per fold and an intermediate group, mesofolds. This measure of families per fold has previously been used to estimate the total number of folds and families we can expect in nature (Zhang and DeLisi, 1998; Wolf et al., 2000; Coulson and Moult, 2002).

(2) The number of copies of a fold per genome

This is the number of genes that occur on a genome classified under the same fold. The distribution for the number of copies of a fold on a single genome is thought to follow a power-law (Qian et al., 2001): most folds have only one copy on each genome, and a very small number of folds have a very large number of copies; the distribution tails off asymptotically to zero. This distribution can be created by self-dependency of a variable, e.g. a fold with many copies is more likely to duplicate again than a fold with only a single appearance on a genome.

Different theoretical models have been developed that can explain the observed power-law distribution. Qian et al. (2001) demonstrated that a power-law distribution can arise when the number of genes on a genome can increase without restriction. Alternatively Karev et al. (2002) provide a birth, death and innovation model (BDIM) for genomes in equilibrium, such that the number of genes on the genome is stable.

In a further study it has been shown that for some superfamilies on bacterial genomes the number of copies is correlated with genome size (Ranea et al., 2004); these superfamilies seem to perform specific cellular functions (Ranea et al., 2004; Konstantinidis and Tiedje, 2004).

(3) The number of genomic occurrences

This is the number of genomes on which a fold occurs. Only a few studies, using a limited number of genomes, have previously investigated this measure (Hegyi et al., 2002; Wolf et al., 1999). We propose that this might be a good measure to estimate the evolutionary age of a fold: e.g. if a fold occurs on all genomes in a set, it is likely that it already

existed on the last common ancestor of this set (not taking into account gene loss and lateral gene transfer). A more thorough discussion of protein folds and age is provided in Chapter 3.

2.1.2 Overview

We investigate how these measures behave on different sets of folds and genomes. Power-law behaviour in all kingdoms and fold classes for both the distribution of fold copies on a genome and the number of superfamilies per fold are found. The distribution of fold occurrences across genomes is more complicated and is dependent on kingdom and fold class. We also investigate whether folds with many occurrences have many copies and/or many sequence families per fold. We find that many folds which occur on a large set of genomes only have a low number of copies and/or superfamilies, whereas folds with many copies usually occur on many genomes. Furthermore we explore whether the chance to create a new superfamily becomes higher when there are many copies of a fold, but find that this is surprisingly not supported by the data. We also show that the distribution of these fold measures is different for the alpha/beta class as compared to all other fold classes. Lastly, we use a simple evolutionary model to explain the observed occurrence distribution. In addition we examine which information can be retrieved by clustering occurrence patterns.

2.2 Methods

2.2.1 Set of completed genomes

We looked for structural domains in the protein sequences of 150 completely sequenced genomes (18 archaeal, 97 bacterial and 35 eukaryotic genomes, see for a full list Appendix A). Some genomes from pathogenic or endosymbiotic species have been removed from this set; particularly those with a largely reduced genomes size. These species are prone

to significant amounts of gene loss and horizontal gene transfer, associated with their living environment in a host species. It has previously been shown through phylogenetic clustering that the gene content of such species has dramatically changed from closely related species in different environments (Winstanley et al., 2005); this makes it hard to compare them in terms of structural domain composition to other complete genomes. In total 1 archaeal, 20 bacterial and 2 eukaryotic genomes were removed.

2.2.2 Assignment sets

Three different assignment methods were used to predict structural domains on the genes of the completed genomes. Note that this work is an update of previously published work (Abeln and Deane, 2005); here we use more recent data.

PSI-BLAST

The first set of assignments was created using PSI-BLAST. Amino acid sequences for domains from the SCOP classification were obtained from a filtered database (Brenner et al., 2000) containing sequences with 95% sequence identity or less for SCOP release 1.69.

The protein sequences for all 150 completed genomes were obtained from the same sources as indicated in the SUPERFAMILY database (Section 1.5.1). Since PSI-BLAST should be used on a non-redundant sequence database, one representative was chosen for genes with identical sequences, and the remaining identical genes were set apart. All unique and representative genes were combined into a non-redundant sequence database.

PSI-BLAST was run with each search sequence from the list of SCOP domains on the non-redundant sequence database. A maximum of five PSI-BLAST iterations per domain were performed and the e-value threshold for hits to be included into the PSSM was set to 10^{-5}. The final e-value threshold for inclusion into the set of assignment was set to 10^{-4} (Abeln et al., 2007).

Different levels of the SCOP classification were are used during the investigation of

fold usage measures. Fold usage measures for SCOP's superfamily and fold level were obtained by clustering the results of all search sequences within this superfamily or fold. To avoid problems with overlapping regions, each gene is assigned at most one copy of a superfamily.

Since PSI-BLAST is able to recognise homologs beyond family boundaries (Lindahl and Elofsson, 2000), stricter rules were used to assign a family to a gene; an additional constraint of a minimum sequence identity of 35% was imposed.

SUPERFAMILY

The second set of assignments were taken from the SUPERFAMILY database. For a detailed description of this database see Section 1.5.1. Assignments with an e-value smaller than 10^{-4} are included. The superfamily and fold level of the SCOP database can be obtained from these assignments, however family assignments cannot be extracted due to the fold recognition method used (`http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/`). Domain assignments in the SUPERFAMILY database have been checked for overlap: no alignment can overlap more than 20% with an alignment to another model. In case of conflict, the assignment with the highest e-value is disregarded. SUPERFAMILY data differs in that respect from our PSI-BLAST assignments. For compatibility, however, the same rules (counting every superfamily once on a gene) as for PSI-BLAST are used to calculated the number of copies for a fold on a genome.

Gene3D

The third set of assignments was obtained from the Gene3D database (`http://cathwww.biochem.ucl.ac.uk:8080/Gene3D/`). The e-value cutoff for assignment inclusion is set by the authors to 10^{-2}. Not all genomes included in the PSI-BLAST and SUPERFAMILY datasets were available (for included genomes see Appendix A); in particular there are fewer eukaryotic genomes in this set, down from 35 to 15. Moreover, some of the genomes, may differ from those used in the other two datasets, as they were taken from different sources. Gene3D assignments are checked for overlapping regions. For compatibility the

same rules as for PSI-BLAST are used to calculated the number of copies for a fold on a genome. Assignments can be used to classify hits to genes at CATH's architecture, topology or homology level; family level assignments are not included.

2.2.3 Power-laws

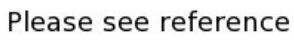


Figure 2.1: Example of a power-law distribution Taken from Cherkasov and Jones (2004).

The distribution of fold copies and families per fold have both been described to follow a power-law on an individual genome. However, demonstrating power-law behaviour is difficult with sparse data. Displaying the distribution in terms of the 'descending rank' provides a neat way to include information of missing observations (Adamic and Huberman, 2002; Troll and beim Graben, 1998).

Using frequency bins or plotting all frequencies to show a power-law distribution can cause problems in regions of very low frequency. In these regions there will be many bins not containing any data (folds). However, this is often not shown when the distribution is plotted, hiding the fact that observations with low frequencies occur less and less. This results in horizontal lines of observations at the lowest frequencies (Figure 2.1). An additional problem occurs when linear regression is used directly on the frequency bins. The majority of data contributes towards bins in the high frequency range (left hand side) whereas most bins, and hence most linear regression points, lie within the low frequency

range (right hand side). This can lead to a strong bias in the linear regression line, as shown in Figure 2.1. Two possible solutions to this problem are: using variable bin sizes (Liu and Rost, 2001) or using the descending rank of the observations rather than the frequency.

It is easy to show that the descending rank of a power-law distribution also follows a power-law, with the power increased by one. A probability distribution function following a power-law has the form: $f(y) = cy^{-\tau}$, where c and τ are constants. The descending rank for element x can be obtained by a summation over all the elements $> x$. This summation can be approximated (for $\tau \neq 1$) by an integral, i.e. the number of observations (N) times the surface under the probability distribution curve from x to ∞ represents the number of observations greater than x. We get:

$$r(x) = N \sum_{x}^{\infty} cy^{-\tau} \approx N \int_{x}^{\infty} cy^{-\tau} dy = Nax^{-\tau+1}$$

where $a = \frac{c}{1-\tau}$ is another constant. Hence the descending rank of a power-law distribution should follow a straight line on a log-log plot with a negative slope. Furthermore we can obtain the power of the distribution function by subtracting one from the power of the descending rank function. For more detailed analysis for the use of ranking methods to show a power-law distributions see Adamic and Huberman (2002) and Troll and beim Graben (1998).

2.2.4 Measures of fold usage

In order to analyse the fold content of each genome we created a table with the number of copies of each fold per genome. Similar tables were created for copies of a superfamily or sequence family as classified by SCOP, and for copies of an architecture, a topology or a homology as classified by CATH . We can calculate the number of distinct folds per genome and the number of genomes on which a fold occurs from these tables. To obtain the average number of copies of a fold per genome, we divide the total number of copies for a fixed set of genomes by the occurrence for the fold on the same set of genomes; the

average number of copies is therefore calculated only over those genomes on which the fold occurs.

2.3 Results

2.3.1 Genome coverage

level	all	archaea	bacteria	eukaryotes
SCOP - PSI-BLAST				
superfamilies	1381	660	972	1156
folds	852	434	624	733
SCOP - SUPERFAMILY				
superfamilies	1421	708	1040	1179
folds	869	468	661	744
CATH - Gene3D				
homologies	1359	617	961	1038
topologies	788	372	572	629
architectures	37	30	35	35

Table 2.1: Fold numbers
The number of classification level entities for each set of assignments that have been found at least once in the given set of genomes (e.g. all, archaea, bacteria, eukaryotes).

set	genes	hits PB	%	hits SF	%	hits G3	%
all genomes	1,032,102	513,248	49.73	547,292	53.03	276,209	45.69
all archaea	40,214	20,274	50.42	23,197	57.68	18,050	44.88
all bacteria	340,113	181,773	53.44	201,017	59.10	148,799	45.64
all eukaryotes	651,775	311,201	47.75	323,078	49.57	101,035	42.40

Table 2.2: Summary of genomic assignments
The number and fraction of genes with one or more assignment (hits) for PSI-BLAST (PB), SUPERFAMILY (SF) and Gene3D (G3) assignments. Note that some genomes were not included in the Gene3D assignments (less than half of the eukaryotic genomes are present).

Table 2.1 shows the number of folds and superfamilies found on the genomes and Table 2.2 shows the number of genes with at least one assignment; a full table including the names of all genomes can be found in the Appendix (Table A). In general, SUPERFAMILY assigns more genes with a structure than either PSI-BLAST or Gene3D. The fraction of genes with an assignment is lower for eukaryotes (Table 2.2), yeast related species show a particularly low coverage (see Table A). In addition the amount of sequence covered

by structural alignments is low for eukaryotes compared to other species (source `http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/gen_list.cgi`), probably due to increased amount of multi domain sequences and disordered regions.

2.3.2 Genome length versus distinct folds

Distinct folds and kingdoms

source		genes					genes with assignment				
		R^2	slope	+/-	intercept	+/-	R^2	slope	+/-	intercept	+/-
PB	sfam	0.80	310	12.9	-562	47	0.89	384	11.2	-749	38
PB	fold	0.82	193	7.5	-311	28	0.89	236	7.0	-417	24
SF	sfam	0.82	316	12.0	-534	44	0.90	361	9.8	-652	34
SF	fold	0.83	190	7.0	-269	26	0.89	216	6.1	-333	22
G3	H	0.67	325	20.2	-673	72	0.90	377	11.4	-769	38
G3	T	0.68	198	12.1	-403	43	0.88	227	7.3	-452	24
G3	A	0.49	7	0.6	2	2	0.54	7	0.6	3	2

Table 2.3: Correlations for distinct folds on genomes
Correlation coefficients for the number of distinct folds (F) versus the number of genes or assignments on the genome (G). Coefficients are determined by linear regression, according to the formula $F = intercept + slope * log10(G)$. The +/- columns indicate standard errors of the estimated variables. Sets of distinct folds are taken from PSI-BLAST (PB) and SUPERFAMILY (SF), at fold and superfamily (sfam) levels, and from Gene3D (G3) at the homology (H), topology (T) and architecture (A) levels.

First we examine how the number of distinct folds found on a genome changes with genome size. Figure 2.2 shows that the correlation between the number of genes (G) and the number of different folds on a genome (F) can be described by a function of the form $F = intercept + slope * log10(G)$. Table 2.3 shows strong correlations for PSI-BLAST and SUPERFAMILY (R-squared > 0.8) and slightly weaker correlations for Gene3D (R-squared > 0.6). Note that the genomes used for the Gene3D assignments have been obtained from different sources than those for SUPERFAMILY or PSI-BLAST. Hence the number of genes on a genome in our data set, may not correspond to those used in Gene3D, which may explain the worse correlation with genes (Table 2.3 first column).

The differences in fold coverage of structural domains on the genomes creates noise in the data. To account for this the number of genes on the genome with a structural hit (hits), rather than the total number of genes on the genome, can be used to model

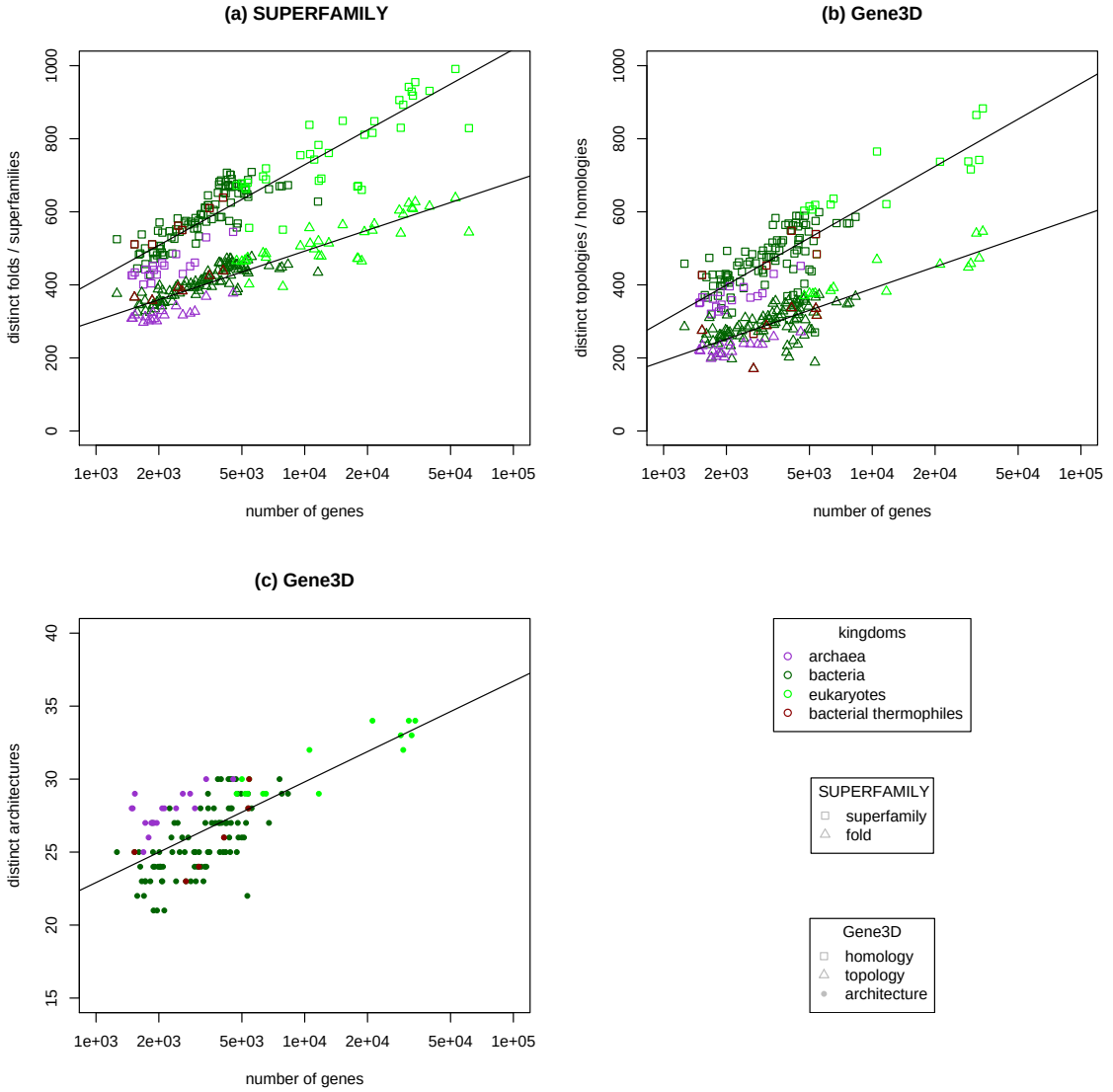


Figure 2.2: Distinct folds on genomes
The number of distinct folds versus genome size, given as the number of genes on a genome, are shown for: SUPERFAMILY (a) and Gene3D (b),(c). (a) Correlations for distinct folds and distinct superfamilies. (b) Correlations for distinct topologies (fold) and distinct homologies. (c) Correlations for distinct architectures. Corresponding correlation coefficients can be found in Table 2.3. Colours indicate the kingdom: archaea (violet), bacteria (dark green) eukaryotes (bright green); bacterial hyperthermophiles are displayed in dark red. Note the log-scale on the x-axes.

the correlation. In this case the correlations are even stronger (Table 2.3 second column). Note that Gene3D correlation coefficients are now in line with those for PSI-BLAST and SUPERFAMILY.

The correlation strengths do not change when instead of folds, the number of distinct superfamilies are correlated with genome size, or in the case of Gene3D, when instead of

topologies, homologies are used. This similarity suggests a similar evolutionary relevance for the fold and superfamily levels in the classification. Nevertheless, the similarity may be solely the effects of clustering superfamilies into folds (see Section 2.3.6). The architecture level shows a significantly weaker correlation.

The trends for genomes in the different kingdoms (Figure 2.2) appear to be different: archaea have relatively fewer distinct folds than similarly-sized bacteria or have larger genomes with the same number of distinct folds. (The points representing archaeal genomes lie below the linear regression line.) In contrast, archaeal genomes tend to have more different architectures for a similar genome length (Figure 2.2 (c)).

The first phenomenon could be due to the extreme living environments of archaea, which may only allow a subset of physiochemically very stable fold structures (Danson and Hough, 1998). However, this effect is not seen for bacterial extremophiles (Figure 2.2 (a),(b)), showing that not all species in extreme environments have necessarily fewer distinct folds. Archaea are known to have relatively few unique folds, i.e. folds only occurring in the archaeal kingdom, compared to bacteria and eukaryotes (Wolf et al., 1999; Caetano-Anollés and Caetano-Anollés, 2003); our data gives similar results. This may suggest that it is hard for archaeal folds to diverge into different folds due to extreme living environments. Extremophilic bacteria on the other hand might have diverged later to extreme environments, so that they were able to develop more distinct folds before moving into extreme conditions. However, the difference in number if distinct folds seems larger in older datasets (Abeln and Deane, 2005) than in the current dataset (Figure 2.2), which might indicate that this is simply caused by an under-representation of folds specific to archaea. Previously, a lower number of distinct folds, was also observed for eukaryotes; there is no evidence for this in the current datasets.

Distinct folds and structural classes

Surprisingly, the main fold classes as defined by SCOP relate differently to genome size (Figure 2.3). Here the genome size is given in the number of distinct folds observed on a genome, to make comparison of structural classes easier; similar trends are found when

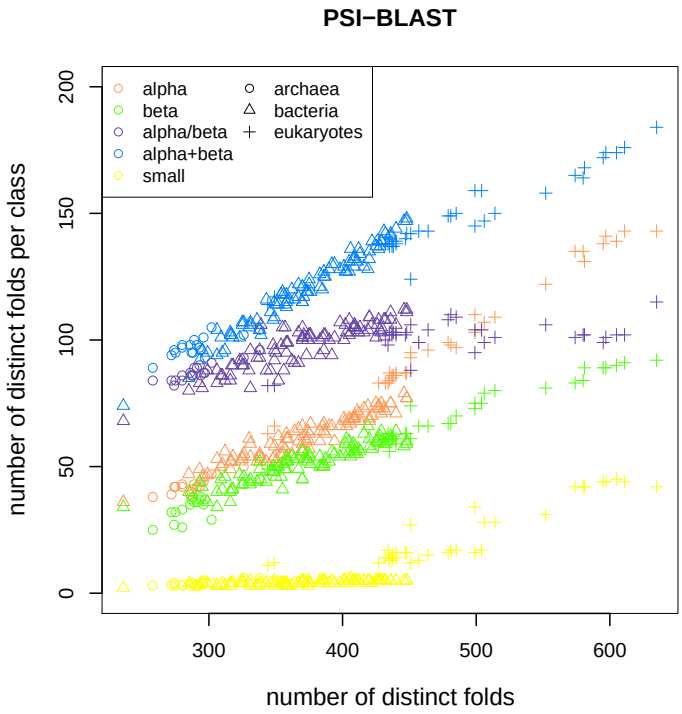


Figure 2.3: Distinct folds and class: PSI-BLAST
The genome size, given as the number of distinct folds found on a genome, versus the number of distinct fold per fold class for PSI-BLAST. Here the number of distinct folds as defined by SCOP is used.

the number of genes on a genome are used to indicate genome size.

The number of distinct all-alpha and small protein folds grows relatively faster on large genomes, i.e. there seems to be more room for developing new folds in these classes on the larger genomes. On the other hand the number of distinct alpha/beta folds seem to reach a maximum after which despite increasing genome size no further increase in distinct folds is seen. Changes in behaviour on larger genomes may be specific to eukaryotes. Very similar patterns as those described above are found for the alpha/beta class, the alpha class and the small proteins class, when we compare superfamilies instead of folds with genome size, or when we consider the SUPERFAMILY dataset.

In CATH, however, topologies (or homologies) do not show such a clear separation between fold classes (Figure 2.4 (a)), perhaps due to insufficient data for larger genomes (Section 2.2.2). Moreover, in CATH the alpha/beta and alpha+beta classes are joined into a single alpha-beta class. Note that the alpha+beta class in SCOP has the largest

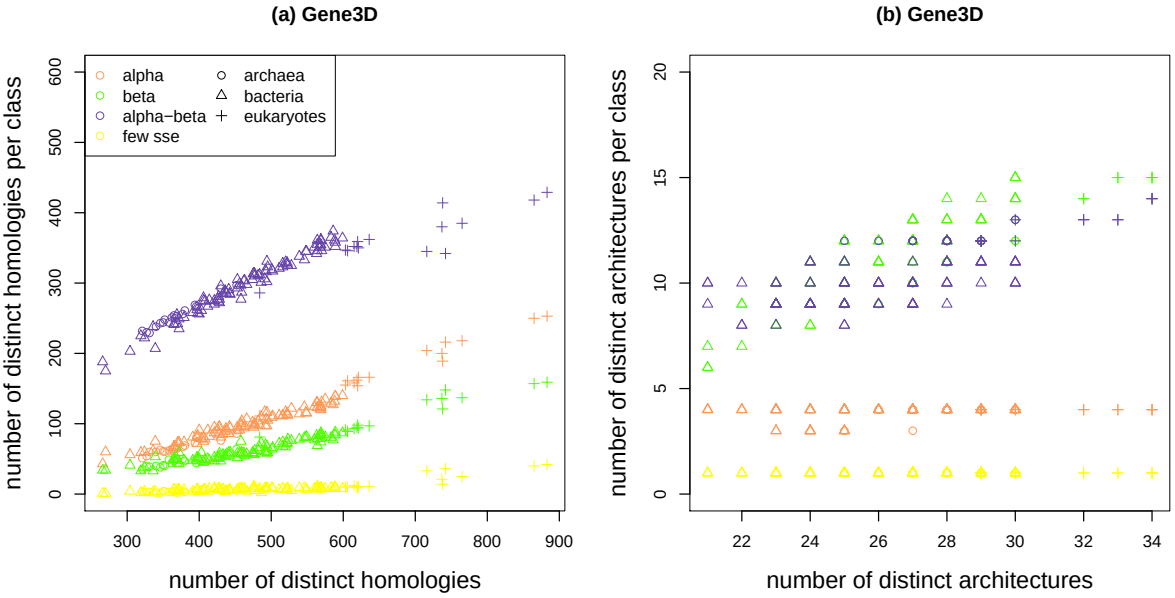


Figure 2.4: Distinct folds and class: Gene3D
The genome size, given as the number of distinct folds found on a genome versus the number of distinct fold per fold class for Gene3D. Here the number of distinct homologies (a) and architectures (b) as defined by CATH are used.

number of different folds, so that any difference might be obscured by the merge of these two classes.

Figure 2.4 (b) shows that the number of architectures for the 'all alpha' and 'few secondary structure elements' are not correlated with genome size, probably due to comparatively few architectures used to describe folds in these classes.

2.3.3 Copies

The distribution of copies of a fold on a single genome has been described to follow a power-law (Qian et al., 2001). Simply binning the data for a power-law will give bad estimates at low frequencies and does not make use of all available data. We therefore plotted the number of copies of a fold against the folds descending in rank according to their number of copies (see Section 2.2.3); this method will give a straight line on a log-log plot, if the underlying distribution is a power-law. In order to use the data from all genomes the average was taken only over those genomes on which a fold occurs. Here we investigate if the distribution of fold copies depends on structural class or kingdom.

Copies and kingdom

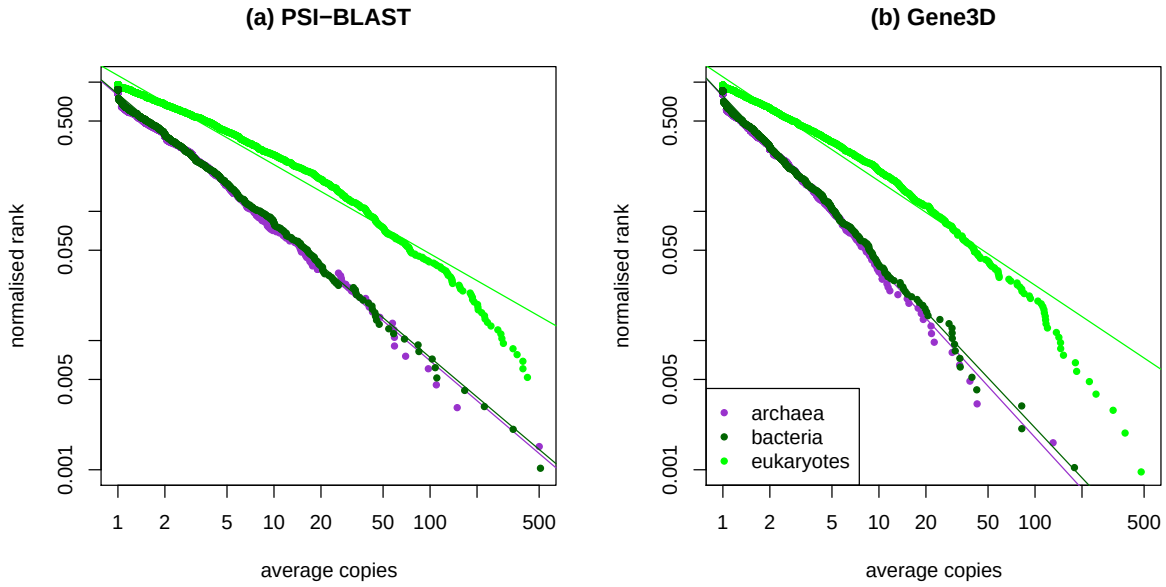


Figure 2.5: Copies and kingdom
Distributions for the average number of copies on genomes for a given kingdom. Results for PSI-BLAST at SCOP's superfamily level (a) and Gene3D at CATH's homology level (b). See Section 2.2.3 for the plotting method used.

The average number of copies of a fold per genome appears to follow a power-law distribution for archaea and bacteria (Figure 2.5). Eukaryotes contain relatively more copies of folds on their genomes than bacteria and archaea. However, the line describing copies on eukaryotes appears to curve more strongly than for archaea or bacteria. This effect seems slightly stronger for PSI-BLAST (Figure 2.5 (a)), and SUPERFAMILY (not shown) than for Gene3D ((Figure 2.5 (b)). Perhaps this is due to different and fewer eukaryotic genomes in the Gene3D set.

This may be caused by the ability of eukaryotes to have longer genomes, allowing genes in eukaryotes to duplicate with less restriction than in either bacteria or archaea. It is possible that the limited genome size actually creates the power-law shape, arguing in favour of the model created by Karev et al. (2002) to model the power-law. In their model the number of genes on a genome is stable (see 2.1.1). Note that fit of the power-law appears to deteriorate especially for larger genomes (not shown).

As with previous correlations, the trends look very similar when calculated at fold or topology level rather than at superfamily or homology level.

Copies and class

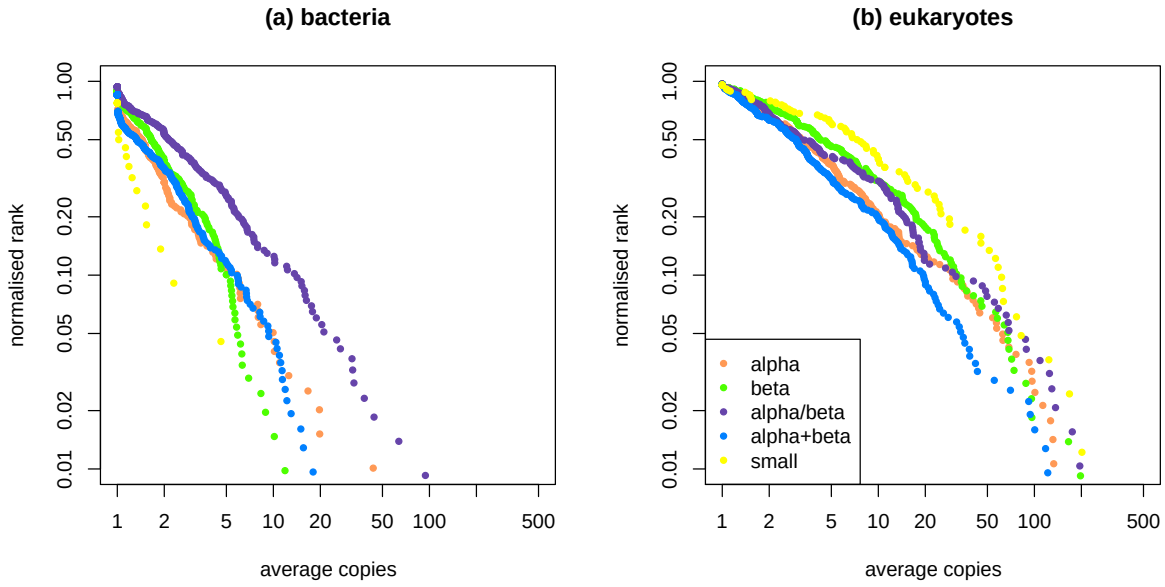


Figure 2.6: Copies and structural classes
Distributions for the average number of copies on genomes for a given kingdom and a given structural class. Results from the SUPERFAMILY database at SCOP's superfamily level are shown for bacteria (a) and eukaryotes (b).

Comparing the different fold classes in Figure 2.6 (b) we can see that the slope of the alpha/beta class in bacteria is more gradual than that of the other fold classes, implying bacteria have relatively more copies of alpha/beta folds than of other classes. On the other hand fewer copies of the small protein class are seen on bacterial genomes. This effect is weaker in archaea (results not shown) and not seen in eukaryotes (Figure 2.6 (c)). Interchanging superfamily and fold level, or replacing the assignment from SUPERFAMILY to PSI-BLAST have no major effects on these trends. For Gene3D's homology and topology level there are no obvious distinctions between structural classes (results not shown). At architecture level, only the copy distribution of the 'all beta' class approximates power-law-like behaviour.

2.3.4 Families per fold

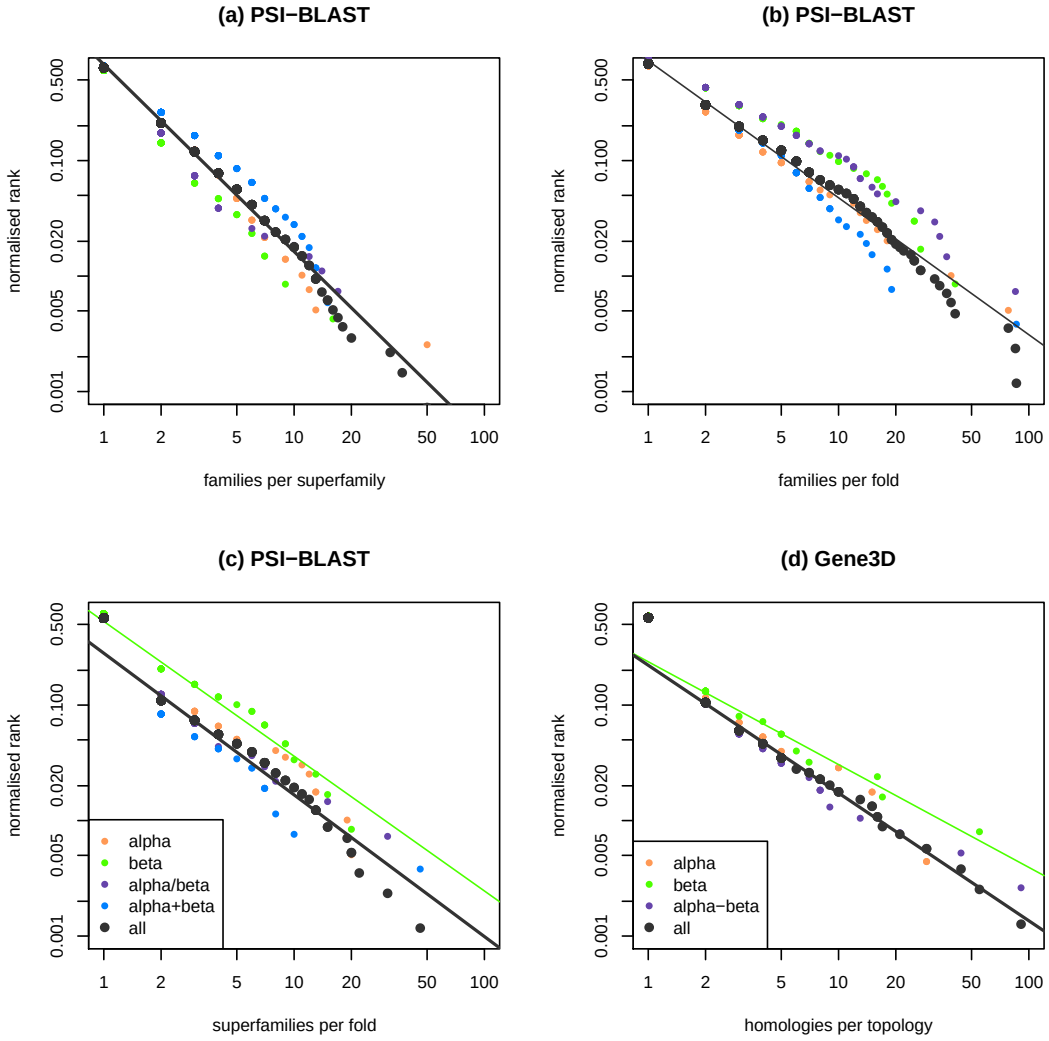


Figure 2.7: Families per fold
Distributions of the number of families per superfamily (a), families per fold (b), superfamilies per fold (c) and homologies per topology (d). Data from *all* genomes was used with assignments of PSI-BLAST in (a),(b) and (c) and from Gene3D in (d). The linear regression lines, are calculated excluding the first point in the distribution, to make the difference in distribution for folds with a single superfamily more apparent. Distributions for sets of data covering single kingdoms follow similar patterns.

To investigate the number of families under a fold, a fixed set of genomes is used, e.g. archaea, bacteria, eukaryotes or all genomes, as opposed to using the number of families available in SCOP. This gives us a more biologically representative set of the classification hierarchy.

Figure 2.7 shows that the number of superfamilies or families under a fold as well as

the number of families under a superfamily follow power-law behaviour within a fixed set of genomes. It has previously been suggested that a power-law is an overly simplistic approximation of the distribution of families under a fold (Wolf et al., 2000; Coulson and Moult, 2002; Liu et al., 2004). However looking at our data using descending rank there is a straight line on a log-log plot (Figure 2.7 (a) and (b)), similar to the descending rank for fold copies.

The power-law-like distribution may be caused by the classification system rather than a biological process. It is probable that in developing a classification system like SCOP or CATH, it is easier to classify a new family under a superfamily already containing many families, since there are more families with which similarity can be observed. A similar argument could be made for superfamilies under a fold. This process of the rich get richer can lead to a power-law distribution (Barabási and Oltvai, 2004).

The power-laws found at different levels, i.e. family, superfamily and fold (Figure 2.7 (a),(b),(c),(d)), may be caused by the property of a power-law distribution to be scale free. This is a fractal property implying that the level of detail in a distribution does not change the overall distribution. This may also explain why power-laws are seen for the distribution of copies at all different classification levels, e.g fold level, superfamily level and for sequence families (Huynen and van Nimwegen, 1998). Furthermore, it could account for the very similar results for both distinct folds and distinct superfamilies versus genome size (section 2.3.2). However, truly scale free behaviour should also result in identical powers for the power-law distribution, which is not observed in our results.

The power-law does not appear to hold for folds with only one superfamily (Figure 2.7 (c)) or for topologies with only one homology (Figure 2.7 (d)). These single superfamily folds seem to be over-represented. It is these data points on the left-hand-side of the figures which represent most data and should therefore be most reliable. Single superfamily folds may correspond to the unifolds as described by Coulson and Moult (2002); a unifold is a fold containing only one sequence family. In our data these folds contain a single superfamily. An analogous outcrop is not seen for family per fold or family per superfamily (Figure 2.7 (a),(b)). Alternatively, the over-representation of folds with only

one superfamily may be caused by a bias in SCOP and CATH, or may be caused by a biological process (see discussion in Section 3.3.7). In contrast all beta folds as defined by SCOP *do* seem to follow a power-law like distribution over the entire range of the number of superfamilies per fold (Figure 2.7 (c), green line).

Distributions for topologies per architecture and homologies per architecture were also investigated. These distributions appear to be much more chaotic (results not shown). Since there is only a small number of architectures, it is difficult to conclude whether these also follow a power-law like distribution or not.

2.3.5 Genomic occurrences

Figure 2.8 shows that the distribution of fold occurrences across the genomes does not follow a power-law, unlike the distribution of number of copies on a genome and the number of families per fold. For most fold classes and kingdoms the occurrence distribution is characterised by a peak for folds occurring on very few genomes, and another peak for folds occurring on many genomes (e.g. Figure 2.8 (e)). This distribution pattern can be explained by combining two sets of folds (1) relatively old folds occurring on most genomes except for a few deletions and (2) relatively new folds occurring on only a few genomes. A model simulating divergent evolution, indicates that such a distribution may indeed be achieved by a combination of young and old folds (see Section 2.4). The correlation between occurrence and age is further investigated in Chapter 3.

When the occurrence distributions of different fold classes and kingdoms are compared, it was observed that the height of the two peaks in the distribution varies with kingdom and fold class. In the case of eukaryotic genome occurrences the peak for folds occurring at very few genomes appears to be much lower than in the other kingdoms. More recent evolution of eukaryotes could well explain this. Alternatively, there might be an under-representation of eukaryote specific structures. Again fold and superfamily occurrence distributions follow very similar patterns. Differences in class and fold age are further investigated in Chapter 3.

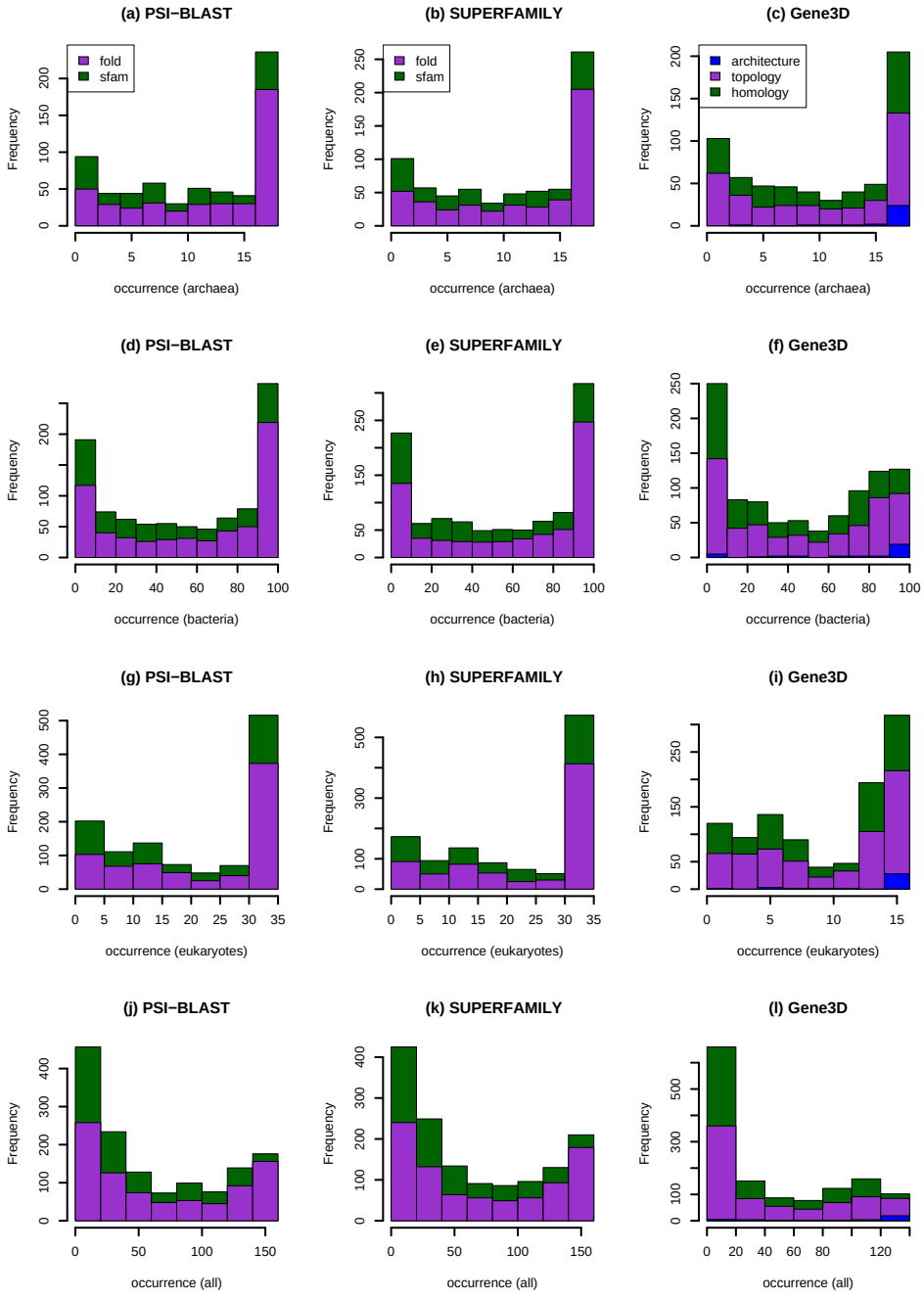


Figure 2.8: Occurrence distribution
Distributions of genomic occurrences of folds, for archaea (a),(b),(c), bacteria (d),(e),(f), eukaryotes (g),(h),(i) and all genomes (j),(k),(l). Note that the distributions here are not cumulative: the fold distributions are depicted on top of the superfamily (sfam) distributions, the topology distributions on top of the homology distributions and the architecture distributions on top of the topology distributions.

Gene3D gives a relatively larger number of folds which occur on only a few genomes (Figure 2.8 (l). This is perhaps due to the CATH classification, which tends to split proteins into more domains than SCOP, creating on average smaller (and more specific)

folds.

2.3.6 Correlations between fold usage measures

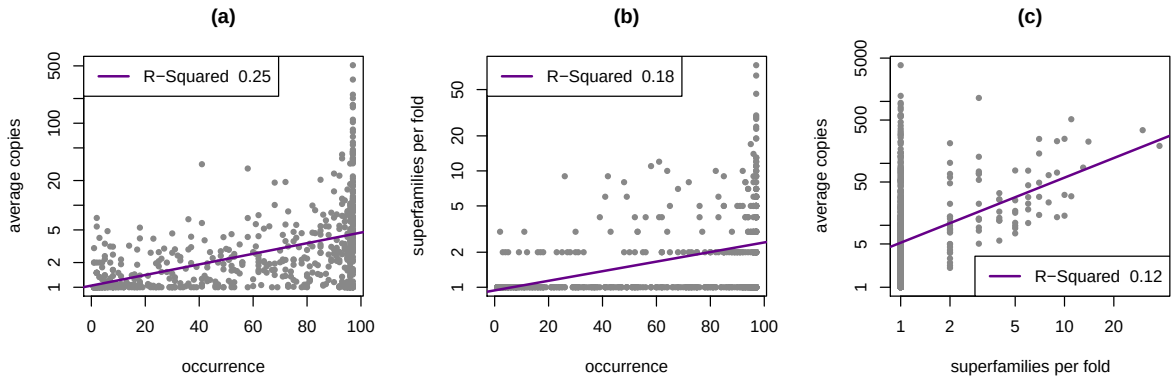


Figure 2.9: Correlations between fold usage measures
(a) Fold occurrence versus the average number of copies. (b) Fold occurrence versus the number of superfamilies per fold. (c) Number of superfamilies per fold versus the average number of copies. All fold usage measures are calculated for bacterial genomes. Please note the log-scales on the axes corresponding to average copies and superfamilies per fold. Here PSI-BLAST assignments are used.

So far we have discussed the distribution for our measures of fold usage, now we will investigate if these measures are related. We expected that the number of copies of a fold per genome will increase if it occurs on more genomes. Similarly it is expected that the number of superfamilies per fold will increase with occurrence and copies (assuming divergent evolution). Since a superfamily has more chance to diverge into a new superfamily, if there exist more copies of a fold. However Figure 2.9 shows that there are no clear relationships between these measures.

Copies vs. fold occurrence

Figure 2.9 (a) shows the relation between fold occurrence and fold copies. The relation appears to be restrictive, such that a fold with a low number of occurrences can not have many copies on a genome. This could be explained by considering that occurrence across genomes can indicate the age of a fold. Only older folds, i.e. folds occurring on many genomes, will have had the time to duplicate significantly. However, it also shows that

old folds do not necessarily have many copies. Hence the age of a fold creates an upper limit for the number of copies of a fold per genome but not a lower limit. This indicates that the number of copies on a genome alone might not be a good estimator for the age of a fold. Within the restriction of time the distribution of copies again appears to follow a power-law. Overall the number of copies on a genome is likely to be more self dependent than related to the fold occurrence. From the power-law distribution it follows that folds with many copies are more likely to duplicate again. In addition, if we assume that occurrence can give us an estimate for the age of a fold, the maximum number of copies for a fold is restricted by the time the fold has had to duplicate.

Superfamilies per fold vs. fold occurrence

A similar restrictive relation appears to occur for the number of superfamilies per fold with fold occurrence across genomes (Figure 2.9 (b)). We could again argue that time limits the number of new superfamilies a fold can create. However, the observed relation might also be caused by the grouping of superfamilies into folds. The larger the group size the higher the chance that there exists a superfamily in the group which has a high number of occurrences. To investigate the effect of grouping superfamilies into folds, we created data sets where superfamilies were assigned to random folds, with the same distribution as in the real data. Using these artificial fold groups does indeed give similar results. This would probably still suggest that the number of superfamilies per fold is self dependent as indicated by the power-law distribution; so that a fold with many superfamilies is more likely to gain a new superfamily, either in SCOP or evolutionarily (see discussion in Section 2.3.4), than a fold with only one superfamily.

These results are very similar if we look at relations between copies, occurrence and families in a superfamily rather than superfamilies in a fold.

Superfamilies per fold vs. fold copies

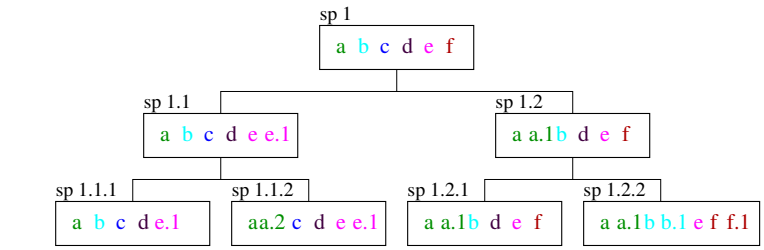
The number of superfamilies per fold compared to the number of copies shows a slight correlation (Figure 2.9 (c)). An obvious correlation may be expected: two superfamilies in

a fold, may provide double the number of copies. If this correlation is removed by dividing the number of copies by the number of superfamilies per fold, no significant correlation remains (not shown).

It seems, once again, that the number of superfamilies in a fold is more self dependent than related to the number of copies. However, there may be small unseen effects due to the grouping of superfamilies.

2.4 Occurrence simulation

2.4.1 Simple model



	sp 1.1.1	sp 1.1.2	sp 1.2.1	sp 1.2.2
a	1	1	1	1
a.1	0	0	1	1
a.2	0	1	0	0
b	1	0	1	1
c	1	1	0	0
d	1	1	1	0
e	0	1	1	1
e.1	1	1	0	0
f	0	0	1	1
f.1	0	0	0	1

Figure 2.10: Simulation model
Top: An example of the model with two speciation events. Bottom: An example of an occurrence table corresponding to the lowest level of speciation in the example above. The genome of the last common ancestor, sp 1, contains 6 distinct folds (a, b, c, d, e, f). It then diverges into sp 1.1 and sp 1.2. An example of a modification event, where a fold diverges into a related fold, occurs from sp 1 to species 1.1 where fold e is modified into fold e.1. An example of a deletion event occurs from sp1 to species 1.2 where fold c is deleted.

Here we investigate with a simple evolutionary model how the observed occurrence

distribution (Figure 2.8) may have developed. Furthermore we examine what possible information can be extracted from an occurrence table, where each matrix cell indicates whether a species occurs on a certain genome. Previously both species trees (Lin and Gerstein, 2000; Yang et al., 2005) and fold trees (Caetano-Anollés and Caetano-Anollés, 2003) have been built from such a table. We inspect with a theoretical model if such clustering is likely to give the desired results. In our model the number of copies of a fold is not explicitly considered.

We propose a simple evolutionary model that modifies and deletes folds on diverging genomes.

Assumptions in the model:

1. The last common ancestor contains N distinct folds.

2. When two species diverge into different species from a common ancestor, some folds will be conserved, others deleted and some modified.
   The *deletion* factor holds the chance that a fold is deleted.
   The $modification$ factor describes the chance that a fold duplicates and gets modified into a related structure, but is unrecognisable as the same fold.

3. The deletion and modification factor do not depend on the fold type. This is a clear simplification, as these factors might depend on the number of copies of a fold as well as on the function of a fold.

4. The deletion and modification factor are independent of the level of speciation n.

5. Species diverge according to a binary tree

6. No horizontal gene transfer is allowed.

Constraints on parameters:

The number of distinct folds on a genome at level n of the species tree is given by:

$$x_n = x_{n-1}(1 + modification - deletion) \tag{2.1}$$

If we assume that the number of distinct folds on a genome remains constant during evolution, we need to put $modification = deletion$. Such an assumption appears roughly feasible, if Figure 2.2 and 2.8 are compared: the number of distinct folds found on bacteria is approximately 300 and around 250 folds are found on almost all bacteria, indicating only a small increase in the number of distinct folds between the last bacterial common ancestor and currently living bacteria. Nevertheless, many more lineage specific folds may be found, increasing the average number of distinct folds for currently living bacteria.

Figure 2.10 gives a graphical example of the model described above, including the observed occurrence table at the lowest level of speciation. Note that folds starting with the same letter are defined as related folds; A child fold (e.g. $a.1$ or $a.2$) is evolutionarily related to a parent fold (e.g. a), but the homology is not recognisable. In the occurrence table the related folds do not in general have a similar occurrence pattern.This would imply that on assumption of this simple model it is challenging to make a phylogenetic tree of folds on basis of the observed occurrence patterns.

2.4.2 Occurrence distribution

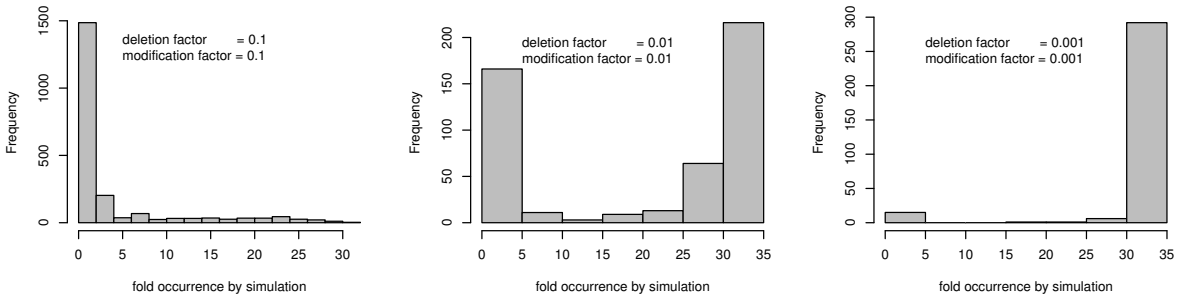


Figure 2.11: Simulation occurrence
Distribution of genomic occurrences obtained by simulation with 300 folds on the last common ancestor, and 5 cycles of speciation, i.e. there is data for 32 genomes.

We simulated the simple evolutionary model for a few generations of speciation to generate occurrence patterns for folds. Figure 2.11 shows the occurrence distribution for folds generated by the simulation for different modification factors. The distribution contains two peaks, one for folds occurring on a few genomes and one for folds occurring

on most genomes, similar to the occurrence distributions for PSI-BLAST and SUPERFAMILY assignments (Figure 2.8). The height of the peaks depends on the modification and deletion factor. If we allow more diverging events, the peak for folds on very few genomes will increase. If we allow more deletion events, the peak for folds occurring on most genomes will decrease.

Comparing the simulated distributions with the distributions from assignment data, we observe that the frequencies of occurrence in the middle of the distribution are much lower for the simulated data than for the data from the assignment exercises. It does not appear possible to create a higher frequency in the middle of the distribution by changing the deletion and modification factors with the current model. Similarly, creating a variable modification and deletion factor based on an underlying power-law distribution (e.g. copies) does not result in higher frequencies in middle. Perhaps, including lateral gene transfer or simulation on a less balanced tree, reflecting non-homogeneous set of genomes, may result in distributions closer to the observed ones.

2.4.3 Species trees

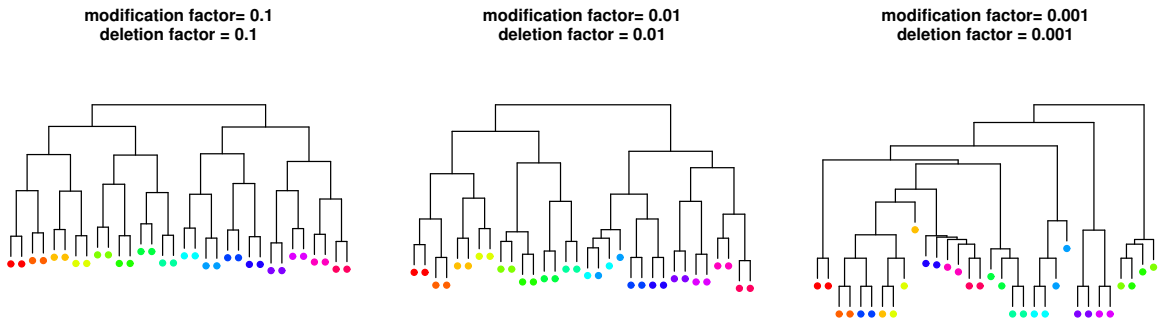


Figure 2.12: Simulation species trees
Species trees from simulation starting with 300 folds on the parental genome and 5 cycles of speciation. Identical colours indicate that genomes should cluster together in the phylogenetic tree. A symmetric distance measure was used to cluster the genomes.

When we build trees of the occurrence patterns from the simulated data, using a standard agglomerative clustering method, 'agnes' (Struyf et al., 1996); related species cluster nicely together (Figure 2.12). Unsurprisingly the accuracy of the trees calculated

from the simulated occurrence data increases with number of modification and deletion events used in the simulation.

2.4.4 Fold trees

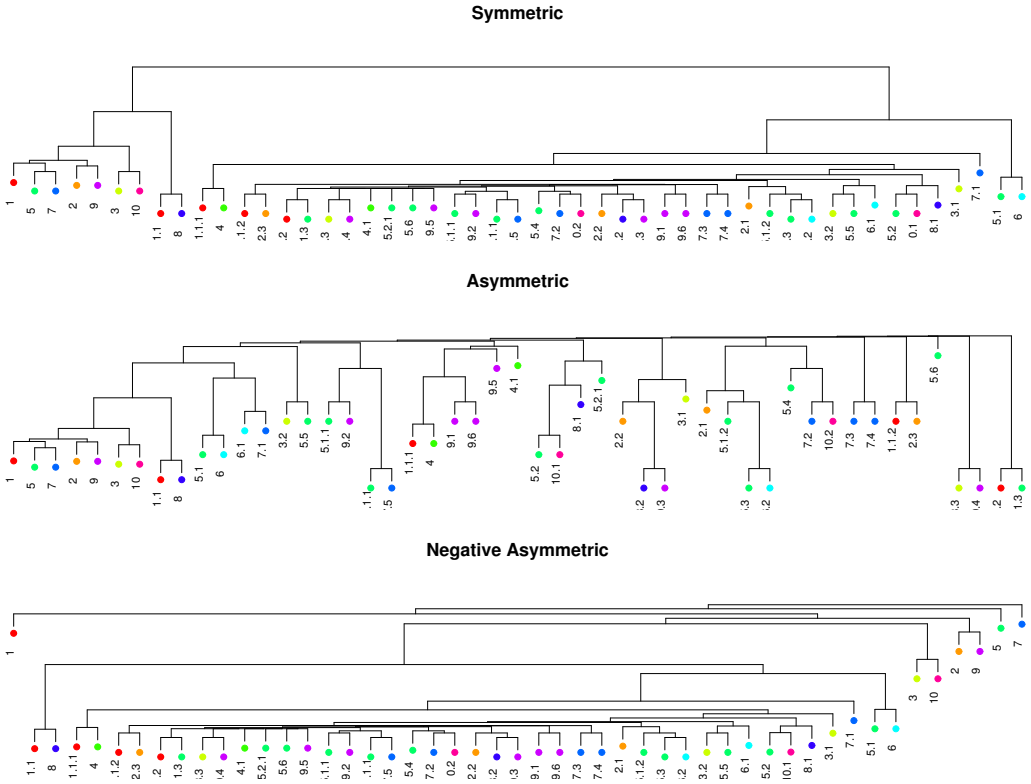


Figure 2.13: Simulation fold trees
Fold trees from simulation starting with 10 folds on the last common ancestor and 5 cycles of speciation. The deletion and modification factor are set to 0.05. Three different distance measures are used for calculating the distances between two fold occurrence patterns; see Table 2.4.

Figure 2.13 shows folds clustered together on the basis of occurrence patterns produced by our model. Here related folds are generally not clustered together. It is easy to see why this happens if we consider an example of the model in detail, e.g. considering a and $a.1$ in Figure 2.10. Folds on the other hand which occurred on the last common ancestor between all species do cluster together. Hence it is likely that fold trees based on occurrence patterns (Caetano-Anollés and Caetano-Anollés, 2003) cluster folds together modified or deleted at a similar speciation event during evolution. Folds, clustered together in this way, are likely to be functionally rather than evolutionarily related, such information

has already been exploited to predict functionally similar proteins (e.g. Pellegrini et al. (1999)).

$$\text{distance} = d(x,y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} \alpha(x_i, y_i)}$$

measure	$\alpha(x_i, y_i)$
symmetric	$= 1$
asymmetric	$= \begin{cases} 0 & \text{if } x_i = y_i = 0 \\ 1 & \text{otherwise} \end{cases}$
inverse asymmetric	$= \begin{cases} 0 & \text{if } x_i = y_i = 1 \\ 1 & \text{otherwise} \end{cases}$

Table 2.4: Distance measures
Above the equation is given to calculate the distance d between two genomes x and y, where x_i and y_i are the occurrence for the i^{th} fold on the genome. In the table three different measures to describe the distance between two binary occurrence vectors are given: a symmetric, an asymmetric and an inverse asymmetric measure. The difference between these methods is the weight of the characters used for the normalisation of the distance. The symmetric measure puts equal weights on both characters. The asymmetric measure puts more weight on the ones in the genome, i.e. occurrence of a fold. The inverse asymmetric measure puts more weight on the zeros, i.e. deletion of a fold.

2.5 Conclusions

We have shown that the fold usage on genomes can give us an idea of the mechanisms behind protein fold evolution. In particular it can give us an indication of the evolution of folds in different kingdoms or in different fold classes.

Comparing different kingdoms showed that archaea have relatively fewer distinct folds on their genomes than bacteria (Figure 2.2). However, this may be due to a bias in the PDB. Furthermore eukaryotes appear to have a different distribution for the number of copies of a fold per genome than archaea and bacteria, which both seem to follow a power-law (Figure 2.5)

The structural class alpha/beta behaves differently from the other fold classes in several ways. The number of distinct folds of the alpha/beta class does not increase with

genome size on eukaryotes (Figure 2.3). The number of copies on bacteria is much higher than that for other fold classes (Figure 2.6). On the other hand the number of distinct folds for the small protein and all alpha classes is shown to increase relatively more rapidly on the larger genomes (Figure 2.3)

It will remain difficult to prove that the distributions for copies and families per fold are true power-laws. However, power-law behaviour where 'the rich get richer' seems likely for copies of a fold on genomes as well as the number of families per fold, although in the latter case it is possible that this effect is caused by the creation of a classification system, rather than by a biological mechanism.

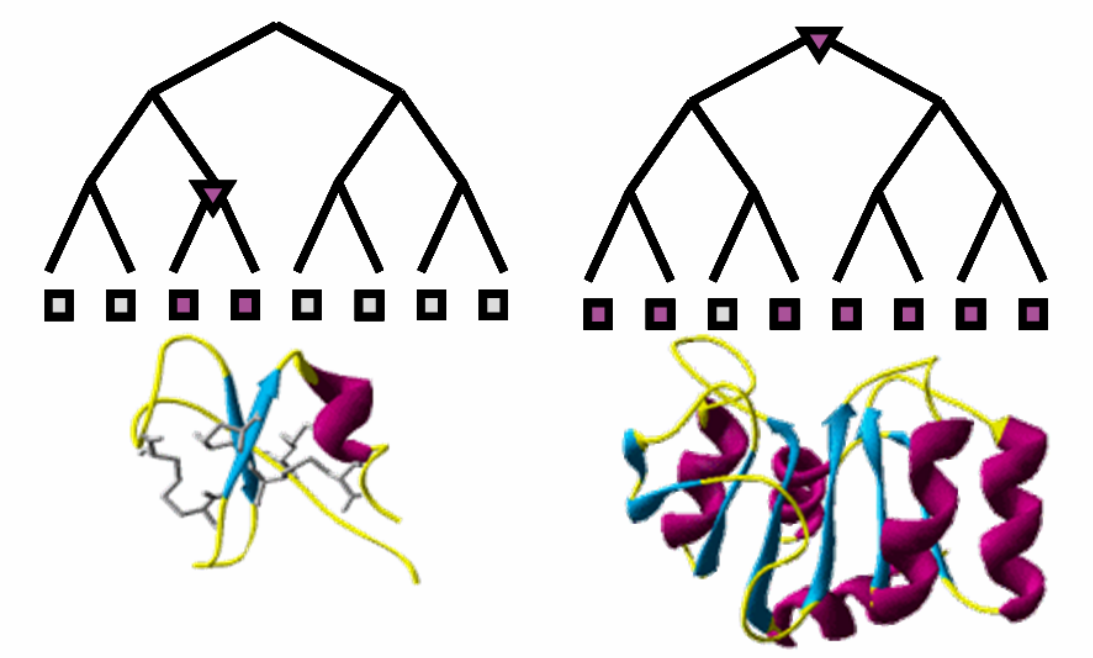
We observe no clear relationship between fold occurrence, copies and the number of sequence families under a fold (Figure 2.9). There exist folds which occur on all genomes, with only one copy per genome on average. Similarly there exist folds with only one superfamily, but many copies of these superfamilies. Although more subtle relations might be hidden due to a grouping effect, it is clear that a high occurrence across genomes does not necessarily imply a high number of copies or a high number of superfamilies.

Similar distributions and correlations are found at different classification levels (e.g. superfamily, homology, fold, topology). CATH's architecture level, however, shows rather different trends, perhaps reflecting its non-evolutionary basis.

Occurrence patterns, are more variable with respect to kingdom, structural class and classification system used. Most distributions show a peak at a high number of genomic occurrences and/or at a low number of genomic occurrences. A simple evolutionary model suggest, that these peaks may be explained by folds that arose relatively early or late in evolution. The model also suggests that genomic occurrence patterns for folds may be used to create a phylogeny of species, predict function or estimate ages for a fold, but that it may be difficult to obtain an evolutionary fold-tree based on such data.

Chapter 3

Relative age estimates for folds and superfamilies



3.1 Introduction

It has become clear from occurrence studies, e.g. Chapter 2, that different folds arose at different points in evolutionary time. An estimation of the ages of different folds would be a starting point for many investigations into protein structure evolution: for example how we arrived at the set of folds we see today.

Here we develop a relative age estimation technique for protein folds. Our method is based on constructing parsimonious scenarios which can describe fold occurrence patterns in a phylogeny of species.

3.1.1 Fold assignments and completed genomes

In this study we use protein fold occurrence data over as wide a range of completed genomes as possible to estimate relative ages of protein folds. Previously, simple age measures have been suggested such as the number of completed genomes possessing a fold or the total number of copies of a fold detected on completed genomes (Abeln and Deane, 2005). We provide a more sophisticated approach that incorporates the phylogenetic distribution of genomes into the analysis. The fold occurrence data is used to construct approximate whole-genome phylogenies of the species. Patterns of occurrence across these trees are then used to estimate relative fold ages.

This investigation of protein structure evolution requires a method of structural classification. Here we use both SCOP and CATH. Divergent evolutionary relationships are presumed to exist at the superfamily level in SCOP and the homology level in CATH. However, the evolutionary relatedness of structures at the level above, fold or topology level, remains in debate, but many including Koonin et al. (2002) argue that most folds are monophyletic. Relative ages are estimated at superfamily (homology) as well as fold (topology) level.

3.1.2 Genome evolution

Using classification systems we can examine the occurrence patterns of folds on genomes. Occurrence patterns can be described as the result of gene or domain duplication, deletion and acquisition events. This leads to a set of processes which create, copy and delete entire folds on genomes. For example duplication and subsequent modification by accumulation of mutations (in a divergent evolutionary scenario) may lead to related structures of the same fold, or unrecognisable structures classified as new folds. Acquisition may be via *ab initio* innovation, or by horizontal gene transfer (HGT) between species. Gene inactivation is equivalent to a deletion event in this context, assuming the probability of reversion is negligible.

This brings into debate the size of ancestral genomes, and whether the genome of the last universal common ancestor (LUCA) contained many genes/folds, with subsequent evolution dominated by losses, or alternatively, LUCA contained a minimal set of genes/folds for independent survival, with current diversity built up by an excess of gains over losses. Snel et al. (2002) and Mirkin et al. (2003) use parsimony arguments with occurrence patterns of Clusters of Orthologous Groups of proteins (COGs) on assumed bacterial phylogenies to relate LUCA inferred genome size to the relative weightings of gene loss and horizontal gene transfer (or independent innovation). Snel et al. (2002) estimate that the last common ancestor of bacteria contained around 2500 genes, and the ancestor of the Archaea around 2050 genes. Mirkin et al. (2003) favour an equal weighting of loss and HGT, yielding a LUCA genome roughly corresponding to a minimal set for functional independence. The work presented here can also be used to estimate the fraction of known folds or superfamilies which is ancestral to all superkingdoms.

The work presented here is an update of previously published work (Winstanley et al., 2005) performed with more recent data. We use a phylogeny, that was previously constructed, to estimates relative ages of folds. All previously constructed whole-genome trees, based on superfamily or fold occurrence data, segregated species into the three superkingdoms and all trees had similar topologies. Here, data from three different struc-

tural domain assignment methods demonstrate the robustness of our age estimates. The distribution of the relative ages for the different fold classes show that alpha/beta folds are relatively older than folds of the other classes, and that small protein folds are relatively younger. Restrictive correlations between our relative age estimate and other possible age indicators, such as protein interactions and genomic abundance, are observed.

3.2 Methods

3.2.1 Genome assignments

The analysis presented in this study is based on PSI-BLAST (PB, Altschul et al. (1997)), SUPERFAMILY (SF, Gough and Chothia (2002)) and Gene3D (G3, Lee et al. (2005)) assignments. In total 150 genomes were used, including 18 archaea, 97 bacteria and 35 eukaryotes. See Section 2.2.1 for more details and Appendix A for a full list of genomes used.

3.2.2 Construction of species trees

A phylogenetic tree of the 150 genomes in our set was built by clustering the occurrence patterns of SCOP's superfamilies. In previous work, multiple tree building approaches were used, resulting in fold ages with little variance. Species trees reconstructed from the occurrence data were able to segregate the three superkingdoms, apart from a few pathogenic or endosymbiotic species, which have already been removed from the genome set in this work. Here a single tree based on a previous set of assignments (SUPERFAMILY 1.65) and built with a neighbour joining algorithm using the Jaccard dissimilarity measure is used (see Figure 3.1). For more details on different tree building methods and their effect on the data presented see Winstanley et al. (2005).

Tree transformation

In order to investigate relative fold age, a rooted tree is required with leaf nodes corresponding to contemporaneous genome observations. For this purpose we rooted the tree reconstructions and applied branch length transformations based on simplifying assumptions.

A root was imposed on the trees at the trifurcation of the superkingdoms. This was consistent with the notion that the current data was of insufficient resolution to conclusively support placing the root on any individual superkingdom branch. The branch lengths are transformed such that the genome observations were at height zero and the root at an arbitrary height one, with nodes distributed over the interval $[0, 1]$. At each internal node, the rate of evolution in the upward (ancestral) branch was assumed to equal the average of the evolutionary rates over the paths to all leaves subtended at the node.

3.2.3 Age measures

The principal aim is to provide an approximate measure of relative – rather than absolute – fold age. Furthermore, no attempt is made to provide a linear timescale, so inferred relative ages are treated purely as an approximate ordering on the time line.

Convergence age

A simple fold age estimate is given by the height of the most recent node which covers all leaves in the tree with an occurrence of that fold; we call this estimate the convergence age A_c. This corresponds to the assumption that fold occurrence patterns are due to a single innovation event followed by lineage-specific losses within the subtree. This age estimate may be considered as our upper bound estimate of the fold age, disregarding possible false negatives in the genome assignments.

Parsimony age

The parsimony age is a more sophisticated fold age estimate, it attempts to account for the possibility of horizontal gene transfer and false positive predictions.

The occurrence pattern of a fold can be explained by the allocation of gain and loss events across the tree according to principles of maximum parsimony. The reconstructed scenario depends on the relative weighting of horizontal gene transfer and gene loss events (Snel et al., 2002).

We make parsimonious allocations of gain and loss events using the method proposed by Mirkin et al. (2003). These authors investigated various methods for parsimonious evolutionary reconstruction of COG occurrence patterns and the effect of varying gain penalty on the inferred genome size and functional independence of a putative LUCA. They found that equal weighting of gain and loss penalties (yielding a minimum total number of gain and loss events) gave LUCA genome reconstructions roughly consistent with functional independence. We use their 'PARS-G algorithm' with equal gain and loss penalties in our parsimonious reconstructions.

The parsimony age (A_p) for a fold is then given by the relative age of the node in the species tree with the highest gain assigned by the parsimony algorithm, given the occurrence pattern. Hence the convergence age estimate presented above effectively represents a method of this type with extreme penalisation of gain events.

3.2.4 Kingdom specific ages

In order to compare superkingdom-dependent features, such as mean copies and interactions, with the relative age, we develop a superkingdom specific age (A_{pK}). This can also correct for lineage specific evolution rates between the superkingdoms. A simple approach would calculate A_{pK} as A_p on the subtree for that superkingdom ($K \in \{A, B, E\}$). However, this would ignore information about occurrence on genomes in other superkingdoms. We would like to distinguish between a fold ancestral to the subtree K and a fold ancestral to the complete tree. To preserve information about occurrence in other superkingdoms we

use the following rules to calculate A_{pK}: if $A_p \neq 1$ than A_{pK} is the height of the highest node in K, otherwise if $A_p = 1$ than $A_{pK} = 1$ unless there is a loss at the highest node in K, in that case A_{pK} is the height of the highest node in K. To understand the special case for a most parsimonious loss at the top node of K, we can consider the example in Figure 3.1. The fold seems ancestral to all archaea and eukaryotes, but it only occurs on a few bacterial genomes. If we try to relate mean copies on bacteria and relative age for bacteria, it seems more plausible to assume that this fold has been obtained later in evolution for some bacterial genomes, for example by HGT.

3.3 Results

3.3.1 Convergence and parsimony age

dataset	total	ancestral to 2 kingdoms		ancestral to all kingdoms		age agreement
		$A_c = 1$	$A_p = 1$	$A_c all = 1$	$A_p all = 1$	$A_p = A_c$
folds						
PB	854	0.64	0.42	0.44	0.33	0.64
SF	869	0.68	0.45	0.47	0.36	0.65
G3	786	0.58	0.36	0.39	0.28	0.61
superfamilies						
PB	1384	0.60	0.37	0.40	0.28	0.61
SF	1418	0.64	0.39	0.42	0.29	0.59
G3	1354	0.55	0.34	0.36	0.25	0.60

Table 3.1: Fraction of folds assigned to LUCA
The fraction of folds and superfamilies which is thought to be ancestral to at least two superkingdoms (second column), the fraction of folds ancestral to all superkingdoms (third column) and the fraction of agreeing age assignments (last column). Here A_c is the convergence age, and A_p the parsimony age. $A_c all = 1$ is defined as $A_{cA} = A_{cB} = A_{cE} = 1$, i.e. folds occurring at least once in each kingdom; $A_p all = 1$ is defined as $A_{pA} = A_{pB} = A_{pE} = 1$, i.e. folds that are ancestral to each kingdom according to the parsimony algorithm. The last column shows the fraction of folds with high confidence age assignments: $A_p = A_c$ is the fraction of folds for which convergence and parsimony age coincide.

We initially compare the convergence (A_c) and parsimony (A_p) age measures (Table 3.1).

Where A_c and A_p are identical there is reasonable confidence in the estimates of the node at which a fold arose, this constitutes around 60% of the folds for all data sets

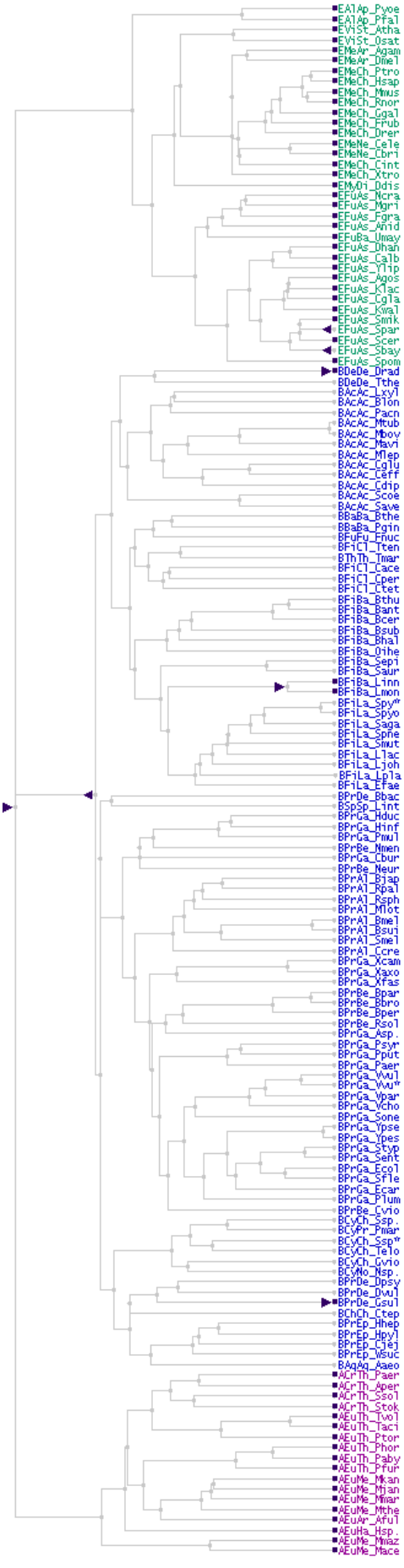


Figure 3.1: Example tree
Species tree with the occurrence pattern for SCOP superfamily a.94.1, ribosomal protein L19 (L19e). The tree was created with SF data on superfamily occurrence patterns, using the Jaccard distance measure. The right pointing triangles indicate gain events, the left pointing triangles loss events. The node at the root determines the relative age for: $A_c = A_p = A_{pA} = A_{pE} = 1.0$. The gain event in the bacterial subtree determines $A_{pB} = 0.15$. Archaeal genomes are depicted in magenta, bacterial genomes in blue and eukaryotic genomes in green. Keys to the genomes can be found in Appendix A.

(Table 3.1). In previous tests it was found that this fraction of folds was much lower for SF than for PB (Winstanley et al. (2005)), it appears that in the more recent version of the SUPERFAMILY database (1.69) the fold occurrence patterns are less dispersed than in the previous version (1.65).

Our method can be used to predict the number of folds or superfamilies present in LUCA (Table 3.1), as previous methods have done for sequence families (Snel et al., 2002; Mirkin et al., 2003). Between 33 and 47% of SCOP's folds are estimated to be ancestral to all kingdoms of life, CATH's topologies show slightly lower fractions. Between 42 and 68% of SCOP's folds are thought to be present at the highest node in our species tree. Fractions are slightly higher in the latter case due to the trifurcation between the three kindoms in the top node. A 'gain' at this node effectively indicates that the fold is present in at least two of the superkingdoms.

3.3.2 Parsimony age

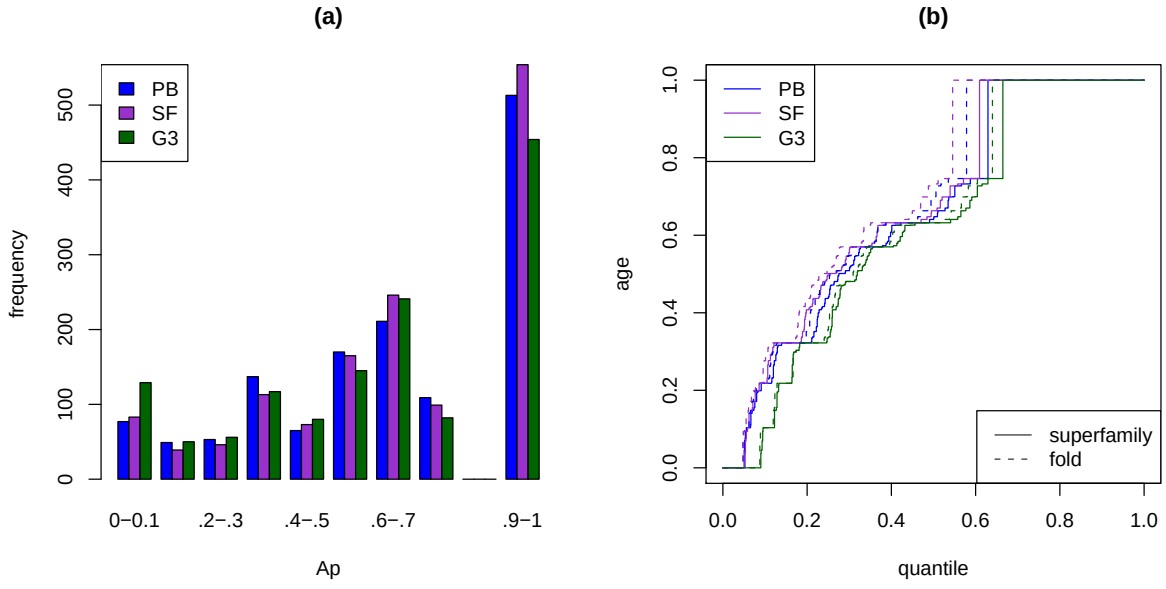
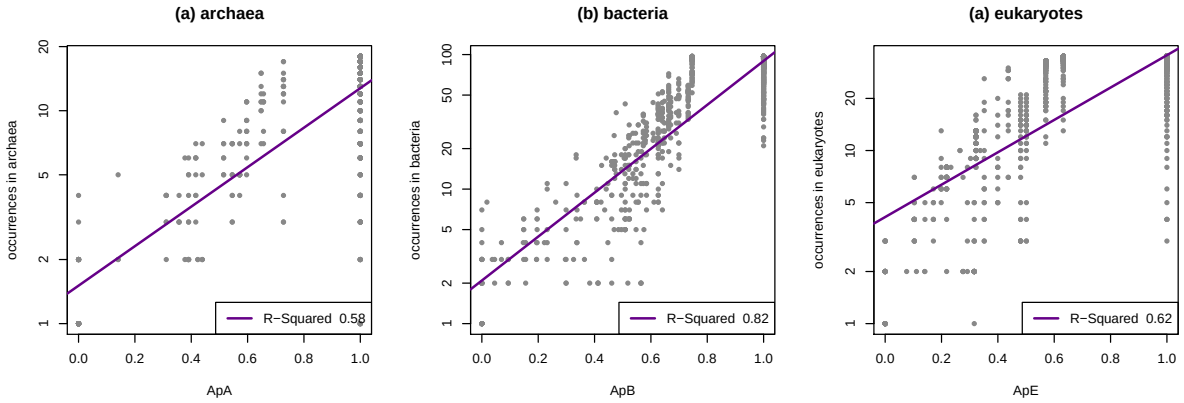


Figure 3.2: Data sources and fold age
(a) Parsimony age A_p distributions from the three different assignment sets. (b) Parsimony age A_p against fold quantile (fraction of folds smaller than the given age, see Section 5.2.5 for more details on the plotting method used). Lines represent the different data sets, with continuous lines showing superfamilies (or homologies) and dashed lines showing folds (or topologies).

Figure 3.1 gives an example of an individual superfamily occurrence pattern, together



ht

Figure 3.3: Superfamily age versus superfamily occurrences
Correlations between kingdom specific parsimony age (A_pK) and the number of genomic occurrence in archaea (a), bacteria (b) and eukaryotes (c) are shown using assignments obtained by PSI-BLAST.

with the assigned gain and loss events. It can be observed that despite a few deletions or false negatives on eukaryotic branches, the superfamily is still estimated to be present in the last common ancestor of all eukaryotes. Despite a few sporadic occurrences on bacterial genomes, the superfamily is not estimated to be present in the last common ancestor of the bacteria (note the loss event). These bacterial occurrences are likely to be due to a either HGT or false positive assignments.

Despite the fact that, the parsimony principle is a generalised approximation of fold evolution and is not necessarily adhered to on an individual basis, Figure 3.2 shows good agreement for the overall Parsimony age (A_p) distributions based on different assignment sets. Different tree building algorithms also show very similar results (Winstanley et al., 2005).

Fold ages, like genomic occurrences, do not follow a normal distribution (Figure 3.3 (a)). In fact, Figure 3.3 shows that the number of occurrences and fold ages are strongly correlated. The correlations are exponential as would be expected for a correlation between the height of a node in a tree, and the number of underlying leaves. The correlation is strongest for Bacteria, probably due to the wider spectrum of genomic sequences for this kingdom.

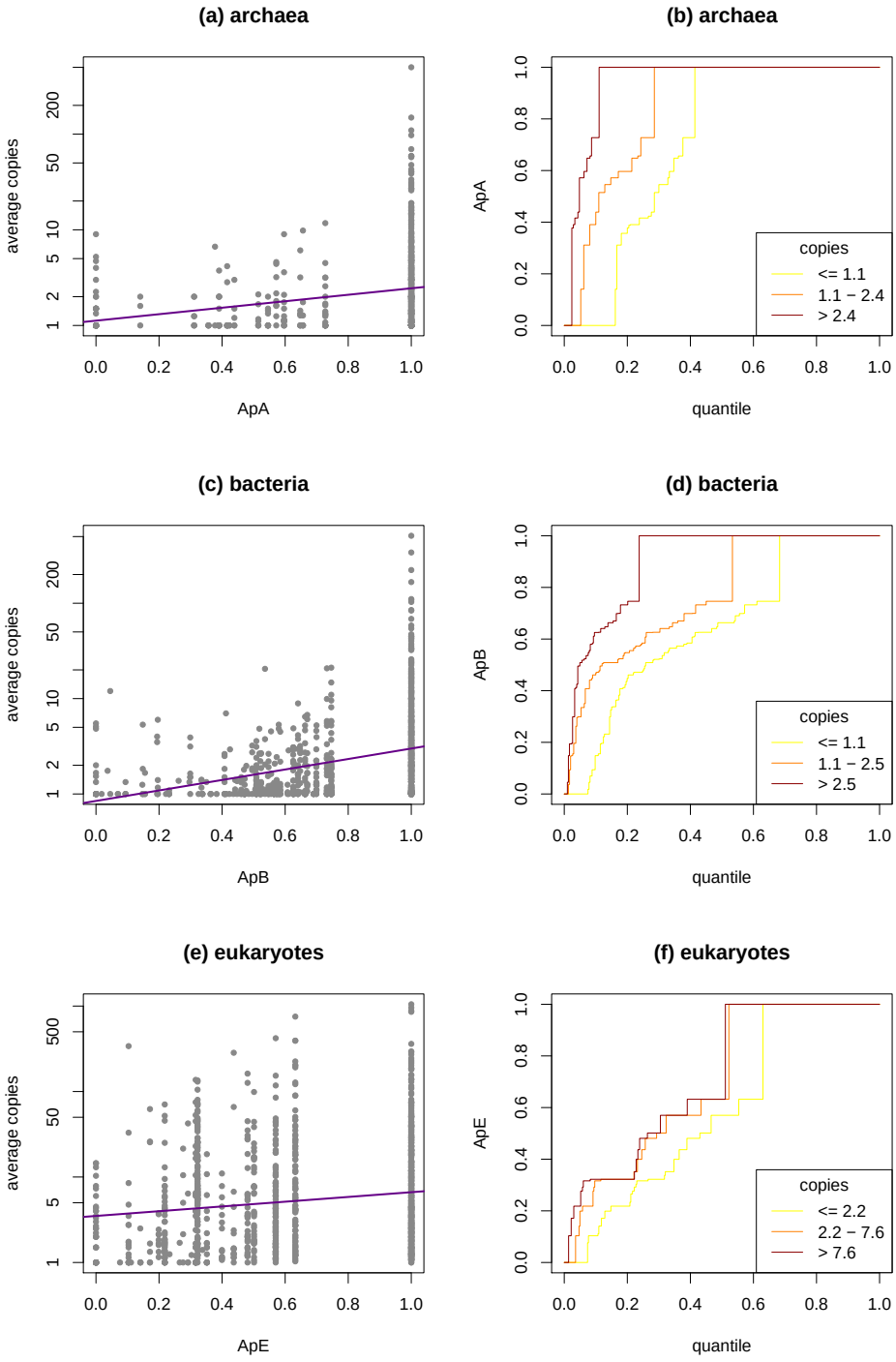


Figure 3.4: Copies and superfamily age
(a), (c), (e): Parsimony age A_pK versus average number of copies on genomes in specific kingdom. (b), (d), (f): Parsimony age A_pK against superfamily quantile; age distributions are split on ranges for the average number of copies in a kingdom . PSI-BLAST assignments are used.

3.3.3 Copies and fold age

It has been shown several times that the number of copies of a fold on a single genome follows an approximate power-law distribution (Qian et al., 2001; Cherkasov and Jones,

2004), with most folds having few copies on a genome and a few folds having many copies. Various authors have suggested evolutionary models at the level of whole genes (Qian et al., 2001; Karev et al., 2002) that lead to such distributions. In particular, the model of Qian et al. (2001) suggests that folds which arose earlier in evolution are able to have more copies on a genome.

Figure 3.4 (a),(c),(e) show restrictive relations between A_p and the copies per genome for the three superkingdoms. In general folds with a low A_p only have a small number of copies per genome, on the other hand folds with A_p close to 1.0 can, but do not necessarily, have many copies per genome. Nevertheless, folds with many copies show a significant shift in age distribution towards older folds (Figure 3.4 (b),(d),(f)).

As expected, Figure 3.4 (e) shows a much higher number of copies in eukaryotes than for the other superkingdoms. This is probably caused by the apparent relative lack of selective pressure on genome size in eukaryotes. The relatively high number of copies for $A_p = 0$ in Eukaryotes might be explained by lineage specific processes. It is evident that the process of duplication is also dependent on genome lineage-specific issues and the functional characteristics of a fold, giving rise to differential selective pressures (Ranea et al., 2004).

3.3.4 Protein-protein interactions and fold age

It has been suggested that the number of protein-protein interactions may correlate well with age, on the basis of evolutionary models of interaction networks (Barabási and Oltvai, 2004). The scale free topology of interaction networks is thought to have arisen from a process of preferential attachment: proteins with many interacting partners are more likely to gain new interactions. In such a model, proteins with more interaction are likely to be older. Eisenberg and Levanon (2003) show support for such a model with biological evidence from cross-genome comparisons of interacting proteins.

This implies that the age of a fold should be correlated with the maximum number of interactions of any member of the fold. Yeast (*S.cerevisiae*) and *E.coli* genome interaction

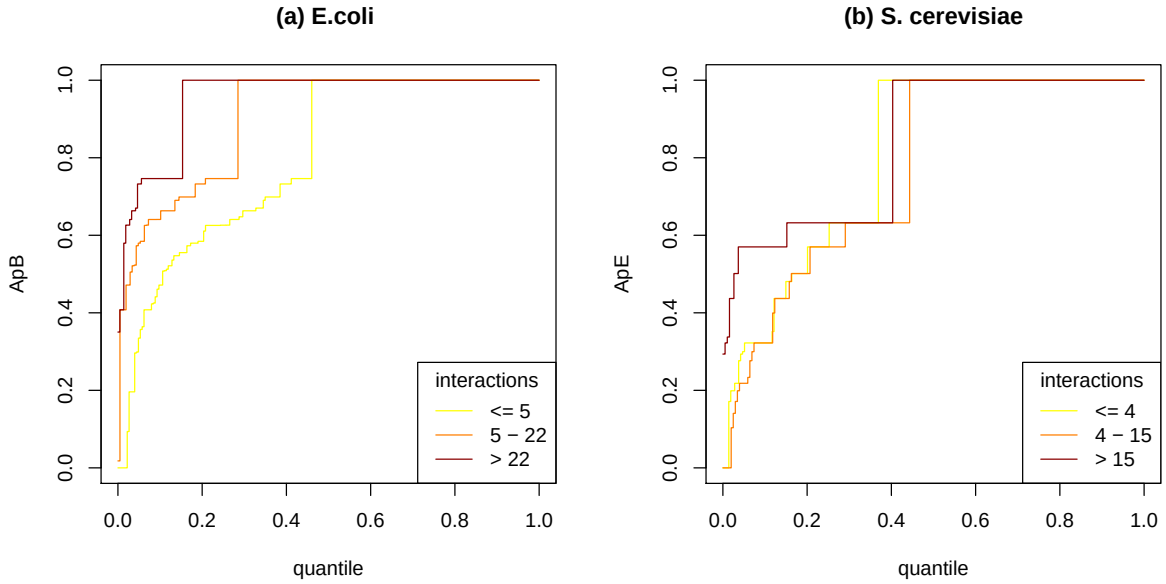


Figure 3.5: Protein-protein interactions and fold age
Parsimony age A_p against fold quantile (fraction of folds smaller than the given age) split on different number of protein-protein interactions on *E.coli* (a) and *S.cerevisiae* (b).

data were taken from the DIP database (Xenarios et al., 2002). Genome assignments using SF were used to predict the superfamilies of the interacting proteins.

Figure 3.5 shows a clear correlation between A_{pB} and protein-protein interactions on *E.coli*: folds that have high interaction numbers tend to be old (but old folds do not necessarily have many interactions). Principally, folds associated with specific essential functions may be ancestral without being highly connected in the interaction network. Such folds are seen with low interaction numbers and age $A_{pK} = 1$.

When considering protein-protein interactions within yeast, the results are less clear, there is still a small shift in age distributions between fold with many and few interactions. Other protein-protein interaction sources such as *H.sapiens*, *H.pylori*, *M.musculus* and *D.melangogaster* do not show significant shifts in age distributions either way; all these interaction dataset are considerably smaller than those for *E.coli* and *S.cerevisiae*.

3.3.5 Fold age and protein function

Figure 3.6 shows the effect of superfamily age on major functional categories (Vogel et al., 2004; Wilson et al., 2007). Superfamilies with functions involved in information and

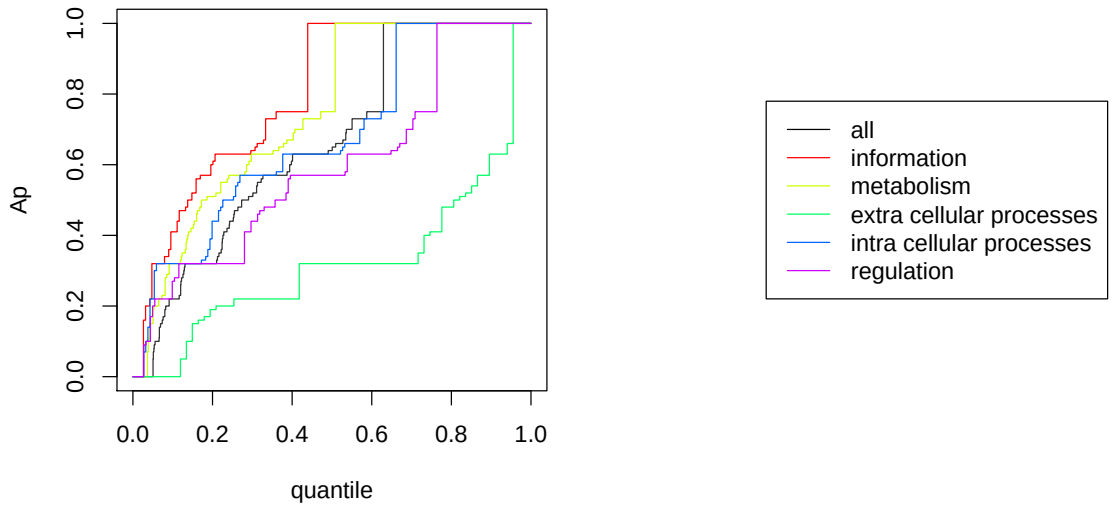


Figure 3.6: Protein function and superfamily age
Parsimony age A_p against superfamily quantile (fraction of superfamilies smaller than the given age) split on superfamilies in different functional categories. Lines represent the different functional categories: information, metabolism, extra cellular processes, inter cellular processes and regulation. Functional assignments were taken from `http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/function.html` (Vogel et al., 2004; Wilson et al., 2007)

metabolism tend to have older ages, whereas superfamilies with functions involved in regulation and extra-cellular processes tend to be younger.

The latter two categories, are likely to be involved in lineage specific expansion on eukaryotes. Vogel and Chothia (2006) investigate correlations between superfamily size expansion (the number of copies of a superfamily on a genome) with the complexity of the eukaryotic species (based on the number of different cell types), and show that superfamilies involved in extracellular processes and regulation, make up close to one-half of the superfamilies for which size expansion is highly correlated to the eukaryotic complexity.

Apart from the extra-cellular category, which is likely to be strongly lineage specific to multi-cellular organisms, these functional categories do not discriminate fold ages as well as structural classes appear to (see below). However, if these categories are split up into smaller functional categories, stronger variation may appear; e.g. GO's functional group (Harris et al., 2004) 'translation' is more strongly associated with older superfamily ages

than the functional category, information, it belongs to (agreeing with results found by Ranea et al. (2006); Marsden et al. (2006)).

3.3.6 Fold age and structural classes

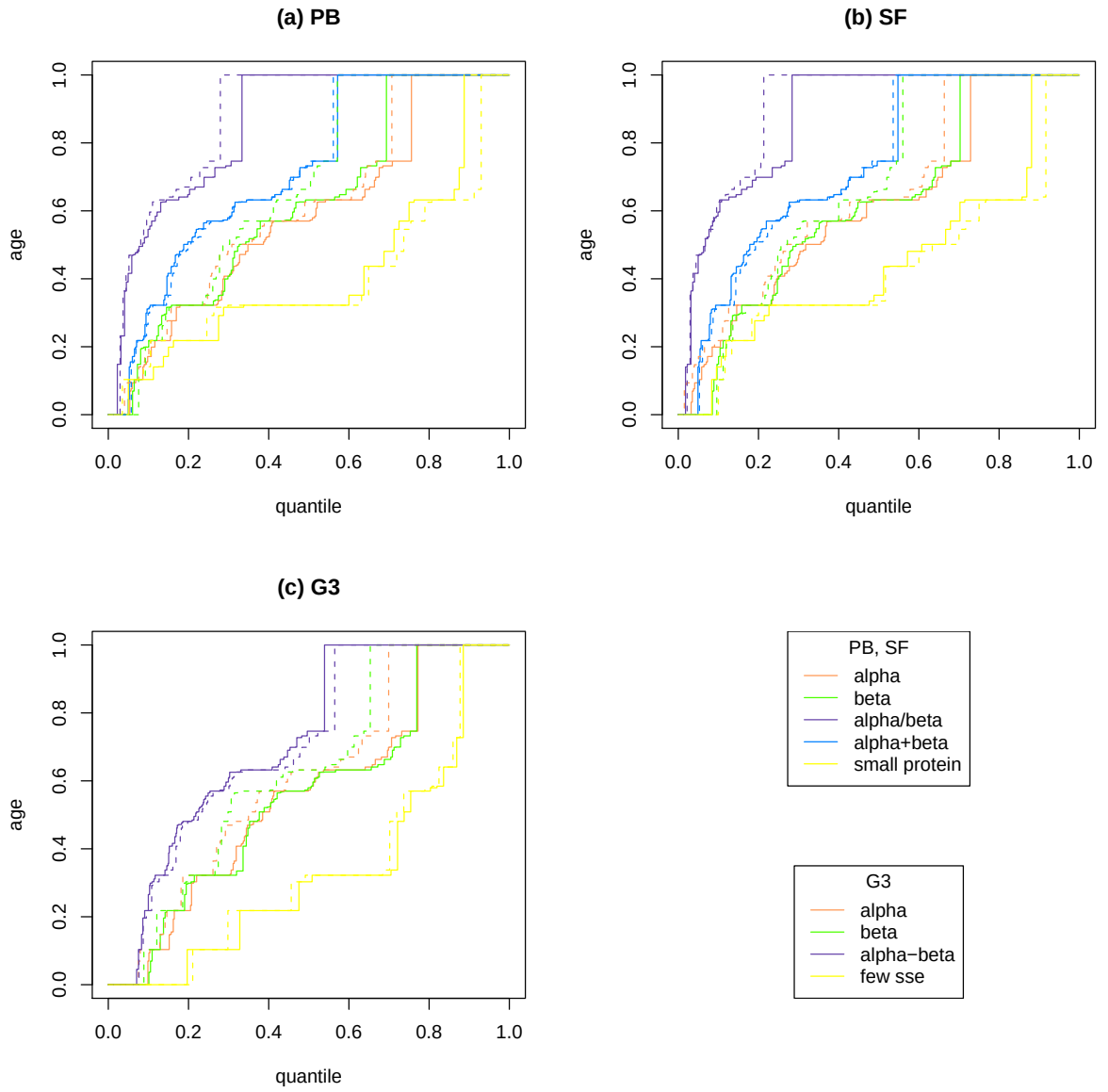


Figure 3.7: Structural classes and fold age
Parsimony age A_p against fold quantile (fraction of folds smaller than the given age) split in folds belonging to different structural classes for PSI-BLAST assignments (a), SUPERFAMILY assignments (b) and Gene3D assignments (c). Lines represent the different fold classes, with continuous lines showing superfamilies (or homologies) and dashed lines showing folds (or topologies).

Fold ages for different structural classes, as defined by SCOP, follow significantly different distributions and are shown in Figure 3.7. Folds identified as ancestral constitute

a major fraction ($> 25\%$) in all fold classes except small proteins for SCOP and folds with few secondary structure elements in CATH. SCOP's alpha/beta folds show the oldest age distribution, with 75% of folds observed in this class estimated to be of ancestral origin. This agrees with other indicators of extreme age for this class based on fold copy numbers in bacteria (Chapter 2) and high occurrence in all three superkingdoms (data not shown).

In general, fold and superfamily age distributions are similar for each individual class, with the exception of all-beta folds. This similarity, however, may be explained through the abundance of single superfamily folds. The beta class shows significantly older fold than superfamily ages, in both SCOP and CATH. A discrepancy between the all-beta class and other structural classes was previously found in the distribution of superfamilies per fold (Section 2.3.4). Both results may indicate a bias in the process of gathering different all-beta superfamilies into a fold, in evolutionary or classificational terms.

3.3.7 Single superfamily folds

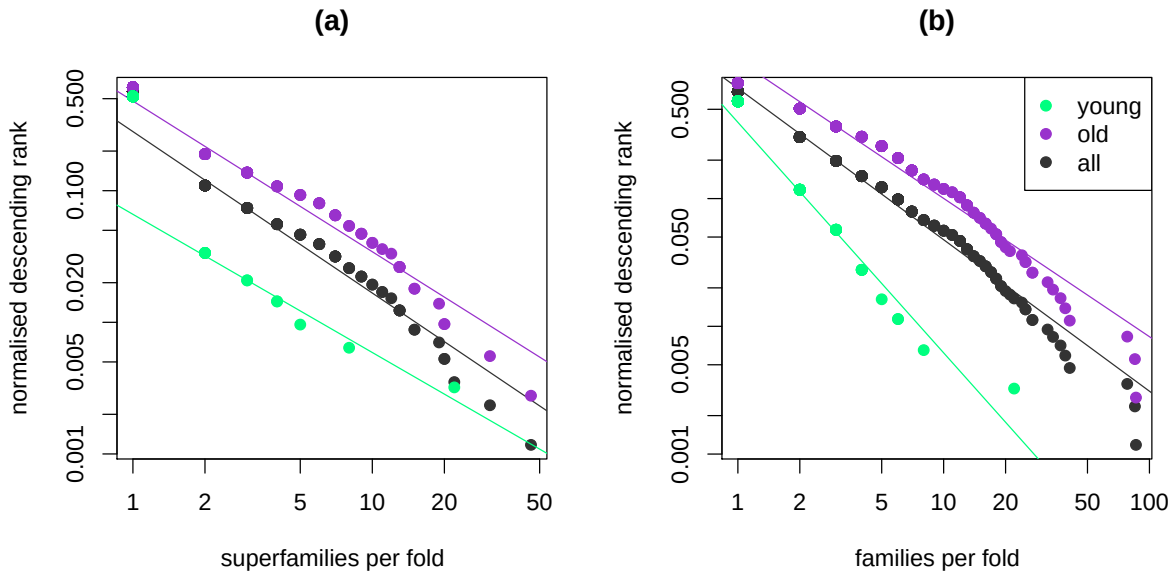


Figure 3.8: Families per fold and age
Distribution of superfamilies per fold (a), and families per fold (b) for PSI-BLAST assignments at different ages. Here young is defined as $A_p < 0.6$ and old as $A_p == 1.0$. Linear regression was performed excluding the lowest bin.

Previously we showed that the number of folds containing a single superfamily is

higher than expected, as compared to distributions for families per fold and families per superfamily (Section 2.3.4) and a restrictive relation between genomic occurrences and superfamilies per fold (Section 2.3.6). Figure 3.8 shows that the number of superfamilies per fold and the number of families per fold become lower for younger folds. As discussed previously, this may be due to a grouping effect: The larger the group size the higher the chance that there exists a superfamily in the group which has a high number of occurrences and hence a higher estimated fold age.

However, Figure 3.8 (a) also shows that the over-representation of single superfamily folds is not present for folds estimated with an old age. In fact, the over-representation becomes stronger for young folds. This discrimination between old and young folds is not seen in the number of families per fold (Figure 3.8 (b)).

The over-representation of single superfamily folds may be due to a bias in SCOP. If relatively more evolutionary relations between sequence families were spotted these folds than in other folds, it would create a single superfamily with a high number of sequence families. Here we assume that many other folds have evolutionarily related sequence families, which could be but are not (yet) classified in the same superfamily, i.e. divergent evolution. In this scenario, the over-representation of single superfamily folds with young ages may be explained through more easily found evolutionary relations between recent families.

3.4 Conclusions

The increasing number of completed genomes provides a rich source of new data for investigating protein structure evolution through their occurrence patterns across multiple genomes. The coupling of such analysis with phylogenetic information (the relationship between the genomes) potentially allows a far better picture of the relevant evolutionary processes to be developed.

Whole-genome trees were constructed using the full fold and superfamily domain assignment data for a large set of genomes. A parsimonious reconstruction of evolutionary

scenarios for each fold yields relative fold age estimates that are surprisingly robust to tree topology variation. No such relative age estimate for protein folds previously existed in the literature to our knowledge.

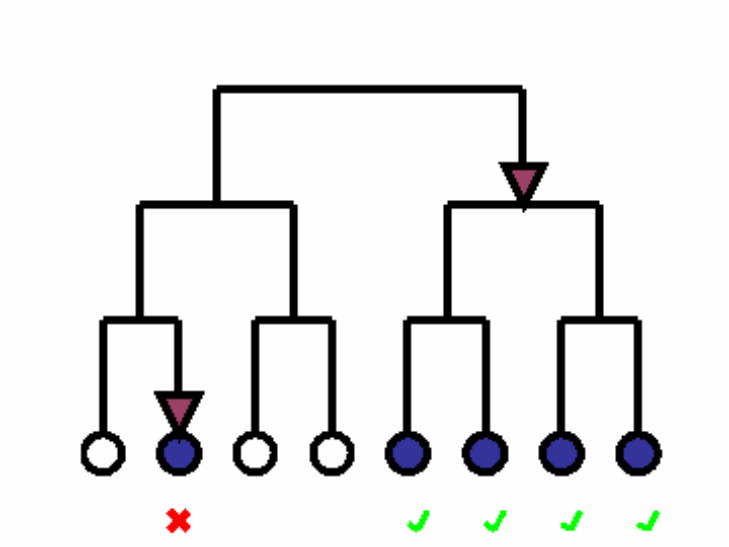
The age measure presented appears to be a more discriminating indicator of fold age than any of the suggested alternatives: simple fold genomic occurrence, genomic abundance, maximum number of protein interactions, or the number of superfamilies under a fold. Results indicate that alpha/beta folds are relatively older than other fold classes, and the class of small proteins is relatively younger. This agrees with non-phylogenetic analyses of fold genomic abundance carried out on the same data.

This kind of analysis also shows that the SCOP and CATH classifications are similar with respect to the major evolutionary trends in fold and superfamily evolution, with CATH showing marginally younger fold ages than SCOP. In the near future, incorporation of data for additional genomes may improve the resolution and reliability of the inferred phylogenies. Additional structural information for folds and superfamilies, and more powerful fold recognition techniques may improve the fold age estimates.

An interactive application of our age estimating method is available at `http://www.stats.ox.ac.uk/~abeln/howold/` for SUPERFAMILY and PSI-BLAST assignments of SCOP version 1.65.

Chapter 4

Improving genome wide fold-recognition



4.1 Introduction

The gap between the number of known protein sequences and structures continues to widen, particularly as a result of sequencing projects for entire genomes. Recently there have been many attempts to generate structural assignments to all genes on sets of completed genomes using fold recognition methods. Here we develop a method that detects false positives made by these genome wide structural assignment experiments by identifying isolated occurrences.

Fold recognition has been a major force behind improvement in structure prediction over recent years (Moult, 2005). In this work sequence-profile based old recognition methods are used, for a more detailed review on such methods see Section 1.4.

Recently many studies have used fold recognition to look at the structural content of entire genomes with aid of PSI-BLAST (Abeln and Deane, 2005), HMMs (Gough and Chothia, 2002; Lee et al., 2005) or threading procedures (Cherkasov and Jones, 2004). These fold recognition assignments can produce an occurrence pattern on a set of species for a given family, superfamily or fold as defined by structural classifications such as SCOP (Andreeva et al., 2004) or CATH (Orengo et al., 1997). Sets of such occurrence patterns have proved to be useful for building a phylogeny of species (Qian et al., 2001; Yang et al., 2005), for grouping proteins within a similar pathway (Marcotte et al., 1999) and for estimating the ages of folds (Winstanley et al., 2005).

A major challenge for all fold recognition techniques is to discriminate a true homologue from a false positive (specificity) using confidence scores such as e-values. E-values (expectation-values) indicate how likely it is that an alignment with the search sequence would occur by chance in a given database, i.e. they should reflect the odds of a false positive assignment.

Previous studies have suggested that analysis of structural assignments on completed genomes may indicate false positives of fold recognition techniques. Yang et al. (2005) showed that the number of hits on completed genomes drastically increased above a certain e-value cut-off, which could be explained by a sudden influx of false positives.

Furthermore we showed in Chapter 3 that occurrence patterns of most superfamilies can explain the phylogeny of the genomes reasonably well. However, a difference in fitness between patterns from two different fold recognition techniques was observed in previous data sets. In particular it appeared that, in one set, more assignments were made which occurred on isolated leaves of the species tree (Winstanley et al., 2005).

We propose that false positives in fold recognition assignments might be identified by considering a phylogeny of species. False assignments in such a set might be expected to occur randomly across the genome tree, whereas true positive assignments to a superfamily should be evolutionary-related. Hence we expect that false occurrences have a stronger tendency to be scattered across the tree than true assignments. This study investigates if isolated occurrence within a phylogenetic occurrence pattern are indeed more likely to be false positives.

Using phylogeny to improve homology searches is not a new idea. It has been known for a long time that by considering the phylogeny of related proteins one can improve sequence alignments. An example is progressive multiple sequence alignment, where an approximate phylogeny of the sequences is used to aid the alignment of multiple sequences (Feng and Doolittle, 1987).

Assignment of function can also be facilitated by phylogenomics: a set of *known* homologs is used to create a phylogeny of proteins in which speciation and duplication events are marked. These can be used to subclassify the proteins in the phylogeny into specific functions. Several protocols as well as automated procedures based on phylogenomics have been able to improve functional annotation (Eisen, 1998; Sjölander, 2004; Engelhardt et al., 2005).

Recently it has become clear that confidence in a modelled structure increases, when homologues of the target sequence give a similar structure prediction (Venclovas and Margelevicius, 2005; Bradley et al., 2005). Here precalculated phylogenies of entire genomes are used rather than phylogenies of individual proteins.

To asses the assignments, occurrence patterns are obtained for every superfamily from two fold recognition sets (PSI-BLAST and SUPERFAMILY). We developed a method

that identifies isolated occurrences within such patterns, by considering if an occurrence causes a gain at leaf level in the phylogeny.

This study demonstrates that false positives in fold recognition assignments can indeed be identified by considering a phylogeny of species: isolated occurrences are shown to have higher e-values (are less reliable) than other occurrences. We formulated criteria to predict false positives based upon these results. The set of predicted false positives were validated by a comparison to overlapping PSI-BLAST assignments and to assignments that changed between different versions of the SUPERFAMILY database. Both tests confirmed that the predicted false positives are far more likely to be falsely assigned than other occurrences.

Analysis of occurrence patterns from genome wide fold recognition also provides a new way to examine the quality of fold assignments. We show that the frequency of occurrences drastically increases for high e-values ($> 10^{-8}$ for SUPERFAMILY and $> 10^{-4}$ for PSI-BLAST) and that this influx is likely to be caused by false positive assignments. In addition, the accuracy of fold recognition is demonstrated to differ significantly for the different structural classes as defined by SCOP.

In principle this technique can screen assignments of any existing fold recognition method for false positives. An extended version of the method is given, which can be applied to assignment sets with a high proportion of false positives. This version might be able to improve the search capacity of any genome wide fold recognition technique.

4.2 Methods

SUPERFAMILY assignments

The first set of assignments is taken from the SUPERFAMILY database (Gough et al., 2001) version 1.65 and covers 1269 different superfamilies as defined by SCOP (Andreeva et al., 2004) on 150 completely sequenced genomes. Around 750,000 structural assignments are made, with the following restrictions imposed: (1) all assignments have an e-value lower than 10^{-4} and (2) no other assignment on the same region of the gene is

made with a lower e-value. Appendix A shows a table with the genomes used, note that the coverage may be slightly different since an earlier version of SCOP (1.65) is used in this Chapter.

4.2.1 PSI-BLAST assignments

The second set consists of assignments obtained by PSI-BLAST (Altschul et al., 1997) searches on the same 150 genomes. Sequences with less than 95% sequence identity from the ASTRAL (Brenner et al., 2000) database were used to search for structural domains in a non-redundant database created from all genes in the 150-genome-set. PSI-BLAST was used with a SEG filter and an e-value cut-off for inclusion in the PSSM of 10^{-5}. Assignments with an e-value smaller than 1.0 were included after the final run. The e-values for the assignments were taken from the PSI-BLAST run in which the assignment first falls below 10^{-5}, i.e. when it is not yet included for scoring in the PSSM. Note that there was no check for overlap of assignments within a gene. In contrast with SUPERFAMILY, a repeat of a domain, from the same superfamily within a gene, was counted as a single copy.

4.2.2 Phylogenies

Phylogenies of the genomes were created using the SUPERFAMILY occurrence data (the number of copies were not included). A neighbour-joining algorithm was used to create a tree. The branch lengths were then normalised, so that all leaves are at an equal distance from the root (1.0) following the method used by Winstanley et al. (2005) and described previously in Chapter 3.

4.2.3 Occurrences

An occurrence in this study is defined as the occurrence of a superfamily on a genome. An occurrence can therefore be caused by more than one assignment. The e-value of an occurrence is defined as the lowest e-value within that set of assignments.

4.2.4 Isolated occurrences - gains at leaf level

A parsimony algorithm is used to find isolated occurrences in a pattern given the phylogeny. This minimises the number of loss and gain events for a superfamily in the species tree (Snel et al., 2002; Mirkin et al., 2003; Winstanley et al., 2005), a detailed description of the algorithm can be found in (Mirkin et al., 2003). Isolated occurrences are identified as occurrences that create a gain at leaf level. The algorithm used a gain penalty which was twice the size of the loss penalty in order to take a high number of false negatives into account. Experimentation with a lower relative gain penalty showed only a small increase in isolated occurrences. This indicates the technique is relatively robust against false negatives. The parsimony algorithm was implemented in Java (J2SE 5.0).

4.2.5 Potential gains at leaf level

The relation between saturation of occurrences in an occurrence pattern and the number of gains at leaf level, that can potentially be detected by our method, is investigated. A 'potential gain at leaf level' is defined as a genome without an occurrence for a given superfamily that would cause a gain at leaf level if an occurrence was added to the existing pattern. To calculate the number of potential gains at leaf level, the parsimony algorithm is run for every genome without an occurrence in the pattern.

4.2.6 Identifying isolated occurrences using a base pattern

An additional method was developed to recognise false positives in an assignment set containing many unreliable predictions. A base pattern is created from assignments with an e-value below a given threshold. Subsequently the parsimony algorithm is run on the base pattern and potential gains at leaf level are predicted. A set of isolated occurrences can be found as the union of (a) the set of occurrences within the base pattern which cause a gain at leaf level and (b) all occurrences caused by assignments above the e-value threshold, with a potential gain at leaf level.

4.2.7 Distance to source

The distance to source for an occurrence is the age of the youngest common ancestor between the occurrence and any genome containing the source domain. A genome is said to contain a source sequence for a SCOP superfamily if it covers 80% of its length and has at least 95% sequence identity. Simple BLAST searches were used to identify the source sequences. No source sequences could be identified for a few superfamilies on our set of genomes.

4.2.8 Cluster distance

The cluster distance reflects how 'far away' an occurrence is from any other occurrence within a pattern, given the phylogeny. The cluster distance is calculated as the sum of distances to every occurrence in the tree divided by the distance to every leaf in the tree. This score is subsequently divided by the average score for each occurrence in the pattern, so that the average cluster distance of a pattern becomes 1.0. A cluster score of 1.0 indicates average clustering, a score < 1.0 indicates tighter clustering. Mediation by the average distance to each leaf is used, since some leaves in the tree lie in tighter clusters and would generally produce lower scores without mediation.

4.2.9 Consensus

A consensus occurrence is an occurrence which is identified by both SUPERFAMILY and PSI-BLAST. Note that the occurrence does not necessarily have to be caused by assignments to the same gene.

4.2.10 Significance test for structural classes

A simple sampling procedure was used to determine if the proportion of false positives for a structural class deviates significantly from the overall false positive rate in Table 4.2. The ratio of predicted false positives was calculated for 500 random samples of the occurrence patterns. For each sample the number of random genome entries was chosen to

match the class size. The resulting distribution of sampled false positives rates was used to determine if the rate for each structural class was significantly lower, falling within the lowest 1% of the sampled distribution or significanlty higher, falling within highest 1%. The number of genome entries for a class depends on the number of superfamilies and ranges from almost 6,000 entries (multi-domain) to just over 50,000 entries (alpha+beta).

4.2.11 PSI-BLAST overlapping region

Overlapping regions of assignments are identified as assignments with a stronger assignment to a different superfamily on the same gene region and with at least 50% of the weaker assignment covered by the stronger assignment. Overlapping occurrences are occurrences which would be taken out of the data set if only the strongest assignment within a region was retained.

4.3 Results

4.3.1 Assignments

	SUPERFAMILY		PSI-BLAST	
	total	fraction of (a)	total	fraction of (a)
(a) assignments	758,437	100.00 %	488,273	100.00 %
(b) occurrences	89,792	11.84 %	80,756	16.53 %
(c) one copy	34,659	4.57 %	32,077	6.77 %
(d) gain at leaf level	1,908	0.25 %	1,488	0.30 %
(e) predicted false positive	1,157	0.15 %	1,023	0.21 %

Table 4.1: The fraction of assignments that can be analysed through occurrence patterns
Rows show for each data set: (a) the total number of assignments; (b) the number of genomes with an occurrence; (c) the number of genomes with only one assignment; (d) the number of isolated occurrences: occurrences causing a gain at leaf level; (e) predicted false positives: isolated occurrences caused by a single copy and with a high e-value ($> 10^{-10}$ for SUPERFAMILY and $> 10^{-4}$ for PSI-BLAST). The large difference between the fraction of occurrences between SUPERFAMILY and PSI-BLAST is probably due repeated assignments to the same gene, which are counted as multiple copies for SUPERFAMILY and as a single copy for PSI-BLAST.

All assignments were generated by searching for SCOP domains on the genes of 150 completed genomes (18 archaea, 97 bacteria and 35 eukaryotes, see Appendix A).

The first set of assignments is taken from the SUPERFAMILY database (Gough et al., 2001) and contains 1269 different superfamilies. The database provides around 750,000 structural assignments for our set of genomes (Table 4.1). One or more structural assignments is made to 48% of the gene sequences in this dataset, with a false positive rate estimated by the authors to be lower than 1% (Gough et al., 2001).

The second set consists of assignments obtained using PSI-BLAST (Altschul et al., 1997). It includes all assignments with an e-value lower than 1.0 for testing purposes. The inclusion threshold to build the PSSM was kept low, at 10^{-5}, to diminish the risk of including false positives within the search profile. Almost 490,000 structural assignments are made with PSI-BLAST (Table 4.1). As expected the coverage of genes is a little lower with one or more structural assignments made to around 36% of the genes.

4.3.2 Parsimony and gains at leaf level

Figure 4.1 shows the precalculated phylogeny (see Section 3.2.2) used in this work. A parsimony algorithm was used to predict an evolutionary scenario of loss and gain events for a superfamily given this phylogeny. The algorithm minimises the number of loss and gain events in the tree for the occurrence pattern of the superfamily.

Our measure for isolatedness indicates whether an occurrence causes a **gain at leaf level** in the set of minimal gains and losses. The parsimony algorithm predicts a gain at leaf level (gain at the lowest level of the tree), when it would be more expensive to cluster the occurrence with other occurrences. Hence an assignment with a gain at leaf level is an isolated occurrence within the pattern. Note that the parsimony algorithm makes the set of loss and gain events quite robust against small changes in the occurrence pattern. A few deletions within an occurrence cluster will usually not cause a gain at leaf level, thus diminishing the effect of deletions and false negatives.

Figure 4.1 shows the full assignment of gains and losses by our parsimony algorithm for superfamily a.126.1 (Serum albumin-like).

Figure 4.1: Parsimony algorithm example
The phylogeny, including all 150 genomes, showing the occurrence pattern for superfamily a.126.1 (Serum albumin-like) for SUPERFAMILY data. Losses (left pointing triangles) and gains (right pointing triangles) are shown as calculated by the parsimony algorithm. The two assignments with a gain at leaf level are isolated within the tree. For each occurrence the log10 e-value (*log(ev)*), the number of copies (*copy*), consensus with PSI-BLAST occurrence (*cons*) and the cluster-distance (*clust*) is shown. The *src* column shows which genomes contain a source sequence for the superfamily by indicating the sequence identity to a SCOP domain.

4.3.3 Detectable false positives

The technique developed here can only observe false positive **occurrences**. However, the same superfamily can be assigned to many different genes on a genome and so one occurrence might cover many assignments. As the number of assignments (or copies) for a superfamily on a genome is known to behave like a power law (Qian et al., 2001; Abeln and Deane, 2005), the number of genomic occurrences is much smaller than the number of assignments. In this technique, the e-value for an occurrence is determined by the lowest e-value in the set of copies on the genome covering the occurrence. In effect, if a superfamily has many copies on a genome, one of the assignments usually has a very low e-value, which will overshadow those with higher e-values. This effect obviously reduces the number of assignments which can be screened by our technique (Table 4.1).

In practise we will restrict the set of predicted false positives to isolated occurrences caused by a single copy with a high e-value (see below).

4.3.4 E-value distributions

In order to understand the e-value distribution of isolated occurrences, we first consider the e-value distribution of all assignments and the e-value distribution of all occurrences.

Figure 4.2 (a) shows the distribution of e-values for all fold assignments. A local minimum is seen at the higher e-value end of the distribution for both the SUPERFAMILY and PSI-BLAST distributions. This may mark the point where false positives assignments begin to play an important role.

Figure 4.2 (b) shows the e-value distributions for all occurrences. This distribution is shifted to the left, with respect to the assignment distribution. The e-value for an occurrence is taken from the assignment on a genome with the lowest e-value, hence assignments with higher e-values from the same genome are not represented in this set.

The left shift is, as expected, not seen for occurrences with a single assignment on a genome (dotted lines 4.2 (b)), since the effect described above cannot occur for occurrences caused by only one assignment. In fact, the 1-copy distribution appears to be very similar

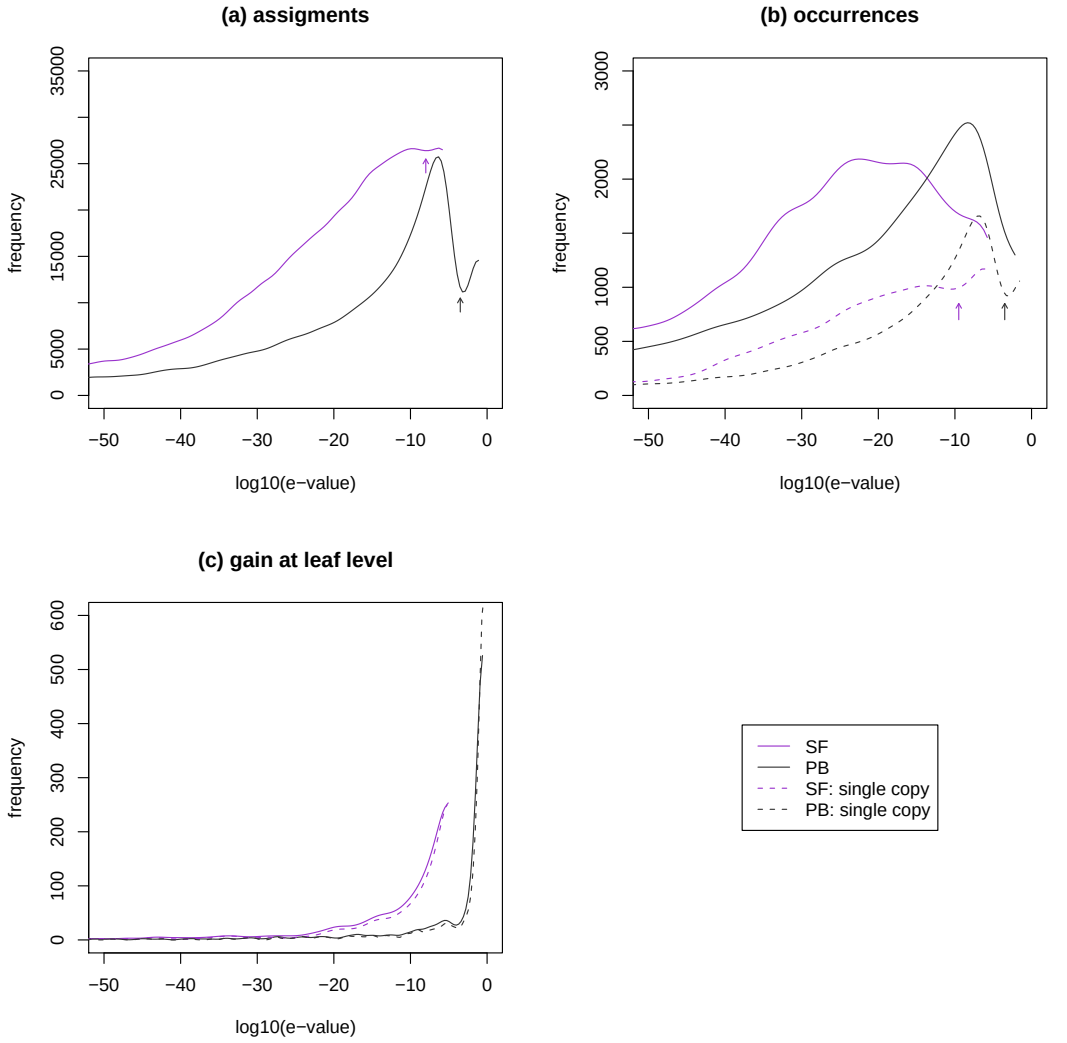


Figure 4.2: E-value distributions
E-value distributions are shown for SUPERFAMILY (SF) and PSI-BLAST (PB) assignments and occurrences. (a) E-value distribution of all assignments. (b) E-value distribution of occurrences. The e-value of an occurrence is defined as the lowest e-value of all assignments to a superfamily on a genome. (c) E-value distribution of occurrences, which cause a gain at leaf level, i.e. predicted false positives. Arrows indicate local minima in the distribution, which might indicate the point where false positive assignments become more dominant.

to the overall assignment distribution.

The fraction of false positive assignments should be very small; false positive occurrences would therefore generally be caused by a single false assignment on a genome. Hence, the fraction of false positives in the 1-copy distributions is expected to be higher than the fraction of false positives for all assignments. Indeed, it is observed that the local minima have become more prominent in these distributions.

4.3.5 False positives

We will now compare two independent sets of observations: 1) a set of binary values indicating if an occurrence causes a gain at leaf level; and 2) a set of continuous e-values for every occurrence (for discussion on independence see Section 4.3.7). E-values should indicate how likely it is that an occurrence is a false positive. We can therefore use the e-values to investigate our expectation that isolated occurrence are more likely to be generated by a false assignment.

Figure 4.2 (c) shows the e-value distribution for occurrences with a gain at leaf level. Comparing it to the other distributions, it is clear that this distribution is shifted to the right. This shift appears to confirm the supposition that occurrences with a gain at leaf level are less reliable. The steep increase in the distribution for the SUPERFAMILY data starts at a significantly lower e-value than for PSI-BLAST, possibly indicating that false positives in the SUPERFAMILY set start occurring at lower e-values. However, the e-values for the two sets of assignments are not necessarily comparable, since they are calculated using different estimation techniques (see Section 1.4).

4.3.6 Ranked distributions

The e-value distributions for the assignments on the genomes are clearly not normally distributed. In order to make comparison between different e-value distributions simpler, we have given a rank to the e-value of each assignment. This ranked e-value distribution of all assignments gives a uniform distribution, which would display as a flat horizontal line (not shown). Similarly, any random subset of this reference distribution is expected to be uniform.

Figures 4.3 (a) and (b) show that the left-hand side of the distribution of occurrences with a gain at leaf level is approximately uniform and hence can be interpreted as a random subset of the assignment distribution. However, for higher e-values a sudden increase is observed; this indicates that relatively more assignments with a high e-value are found in occurrences with a gain at leaf level than in the reference distribution. We

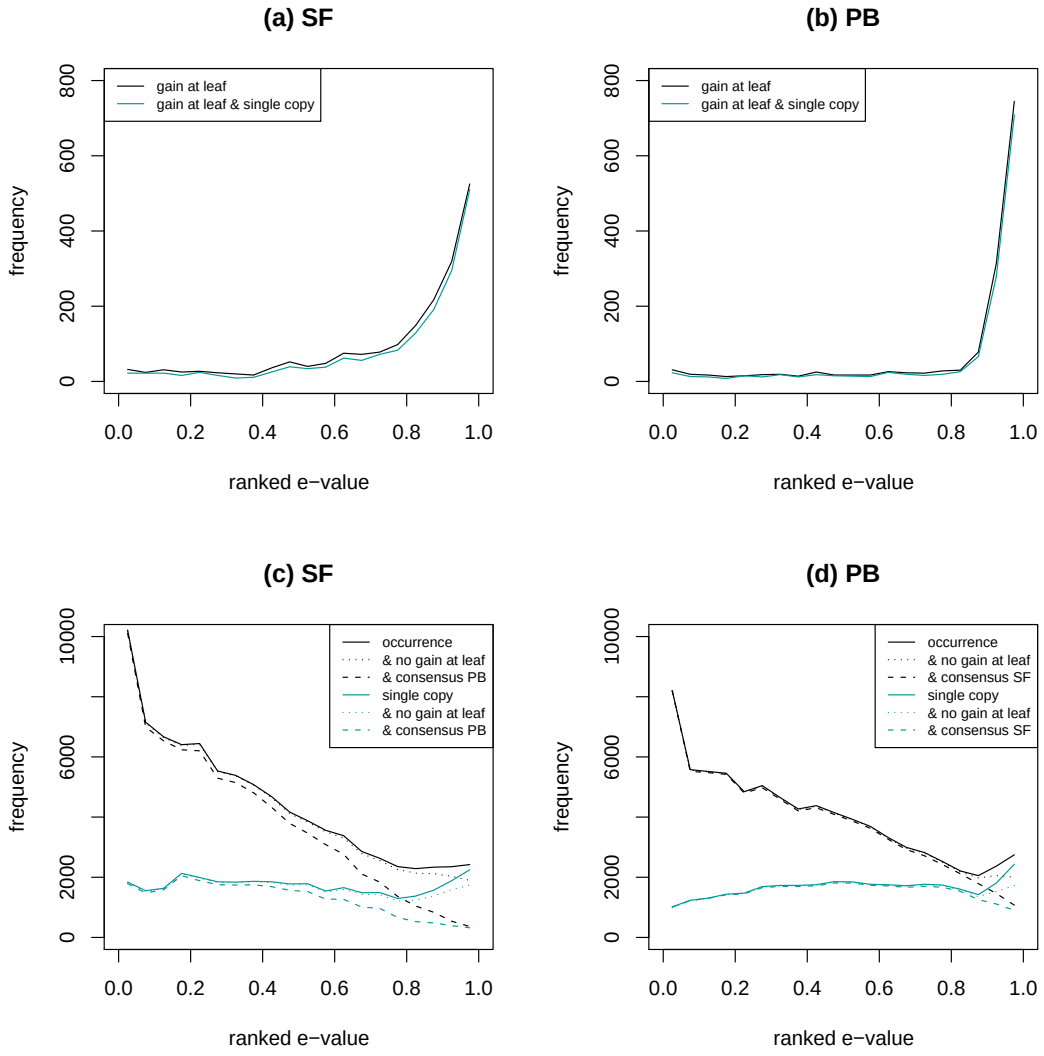


Figure 4.3: Ranked e-value distributions
Distribution are shown for SUPERFAMILY (a), (c) and PSI-BLAST (b), (d) occurrences. The ranked e-value is obtained by ranking the e-values in the set of all assignments (including multiple copies) and is normalised between 0 and 1. (a) and (b) show the ranked e-value distributions for occurrences with a gain at leaf level. The peak at the right hand side might indicate a sudden increase in false positives. (c) and (d) show the ranked e-value distribution for all occurrences. The dotted line indicates that occurrences with a gain at leaf level have been removed and the striped line shows only consensus occurrences, which have been assigned by both SUPERFAMILY and PSI-BLAST. Occurrences caused by a single assignment (one copy) are shown in colour. Both the distributions for all occurrences and for occurrences caused by a single assignment show an increase at the right-hand side, probably caused by false positives. This increase is not seen in the distribution of the consensus occurrences or in the distribution from which predicted false positives have been removed.

believe that this set of additional isolated occurrences at high e-values is caused by false positive assignments.

The ranked e-value distribution for the set of all occurrences (black line) in Figures 4.3 (c)and (d) is as expected shifted to the left with respect to the reference distribution;

the e-value of an occurrence is defined as the best e-value on the genome.

However, at higher e-values the distribution stops decreasing, and shows for PSI-BLAST even an increase. Moreover the distribution of consensus occurrences, which are predicted by both SUPERFAMILY and PSI-BLAST assignments, does not show a sudden increase at high e-values (striped line). The consensus of two or more fold recognition methods can significantly improve specificity over individual methods (Rychlewski et al., 2003), thus the observed increase is likely to be an artefact due to the sudden appearance of false positives for high e-values.

If the set of 'gains at leaf level' is taken away from this ranked occurrence distribution (dotted line), the distribution continues to decrease as expected. This shows that the set of isolated occurrences corresponds in size and e-value distribution to a set of likely false positive occurrences.

As described above, occurrences caused by only one assignment should have a similar distribution to that of all assignments and hence the ranked distribution should appear uniform. The coloured line in Figure 4.3 (C) shows that this is roughly the case for both the SUPERFAMILY and PSI-BLAST occurrences. However, both distributions show a sudden increase at high e-values. Once again, the distribution of consensus occurrences does not show such an increase (coloured striped line). The set of single occurrences from which isolated occurrence have been removed (coloured dotted line) shows less of an increase than the occurrence distribution.

These results all support that the observed change in distribution for high e-values is created by a sudden increase in false positives and that our set of isolated occurrences corresponds to these false positives. The results also demonstrate that occurrences with a gain at leaf level have in general worse e-values than other assignments.

4.3.7 Evolutionary caveats

We are using fold recognition methods to find members of a superfamily. Hence all assignments and occurrences within a pattern should be evolutionary-related. We propose

that an isolated occurrence on a species tree might indicate that that occurrence is a false positive. Below we consider what other mechanisms might cause a bad fit within an occurrence pattern for a superfamily.

Horizontal Gene Transfer

An isolated occurrence within the tree might be due to horizontal gene transfer between species. This is known to be an important process for the enrichment of diversity on genomes (Pál et al., 2005). However, we currently have no reason to believe that instances of HGT would give higher e-values for true superfamily assignments than their non-HGT-counterparts. We will in practise only consider isolated occurrences as a false positive when the e-value of the occurrence lies above a certain threshold. These thresholds should match values at which we observe a sudden increase in frequency for the ranked e-values (10^{-10} for SUPERFAMILY and 10^{-4} for PSI-BLAST).

Vice versa the extent of horizontal gene transfer might be estimated by the part of the distribution for occurrences with a gain at leaf level, which is independent of e-values. In the ranked e-value distributions this is the uniform section (Figure 4.3 (a),(b)).

Distance to source sequence

Another possible source of error in the technique is a dependency of e-values on the evolutionary distance between the search sequence and an assignment. The search sequence is the initial sequence from which a search profile (HMM/PSSM) is created. In this study, the search sequences are the 'source' domains taken from SCOP.

We have so far assumed independence of e-values and phylogenetic fitness of an occurrence. The above effect would not cause a direct dependence, but together with deletion of domains, false negatives and random noise in e-values, it could result in higher e-values close to the edges of an occurrence cluster. Two different tests were performed to see if our independence assumption would hold.

Firstly, we examined the relationship between e-values and the distance from an occurrence to its source sequence. A genome is said to contain a source sequence for a

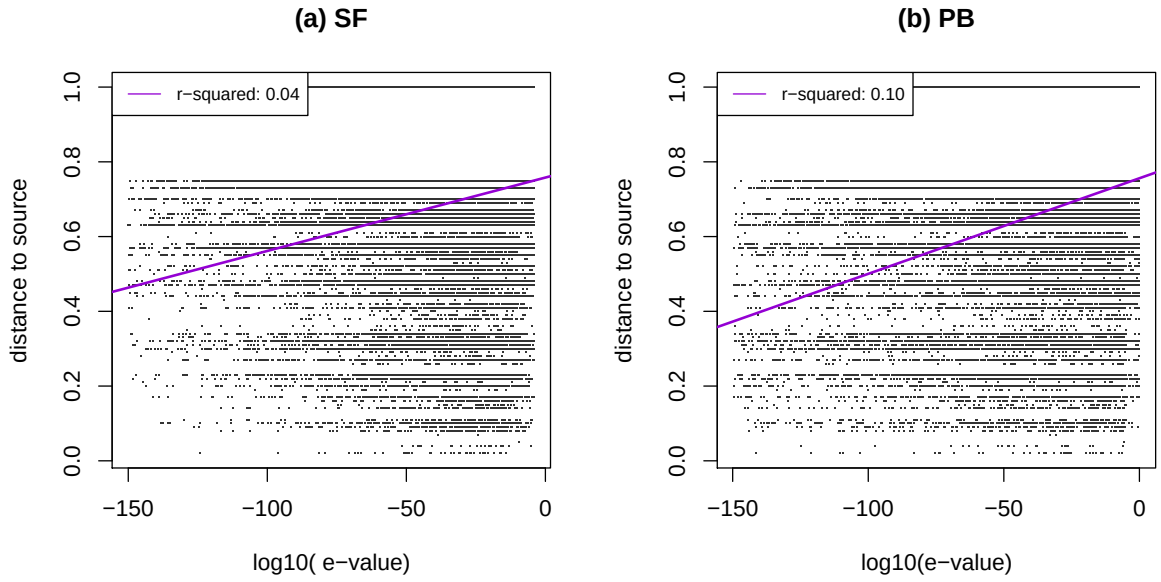


Figure 4.4: Distance to source
Correlation between e-values and the distance to a source sequence for both SUPERFAMILY (a) and PSI-BLAST (b). The distance to a source sequence for an occurrence is calculated as the distance to the youngest common ancestor between the occurrence and a genome containing a source sequence. The correlation shown is very weak (r-squared 0.05 for SUPERFAMILY and 0.15 for PSI-BLAST).

superfamily when a very close match to a sequence of a known 3-dimensional structure is found in its genes (see Section 4.2.7). The distance to a source was calculated as the height of the lowest common ancestor in the tree between the genome of the occurrence and any of the genomes containing a source sequence for the superfamily.

There exists a very weak correlation between the distance to a source sequence and the e-value of an occurrence, with r-squared values of 0.05 for SUPERFAMILY and 0.14 for PSI-BLAST (see Figure 4.4). The correlation becomes even weaker (r-squared 0.04 and 0.10) when the set of source occurrences is removed. Despite the correlations being very weak, they are significant (with p-values $< 2e^{-16}$) for both PSI-BLAST and SUPERFAMILY.

To assess the influence of this very weak correlation on our method, we examined whether e-values deteriorated towards the edges of a cluster. The cluster distance of an occurrence is calculated as the distance to every leaf in the occurrence pattern divided by the distance to every leaf in the tree. It is then mediated so that the cluster distance of an occurrence is relative to its pattern. Scores significantly higher than 1.0 reflect bad

clustering of an occurrence in the tree.

Figures 4.5 (a) and (b) show that there is a negligible correlation between the cluster distance and the e-value of an occurrence, and in general we can assume independence of these entities (see Figure 4.5 (c),(d)). This allows us to rule out the possibility that a general correlation between cluster distance and e-values causes the previously observed increase in isolated occurrence at very high e-values.

Saturation effects on patterns

The more occurrences there are in a pattern, the lower the chance that an isolated occurrence can appear. Therefore, if a pattern saturates with occurrences, fewer false positives can be predicted through our gain at leaf level technique. There is a strong anti-correlation (r-squared 0.95) between the number of occurrences in a pattern and potential gains at leaf level (Figure 4.6).

Occurrence patterns can become saturated through an increase in either true or false positive assignments (see Section 4.4 for consequences of additional true assignments). An increase of false positive occurrences could be created by the inclusion of very high e-values into the set of assignments. This might be desirable to enlarge the capacity of existing fold recognition methods. However, if false occurrences start to dominate the pattern, our technique would begin to fail.

The problem can be overcome using a slight modification to the technique. The parsimony algorithm is first run on a base pattern, created from assignments below a strict e-value threshold. Genomes without an occurrence in this base pattern, are checked for potential gains at leaf level. Then the set of isolated occurrences becomes the union of all gains at leaf level in the base pattern and all potential gains at leaf level, which have an occurrence with an e-value above the threshold. This modification could cause the algorithm to over-predict the number of false positives if the e-value cutoff is set too low.

Figure 4.7 shows the results of running this procedure with different e-value cutoffs for the base pattern. Only when using extremely low e-value cutoffs (e.g 10^{-20}), are a

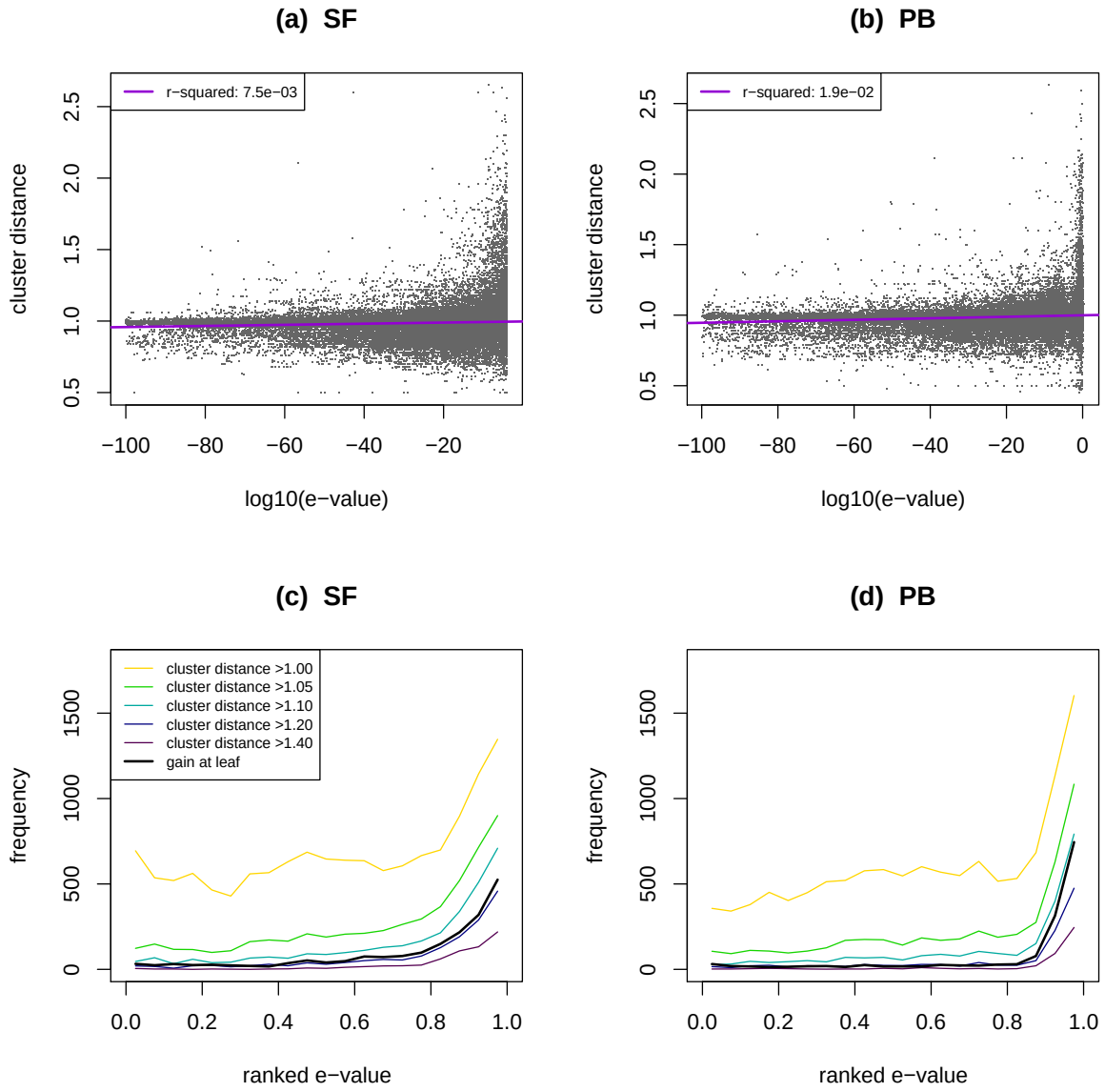


Figure 4.5: Cluster distance
E-values and cluster distance for SUPERFAMILY (a),(c) and PSI-BLAST (b),(d). (a),(b) log10(e-value) against cluster distance of occurrences. (c),(d) e-value distributions for occurrences in a specific range of cluster distances. (a) and (b) show there is no correlation between the e-value and cluster distance. The cluster distance indicates whether an occurrence lies within a cluster of occurrences (see Section 4.2 for more details). (c) and (d) show that occurrences with a relatively high cluster distance (> 1.00) have a distribution similar to the assignment distribution for most of the e-value spectrum, since the ranked distribution is near uniform. Only the right-hand side of the distribution shows a drastic increase in frequency compared to the reference distribution. This sudden increase is likely to be caused by false positives. Occurrences with a cluster distance > 1.20 give a similar distribution of e-values to occurrences with a gain at leaf level.

considerable number of isolated occurrences detected for middle-range e-values. Hence the above method can potentially be used without a huge over-prediction of false positives. In fact, comparing the distributions (coloured lines) to the consensus data (striped line),

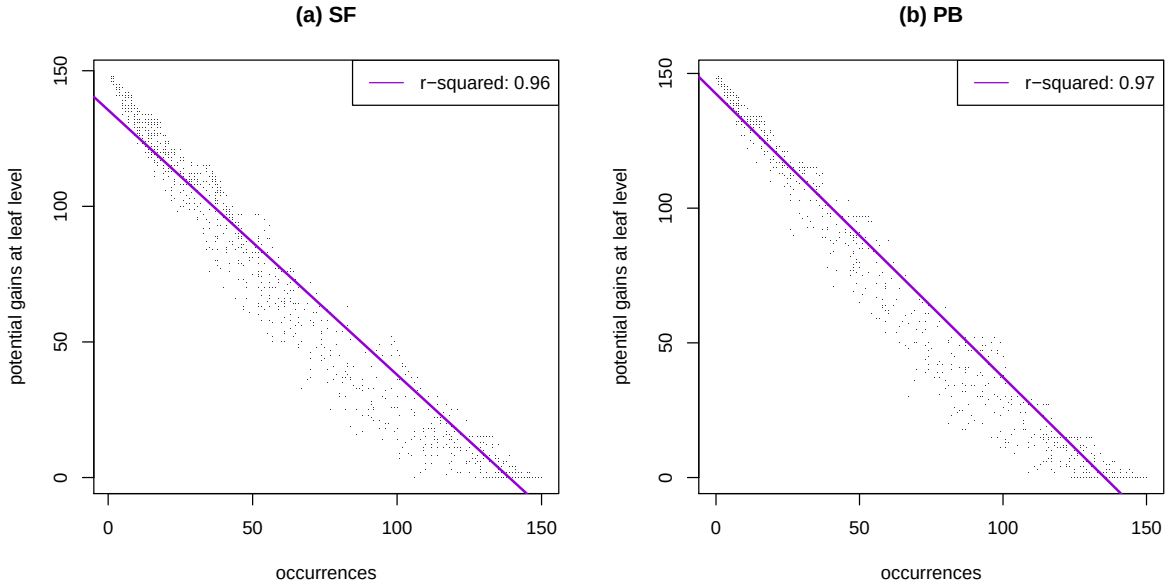


Figure 4.6: Saturation effect
Correlation between the number of occurrences and potential gains at leaf level for SUPERFAMILY (a) and PSI-BLAST (b). A potential gain at leaf level is defined as a genome without an occurrence, which would cause a gain at leaf level if an occurrence would be added

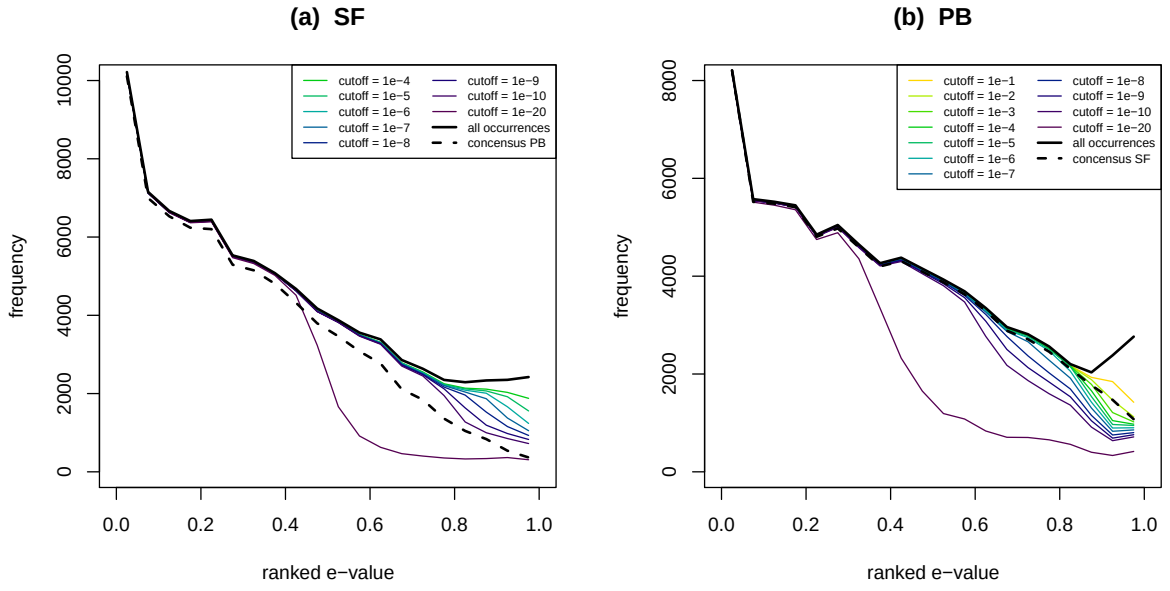


Figure 4.7: Base patterns
Ranked e-value distributions for occurrences without isolated occurrences. The isolated occurrences are calculated through a base pattern created from assignments with an e-value below the given threshold (see legend). Here an isolated occurrence is either a gain at leaf level in the base pattern or a potential gain at leaf level coinciding with an occurrence that has an e-value above the threshold. A clear over prediction of isolated occurrences is only seen in the set with a base pattern cutoff of 10^{-20}.

a sensible threshold can be determined with an e-value cutoff of 10^{-8} for SUPERFAMILY and 10^{-2} for PSI-BLAST. This version of our method can be used on occurrence pattern sets with a high proportion of false positives.

4.3.8 Fold recognition and structural classes

proportion of false positives				
class	gain at leaf		high e-value	
	SUPERFAMILY			
all	**0.014**		**0.659**	
alpha	0.023	>>	0.758	>>
beta	0.019	>>	0.573	<<
alpha/beta	0.001	<<	0.291	<<
alpha+beta	0.012	<<	0.657	
multi-domain	0.007	<<	0.375	<<
membrane	0.027	>>	0.717	
small protein	0.049	>>	0.877	>>
	PSI-BLAST			
all	**0.014**		**0.741**	
alpha	0.023	>>	0.836	>>
beta	0.021	>>	0.692	
alpha/beta	0.004	<<	0.526	<<
alpha+beta	0.012	<<	0.749	
multi-domain	0.007	<<	0.714	
membrane	0.031	>>	0.811	
small protein	0.021	>>	0.857	>>

Table 4.2: Predicted false positive assignments for structural classes
The first column shows the proportion of all occurrences with a gain at leaf level. The second column indicates the proportion of occurrences with a gain at leaf level that have a high e-value. A high e-value is here defined as higher than 10^{-10} for SUPERFAMILY and 10^{-4} for PSI-BLAST. Double left pointing arrows (<<) indicate that the proportion of false positives for the specific class is significantly lower than would be expected for a random subset of occurrences in **all** classes, right pointing arrows (>>) indicate that this proportion is significantly higher.

At the highest level of the SCOP hierarchy, domains are grouped into 7 different classes, predominantly based on secondary structure content. In our analysis predicted false positives rates for these structural classes differ significantly from one another. For example, in Table 4.2 the proportion of occurrences with a gain at leaf level is lower for the alpha/beta and multi-domain classes than for the all-alpha, all-beta, trans-membrane

and small protein classes (SUPERFAMILY).

Some of these differences, although not all, might be explained by the average chain length of the proteins. Domains from the alpha/beta and multi-domain classes have on average longer chain lengths than domains from the all-alpha and small protein classes. Karplus and coworkers Karplus et al. (1998) previously observed that HMM based methods might be less accurate at estimating correct e-values for small sequences; this could result in a higher false positive rate. Note that the multi-domain class is different from the other six, as it contains proteins which can not yet be split up into separate domains based on SCOP classification rules. This class is included here, as it shows some evidence for sequence-length dependence.

To investigate in more detail why certain structural classes have higher rates of predicted false positives, we submitted a small number of predicted false positive gene-regions to the meta servers (`http://bioinfo.pl/meta/` and `http://genesilico.pl/meta`). The servers predicted quite a few of the predicted alpha class regions to be in coil like structure. On inspection of the source domain, we observed that the region, which generated the assignment, often contained a very long alpha helix. This strong helical-helical scoring may explain the slightly higher false positive rate for the all alpha class.

These results must be interpreted with a little care, as the proportions of isolated occurrences might correlate with saturation of the occurrence patterns. As described above, more saturated occurrence patterns result in fewer potential gains at leaf level. Domains from the class of small proteins have on average a lower age than domains from the alpha/beta class (see Chapter 3), and will therefore have on average emptier occurrence patterns.

To correct for this Table 4.2 also shows the proportion of occurrences with a gain at leaf level that have a high e-value ($> 10^{-10}$ for SUPERFAMILY and $> 10^{-4}$ for PSI-BLAST). These proportions indicate that the all-alpha and small protein classes still display a higher proportion of false positives, than the alpha/beta class for both the SUPERFAMILY and PSI-BLAST assignments.

The second score seems to deviate more per class for SUPERFAMILY than for PSI-BLAST, while the standard deviation for the sample distributions are very similar. This may indicate that there is a higher e-value dependence on secondary structure content or domain length for the former method.

4.3.9 Validation

overlapping assignments in PSI-BLAST			
	total	overlap	fraction
occurrences	80778	1707	2%
predicted false positives	1035	198	19%

Table 4.3: Overlapping regions in PSI-BLAST assignments
Occurrences for PSI-BLAST that are generated by assignments which have a stronger hit in the same region to a different superfamily. The fraction of predicted false positives, which have an overlap, is about 9 times larger than the the fraction of occurrences with an overlap with respect to all occurrences. A predicted false positive occurrence is a single copy occurrence with a gain at leaf level and an e-value larger than 10^{-4}.

Conventional benchmarking of fold recognition methods involves blind tests on subsets of protein-domains with a known structure (Madera and Gough, 2002; Park et al., 2000; Pearl et al., 2002) or on sets of sequences annotated with expert knowledge (Schäffer et al., 2001). Using such a procedure the ratio of false predictions can be estimated as well as the ratio of false negatives. A similar fashion of benchmarking would be difficult for our method, since the majority of genes have unknown structures and using all genes on the genomes is essential to the method. We have already shown through e-value distributions that isolated occurrences are less reliable. Below we describe two other independent tests to verify our set of predicted false positives. Predicted false positive are defined as an occurrence with a gain at leaf level, a single copy and a high e-value ($> 10^{-10}$ for SUPERFAMILY and $> 10^{-4}$ for PSI-BLAST).

Although no actual structures are available for the majority of the genes, a set of likely false positive assignments can be obtained by checking for overlap. For a given gene region more than one assignment may be obtained. If the assignments are to two different superfamilies then the assignment with the worse e-value is likely to be false.

Translating this to occurrence patterns, we can say that occurrences, which are solely

generated by assignments with a stronger assignment in the same region, are likely to be false; we selected these occurrences and compared them to our sets of predicted false positives. Table 4.3 shows that the overlapping occurrences are found 9 times more often in our set of predicted false positives than in general occurrences. Fisher's exact test confirms that this difference in numbers is highly significant (p-value $< 2.2 \cdot 10^{-16}$). Moreover if the ratio of overlapping regions is measured within a set of occurrences above the same e-value threshold as the set of predicted false positives and is compared to this ratio for the set of predicted false positives, the fraction of overlapping regions remains significantly higher for the predicted false positives (p-value $= 2.7 \cdot 10^{-5}$).

occurrence changes SUPERFAMILY 1.65 to 1.69			
	total	removed (1.69)	fraction
occurrences (1.65)	89792	12032	13%
predicted false positives (1.65)	1183	1099	93%

Table 4.4: Changes in SUPERFAMILY from 1.65 to 1.69
Occurrences removed from SUPERFAMILY version 1.69, which were present in version 1.65. 93% of the predicted flase positives in version 1.65, have been removed in 1.69. A predicted false positive occurrence is a single copy occurrence with a gain at leaf level and an e-value larger than 10^{-10}.

Unfortunately this test can only be carried out on PSI-BLAST assignments as in the SUPERFAMILY database overlapping regions have already been removed. Instead we can compare two different versions of the database. Our study has been carried out on SUPERFAMILY version 1.65. The more recent version (1.69) is deemed to be more reliable by the authors. We assess assignments which were present in SUPERFAMILY version 1.65, but have been removed in version 1.69. Again we select occurrences for which all assignments on the genome were removed (Table 4.4). Over 90% of our predicted false positives in SUPERFAMILY 1.65 have been removed in version 1.69. The fraction of removed occurrences is signifcantly higher in the set of false positives than in general ooccurrence (p-value $< 2.2 \cdot 10^{-16}$) and significantly higher than in the set of occurrences with a similar e-value (p-value $< 2.2 \cdot 10^{-16}$).

Both these results confirm that our set of predicted false positives, are significantly more likely to be false than general occurrences.

4.4 Discussion and conclusions

4.4.1 An increase in true positives

As described above, fewer isolated occurrences can be detected as occurrence-patterns become more saturated. Such saturation could be caused by an increase in true positive assignments, when fold recognition techniques become far more sensitive or when many more protein structures become available.

In chapter 2 we showed that the number of occurrences is not uniformly distributed. The distribution peaks at a low and a high number of occurrences, with a minimum in the middle. In addition recent work by Yan and Moult (2005); Sadreyev and Grishin (2006) shows that the number of known (super) families, which occur on very few genomes, is expected to grow. This implies that although some superfamily-patterns might become more saturated, the number of superfamilies with emptier patterns will also grow. Hence this method will remain applicable to a large number of superfamilies.

4.4.2 Biological relevance

When interested in the fold assignment of a single protein, it is important to keep in mind that an isolated occurrence may appear due to lateral gene transfer or extensive gene loss. Nevertheless, it can be advantageous to visually inspect predicted occurrence patterns of homologous sequences with a weak hit to the protein of interest. Anomalies in such a pattern can give an idea about the reliability of the assignments, and may also indicate false negatives or deletions, in cases of a 'loss at leaf level'.

When working on a specific protein it is also important to be aware of the difference in aim and methodology between this method and phylogenomics methods used for functional annotation. The aim of this study is to find true homologs in a set of likely homologs, whereas phylogenomic methods aim to subclassify a set of known homologs. The difference in aim, results in a distinctively different methodology. For known (close) homologs, it is feasible to calculate a phylogeny of proteins and subsequently as-

sign duplication and speciation events (Engelhardt et al., 2005), whereas in this work the proteins of interest have very distant (if any) evolutionary relationships and a precalculated phylogeny of genomes is used instead. Hence our method is not a substitute for a phylogenomics method, but could perhaps be used as a prefilter in cases where structural assignments are involved (see for example Sjölander (2004)).

4.4.3 Consensus data and meta servers

Previously we described how the method could be extended to assess assignment sets with a large proportion of false positives using a base pattern. However, rather than assignments below an e-value threshold, consensus occurrences could be used to create the base pattern. In this study two profile based searching methods were analysed. Although a large proportion of occurrences agree (84% for SUPERFAMILY and 95% for PSI-BLAST), only around 50% of these occurrences were caused by the same number of assignments. The two fold recognition methods might recognise different homologues on the same genome. Consensus occurrences appear to be more reliable on evaluation by our method: the proportion of isolated occurrences was significantly lower than for occurrences obtained by a single method, even though consensus occurrence patterns are naturally sparser. The usage of consensus occurrences rather than consensus assignments could therefore provide additional information about the reliability of the assignments. This technique might be used as an additional quality check for meta servers, using consensus data of several genome wide fold recognition methods.

4.4.4 Conclusions

This study shows that, false positives, assigned by fold recognition methods on completed genomes, can be detected by determining isolated occurrences in a phylogeny. We developed a method to identify these isolated occurrences by applying a parsimony algorithm to a phylogenetic occurrence pattern. An occurrence is said to be isolated if it causes a 'gain at leaf level' in the most parsimonious evolutionary scenario.

It is shown that in principle isolated occurrences are less reliable than other assignments: the majority of isolated occurrences have a high cut-off ($> 10^{-8}$ for SUPERFAMILY ,$> 10^{-4}$ for PSI-BLAST). E-values are shown to be almost independent of evolutionary distance between the source sequence and the genome of assignment. Deletions and/or false negatives are therefore unlikely to cause the observed high e-values of isolated occurrences.

To predict false positives in practise additional constraints should be imposed. The e-value of the occurrence should be higher than a given e-value ($> 10^{-8}$ for SUPERFAMILY ,$> 10^{-4}$ for PSI-BLAST), to minimise the number of isolated occurrence caused by lateral gene transfer. In addition the occurrence (of a superfamily) should be caused by a single assignment to the genome. Using this technique more than 1000 false positives can be predicted for both SUPERFAMILY and PSI-BLAST. Independent tests were performed to validate our predicted false positives; one test was based on the overlap of PSI-BLAST assignments, the other considered changes between different versions of the SUPERFAMILY database. Both tests confirmed that isolated occurrences are more likely to be falsely assigned. The method can be extended to assess sets of assignments with a large proportion of false positives and could be used to enhance the searching power of any existing fold recognition techniques. This technique could therefore provide a way to fundamentally improve the assignment sets of genome wide fold recognition.

Considering occurrence patterns from genome wide fold recognition also gives a new way to examine the quality of fold assignments. It was observed that the number of assignments and occurrences on genomes drastically increases for high e-values. The most likely explanation for this phenomenon is a sudden increase in false positive assignments above certain e-values. Since the sudden increase is not observed in consensus data, with occurrences predicted by both SUPERFAMILY and PSI-BLAST assignment, or for data where isolated occurrences were removed.

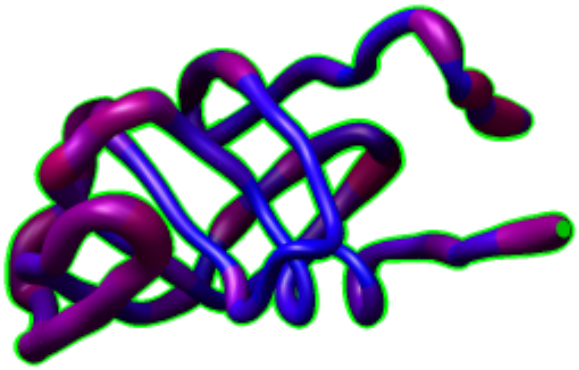
When examining the rate of predicted false positives for different structural classes, a significant variance was observed. Domain length and secondary structure content might

cause this dependency between false positive rate and structural class.

Here we have shown that it is possible to asses and improve genome wide fold recognition experiments by considering occurrence patterns on a phylogeny of species.

Chapter 5

Structural domain properties and superfamily age



5.1 Introduction

In Chapter 3 we showed that fold ages are related to specific structural properties: domains from different structural classes correspond to significantly different age distributions. Here we will explore this further by examining the correlations between structural domain properties and age.

Several structural properties potentially affect, or are affected by, fold ages. Based on previous results, we may expect that both secondary structure content and domain length correlate with age. We also investigate structural properties which are associated with the stability, designability and folding rate of protein domains, such as compactness, contact density and contact order. In addition amino acid composition and locations of termini within the domain are examined for dependencies.

Any correlation between domain specific properties and superfamily ages may be caused by several factors, e.g. the fold recognition method, the hierarchical classification system used or evolutionary pressure. To investigate the first of these factors, all structural domain properties are also checked for their effect on the reliability of the fold recognition methods.

5.1.1 Amino acid composition

Amino acids are known to have different propensities for different secondary structure types (e.g. Chou and Fasman (1978); Swindells et al. (1995)). There are also significant differences in amino acid content for proteins from different species; this may be due to living environment and/or codon usage biases. For example Lieph et al. (2006) show that there is a significant correlation between GC-content in the DNA and codons expressing hydrophobic amino acids (in particular Phe, Leu, Ile, Met and Val). High GC-content stabilises the DNA and is often found in (hyper)thermophilic species.

Not all amino acids are thought to have been present in primitive life. Trifonov (2004) proposed the following chronological order for appearance of amino acids during evolution: Gly, Ala, Asp, Val, Pro, Ser, Glu, (Leu, Thr), Arg, (Ile, Gln, Asn), His, Lys, Cys, Phe,

Tyr, Met, Trp. These results are based on a combination of codon usage, amino acids found on the Murchiso Meteorite and the generation of amino acids from non-organic compounds (Miller's experiment, Miller (1953, 1987)).

Several studies that have tried to determine amino acids which are abundant in 'ancient' proteins and compare these results to amino acids found by Miller's experiment; in most studies some overlap is detected. However, methods to determine the sequences of such ancient proteins remain somewhat controversial (Brooks et al., 2002; Jordan et al., 2005; Goldstein and Pollock, 2006; Hurst et al., 2006).

In this study will we use amino acid propensities to investigate differences between young and old superfamilies.

5.1.2 Fold stability

Protein folds must have an adequate difference in free energy between the folded and unfolded state in order to be stable. It is thought that most proteins are only marginally stable: there is little or no pressure to gain stability beyond the minimum required to fold. Thermodynamic stability of proteins (or unfolding free energies) tend to be similar in close homologues (Serrano et al., 1993) and during evolution a comparable thermodynamic stability tends to be maintained (Sanchez-Ruiz, 1995). However, for most proteins it is possible to drastically improve their stability: for example Perl et al. (2000) show the thermostability of a cold shock protein can be improved by only a couple of strategic amino acid changes.

Methods to improve thermostability have been extensively studied. Structural properties that increase thermostability can be identified by comparing homologous protein structures expressed in mesophiles (organisms living at room temperature), thermophiles (organisms living at higher temperatures, e.g. 50°) or hyperthermophiles (organisms living at extremely high temperatures, e.g. 100°). Proteins that are stable at high temperatures tend to be longer (Ganesh et al., 1999), more compact (Russell et al., 1997), have decreased lengths of surface loops (Kumar and Nussinov, 2001), have a higher Glutamic

Acid content (Tekaia et al., 2002), an increased number of ion pairs (Russell et al., 1997) and a larger fraction of hydrophobic residues (Kumar and Nussinov, 2001; Lieph et al., 2006).

We show that many of the properties associated with stability at high temperatures are also associated with old fold ages.

5.1.3 Designability

The number of copies for a family, superfamily or fold on a genome is highly variable for different families (Figure 2.5).

One explanation for this phenomenon is that some structures are able to harbour a greater number of different sequences than others, this is also referred to as the designability of a fold. Using energy distribution functions, Finkelstein et al. (1993) argued that small energy defects in protein folds may cause structures to be less adaptable to a wide variation of sequences. Since then it has been shown for theoretical protein models (on a square or cubic lattice), that designability can create variation in family sizes (Finkelstein et al., 1995; Govindarajan and Goldstein, 1996; Taverna and Goldstein, 2000).

England and Shakhnovich (2003) predict with a similar theoretical model that the trace of a power of the contact matrix for a structure may determine the number of different sequences that will fit onto the structure while producing a low energy (i.e. displaying the ability to fold). Shakhnovich et al. (2005) showed that the contact density, related to this trace measure, indeed correlates with family size for real protein domains. They also suggested that families present in the LUCA have relatively higher contact densities.

Here we will investigate the relation between age and these properties, rather than family size and these properties.

Please see reference

Table 5.1: Accuracy of fold recognition
Fraction of true positives in all possible positive pairs for different fold recognition methods taken from Shi et al. (2001). A true positive prediction is given as the best prediction (Top 1) by the fold recognition method or in the best five predictions (Top 5). Positives are pairs belonging to the same family (Family only), superfamily but not family (Superfamily only) or fold but not superfamily (Fold only).

5.1.4 False positive and false negative rates in profile based fold recognition methods

Profile based fold recognition methods, such as PSI-BLAST and SAM-T99 were used to generate the estimated fold ages from occurrence patterns on completed genomes (see Chapter 3). Since we will be looking at relatively weak correlations between ages and (structural) domain properties in this Chapter, it is important to investigate the effect of structural properties on the false positive and false negative rates of these fold recognition methods.

To determine the reliability of a fold recognition method, a set of known true positives is required. Usually an annotated set of known structures is used (Park et al., 1998; Madera and Gough, 2002; Pearl et al., 2002). A method's sensitivity is given by the fraction of predicted true positives in the set of all positives, and its specificity is given by the fraction of false negatives in all possible unrelated pairs. Furthermore, a 'true positive' can be defined at different levels, e.g. a domain that belongs to the same family, same superfamily or same fold. Note that any benchmark based on a structural classification may be highly affected by biases in known 'true positive' pairs.

To gain insight into the sensitivity of a fold recognition method at different levels of the structural hierarchy, an exclusive approach needs to be taken. Table 5.1 shows that it

is considerably easier to recognise two domains which belong to the same family, than to recognise two domains that belong to the same superfamily, but not to the same family.

Here we are particularly interested in any correlation between (structural) domain properties and the accuracy of the fold recognition methods. It has been shown that adding structural information to a sequence profile can improve distant homology detection and fold recognition. For example, Shi et al. (2001) show that local structural features, such as hydrogen bonding and solvent accessibility, can aid the detection of remote homologs. Moreover, (predicted) secondary structure can significantly improve the accuracy of sequence profile based methods (Ginalski et al., 2003)

However, little is known about the effect of structural properties on the accuracy of fold recognition solely based on sequence. Some anecdotal evidence for dependencies is available; Karplus et al. (1998) note that SAM-T98 does not perform as well for smaller domains; false positives with high confidence scores are frequently found for such domains.

In addition, Casbon and Saqi (2004) show in a small study that profile-profile recognition is superfamily dependent: there are very large differences between the performance of sequence profile recognition algorithms between different superfamilies. A weak correlation also appears to exist between the accuracy of detection and conservation of structure during evolution.

Since little is known about structural effects on sequence profile methods we used the false positives and false negatives produced by PSI-BLAST in our own test to explore any dependencies of the used structural properties. We show that properties such as domain length and secondary structure content do indeed affect the reliability of profile based fold recognition methods. Most of these effects, however, are too weak to fully explain the observed correlations with fold ages.

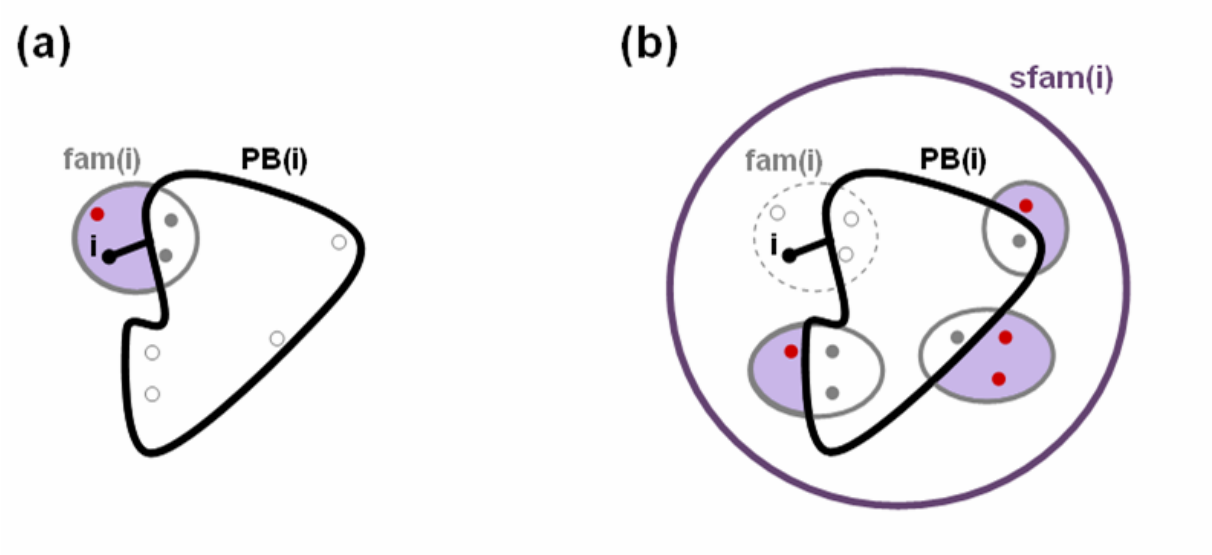


Figure 5.1: False negative rate - example
Sets of domains in the calculation of false negative rates at family level (a) and superfamily level (b). Here i indicates the domain for which the false negative rate is calculated, $PB(i)$ the set of assignments found by PSI-BLAST for domain i, $fam(i)$ the set of domains in the same family as i and $sfam(i)$ the set of domains in the same superfamily as i. Domains excluded from the calculations are depicted by small empty circles, whereas domains incorporated in the calculations as 'positives' are depicted as small dots. False negatives (fn) are depicted in red.

5.2 Methods

5.2.1 False negative rates

Using PSI-BLAST assignments, we estimate false positive and false negative rates for individual domains to simplify the comparison between structural properties and reliability rates. However, estimating individual reliability rates for domains remains difficult due to highly skewed structural classifications. Nevertheless, we may be able to explore major trends.

False negatives rates for domains are essentially estimated by the fraction of domains with a similar classification group (positives) that are not found by PSI-BLAST. All domains in these calculations are representatives of domains defined by SCOP (see below). The rates are identified at different classification levels using an exclusive approach: the set of positives is defined by domains in the same classification group (e.g. superfamily) as the domain, but only if they do not belong to the same subgroup (e.g. family); see Figure 5.1 and equations below.

For a domain i the false negative rate $fnr_{cl}(i)$ is given as the fraction of false negatives (fn) in all positives (pos) for a specific classification group (cl):

$$fnr_{cl}(i) = \frac{\#\{fn_{cl}(i)\}}{\#\{pos_{cl}(i)\}} \tag{5.1}$$

where $cl \in \{fam, sfam, fold\}$ (family, superfamily, fold) and $\#\{\}$ gives the number of members in a set (or cardinality). At family level $fn_{fam}(i)$ and $pos_{fam}(i)$ are given by:

$$\{fn_{fam}(i)\} = \{x \,|\, x \in fam(i) \ \wedge \ x \neq i \ \wedge \ x \notin PB(i)\} \tag{5.2}$$

$$\{pos_{fam}(i)\} = \{x \,|\, x \in fam(i) \ \wedge \ x \neq i\} \tag{5.3}$$

where $fam(i)$ is the set of domains in the same family as i and $PB(i)$ is the set of domains found by PSI-BLAST. At superfamily level $fn_{sfam}(i)$ and $pos_{sfam}(i)$ are given by:

$$\{fn_{sfam}(i)\} = \{x \,|\, x \in sfam(i) \ \wedge \ x \notin fam(i) \ \wedge \ x \notin PB(i)\} \tag{5.4}$$

$$\{pos_{sfam}(i)\} = \{x \,|\, x \in sfam(i) \ \wedge \ x \notin fam(i)\} \tag{5.5}$$

where $sfam(i)$ is the set of domains in the same superfamily as i. A similar expression can be obtained for fold level false negative rates, by replacing fam by $sfam$ and $sfam$ by $fold$.

Here all domains i are 'search domains' (i.e. SCOP domains) used for generating the PSI-BLAST assignments (Chapter 2). To find $PB(i)$, genomic assignments (Chapter 2) were used directly by identifying a set of genes that represent the search domains. The gene with the highest sequence identity to the search sequence was chosen as the representative for a search domain. In addition such a representative must have at least 90% sequence identity with the search domain and at least a 90% overlap. Search domains without representatives were disregarded.

Note that in this work the direction of one domain finding another is reflected in the mean scores of the false negative rates: if a classification group contains two domains,

both domains need to find each other in order to generate a mean false negative rate of 0.

Unfortunately, we can not apply a similar method to the SUPERFAMILY assignments, as multiple assignments to the same region have been removed from the database.

5.2.2 False positive rates

Even though Chapter 4 deals with a novel method to detect false positives in our set of assignments, this technique is not suitable to verify whether our age estimates are affected by false positive rates, since the detection procedure is dependent on the number of occurrences in an occurrence pattern as is the fold age.

False positive rates are in this work estimated by overlapping regions; to suggest a false positive, the overlap needs to be at least 50% of both alignment regions. The hit with the highest e-value is counted as a false positive if the search sequences of the overlapping alignments belong to different folds. The false positive rate is the fraction of false positives in the total number of hits for the search domain. Note that the true false positive rate is likely to be underestimated by this approach, as false positives to sequences without homology to a known structure will not be identified. Again it is not possible to use a similar approach for the SUPERFAMILY assignments as overlapping regions have been removed.

5.2.3 Structural properties of domains

To calculate structural properties, representatives of SCOP domains with a sequence identity smaller than 95% and an aero-spaci score > 0.4 were taken from the ASTRAL database (Brenner et al., 2000). The authors of the database indicate that structures with such an aero-spaci score are adequately accurate.

Length

The domain length was calculated on the length of the search sequences.

Secondary structure assignment

All secondary structure is assigned with DSSP (Kabsch and Sander, 1983).

Surface accessibility

The surface accessibility (ACC) of each residue is also obtained with DSSP. DSSP gives the ACC as an estimate of the number of water molecules in contact with a residue times 10, or in other words the residue water exposed surface in Å^2 (`http://swift.cmbi.ru.nl/gv/dssp/descrip.html`).

Buried residues

In this work a residue is said to be 'buried' when less than than 7% of its surface is exposed to water (Hubbard and Blundell, 1987).

Contacts

A contact between two residues was said to be made when the two $C\alpha$ atoms are closer than 7.5 Å. Most studies considering contact-based measures such as contact density and contact order use a threshold between 6.0 and 8.0 Å.

Contact Order

Contact order is the sum of the distances, in sequence, between all residues which make a contact. Contact order is associated with folding rates of single domain proteins (Plaxco et al., 1998).

Squared radius of gyration

The centre of mass and the squared radius of gyration were both calculated from the coordinates of all $C\alpha$ atoms within a domain. The centre of mass (C_{mass}), or mean position coordinate vector ($\mathbf{r}_{mean}$) is given by:

$$C_{mass} = \mathbf{r}_{mean} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{r}_k \tag{5.6}$$

where N is the total number of residues in the domain and $\mathbf{r}_k$ is the Cα coordinate vector of residue k. The squared radius of gyration (R_g^2) is then given by:

$$R_g^2 = \frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{r}_k - \mathbf{r}_{mean}\right)^2 \tag{5.7}$$

which can also be written as:

$$R_g^2 = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{r}_i - \mathbf{r}_j\right)^2 \tag{5.8}$$

and is therefore proportional to the root mean squared distance between all particles (Flory, 1953; Grosberg and Khokhlov, 1994).

If the coordinates were evenly distributed in a sphere, the radius of gyration (R_g) is proportional to the radius of that sphere.

5.2.4 Amino acid propensities for young and old folds

The propensity for an amino acid in a given age group $P(aa)_g$ is given as:

$$P(aa)_g = \frac{N(aa)_g/N(aa)}{N(total)_g/N(total)} \tag{5.9}$$

where $g \in \{young|old\}$, aa is any of the 20 amino acids, $N(aa)_g$ represents the total number of residues with amino acid aa. Domains with a superfamily age equal to 1.0 are counted in the group of *old* domains and domains with a superfamily age < 0.2 are counted in the group of *young* domains.

For each propensity a chi-square test was performed to compare $N(aa)_g, N(total)_g$ with $N(aa), N(total)$, if the corresponding p-value was smaller than 0.001 to propensity was counted as significant (see Table 5.2).

5.2.5 Quantile plots

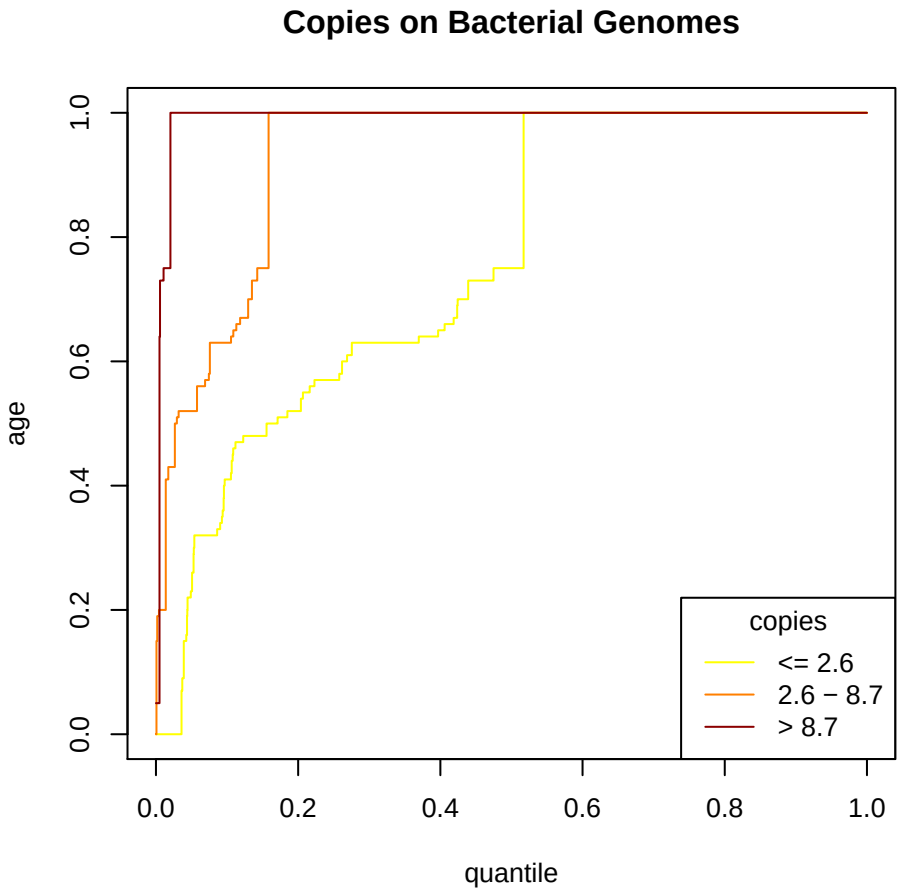


Figure 5.2: Example of a quantile plot
The dependence of superfamily age on the number of copies in Bacterial genomes. The shift in age distribution between superfamilies with few and many copies is significant (p-value $< 2.2 \cdot 10^{-16}$).

In order to investigate shifts in age distributions associated with different domain properties we used a rank based method, since superfamily ages are not normally distributed (see Figure 3.2 (a)).

The method used for plotting such distributions is shown in Figure 5.2. Here the x-axis represents the x-quantile of the data Y, or the fraction of observed values from Y smaller than the y-value; the y-axis represents each $y_i \in Y$. For example at $x = 0.5$ the corresponding y-value represents the value of the median of the distribution of Y.

In this work x-quantiles for an ordered set of values $Y = [y_1, y_2, ..., y_n]$ are estimated

as

$$x_i = \frac{rank(y_i)}{n} \tag{5.10}$$

where n is the total number of values in Y; a random $rank$ is assigned in the case of ties.

To make the process of partitioning distributions unbiased, all variables (Z) aimed at dividing the distributions are grouped into three intervals (low, middle, high) with roughly the same number of observations in each group. This three way partitioning is used in each quantile figure below, unless otherwise stated.

In Figure 5.2 the age-distributions are split into three groups corresponding to superfamilies with a low (yellow line), middle (orange line) and high (red line) number of copies. The closer the distribution lies to the left-top-corner the older the corresponding age distribution. Hence the figure shows that superfamilies with more copies are generally older.

To test if there is a significant shift in the distribution of Y, a Wilcoxon (or Mann-Whitney) test is performed between the distribution of Y for high values of Z and low values of Z. The p-values mentioned in the legends correspond to these tests, unless otherwise stated.

From the test in Figure 5.2 it may be concluded that the shift in age distribution between superfamilies with few and many copies is highly significant.

5.3 Results

5.3.1 Reliability of assignments and age

False negatives

Figure 5.3 shows the distribution of false negative rates for individual search domains using PSI-BLAST assignments. Rates for individual domains vary widely and do not follow a normal distribution. As mentioned above, false negative rates for individual domains are likely to be unreliable and are here used to investigate general trends.

The mean false positive rate at family level is 0.292, at superfamily level 0.903 and

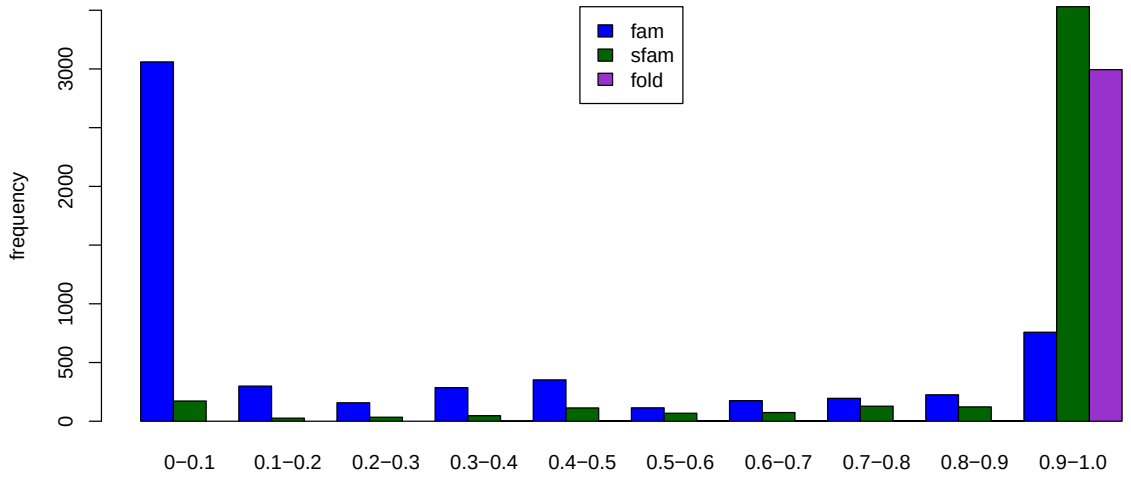


Figure 5.3: Distribution of false negative rates
The false negatives rate for a search domain is given by the fraction of false negatives found in the set of possible true positives (i.e. $1 - sensitivity$). The average values are given by: 0.292 at family level (fam), 0.903 at superfamily level (sfam) and 0.997 at fold level (fold).

at fold level 0.997. These numbers appear high when they are compared to previously calculated true positive rates in Table 5.1. However, Table 5.1 shows results for the top hit found by PSI-BLAST, in our work assignments are based on a strict e-value cutoff (10^{-4}). If the mean rates in our work are compared to the sensitivity scores at a similar specificity (99%) in work by Shi et al. (2001), our false negatives rates agree reasonably well; slightly lower false negative rates at family level are observed in our study.

Figure 5.4 shows the effect of false negative rates at family (a) and superfamily (b) level on the superfamily age of a domain: domains with a low false negative rate have generally older superfamily ages. This suggests that the sensitivity of PSI-BLAST may affect the age estimate of a superfamily. However, cause and effect are not immediately obvious (see below).

Even though the correlation between superfamily age and the false negative rates is relatively small, it may induce similarly weak correlations between structural properties and age. We therefore test whether any of the structural properties investigated in this Chapter influence the sensitivity of PSI-BLAST.

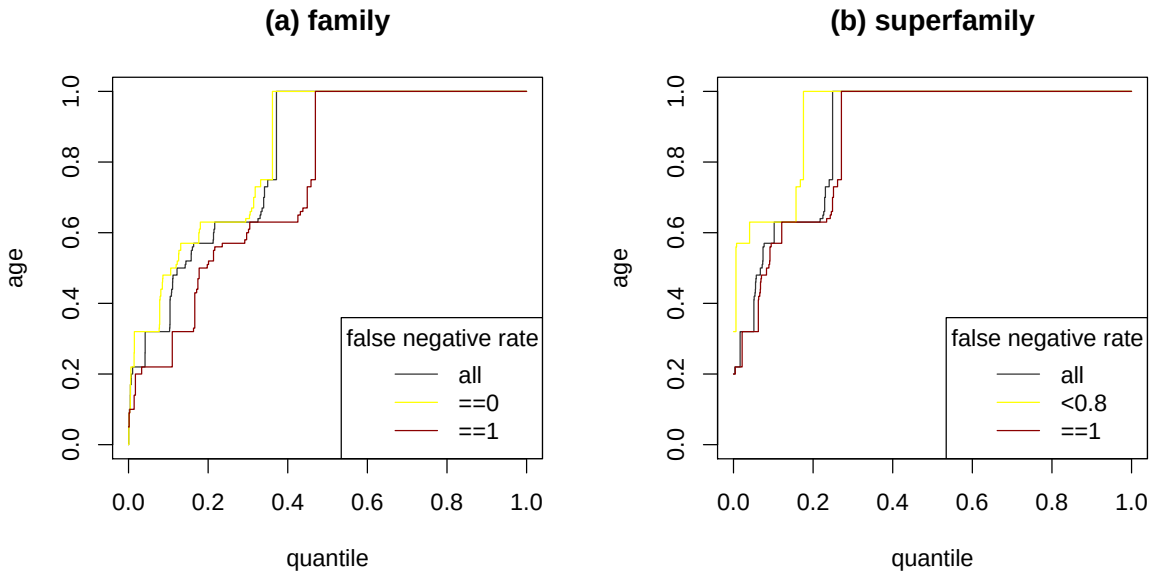


Figure 5.4: Superfamily age and false negative rates
The dependence of superfamily age on false negative rates at family level (a) and superfamily level (b). Age distributions are significantly shifted to older superfamily ages for low false negative rates ((a) p-value $= 2.95 \cdot 10^{-11}$, (b) p-value $= 2.09 \cdot 10^{-8}$). Here age distributions are split into two groups, since the distributions of false negative rates are highly skewed. A reference age-distribution including 'all' data is also shown.

If false negative rates determined by PSI-BLAST are compared with superfamily ages as determined by SUPERFAMILY assignments, very similar results are obtained (not shown). In addition, high false negative rates at superfamily level may indicate that some superfamily ages are significantly underestimated (see Section 5.4.1).

Figure 5.5 shows that there is a strong correlation between the number of hits found for a domain in the set of completed genomes and the false negative rate. However, cause and effect are not clear: either search domains that generate low false negative rates produce more hits, or the false negative rates for domains with many homologs (and hence more hits) are lower since the sequence profiles contain more information, enabling PSI-BLAST to detect more distantly related homologs. The latter effect is known to be important (Sadreyev and Grishin, 2004) and may also affect the correlation between false negative rate and superfamily age described above.

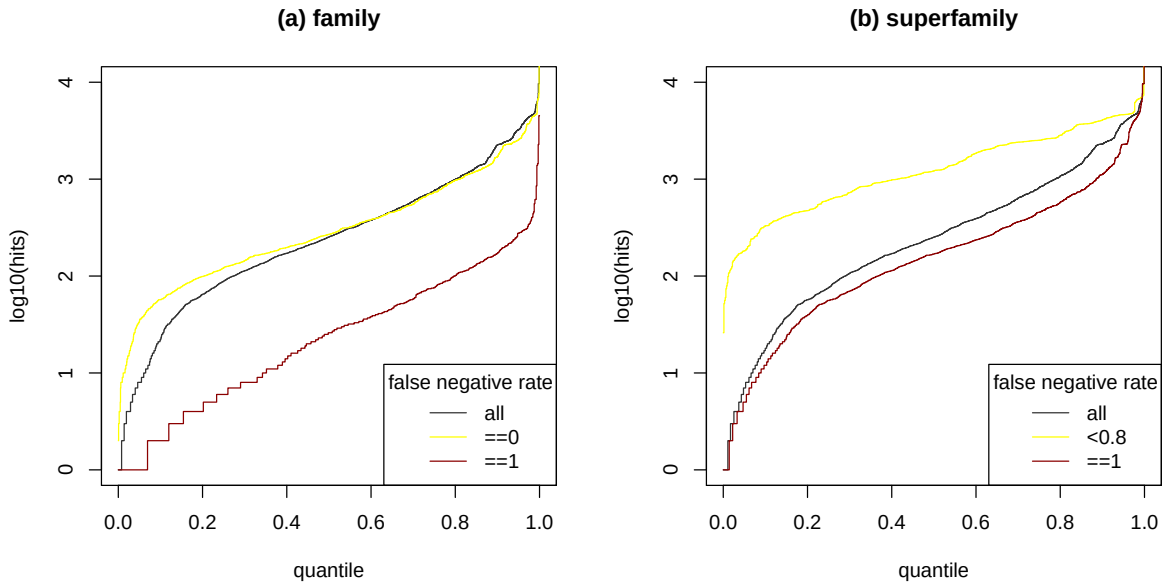


Figure 5.5: False negative rates and genomic hits
Dependence of search domain hits on false negative rates at family level (a) and superfamily level (b). Domains with a low false positive rate (yellow line) produce significantly more hits (both with p-values $< 2.2 \cdot 10^{-16}$). Here the age distributions are split into two groups, based on false negative rates.

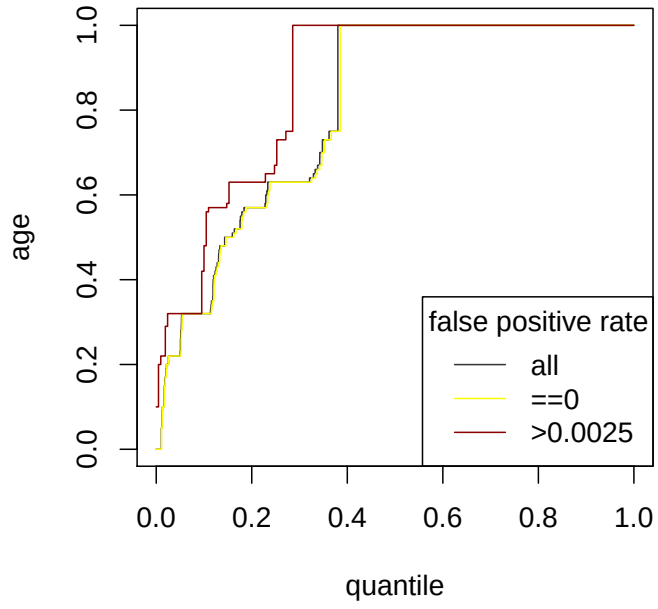


Figure 5.6: False positive rates and superfamily age
Correlation between false positive rate and age. Domains with a high false positive rate (yellow line) have significantly older ages (p-value $= 9.8 \cdot 10^{-4}$). Here the age distributions are split into two groups, based on false positive rates.

False positives

Search domains have relatively low false positive rates, with a mean value of $2.5 \cdot 10^{-3}$. The false positive rates are likely to be underestimated however (see Section 5.2.2). Even so, relative differences of these false positive rates between the domains are likely to be informative. As shown in Figure 5.6, false positive rates have a small but significant effect on the superfamily age: search domains with a higher false positive rate have generally older ages. This suggests that the (structural) domain properties investigated here should also be assessed for any correlation with the false positive rate.

5.3.2 Length

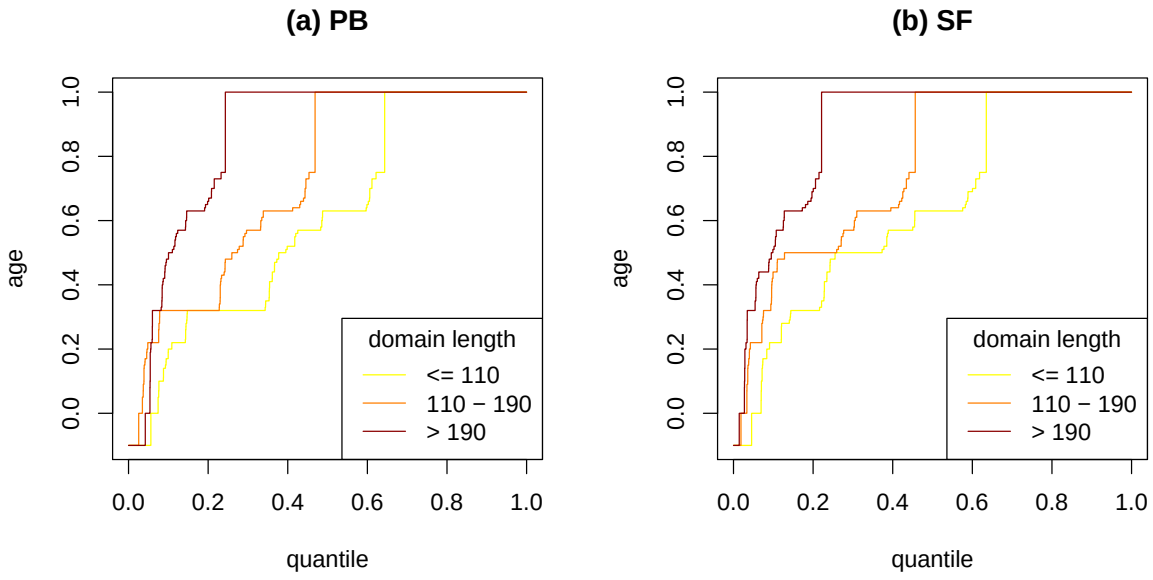


Figure 5.7: Domain length and age
Dependence of superfamily age on domain length with ages for PSI-BLAST assignments (a) and SUPERFAMILY assignments (b). For both datasets longer domains (red line) have significantly older ages than shorter domains (yellow line), both have p-values $< 2.2 \cdot 10^{-16}$.

Previous work has suggested that there may be a relationship between age and domain length, since (a) domains of the 'small protein' class have generally younger superfamily ages than domains of other classes and (b) alpha/beta folds tend to be longer as well as older than domains of other classes (see Figure 5.10).

Figure 5.7 shows a relatively strong correlation between domain length and age: longer

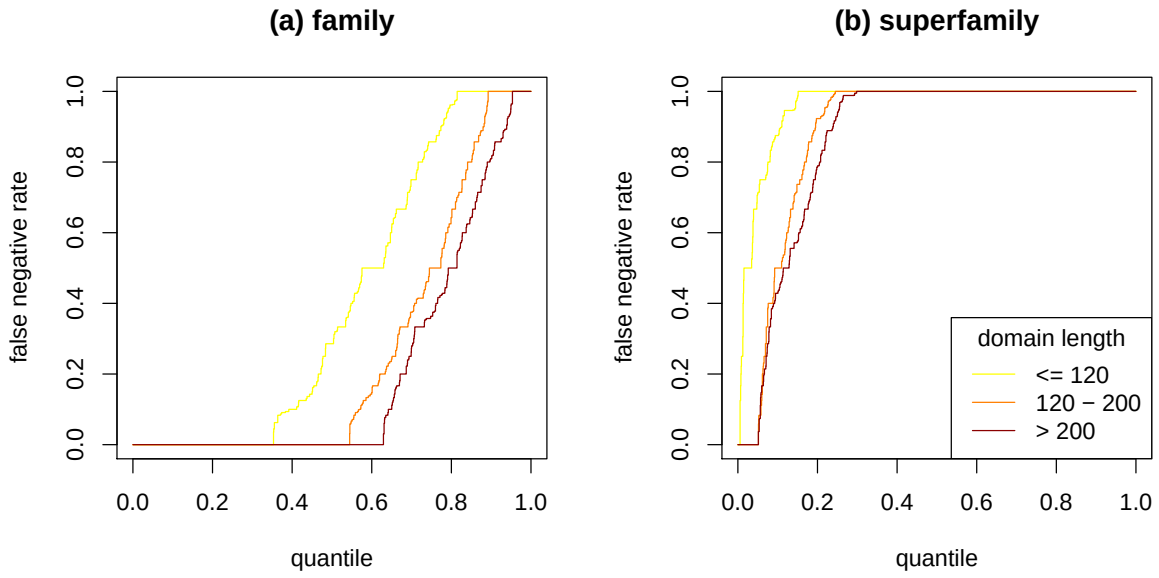


Figure 5.8: Domain length and false negative rate
Dependence of the false negative rate and domain length at family level (a) and superfamily level (b). The difference in false negative rate distributions between the lowest and middle length domains is significant at both family level and superfamily level, with p-values $< 2.2 \cdot 10^{-16}$.

domains are generally older. This could potentially be caused by increased accuracy of fold recognition methods on longer domains. Longer alignments usually obtain lower e-value scores.

The effect of domain length on false negative rates is shown in Figure 5.8. At family level, false negative rates for small domains are significantly higher (p-value $< 2.2 \cdot 10^{-16}$, yellow line Figure 5.8 (a)). Hence young ages of the small protein class may partly be explained through a lack of recognition of remote homologs in distantly related species.

Recently Cheek et al. (2006) showed that previously unidentified remote homologs of the small protein class could be detected through close observation, indicating possible merges for some SCOP superfamilies. This may suggest that superfamilies as defined by SCOP have indeed a 'younger' origin: if two related superfamilies are not classified as one, the ages of the two separate superfamilies may be estimated younger than the age of their union.

Figures 5.8 (a) and (b) also show that there is no difference between middle sized and large domains with respect to the false negative rate, although Figure 5.7 shows a clear

difference in age distribution between middle and large sized domains. It therefore seems unlikely that the difference in age distribution can be attributed to the fold recognition methods used. Moreover, the magnitude of difference in age-distributions shown in 5.7 can not be explained through the variations in the sensitivity of PSI-BLAST.

Surprisingly, false positive rates did not seem to be affected by the domain length in our PSI-BLAST assignments. This is at odds with work on HMM based models where a correlation between false positive rates and small sized domains was observed (Karplus et al. (1998), see also Table 4.2).

5.3.3 Secondary structure content

Previous results showing asymmetry in the age distribution of different structural classes suggest an age dependence on secondary structure content (see Figure 5.9 (d)). Figure 5.9 (a) and (b) show that domains with a large proportion of helices or parallel beta-bridges are older. Generally, the larger the fraction of residues in secondary structure the older the superfamily age of a domain (not shown). However, anti-parallel bridges are associated with younger ages (Figure 5.9 (c)). Moreover, splitting domains on the number of alpha helices, parallel beta-strands and anti parallel beta-strands gives rise to similar age distributions; splitting on the number of residues involved in the secondary structure elements also gives similar correlations. All these results are highly significant.

At this point it is important to verify how secondary structure content relates to structural classes as defined by SCOP. Figure 5.10 shows that the alpha/beta class can be discriminated through a high proportion of parallel-beta bridges (b) and a high domain length (d). Domains in the 'all beta', 'alpha+beta' and 'small protein' classes contain relatively more anti-parallel beta bridges.

Proportions of secondary structure do not appear to be quite as discriminating with respect to the age distributions as structural classes, comparing Figure 5.9 (a),(b) and (c) to Figure 5.9 (d). However, in the first three of these Figures, the distribution of secondary structure proportions is split into thee equal subsets (yellow, orange and red

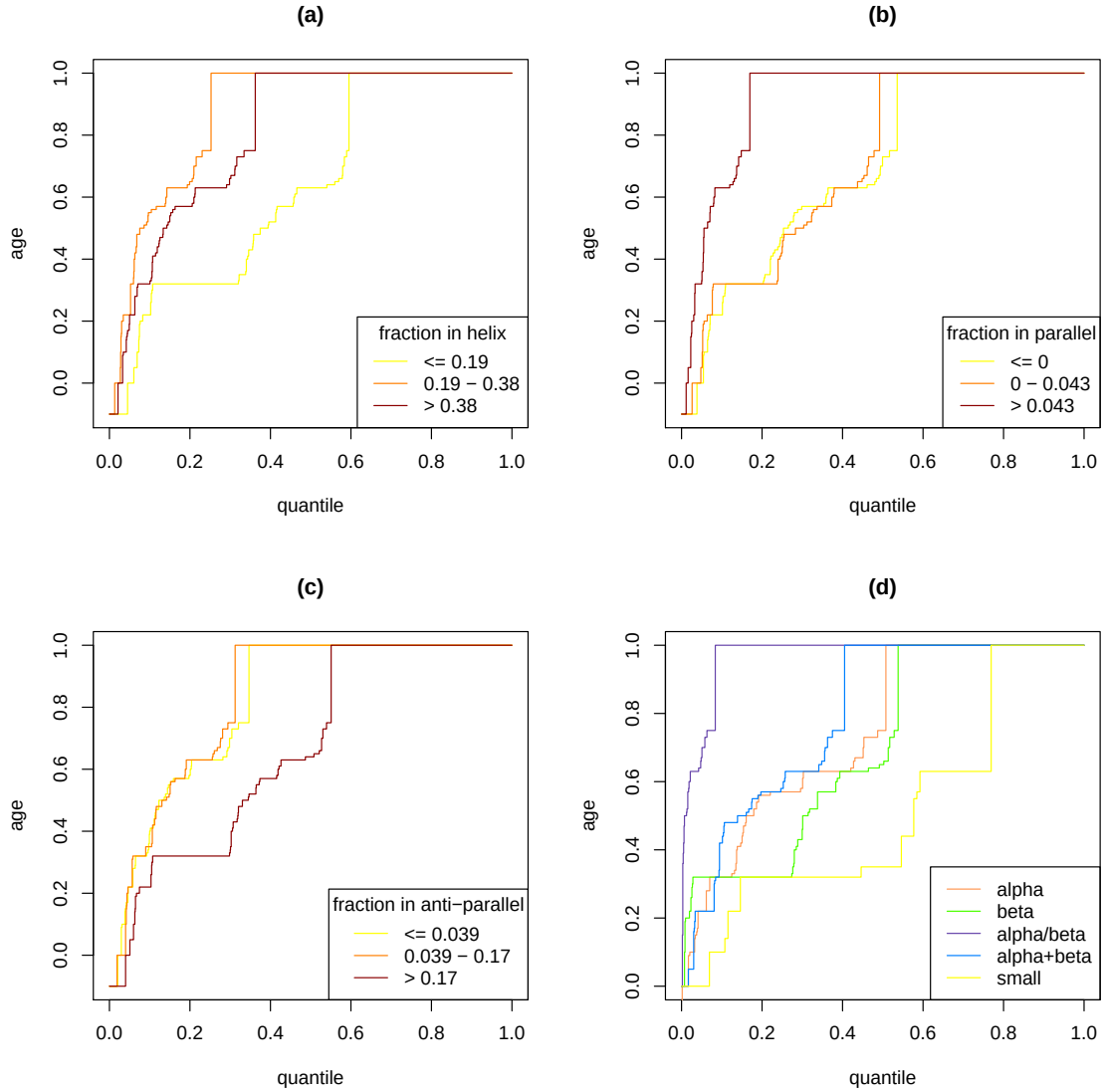


Figure 5.9: Secondary structure content and age
(a) Domains with a large fraction of helices are significantly older than domains with a low fraction of helices (p-value $< 2.2 \cdot 10^{-16}$). (b) Domains with a large fraction of parallel beta bridges are significantly older than domains with a low fraction of parallel beta bridges (p-value $< 2.2 \cdot 10^{-16}$). (c) Domains with a large fraction of anti-parallel beta bridges are significantly younger than domains with a low fraction of anti-parallel beta bridges (p-value $< 2.2 \cdot 10^{-16}$). (d) Superfamily age distributions for the different fold classes on the same set of domains as Figure (a), (b) and (c).

lines), whereas Figure 5.9 (d) is split into five subsets corresponding to the structural classes.

The median value of the fraction of parallel beta-bridges for alpha/beta domains lies considerably higher ($\approx$ 0.09 ,Figure 5.10 (b)) than the cutoffs used in Figure 5.9 (b) (0.043). If we use a lower bound for the fraction of parallel beta bridges, similar to this alpha/beta median, an age distribution comparable to that of the alpha/beta class

is obtained (not shown). This suggests that the fraction of parallel beta-bridges has a similar discriminative effect as SCOP's alpha/beta class. Extreme limits for the proportion of helical residues or anti parallel beta-bridges do not result in more extreme age distributions.

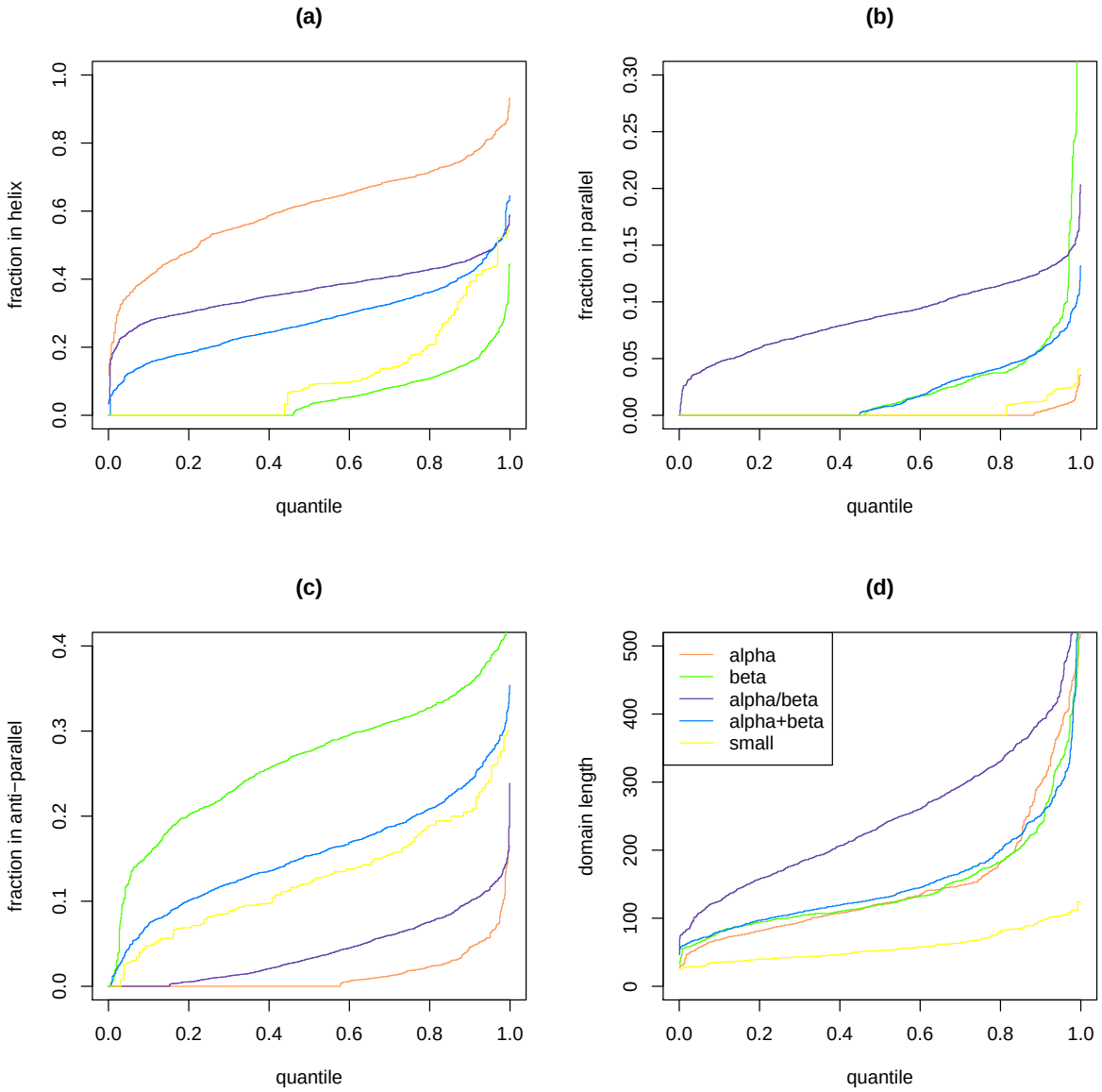


Figure 5.10: Secondary structure content and structural class
(a) Proportion of alpha helical residues. (b) Proportion of parallel-beta bridges. (c) Proportion of anti parallel beta-bridges. (d) Correlation between domain length and structural classes. In all figures the distributions are split on the basis of structural class of the search domain.

False negative rates do not show any correlation with secondary structure content. Slightly higher false positive rates are seen for domains with few alpha helical residues or

parallel-beta bridges, and for domains with more anti-parallel beta strands. The difference in distribution are however only just significant, and may be due to the normalisation process.

5.3.4 Location of termini

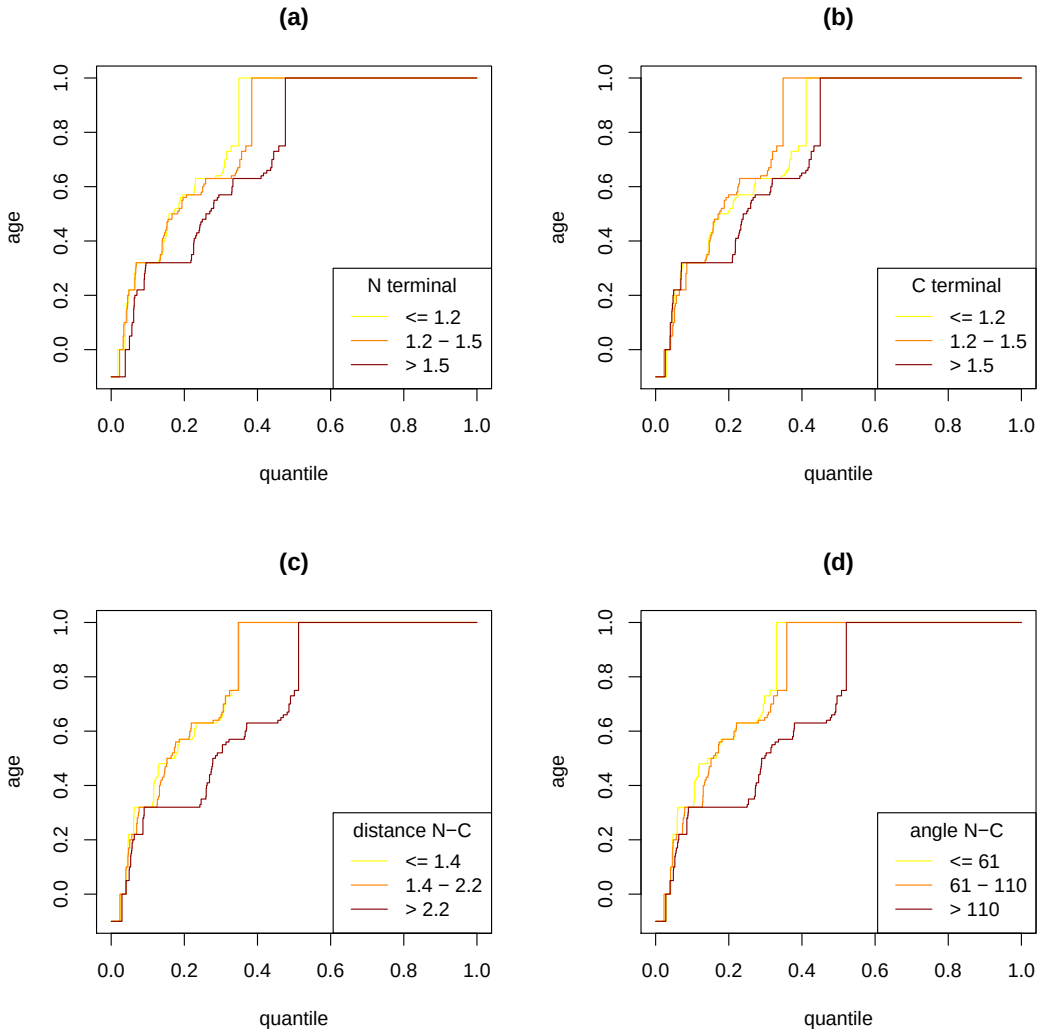


Figure 5.11: Location of domain termini and superfamily age
(a) Domains with an N-terminus further away from the centre of mass correspond to a younger superfamily age distribution (p-value $= 9.22 \cdot 10^{-12}$) (b) The shift towards young superfamily ages for domains with an C-terminus further away from the centre of mass is only just significant (p-value $= 4.3 \cdot 10^{-3}$) (c) Domains with a long distance between the N- and C-terminus are generally younger (p-value $< 2.2 \cdot 10^{-16}$). (d) Domains with a large angle between the N- and C-terminus are generally younger (p-value $< 2.2 \cdot 10^{-16}$); the angle is given in degrees. Distances in (a),(b) and (c) are normalised by the radius of gyration (R_g).

Figure 5.11 shows that the locations of the N- and C- termini follow slightly different trends during evolution. If Figure 5.11 (a) and (b) are compared it can be seen that

domains with an N-terminus further away from the centre of mass correspond to younger superfamily ages, but that this relation is much weaker for the C-terminus. In fact the C-terminus tends to show a typical pattern for a variable that is not linked to age: domains with a value around the mean tend to be the oldest.

Such an asymmetry between the N- and C-termini may reflect asymmetrical folding. Bhattacharyya et al. (2002) found that C-terminal regions appear more often in parallel beta strand than N-terminal regions. They suggest that N-terminal sequential folding is a likely explanation. Recently Taylor (2006) showed that alpha-beta domains are more asymmetrical than domains of other structural classes, and argues that this may be due to N-terminal folding. In addition he suggests that this may be typical for older domains. The results above - even though the difference is small - fit in well with this explanation: if the folding process starts at the N-terminus, the core of the protein is likely to be close to this terminus, and there would be no such restriction on the C-terminus.

Figure 5.11 (c) and (d) consider the location of the N- and C- termini relative to each other. Both Figures show that the further away the termini are from each other, the younger the superfamily ages.

It has previously been suggested that the termini of domains lie closer together than expected (Thornton and Sibanda, 1983; Christopher and Baldwin, 1996) due to the effect of circular permutations (Ponting and Russell, 1995): any circular permutation is likely to result in nearby endpoints. However, if only the *results* of circular permutations give close endpoints, one would perhaps expect the opposite effect: young folds with close termini. On the other hand domains with close terminal residues are also more likely to be suitable for circular permutation, such folds may be more adaptable during evolution. Alternatively, some of the structures with large angles between the termini may reflect non-compact, stretched domains, with termini on opposite ends (see Section 5.3.6). At this point it is not clear what effect causes younger domains to have more distant termini.

5.3.5 Contacts density and age

Contact Density

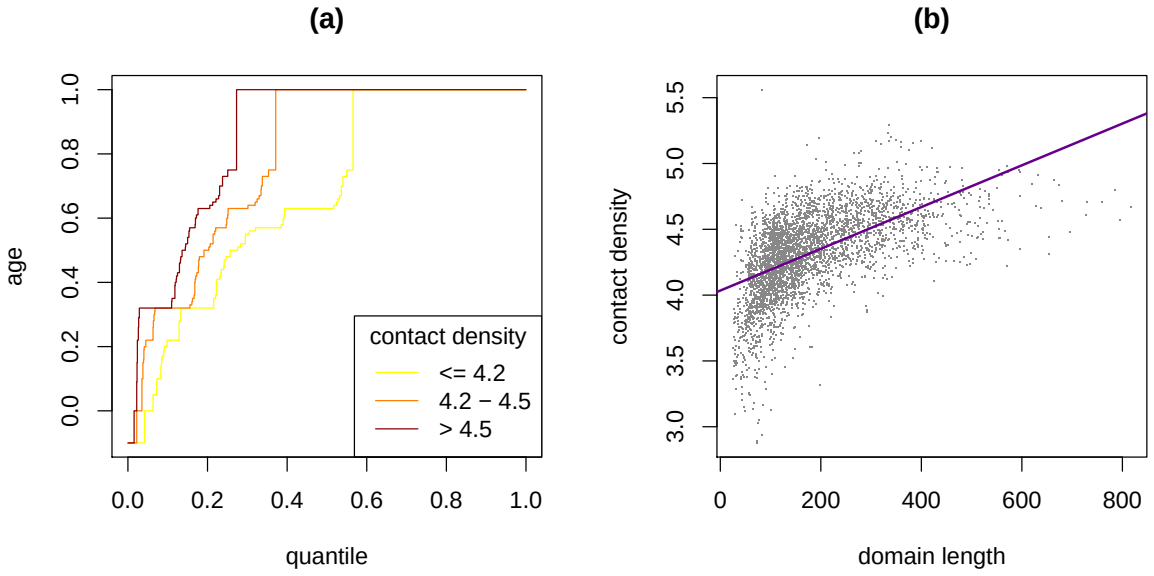


Figure 5.12: Contact density and age
(a) Dependence of superfamily age on contact density; domains with a high contact density correspond to superfamilies with older ages than domains with lower contact densities (p-value $< 2.2 \cdot 10^{-16}$). (b) Correlation between domain length and contact density (r-squared = 0.31).

Shakhnovich et al. (2005) proposed that Contact Density (CD) is higher for domains from older families; $CD = \frac{C}{N}$, where C is the number of contacts and N is the number of residues in a domain.

Figure 5.12 (a) shows that for our test set CD is indeed positively correlated with superfamily age. However, Figure 5.12 (b) shows that even though the number of contacts (C) is normalised by the domain length N, there is still a weak (non-linear) correlation between CD and N. As the correlation between domain length and superfamily age (Figure 5.7) is stronger, it is likely that Shachnovich's findings are merely due to domain length effects. Below we show that the expected number of contacts for a given domain length is not appropriately modelled by a linear correlation and we show how to calculate a length independent CD.

Normalisation

For a globular domain, we expect that the more residues a domain contains the larger the hydrophobic core and the smaller the fraction of surface residues. Surface residues make in general fewer inter-residue contacts than residues in the core of a fold. Therefore, the expected number of contacts for a protein of a given length should grow faster than linear.

To estimate the total number of contacts (C_{tot}), we can express it in terms of the length of a domain (N), the average number of contacts for a buried residue (C_{res}), the number of surface residues (N_{surf}), and the fraction of contacts lost for surface residues (f_{ext}):

$$C_{total} = C_{res}(N - f_{ext}N_{surf}) \tag{5.11}$$

For the purpose of normalisation it is desirable to obtain an expression for C_{total} in terms of N only. We will assume that a globular domain is approximately spherical. The volume of a sphere (V) with radius (R) is given by:

$$V = \alpha R^3 \qquad \text{with} \qquad \alpha = \frac{4}{3}\pi \tag{5.12}$$

Let V_{surf} be the volume of the outermost layer, containing residues which are in contact

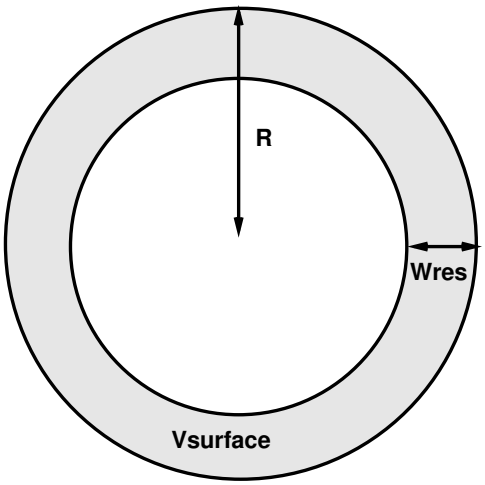


Figure 5.13: Spherical model of a protein.

with the surface and let the width of this layer be W_{res} (as shown in Figure 5.13). We

can then calculate the number of surface residues (N_{surf}) from the fraction of the surface layer volume with respect to the total volume of the entire sphere (V_{tot}).

$$N_{surf} = N \cdot \frac{V_{surf}}{V_{total}} \tag{5.13}$$

The total volume of the sphere V_{total} is given by:

$$V_{total} = V_{res} \cdot N \tag{5.14}$$

where V_{res} is the average volume for a residue. The volume of the outer layer V_{surf} is given by:

$$\begin{aligned} V_{surf} &= \text{volume outer sphere} - \text{volume inner sphere} \\ &= \alpha R^3 - \alpha(R - W_{res})^3 \\ &= \alpha(3R^2 W_{res} - 3RW_{res}^2 + W_{res}^3) \end{aligned} \tag{5.15}$$

The volume and width of a residue can be approximated by a cubic relation of the form:

$$V_{res} = \beta W_{res}^3 \tag{5.16}$$

From (5.12), (5.14) and (5.16) we have:

$$R = \sqrt[3]{\frac{V_{res}N}{\alpha}} \qquad\qquad = W_{res}\sqrt[3]{\frac{\beta N}{\alpha}} \tag{5.17}$$

Now substituting this result into (5.15), we get for the fraction of surface residues:

$$V_{surf} = \alpha W_{res}^3 (3(\frac{\beta N}{\alpha})^{\frac{2}{3}} - 3(\frac{\beta N}{\alpha})^{\frac{1}{3}} + 1) \tag{5.18}$$

$$\begin{aligned} N_{surf} &= N(\frac{V_{surf}}{V_{res}N}) = \frac{V_{surf}}{\beta W_{res}^3} \\ &= \frac{\alpha}{\beta}(3(\frac{\beta N}{\alpha})^{\frac{2}{3}} - 3(\frac{\beta N}{\alpha})^{\frac{1}{3}} + 1) \end{aligned} \tag{5.19}$$

Equation (5.11) becomes:

$$C_{total} = C_{res}(N - f_{ext}\frac{\alpha}{\beta}(3(\frac{\beta N}{\alpha})^{\frac{2}{3}} - 3(\frac{\beta N}{\alpha})^{\frac{1}{3}} + 1) \tag{5.20}$$

Hence the total number of contacts is dependent on N, with a non-linear relation. We can use this function to normalise C_{total}.

Normalised contact density

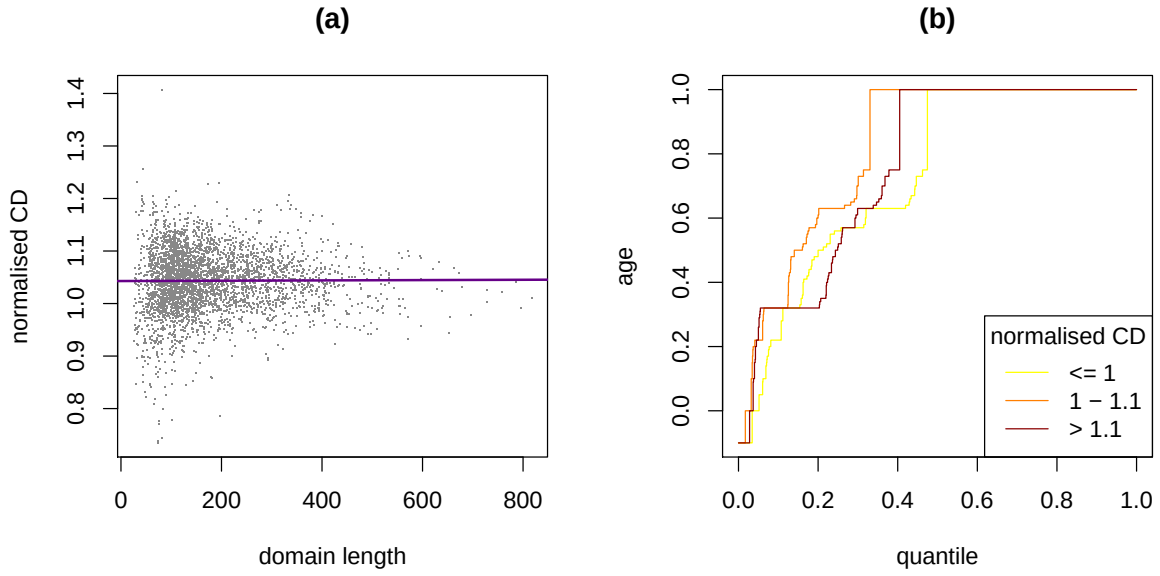


Figure 5.14: Normalised contact density
(a) Correlation between domain length and normalised contact density (r-squared = $3.1 \cdot 10^{-5}$). (b) Superfamily-age dependence of normalised contact density. Here equation 5.20 is used to normalise the total number of contacts, with $\beta = 1.0$, $C_{res} = 5.5$ as determined from the average number of contacts a buried residue makes and with $f_{ext} = 0.38$ (optimised to give a small length correlation).

Figure 5.14 (a) shows that after non-linear normalisation, the contact density has a negligible correlation with domain length. The correlation between contact density and age has also been significantly reduced (Figure 5.14 (b)), and it therefore seems likely that the correlation previously observed is mainly due to domain length.

Another way to create a length-independent measure of contact density, is by considering only the contacts of buried residues and normalising this (in a linear fashion) by the number of buried residues (Figure 5.15 (a)). Figure 5.15 (b) shows that there is in fact

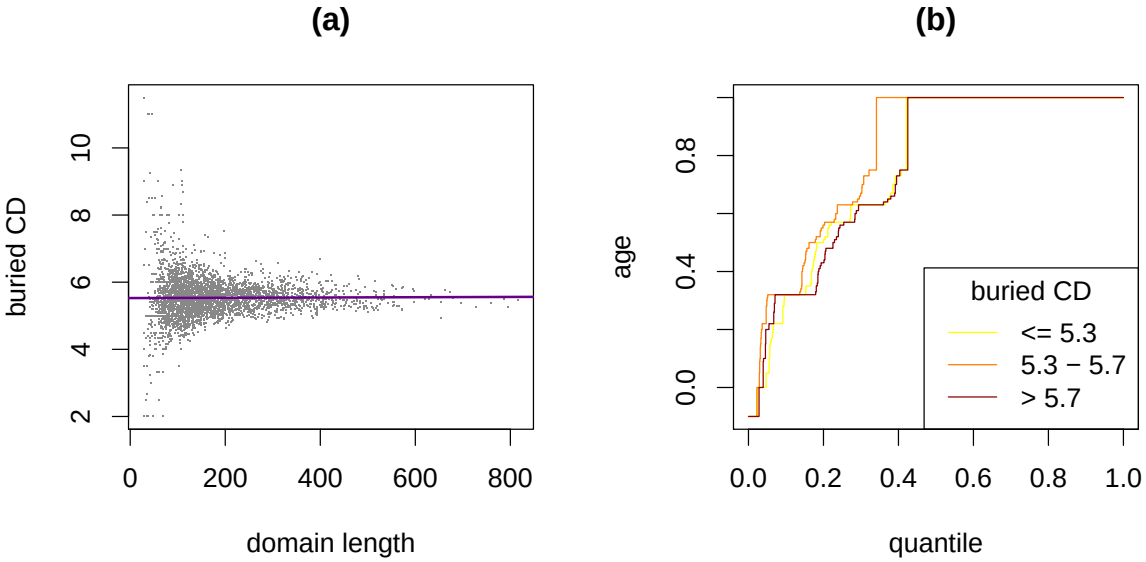


Figure 5.15: Buried contact density
(a) Correlation between domain length and buried contact density (r-squared = $5.36 \cdot 10^{-05}$). (b) Superfamily-age dependence of buried contact density.

no correlation between buried contact density and superfamily age. The age distributions behave as many uncorrelated variables: domains with buried contact density values close to the mean tend to be slightly older than domains at the more extreme ends of the distribution.

When considering the total number of contacts, a strong age-dependency is seen; the shift between age distributions is comparable to those divided by domain length. Similar shifts are also obtained for the following domain properties:

- the total number of buried residues
- the total number of buried contacts
- the proportion of buried residues in the total number of residues
- the sum of the total accessible surface
- contact order, normalised by length or not

Note that all these measures are strongly dependent on domain length and hence it is difficult to decide which of these measures, or in fact length itself, is causative.

5.3.6 Compactness

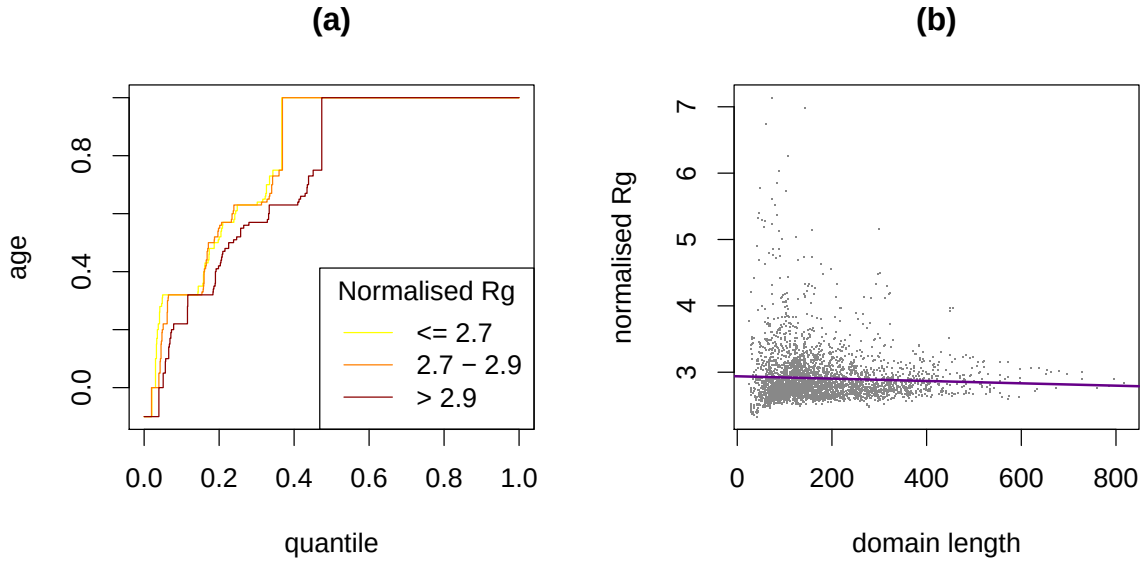


Figure 5.16: Compactness and age
(a) Age dependence of normalised radius of gyration: less compact domains tend to be younger (p-value $= 1.58 \cdot 10^{-08}$). (b) Correlation between length and normalised radius of gyration (r-squared $= 2.5 \cdot 10^{-3}$). The normalised radius of gyration is given as $\frac{R_g}{N^{1/3}}$.

The compactness or globularity of a domain can be investigated by comparing the radius of gyration of the domain to the expected radius of a perfect solid sphere. We can therefore estimate compactness for a given domain length (N) as:

$$normR_g = \frac{R_g}{N^{1/3}} \tag{5.21}$$

Figure 5.16 (a) shows that less globular domains tend to be younger. In addition Figure 5.16 (b) shows there is no dependence on length, but that some small domains seem to be extremely non-globular. The three domains with the highest normalised R_g score are all extended alpha helical domains, containing a long single alpha helix as their major structural component. The observed tendency of non-compact domains to be young may be due to such small stretched domains.

amino acid	propensity young	p-value	propensity old	p-value
Ala	**0.94**	1.8e-06	**1.06**	1.1e-50
Cys	**1.65**	1.3e-109	**0.78**	1.1e-152
Asp	0.95	0.0013	1.01	0.032
Glu	**0.82**	2.1e-34	**1.02**	8.9e-08
Phe	0.97	0.13	0.99	0.28
Gly	**0.93**	5.1e-08	**1.03**	1.1e-11
His	**0.81**	1e-15	**1.03**	8.7e-05
Ile	0.97	0.1	**1.04**	7.8e-20
Lys	1.00	0.9	0.99	0.025
Leu	**0.90**	4.2e-14	**1.02**	6.5e-07
Met	**0.89**	5e-06	**1.04**	3e-06
Asn	**1.30**	8.6e-62	**0.97**	2.6e-08
Pro	1.05	0.0045	0.99	0.071
Gln	1.06	0.0034	**0.94**	1e-22
Arg	**0.92**	3.1e-06	1.02	0.0014
Ser	**1.13**	6.5e-17	**0.92**	6e-77
Thr	**1.22**	2e-42	**0.94**	2.3e-32
Val	**0.94**	4.2e-05	**1.04**	1.2e-18
Trp	1.06	0.038	**0.95**	3.2e-08
Tyr	**1.07**	0.00029	**0.97**	7.3e-08

Table 5.2: Amino acid propensity and superfamily age
Propensity values that are typeset in bold font are statistically significant. Propensities > 1.0 indicate that the amino acid is more likely to occur in the given age group, whereas propensities < 1.0 indicate that the amino acid is less likely to occur in the given age group. Domains with a superfamily age equal to 1.0 are considered to be *old* and domains with a superfamily age < 0.2 are considered to be *young*. See Section 5.2.4 for further details.

5.3.7 Sequence composition

Table 5.2 shows amino acid propensities for amino acids in different age groups. Cys, Asp, Thr, Ser and Tyr are associated with young superfamilies. All these amino acids belong to the class of amino acids that are polar but not charged. The drastic over-representation of cysteine may be explained through the need for disulphide bridges to stabilise domains in the class of 'small proteins'. However, other amino acids with properties that are typically associated with this class such as metal-binding, do not show higher propensities for young superfamilies.

On the other hand, Ala, Val, Met, Gly ,Ile, His, Glu and Leu are associated with older superfamilies. Most of these amino acids are hydrophobic, with exception of Glu and His.

Three polar amino acids (Asp, Glu and Arg) show a slight tendency towards older

superfamilies, but this shift is only significant for Glu; no shift either way is observed for Lys.

No obvious links between these results and amino acid propensities for secondary structure elements can be made (helix, parallel strand, anti-parallel). Amino acids which have previously been suggested to have arisen first during the evolution of life (Trifonov, 2004) do tend to be associated with older superfamilies, but this correlation is in no sense strict. Hydrophobic residues associated with high GC-content (Perl et al., 2000) tend to have propensities for old superfamilies. Perhaps the strongest discriminative property for amino acids associated with younger superfamilies is their fragility in high temperature, high pressure environments (Bernhardt et al., 1984).

5.4 Discussion and conclusions

5.4.1 Sensitivity and superfamily age

Very high false negative rates for PSI-BLAST searches at superfamily level may be cause for some concern, as this shows that very few superfamily relations are found by PSI-BLAST without a family member in the set of search domains. This may bias superfamily ages strongly towards those families present in SCOP.

On the other hand, superfamily ages are calculated from occurrence patterns. It is therefore more important to correctly predict if a superfamily occurs on a genome than how many copies (or different families) the superfamily has on the genome. It has been observed that genome occurrences of superfamilies are more reliable than individual assignments (Chapter 4). In addition, SAM-T98, used for both SUPERFAMILY and Gene3D assignments, is thought to be more sensitive than PSI-BLAST. Ages estimated by these assignment sets correspond well to those estimated by PSI-BLAST.

Even so, we may have to consider superfamily occurrence patterns, as a simple union of occurrence patterns from families found in SCOP.

5.4.2 What does it mean to be an 'old' superfamily?

So far we have considered a superfamily to be 'old' if it was already present early on in evolution, or more precisely in the last common ancestor of at least two of the three kingdoms (archaea, bacteria, eukaryotes).

However many domain properties which are found to be associated with an old superfamily age are at first sight counter intuitive: why would the first domains be longer and more complex (parallel beta strands, are usually less local than anti-parallel beta strands)?

Looking at the description of an old superfamily from the perspective of the domain, we can turn the previous statement around and say that it is possible to trace its superfamily back to early on in evolution; in other words it belongs to a superfamily for which many evolutionary links could be generated. Whereas for a domain associated with a younger superfamily age, its evolutionary path can not be traced back further than itself or a recent ancestor. The question of what the differences are between young and old folds may be rephrased thus: why can the origins of some domains be traced further back in evolution than others?

Firstly, it may be, that an entirely new structural domain was formed at some point in evolution. In this scenario, the new domain is likely to be simpler and shorter than most existing domains.

Secondly, there may be an evolutionary path of a domain with a young superfamily age all the way up to the last universal common ancestor, but this has not been found. This may be because the fold recognition software has not found homologs in distantly related species. False negative rates of search domains indeed show a weak anti-correlation with superfamily ages (Figure 5.4), though these can not explain all the differences observed between young and old superfamilies. Alternatively, this may be due to an unfound evolutionary link between two superfamilies in SCOP. It is quite reasonable to argue that it is easier to find (structural) evidence for homology in larger domains.

5.4.3 Stability

Many of the domain properties associated with an old superfamily age are those which give rise to a more stable fold. It is plausible that evolutionary links between domains can more easily be detected for domains that remain relatively stable during evolution.

For example, longer domains have more contacts and relatively larger hydrophobic cores, increasing thermodynamic stability. In addition parallel beta-strands provide a very stable scaffold (Chou et al., 1983). Increased secondary structure content is also associated with both stability and older ages. Furthermore, non-globular structures are likely to be less stable and are associated with younger ages.

An alternative explanation for the apparent link between stability and older age may be the living environment of LUCA. This is thought to have been considerably warmer than the environment of most currently living species (Di Giulio, 2000).

A last explanation may be that newly developed superfamilies are in fact 'a quick fix', whereas domains belonging to older superfamilies have had time to find the most optimal confirmation. This would agree with the finding that the values around the mean of many structural properties are associated with an older age.

5.4.4 Conclusions

We compared the superfamily ages of domains to several structural properties while considering if the reliability of the fold recognition method, used to estimate the ages, caused the effect.

A longer domain length is strongly associated with an older superfamily age, whereas short domain length is associated with a young superfamily age. This difference cannot be explained entirely by the differences in false negative rates for different length domains. At this point it is unclear if the dependence on age is due to domain length, the number of contacts, the size of the surface or the relative size of the hydrophobic core.

Contact density, in contrast with previously reported work, has not been found to correlate with superfamily age, other than the correlation caused by domain length de-

pendence.

Secondary structure content is also strongly associated with age. A large fraction of parallel beta strands is associated with an old superfamily age; anti-parallel beta strands on the other hand are associated with young superfamilies. The fraction of helices tends to be higher in older superfamilies.

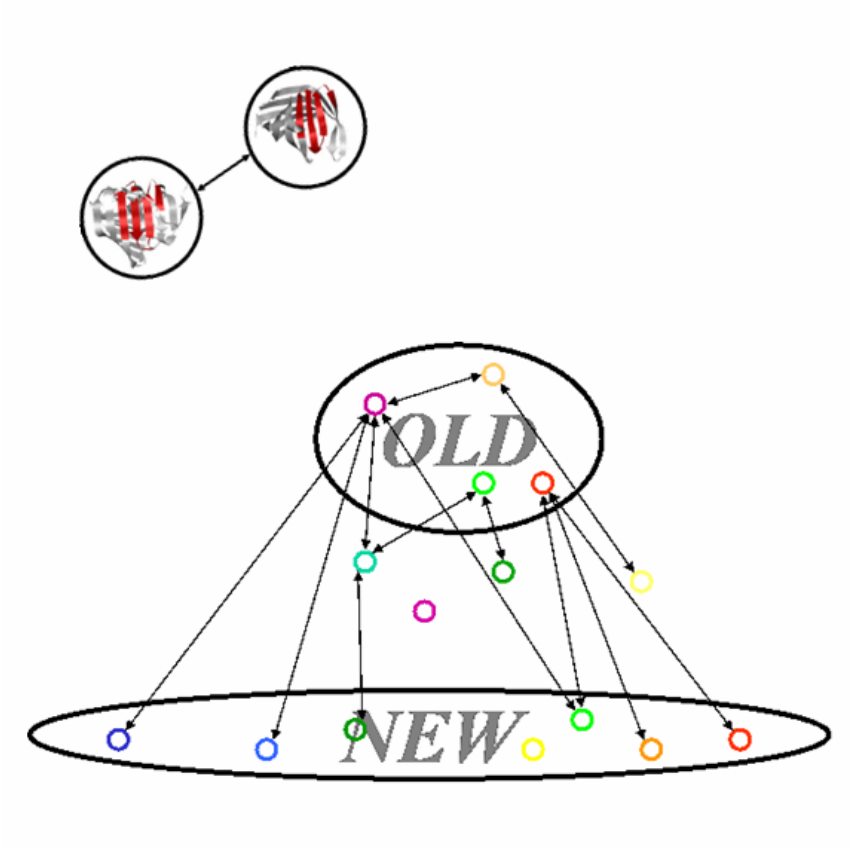
The position of the endpoints of the peptide chain appear to have changed slightly over evolution: interestingly this effect is seen more strongly for the N-terminus than for the C-terminus. In addition, a large distance between the termini is associated with young superfamilies. These results may correspond to N-terminal folding.

Lastly older superfamilies tend to contain relatively more hydrophobic residues, and young superfamilies relatively more polar (but non-charged) residues.

Most of these results indicate that the thermodynamic stability of a fold is increased in older superfamilies. At this point it is not clear whether old superfamilies have had more time to find an optimal confirmation, if the stability results in fewer changes during evolution or if perhaps the last universal ancestor lived in an environment that required more stable protein structures.

Chapter 6

Evolutionary relevance of structural fragments



6.1 Background

At present there is no universal understanding of how proteins can change topology during evolution, and how such pathways can be determined in a systematic way. The ability to create links between fold topologies would have important consequences for structural classification, structure prediction and homology modelling.

Here we consider methods to create links between domains above SCOP's fold level. In particular, we investigate links generated by a method that identifies similar structural fragments between domains. We correlate the number of links generated by this method with the superfamily age of a domain and discuss the evolutionary relevance. The results suggest that exchange, insertion and deletion of structural fragments is a likely evolutionary pathway to change the topology of a fold.

6.1.1 Geometric links

The authors of SCOP have recently indicated that a small proportion of structural and evolutionary relationships cannot accurately be reflected by the hierarchical structure of SCOP. Andreeva et al. (2007) have collected a small database, SISYPHUS, of examples where SCOP's classification alone may not represent structural relations between domains adequately. SISYPHUS contains examples of circular permutations, chameleon sequences (folding into different structures) and shared fragments/subdomains between domains of different folds.

Several methods based on geometrical measures have been proposed to indicate links between topologies. CATH's architecture level (Orengo et al., 1997) is an example of such an indicator. There is, however, for this measure currently no reason to believe that it reflects evolutionary relatedness between proteins. The structure of a protein is thought to be more conserved than its sequence, because the protein needs to fold into a stable structure to perform its function (Murzin, 1998). At the architecture level connections between secondary structure elements are not taken into account; folds in the same architecture with different connectors may well follow different folding pathways.

There are several geometrical measures that do keep the sequential information. To examine structural similarities between different folds in CATH, Harrison et al. (2002) developed a graph-based structure comparison algorithm (GRATH). In GRATH, graphs are built that consist of nodes representing secondary structure elements, and labelled edges representing their relative orientations. Matches are found by comparing the two largest common cliques between the folds, where the sequential topology of the nodes match. After normalising clique sizes, it was found that the most gregarious folds are from CATH's alpha-beta class.

Other fold comparison measures can be generated by general structural alignment programs. Holm and Sander (1993) used the structural alignment program, DALI, to generate an all against all comparison of protein structures. They showed that it was possible to separate folds into different structural classes, by using two principal eigenvectors of the similarities within the set of structures. Later it was shown, that in fact the four main classes as defined by SCOP (alpha, beta, alpha/beta, alpha+beta) could be separated by adding a third component (Hou et al., 2003). In the same comparison Holm and Sander (1993) also noted the existence of several folds which show gregarious behaviour, sharing super-secondary structure elements with many other folds. Shindyalov and Bourne (2000) used another structural alignment program, CE (Shindyalov and Bourne, 1998), and found that many sub-domains are shared between protein structures belonging to different folds.

By examining fold topologies in more detail, schemes have been developed that can sub-classify folds. Most of these schemes, work by adding or removing a few secondary structure elements to walk from one fold to another. Such schemes can create hierarchies for protein folds (Efimov, 1995, 1997; Matsuda et al., 2003) or a table of protein structures, that lists all topological possibilities in a systematic way (Taylor et al., 2002).

More recently, Friedberg and Godzik (2005) developed a method, Fragnostic, based on similarity between structural fragments of domains to create links between folds. In their work connections between folds form a network. Different networks were created with fragments of 5, 10 , 15 and 20 residues. The authors showed that folds which share

fragments are also likely to belong to similar functional categories.

Despite most of these measures being based on the idea of structure evolution, it is extremely difficult to show if such measures indeed reflect evolutionary relatedness. Here we use our previously developed age measure for protein superfamilies (Winstanley et al. (2005), Chapter 3) to investigate the relationship between structural fragments and protein structure evolution. Initially we show that the number of links with other folds as determined by Fragnostic correlates with the age of a fold. Folds with many links are generally older.

However, part of the method used to create the Fragnostic network relies on sequence identity. Since we want to make sure that no bias is caused by commonly used sequences in the fragments and the recognition of folds on the genomes to determine the age, we develop a new protocol to create fragments based on structure only. Links from these structural fragments show slightly weaker correlations with fold age; the correlation becomes stronger the longer the fragment length, indicating that structural fragments may indeed reveal evolutionary links between domains.

6.1.2 Fragnostic

Fragnostic (Friedberg and Godzik, 2005) uses a set of PDB structures with less than 25% sequence identity. All fragments (containing 5, 10, 15 or 20 residues) within these structures are compared in a pairwise fashion.

A prefilter was used to make the computational task of scanning all fragment pairs in the set possible. For each protein a sequence profile was created using the non-redundant sequence database (NR). The sequence profile of each fragment was tested for similarity against that of every other fragment in the set with the profile-profile scoring function of FFAS (Jaroszewski et al., 2005); fragment pairs with a p-value < 0.001 passed this pre-filter test. Finally, the minimal Root Mean Square Distance (RMSD) between the Cα atoms of these pairs were calculated. The pair is added to the fragment database if the RMSD was smaller than 1.0 Å.

To establish links between folds based on these pairwise fragments a normalisation procedure was used, to account for frequently occurring structural patterns shared between many folds (e.g. secondary structure). Multiple networks were created by using different cut-off levels on $f(i, j)$ for folds i and j as defined below (Friedberg and Godzik, 2005):

$$f(I, J) = \frac{Sim(I, J)}{min(Sim(A, I), Sim(A, J))} \text{ if } I \neq J \tag{6.1}$$

Here Sim(I,J) is the number of shared pairwise fragments between two sets of domains (e.g. superfamilies), and A is the set of all domains.

6.1.3 Fragment based structure modelling

Fragnostic is certainly not the first database to describe fragments of protein structure. Fragments have been used extensively to model protein structures by recreating structures from fragments of other protein structures. This can be achieved in many different ways with many different sizes of fragments (Jones and Thirup, 1986; Summers and Karplus, 1990; Levitt, 1992). Du et al. (2003) show that in a fixed set of PDB structures , filtered at 96% sequence similarity, 91% of all fragments with 15 residues can be represented by at least one other fragment in the set, with a structural similarity of 2.0 Å RMSD. The smaller the fragment size the higher the coverage. This shows that in principle it should be possible to model new folds with fragments of existing structures.

Fragments have been particularly useful for modelling protein (regions) without a known homologous structure. Initially fragments of known structures were used to model small regions, which lack a good alignment to a known structure (e.g. loops) (Rufino et al., 1996; van Vlijmen and Karplus, 1997). Now, the most successful methods for 'ab initio' structure prediction use a fragment based approach, e.g. ROSETTA (Simons et al., 1999). ROSETTA generates many decoy structures by combining fragments of known structure while minimising an energy function. The best decoys, with the lowest energy scores and broadest energy minima, are selected from the large set of structures,

to be tested by further knowledge based scoring functions.

Many of these fragment-based methods search for fragments with similarities to those of the modelled sequence; e.g. ROSETTA creates a sequence profile and secondary structure prediction for the entire chain and matches fragments of the profile to sequence and secondary structure of known proteins.

6.1.4 Fragments and protein fold evolution

It may be useful to question why fragment based modelling approaches are so successful. For small fragment lengths, existing fragments may indicate favourable confirmations. Moreover, it has proved difficult to develop effective energy functions that correctly predict the folding of protein structures. Using short fragments can be seen as a knowledge-based way of implementing an energy function that favours frequently observed confirmations (due to physical forces).

On the other hand, for longer fragments there may be some evolutionary relevance. It has been proposed that protein structures may evolve through insertions of structural fragments (Grishin, 2001) or that in fact modern protein structures are built from ancient short peptide fragments (Lupas et al., 2001; Trifonov and Berezovsky, 2003; Sobolevsky and Trifonov, 2006). Considering, that a general trend in genome evolution consists of copying fragments of the genome, it is not unlikely that a similar process occurs at all fragment sizes of protein sequence. In fact one may expect a similar correlation between fold age and the number of fragments a domain shares with other domains, as previously has been shown between fold age and copies of the entire domain. Figure 3.4 shows that superfamilies with more homologs on a set of completed genomes are generally older than superfamilies with fewer homologs, but that old superfamilies do not necessarily have many homologs. The correlation may be explained by a process where the rich get richer: older structures have had more time to duplicate and may create a large number of copies similar to the domain (or part of it) (Qian et al., 2001; Abeln and Deane, 2005).

Currently the most widely used structural classifications are hierarchical. As discussed

above several studies have suggested that this may not be the best representation to show structural similarities. If proteins can indeed change their topology during evolution through various fragment insertions and deletions, a network representation may be more suitable. One can compare this to the classical view of representing the evolution of species by a (phylogenetic) tree. Lateral gene transfer between species can create additional links between nodes in the tree, creating a network instead. Lateral gene transfer is relatively rare, as DNA of different species does not usually mix. Proteins on the same genome, however, are constantly in close proximity, and insertions and deletions of DNA fragments are therefore not unlikely as long as the protein retains its ability to fold. Hence a network topology may be a more suitable representation of structural and evolutionary similarities between folds than a strict hierarchy.

6.2 Methods

In this study fragments of 10, 15, 20 and 30 residues are used; pairwise fragments indicate structural similarity between SCOP domains (see Section 6.2.5) detected in an all-against-all comparison. The method differs on two major points from the one used to produce Fragnostic. Firstly, no sequence based pre-filter is used to estimate structural similarity in a pair of fragments. Secondly, for evolutionary reasons we are especially interested in longer pairwise fragments, of 20 and 30 residues. It may be difficult to look for similarity between such fragments without considering gaps, since they are likely to contain more than one secondary structure element. Loop regions, connecting the secondary structure elements, are prone to insertion and deletion.

As RMSD superposition is computationally expensive, it is desirable to create a pre-filter that can estimate structural similarity between two fragments and that can also give an alignment of fragments allowing small gaps. The structural alignment program Mammoth (Ortiz et al., 2002) already contains a prefilter with these properties; this filter formed the basis for the first stage of our search for pairwise fragments with structural similarity.

6.2.1 Mammoth

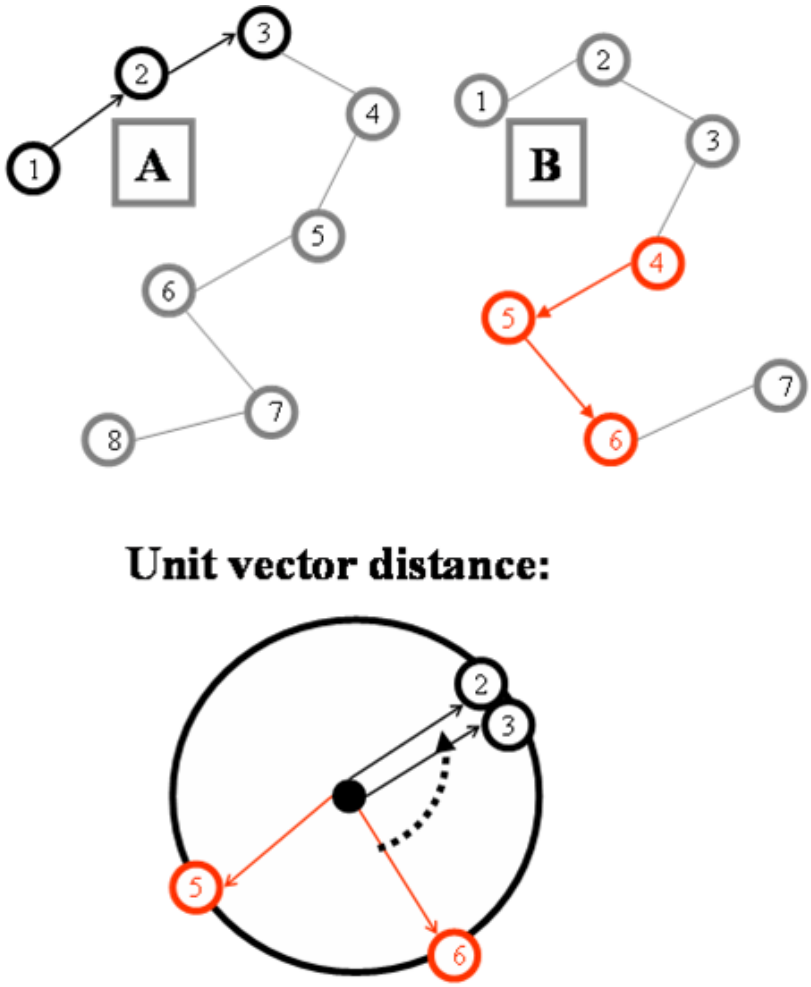


Figure 6.1: Schematic URMS distance
A schematic representation of the URMS distance in two dimensions. For each center residue a set of Cα vectors of neighbouring residues is generated (2 neighbouring residues in the Figure, 6 in Mammoth). The minimum URMSD between two sets of Cα vectors is calculated through a simple rotation procedure, superimposing the unit vectors on a unit circle (or sphere in three-dimensions).

Mammoth (Ortiz et al., 2002) is a structural alignment program originally designed to compare experimental and predicted protein structures. However, many groups have since used it for general structure comparison. In particular, Krishna and Grishin (2005) use Mammoth to find structural similarities with a focus on evolutionary relationships.

The structural superposition of two proteins is generated in three main steps. First a similarity matrix between all residues of the two proteins is calculated based on a local distance measure: the unit-vector root mean square distance (URMSD, see Figure 6.1). Next the URMSD similarity matrix is used to globally align the two proteins. In the

final stage a maximum subset of similar local structures is searched for by the MaxSub algorithm (Siew et al., 2000).

To generate our set of fragments, only the first stage of mammoth is used, which creates the similarity matrix. This matrix is then used to align residues within a given fragment window.

The values in the URMSD similarity matrix indicate the pairwise distances between two sets of Cα unit vectors, covering a heptapeptide around the central residues. The URMS distance has a useful property: it is bound by a natural upper limit (Chew et al., 1999). Furthermore, it is possible to theoretically obtain an expected minimum distance, $URMS^R$ (Chew et al., 1999):

$$URMS^R = -\sqrt{2.0 - \frac{2.84}{\sqrt{n}}} \qquad \text{here} \qquad n = 7$$

A similarity score S_{AB} between between the two sets of Cα unit vectors A and B can then be calculated as (Ortiz et al., 2002):

$$S_{AB} = \frac{(URMS^R - URMS^{AB})}{URMS^R} \Delta(URMS^R, URMS^{AB}) \tag{6.2}$$

where

$$\Delta(URMS^R, URMS^{AB}) = \begin{cases} 10 & \text{if} \quad URMS^R > URMS^{AB} \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

We used the FORTRAN code from the Mammoth program to generate the URMSD based similarity matrix. This matrix was then applied to find optimal local alignments for all possible pairwise fragments between two given protein structures.

6.2.2 Alignment of fragments

Classical global alignment is a well defined problem: find the highest scoring alignment between two sequences in which all residues are part of the alignment, given a similarity measure and gap penalty. Global alignment can be solved exactly by the algorithm of Needleman and Wunsch (1970). Similarly, the solution to a local alignment is defined as the highest scoring sub-alignment of two sequences and can be solved by the algorithm of Smith and Waterman (1981). Both problems are solved using a dynamic programming approach. In the process of finding the optimal solution, sub alignments for each possible starting point are dynamically calculated and stored in an alignment matrix. The optimal alignment can be found by tracing back the path that resulted in the optimal score for local alignment, and by tracing back from the bottom-right corner towards the top-left corner for global alignment.

Our alignment problem, however, is slightly different as we want to find an alignment of fixed length (N) and a maximum number of gaps (G), higher than a given score (T). Since it is desirable to limit the number of total alignments, we would also aim to search for locally optimal alignments so that not every single variant of a gap displacement is included in the set of pairwise fragments.

In order to achieve this, local alignment is performed on windows associated with the fragment length (N) and maximum number of gaps (G); see below for more details. Gotoh's algorithm for affine gap penalties is used to improve speed (Gotoh, 1982).

Since the URMSD similarity matrix gives us (semi) continuous values, rather than discrete values as for DNA or protein sequences, the chance of multiple optimal paths is relatively small. Hence by tracing the matrix for a given fragment we ignore such multiple paths and always give preference to directions in a specific order (Diagonal, Up, Left).

For each possible fragment pair window, an optimal alignment is given based on the URMSD matrix. Penalties for gaps in the alignment are subtracted from this score. Following Mammoth, gap penalties were set with the cost of opening a gap at 7.0 and extending a gap at 0.45. The maximum allowed number of gaps (G) is 20% of the

fragment length (N). Only pairwise fragments with a score higher than a given threshold (see Section 6.2.4), are considered for RMSD superposition.

Window length

The alignment procedure is calculated on a set of fragment windows, which are in effect sub-matrices of the alignment matrix over the entire length of the two proteins. In a simple approach, all possible fragment windows for a given fragment length (N) could be used. For two proteins of length A and B the number of such windows (W) is then given by $W = (A-N)(B-N)$. Each window would have $E = N^2$ elements. The total number of elements to be computed (E_{tot}) for the two proteins is then given by:

$$E_{tot} = W \cdot E = (A - N)(B - N)N^2 \tag{6.4}$$

To save time on this computationally expensive task, the number of elements and the number of fragment windows used were limited by the following approach.

Firstly, a simple reduction for the number of elements to be calculated can be obtained by limiting the search space to the maximum allowed gap length (G). It is impossible to find any allowed alignments further away from the diagonal than G steps, hence the total number of elements in each alignment window can be reduced to:

$$E = N(2G + 1) \tag{6.5}$$

The total number of elements to be calculated is now given by:

$$E_{tot} = W \cdot E = (A - N)(B - N)(2G + 1)N \tag{6.6}$$

Since we have set $G = \frac{1}{5} \cdot N$ for our fragments, equation 6.6 gives a smaller value for E_{tot} than equation 6.4.

The number of alignment windows can also be limited. However, we do not want to lose any optimal fragment pairs by doing so, hence the following condition is imposed:

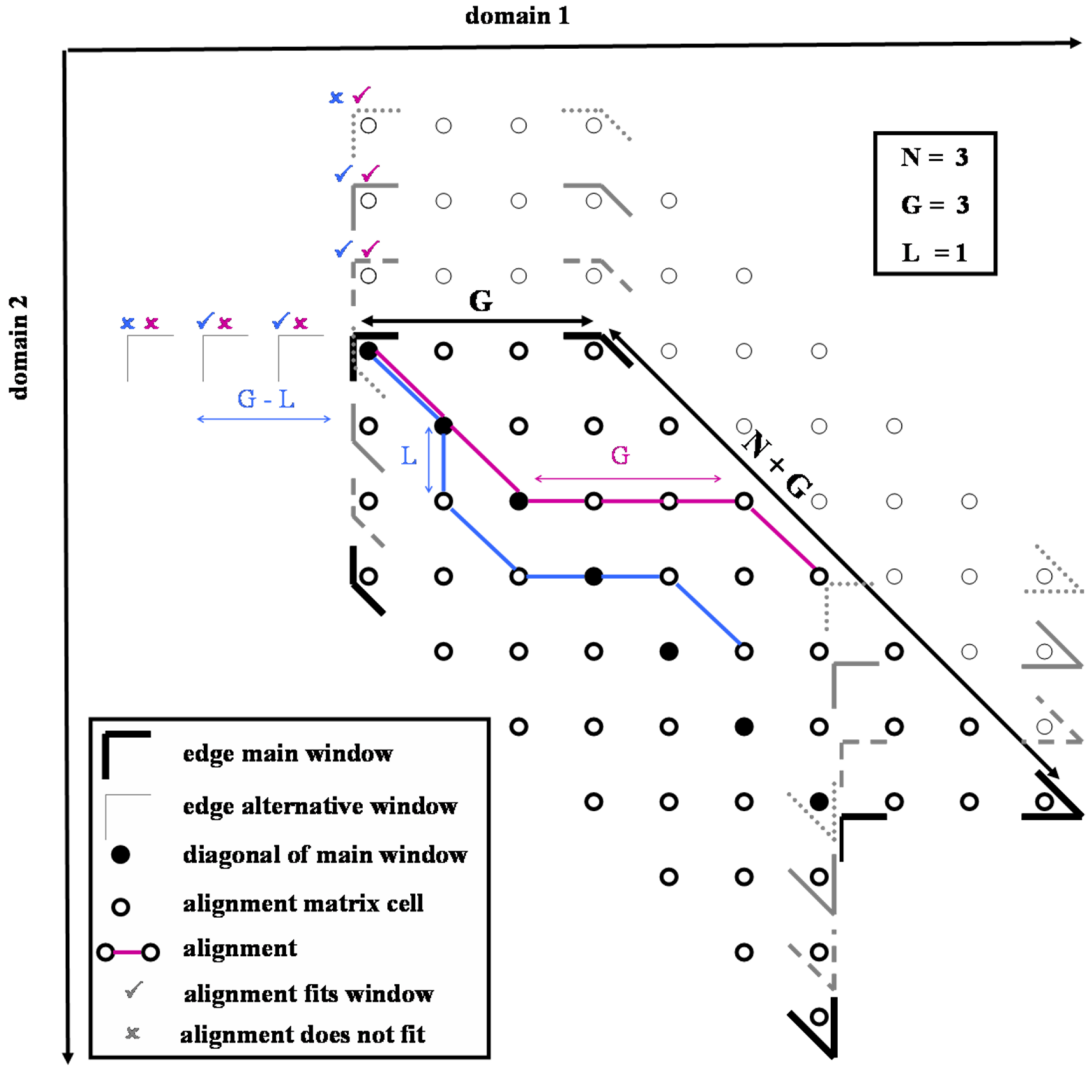


Figure 6.2: Pairwise fragment window. A typical pairwise fragment window is depicted by black arrows (G and $N+G$) and bold black lines mark the corners of the window, the diagonal is represented by black dots. Parameters for the fragment window are given by fragment length (N=3), maximum gap length (G=3) and maximum off-diagonal distance on smallest side (L=1). The fragment window may be considered as a submatrix of the alignment matrix over the entire length of the two domains.
The figure shows two possible alignments both with the maximum number of gaps; one extends in a single direction off the diagonal (violet) and one extends in both directions of the diagonal (blue). Both alignments fit within the given fragment window. Diagonal connections in the alignment represent match states, whereas horizontal and vertical connections represent gaps in the alignment.
Now consider a fragment window one step above the main window, the corners are depicted in a grey striped line. Both alignments still fit in this fragment window. The next window (grey continuous line) also contains both alignments. The third window up from the main window (grey dotted line) does contain the alignment extended into one direction (violet), but the alignment extended into two directions (blue) does not fit in this window. A similar approach can be taken by moving leftwards or diagonal. $G-L$ is the limiting distance, if we would like to skip as many windows a possible.

Condition:

Each possible alignment, given N and G, should occur in at least one of the alignment windows.

The following observations can then be made:

Observation 1:

Each alignment window is limited by distance G from the diagonal.

Observation 2:

If an alignment would only defer from the diagonal in a single direction, the alignment could also be found in the next window, starting at $i + G + 1$

Observation 3:

The side with the smallest maximum distance to the diagonal is the limiting factor, say L, and so we can start the next window at $i + (G - L) + 1$ *(see Figure 6.2)*

Using Observation 3 above, we can describe L as follows:

$$L = \max \{x : \min \{|a - b|, b\} \, | a + b = G\} \tag{6.7}$$

where a and b are the number of gaps in a single direction . We get a maximal L if $|a - b| = b$, hence if $G = 3b = 3a$. We can now define L for a given G through integer division:

$$G = 3L + R \qquad 0 \leq R < |L| \qquad G, L, R \in \mathbb{Z} \tag{6.8}$$

The number of windows (W) used in this new approach is given by:

$$W = \frac{(A - (N + G))(B - (N + G))}{(G - L + 1)^2} \tag{6.9}$$

A drawback of this method is that we need to extend the search along the diagonal by G such that the total points in the fragment alignment matrix becomes $(2G+1)(G+N)$, where previously we had, $(2G + 1)N$.

$$E_{tot} = W \cdot E = \frac{(A - (N + G))(B - (N + G))}{(G - L + 1)^2}(2G + 1)(N + G) \tag{6.10}$$

Since $G = 1/5N$ is used, the new method saves computations if we compare E_{tot} as

given by equation 6.10 to E_{tot} as given by equation 6.6.

6.2.3 Fragment superposition

In the second stage of the structural fragment comparison, each fragment pair with a satisfactory URMSD alignment score (see Section 6.2.4) is passed on to a superposition algorithm that minimises the RMSD between the two fragments. Since the alignment is already given, the problem reduces to minimising the RMSD of paired three-dimensional coordinates, in this case the aligned Cα atoms. A transformation needs to be found that retains the relative positions of the coordinates in both fragments. This transformation can be described as a translocation and a subsequent rotation. The translocation is simply found by superimposing the centres of mass of each fragment.

The second transformation can be solved as a least squares problem, with the additional restriction that the transformation matrix needs to describe a rotation. However, these rotational restraint are non-linear. Kearsley (1989) shows that by adding an extra dimension (using quaternion algebra), this problem can be rewritten as an eigenvalue problem in four-dimensions. The eigenvalue problem that arises from this method was solved by the Jacobi algorithm (Press et al., 1992). The program was implemented by using code written by Charlotte M. Deane.

6.2.4 Prefilter URMSD alignment scores

N	**10**	**15**	**20**	**30**
G (maximum gap)	2	3	4	6
URMSD alignment score	40	70	100	150
RMSD - initial (Å)	1.0	1.5	2.0	3.0
RMSD - final (Å)	0.8	1.0	1.5	2.0

Table 6.1: Fragment thresholds
Thresholds used in the process of selecting pairwise structural fragments. URMSD alignment score reflects the alignment score above which a fragment pair passes the prefilter test. Information for all fragments with an RMSD smaller than the 'initial' cut-off fragment are stored. The 'final' RMSD cut-off is used to generate the set of fragments that create links between domains.

To investigate the evolutionary relevance of the URMSD measure, 50 domain pairs at

family level, 50 pairs at superfamily level, 50 pairs at fold level and 50 at class level and 50 random pairs were chosen. For each set of pairs local optimal scores for all fragment windows were calculated by the prefilter, and all aligned fragments generated by the prefilter were subsequently superimposed to obtain a minimal RMSD.

Figure 6.3 shows that the correlation between the URMSD alignment score and the RMSD becomes stronger at lower levels in SCOP's hierarchy. Moreover, the correlation becomes stronger at low RMSD scores.

The cut-offs for the URMSD alignment scores from the pre-filter were chosen, so that more than 95% of all fragments pairs with an RMSD lower than the initial RMSD cut-off would be included (see Table 6.1).

6.2.5 Data Set

The number of domains (D) in the dataset affects the computational time with $O(D^2)$, hence a reduction in the number of domains will have a great impact on the computational time.

The objective of the fragment set is to compare fragment based links to fold or superfamily ages. These ages are determined at domain level and to make comparison more straightforward structures of SCOP domains are used in the dataset, rather than entire PDB files.

A subset of protein domains, as defined by SCOP 1.69, was extracted from the ASTRAL database (Brenner et al., 2000). The authors of the ASTRAL database indicate that a good structure will have an aero-spaci score > 0.4. Figure 6.4 shows the effect of an aero-spaci cutoff on the diversity within the structural set.

To create a minimal representative set of domain structures, the following conditions were imposed:

1. structures have less than 95% sequence identity
2. structures have an aero-spaci > 0.5
3. only the structures with the highest aero-spaci score in each SCOP family are kept

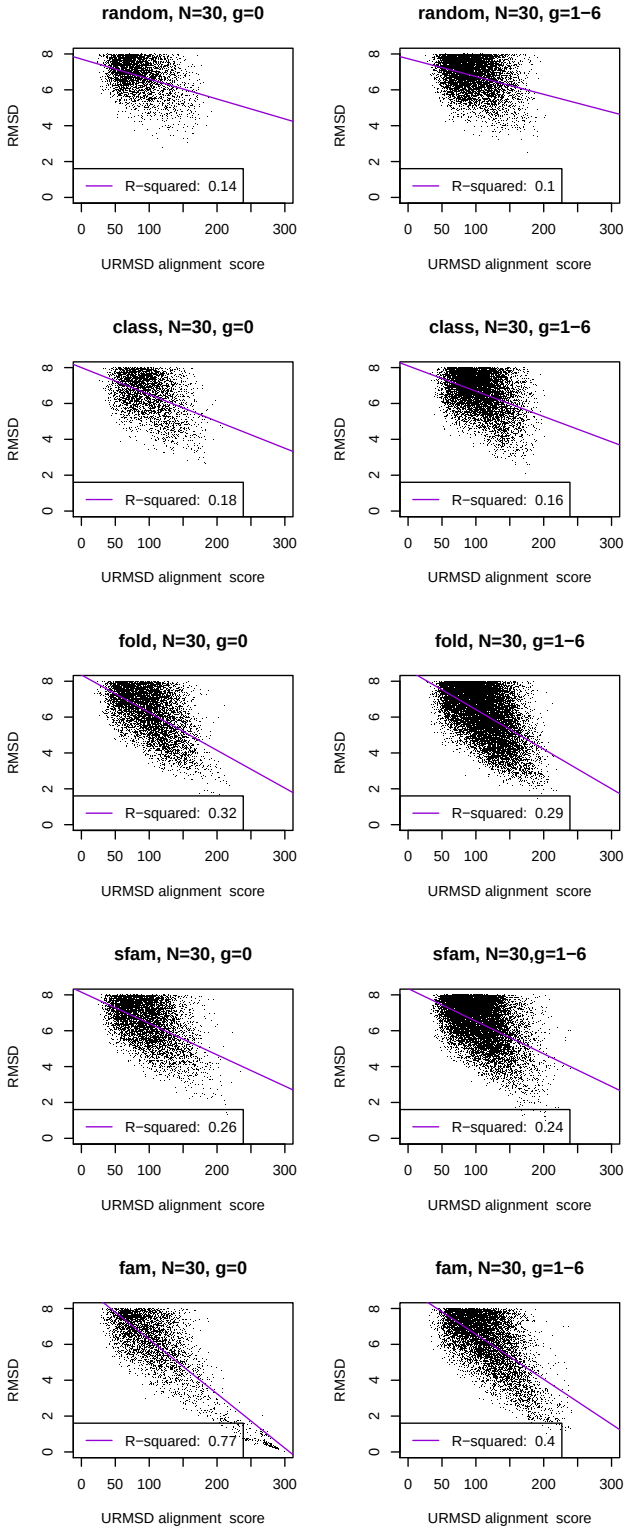


Figure 6.3: URMSD score
The URMSD alignment score versus the RMSD of pairwise fragments of 30 residues. Pairs are subdivided into categories of family-family (fam), superfamily-superfamily (sfam), fold-fold, class-class, random-random relationships. Columns 1 and 2 show the alignments which contain 0 ($g = 0$) gaps and 1-6 ($g = 1 - 6$) gaps respectively.

This gives a total of 1107 domains to be used for pairwise fragment comparison.

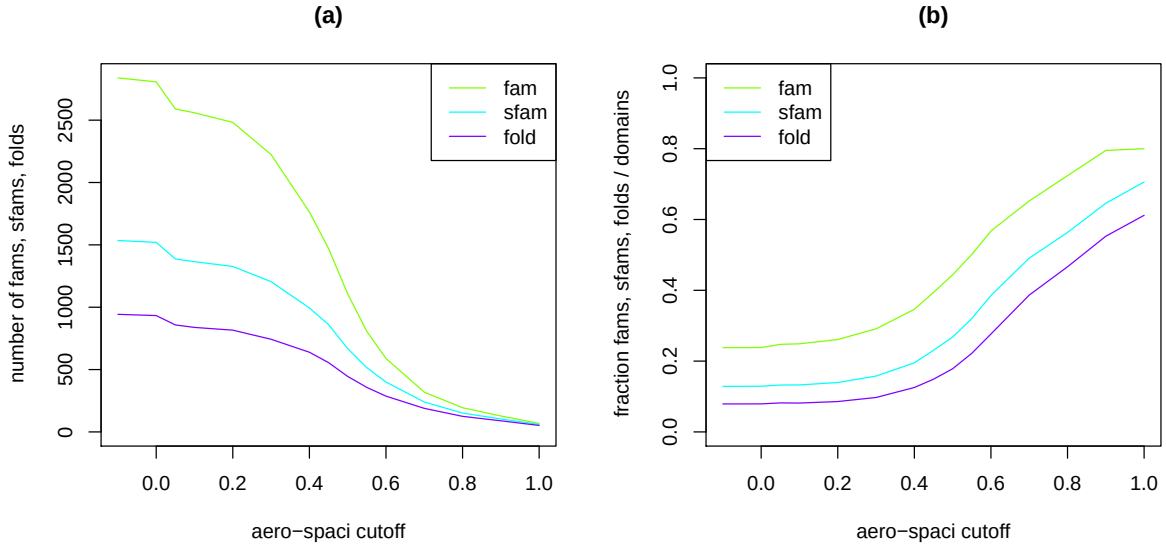


Figure 6.4: Aero-spaci thresholds
The effect of the aero-spaci cut-off with respect to the number of domains in the set. (a) The number of folds, superfamilies and families for a given aero-spaci threshold. (b) The number of folds superfamilies and families relative to the total number of domains in the set for a given aero-spaci cut-off.

6.2.6 Computation

The pairwise fragment program was run on Zuse, a 96 processor Super Computer from the Oxford Supercomputer Centre (`http://www.osc.ox.ac.uk/?area=ZUSE_Further_Detail`). The task took around 85 computing hours with a Processor Speed of 3.6GHz. Table 6.2 shows the number of fragments produced for each fragment length.

6.2.7 Age estimates

Ages for protein folds or superfamilies are estimated by searching for distant homologs of known structural domains on a set of completed genomes. The occurrence patterns of such predictions are analysed with a parsimony algorithm to estimate an age for a superfamily; for more details see Chapter 3.

The age of a superfamily is based on a score between [0.0,1.0], with 1.0 indicating the superfamily was estimated to be present at the root of the species tree (oldest) and 0.0 estimating that the superfamily was created at the leaf level (youngest).

6.3 Results

6.3.1 Fragnostic and superfamily age

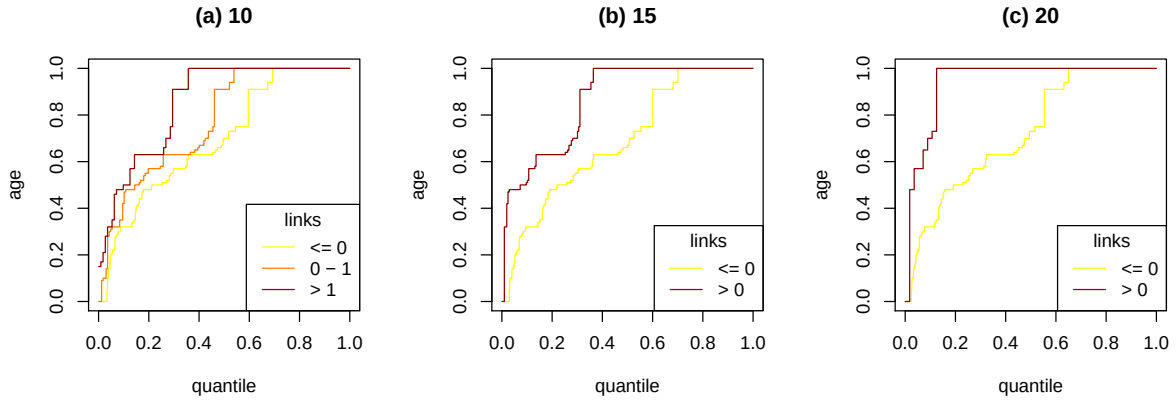


Figure 6.5: Fragnostic links and fold age
Fold age against fold quantile (fraction of folds smaller than the given age). The distributions are split on the numbers of links per fold as defined by Fragnostic. The fragment length used to obtain the links between folds was fixed at 10 for Figure (a), 15 for Figure (b) and 20 for Figure (c). In all these a link between two proteins was established when $f(i, j) > 0.2$ (as defined by equation 6.1). The differences in distribution between the set with the lowest number of links and highest number of links are significant. A Wilcoxon rank sum test estimated p-values of 1.9e-09 for Figure (a), 3.4e-16 for Figure (b) and 8.8e-12 for Figure (c).

The correlation between fold age and the number of links as defined by Fragnostic will be examined in order to compare these results to fold links based solely on structural fragments. Figure 6.5 shows that folds with no links to other folds are generally younger than folds that have more links. The difference in distributions are statistically significant. Moreover, the difference becomes more pronounced at longer fragment lengths.

A large number of shared fragments for older folds may be explained by a similar process that explains the large number of copies found for older folds: older folds have had more time to duplicate (parts of) the domain during evolution. In this case the results are explained with a divergent evolutionary scenario of fragment evolution. Conversely, older folds may share more fragments, since the confirmations are extremely favourable, and have been found many times independently over evolution, i.e. convergent evolution. For a more in depth discussion about divergent and convergent fragment evolution see Section 6.4.

An alternative explanation may be the strong sequence signal found between fragments in the prefilter of Fragnostic. A sequence signal that is shared with many other proteins may also cause fold recognition methods such as PSI-BLAST or SUPERFAMILY to find many instances of the fold, especially for sequence profiles of longer fragment lengths (15 or 20 residues). Since our age estimates are based on such fold recognition methods, this may bias the results. In order to rule out this issue, we will investigate the pairwise fragments based purely on structural comparison.

6.3.2 Structural fragments

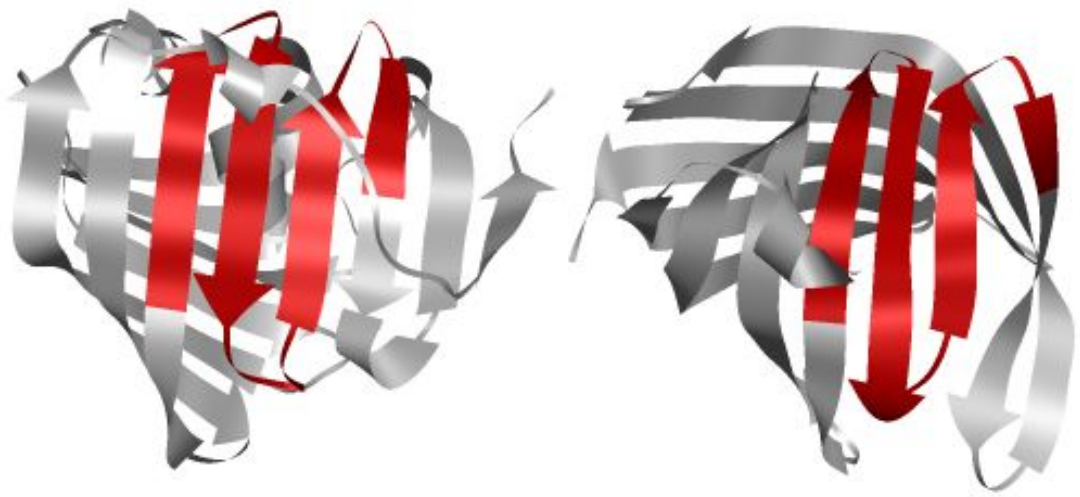


Figure 6.6: Example of a pairwise fragment
Left: chain A of IWLA . Right: chain A of R0UA. The domains shown both belong to the 'all beta' class but belong to different folds, b.125.1.1 and b.60.1.5 respectively. A pairwise fragment of length 30 is displayed in red ribbon for both molecules. For IWLA residues 37-67 and for R0UA residues 54-83 take part in the pairwise fragment. There is a single gap in the alignment, residue 44 of IWLA is not aligned (first residue in the loop region after the first beta strand included in the fragment). The fragment has a URMSD alignment score of 185 and an RMSD of 1.97 Å.

More than 130 million pairwise fragments passed both our structure-based prefilter and the 'final' RMSD thresholds as given in Table 6.1, and were used to create links between domains. See Table 6.2 for the number of fragments produced for each fragment length.

Figure 6.6 shows an example of a pairwise fragment of length 30. In this case the fragment shows a typical supersecondary structure motif: an anti parallel beta-sheet including four strands.

fragment length (N)	**10**	**15**	**20**	**30**
number of fragments	$1.2 \cdot 10^{-8}$	$1.5 \cdot 10^{-7}$	$2.7 \cdot 10^{-6}$	$1.1 \cdot 10^{-5}$

Table 6.2: Fragment numbersThe number of fragments produced for each fragment length.

6.3.3 Structural fragments normalised by Fragnostic's method

Some fragments may be over-represented (e.g. secondary structure is not considered) therefore the number of shared fragments needs to be normalised for the number of times a fragment occurs. To normalise the structure based fragments we used a similar approach to Friedberg and Godzik (2005):

$$f(I,J) = \frac{Sim(I,J)}{min(Sim(A-I,I), Sim(A-J,J))} \text{ if } I \neq J \tag{6.11}$$

Here Sim(A,B) is the number of shared fragments between two sets of domains (e.g. superfamilies), and A is the set of all domains. For this particular analysis we do not consider self-similarity of superfamilies. Figure 6.7 shows that these structure-based fragments produce a correlation between superfamily links and age for long fragment lengths ($N = 30$).

It may be that the normalisation procedure of Fragnostic is not suitable to our method. Fragnostic contains significantly fewer pairwise fragments than our dataset: only 2.5e-7 fragment pairs (including fragment lengths 5, 10, 15 and 20), whereas our dataset contain 1.2e-8 fragment pairs for fragment length 10 alone (see Table 6.2). Furthermore, our

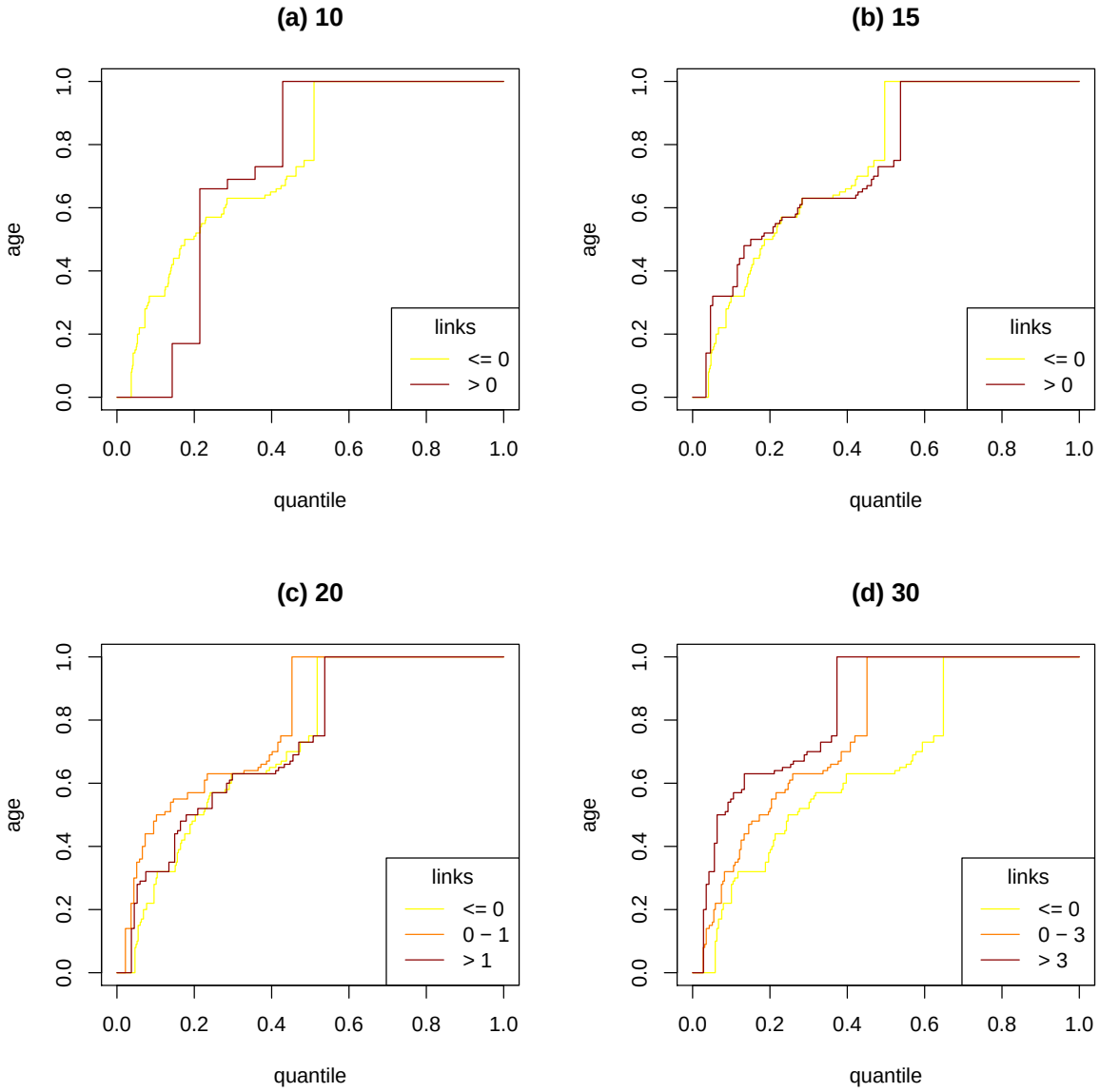


Figure 6.7: Fragments and age - normalised as Fragnostic
Superfamily age against superfamily quantile. The distributions are split on the numbers of links per superfamily as defined by pairwise structural fragments. The fragment lengths used to obtain the links between folds were fixed at 10 for Figure (a), 15 for Figure (b) , 20 for Figure (c) and 30 for Figure (d). In all these a link between two superfamilies was established when $f(i, j) > 0.1$ (as defined by equation 6.11). The difference in distribution is only significant in Figure (d) with a p-value of 1.38e-09.

RMSD cut-off at fragment length 10 is set to 0.8 Å, whereas in Fragnostic 1.0 Å is used. The fraction of pairwise fragments in our set with one or two gaps is smaller than 0.1% at fragment lenth 10, and thus does not explain this difference.

Firstly, the difference in fragment numbers signifies the huge effect of the sequence-based pre-filter in Fragnostic, which reduces the total number of fragments by an order

of magnitude at fragment length 10.

Secondly, at small fragment lengths the number of fragments produced by our method is very large. This is not surprising, since regular secondary structure elements are allowed to form fragments. In Fragnostic's method, pairwise fragments covering a secondary structure element still need to show a similar sequence profile, which may reflect a similar packing or similar functional unit. The sequence profiles in Fragnostic's prefilter were generated with the entire sequence of a domain and a fragment pair needed to give a significant FFAS score to pass the prefilter. It seems therefore reasonable to use a different normalisation procedure on our fragments that penalises commonly seen fragments.

6.3.4 Normalisation of pairwise structural fragments

Below a normalisation scheme for fragment usage is established. It aims is to derive a score which reflects the importance and uniqueness of the pairwise fragments shared between two structural domains. Here, the usage score $U(a)$ for a fragment piece in a single domain (a) is defined as:

$$U(a) = \sum_{i \in res(a)} u(i) \tag{6.12}$$

where $res(a)$ is the set of residues of fragment a and $u(i)$ is the number of times residue i is used within any of structurally similar pairwise fragments in the entire dataset. Then the linkage score between two domains A and B can be written as:

$$linkage(A, B) = \sum_{(a,b) \in Fr(A,B)} \frac{2N}{U(a) + U(b)} \tag{6.13}$$

where $Fr(A, B)$ is the set of pairwise fragments between domain A and B, with fragments pieces a and b respectively and N is the fragment length.

To investigate the evolutionary and structural significance of these linkage scores, scores between two domains of the same superfamily or fold are investigated. Figure 6.8 show Receiver Operator Characteristic (ROC) curves for linkage scores, where a true

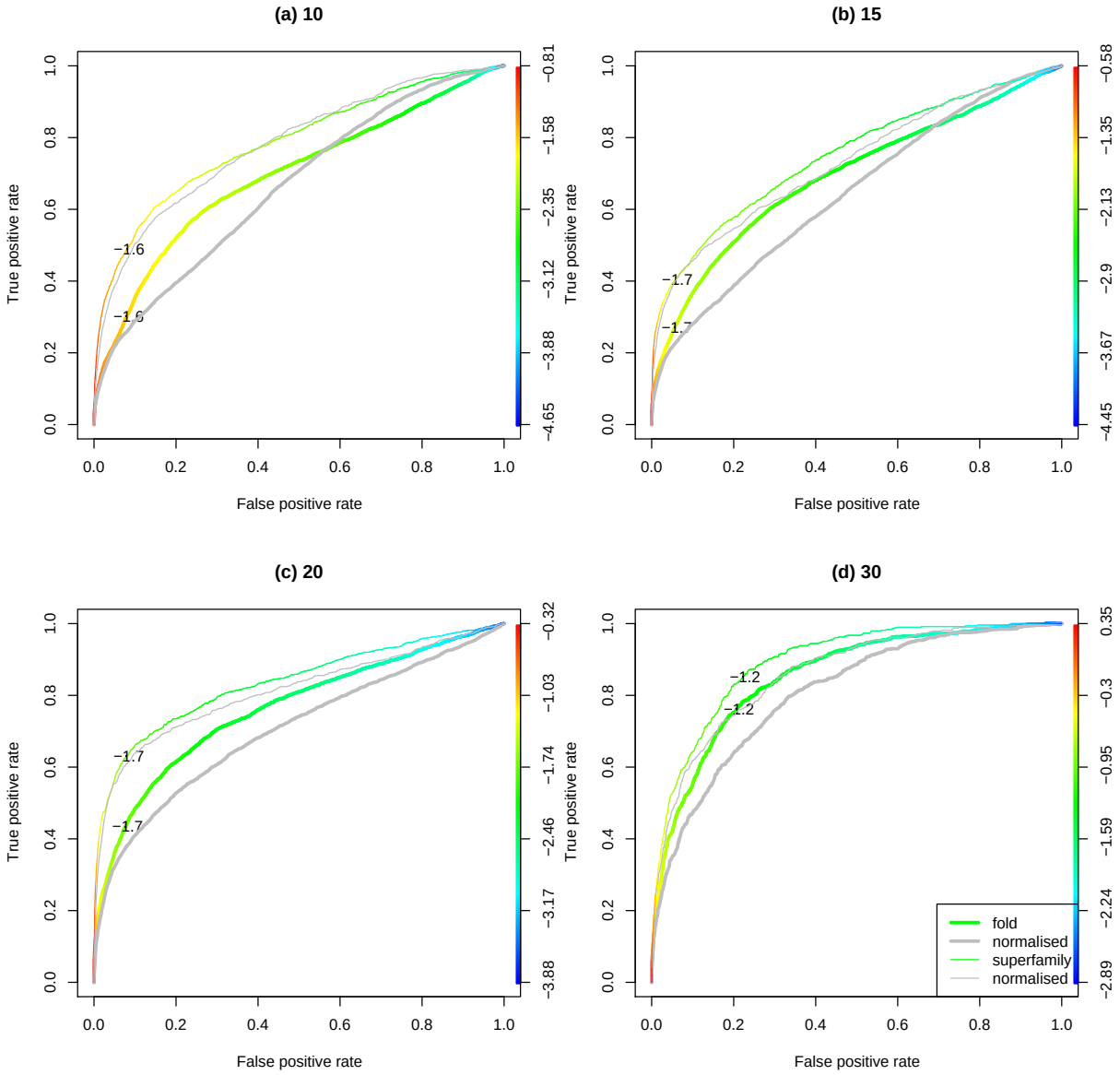


Figure 6.8: ROC-curves for fragment linkage scores
Selectivity against sensitivity is shown for fragment length 10 (a), 15 (b), 20 (c) and 30 (d). A function of the fraction of false positives against true positives is parametrised by $log10(linkage)$, where $linkage$ is given by equation 6.13, for the coloured lines. The colours reflect the value of $log10(linkage)$. The linkage thresholds are shown on the coloured curves. Wide curves indicate that a true positive is defined as two domains in the same fold, thin curves indicate a true positive to be defined as domains in the same superfamily. The corresponding grey curves are parametrised by a normalised linkage score (equation 6.14). These curves perform considerably worse at discriminating domains at fold and superfamily level.

positive reflects two domains of the same superfamily or fold and a false positive any other two domains. It can be observed that the longer the fragment length, the better the linkage score can discriminate between related structures. This is not surprising, as a high linkage score for short fragments may simply correspond to a large amount of shared

secondary structure, where longer fragment lengths may correspond to supersecondary structure motives or subdomains of a fold.

A new threshold for the linkage score above which two domains form a 'link' needs to be determined. Since we are not actually trying to distinguish between members of the same fold but trying to establish links above the fold level, our main aim is not to have a small as possible false positive rate, but to find an optimal balance between true and false positives. This can be found by considering the part of the ROC-curve that has a 45° slope and is as close to the top-left corner as possible. In fact most of the ROC-curves show a distinctive 'kink', after which the fraction of false positives increases at a much higher rate. The values before this kink may be viewed as evolutionary relevant and the threshold are chosen within this range of linkage scores. The thresholds for each fragment length are indicated within Figure 6.8.

6.3.5 Links and length

In Chapter 5 it was shown that domain length correlates with the age of a superfamily, and that long domains tend to be old. It is therefore important to investigate a possible length-dependence of any variable that is compared with age with the length of a domain.

Figure 6.9 shows that there is quite a strong correlation between the number of links and domain length at fragment length 10. To correct this we tried normalising the linkage score by the domain length. Since the total number of *possible* fragments is dependent on the length of two domains A and B by $length(A) \cdot length(B)$, we can normalise the linkage score for domains as follows:

$$linkage_{norm}(A,B) = \frac{\sum_{(a,b)\in Fr(A,B)} \frac{2N}{U(a)+U(b)}}{length(A) \cdot length(B)} \tag{6.14}$$

where $Fr(A,B)$ is the set of pairwise fragments between domain A and B, with fragments a and b respectively and N is the fragment length. $U(a)$ is given by equation 6.12.

However, Figure 6.8 shows that these normalised scores are significantly worse at predicting if two domains come from the same fold or superfamily. No further attempts

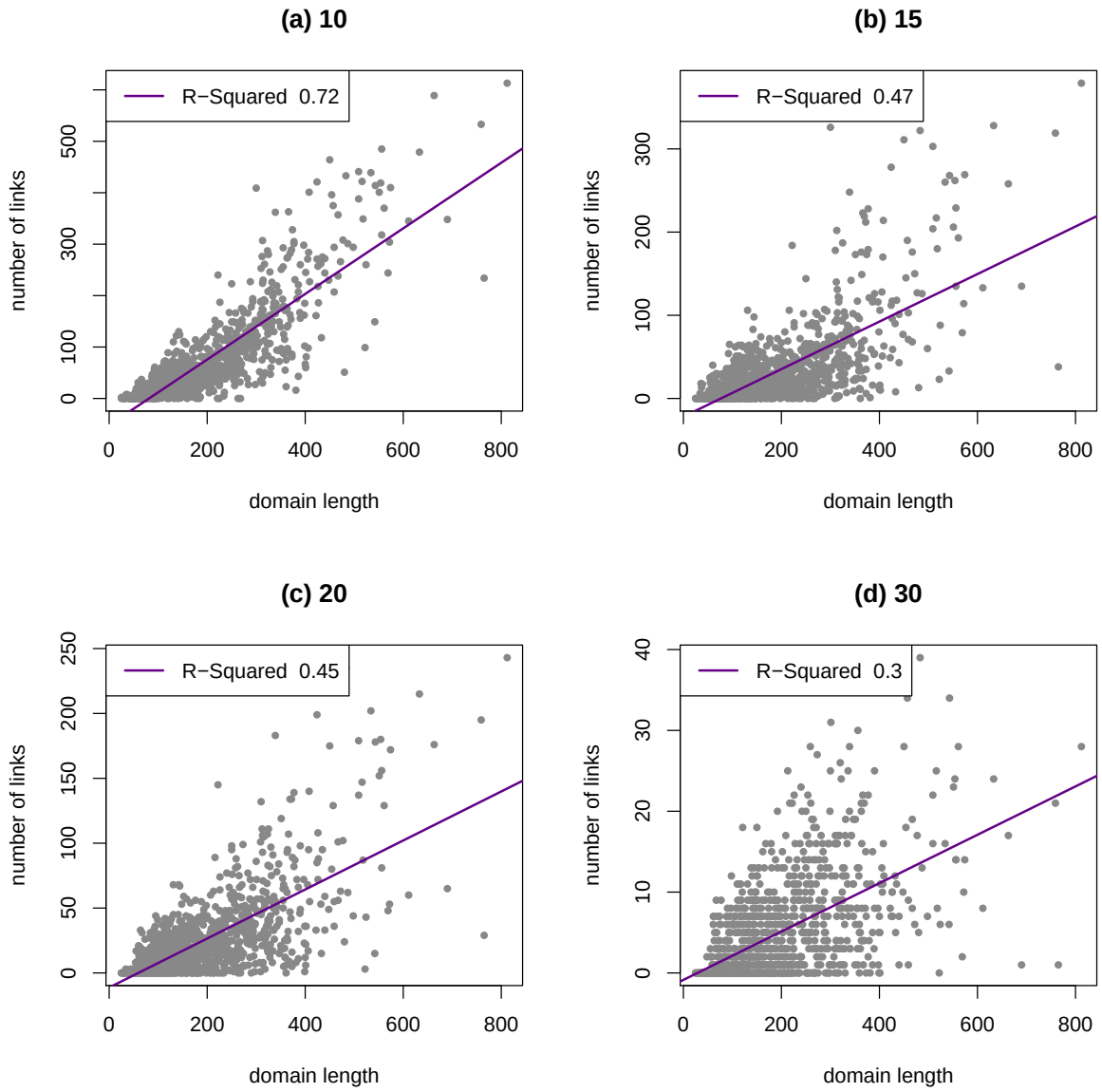


Figure 6.9: Domain length and links
Domain length versus links based on structural fragments of length 10 (a), 15 (b), 20 (c) and 30 (d). A link between two domains was established when the linkage score was greater than the threshold as given in Figure 6.8

at normalisation are made, but it is important to be aware of the dependence of fragment-based links on length.

6.3.6 Structural fragment links and age

Using the linkage score thresholds in Figure 6.8, it is possible to create fragment based links between domains. Figure 6.10 shows that the number of links for a domain correlates

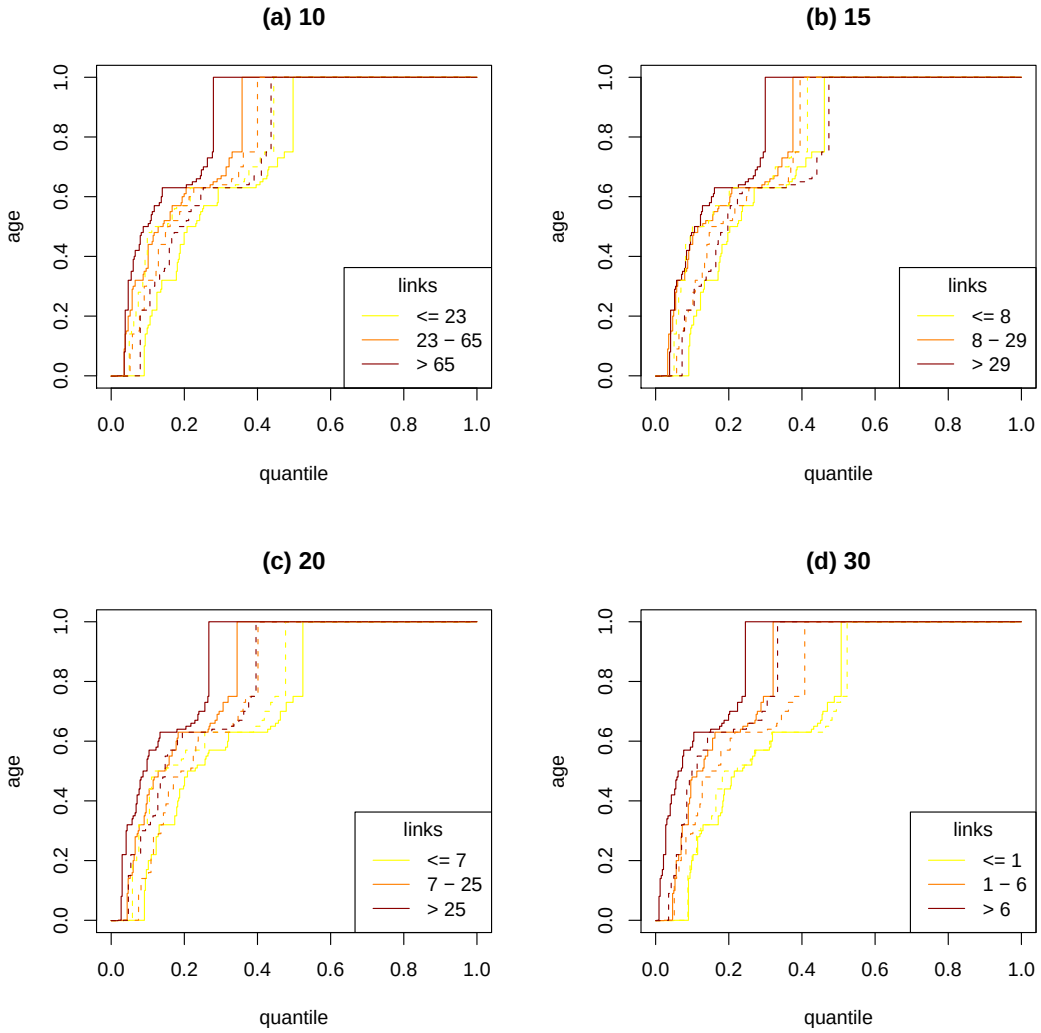


Figure 6.10: Structural fragment links and age
Superfamily age against domain quantile. The distributions are split into roughly three similar sized groups based on the numbers of links per domain. The fragment lengths used to obtain the links between domains were fixed at 10 for Figure (a), 15 for Figure (b) , 20 for Figure (c) and 30 for Figure (d). In all these a link between two domains was established when the linkage score was greater than the threshold as given in Figure 6.8. The continuous lines describe the age distribution for all domains, the dotted lines the age distribution for domains with a length ranging between $(100, 200)$

with the superfamily age (continuous lines): younger superfamilies have relatively fewer links for all lengths of fragments. All separations between distributions are significant (Figure 6.11 (a)).

However, these results may be due to domain length, especially for links based on small fragments. To see if the domain length has an effect on the age distribution, we narrow the dataset down to domains with a length ranging between 100 and 200 residues (dotted lines, Figure 6.10). Now the difference in age distribution is only significant for

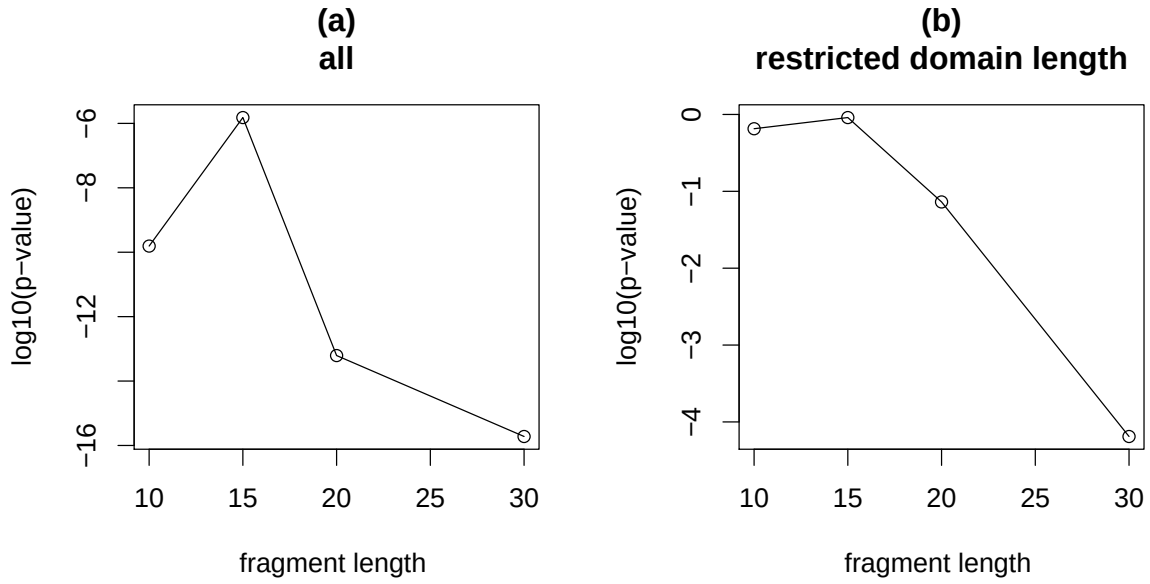


Figure 6.11: Significance of fragment separation and fragment length
Fragment length against the p-value of a Wilcoxon sum-rank test between age distributions with few and many links in Figure 6.10. (a) P-values for the continuous lines of Figure 6.10, containing all domains, and. (b) P-values for the striped lines in Figure 6.10, containing only domains with lengths between 100 and 200 residues. The separation in distributions is only significant for fragments of length >= 20 if a length restriction is imposed (b).

links based on a fragment length of 20 or 30 (Figure 6.11 (b)).

Another point of caution is the number of examples in the subset of each fold. In Chapter 2 and 3 it was shown that the number of families under a fold is correlated with age. This may bias our results, since it becomes more likely that a large number of links are observed for a domain if there are more domains of the same fold within the dataset.

Figure 6.12 (a) shows that there is only a very weak correlation between the number of links per domain and the number of domains that have the same fold for fragments of length 30; the correlations of other fragment lengths were even weaker. However, very weak correlations have previously been shown to have a significant effect on the age distribution. Therefore, the age distributions are shown in Figure 6.12 (b) for domains that have only one example of that fold within the dataset. Clearly the results as shown in Figure 6.10 (d) still hold in 6.12 (b).

Thus we can say the structural fragments indeed show a significant correlation with superfamily age, and that this dependence is unlikely to be caused by domain length, su-

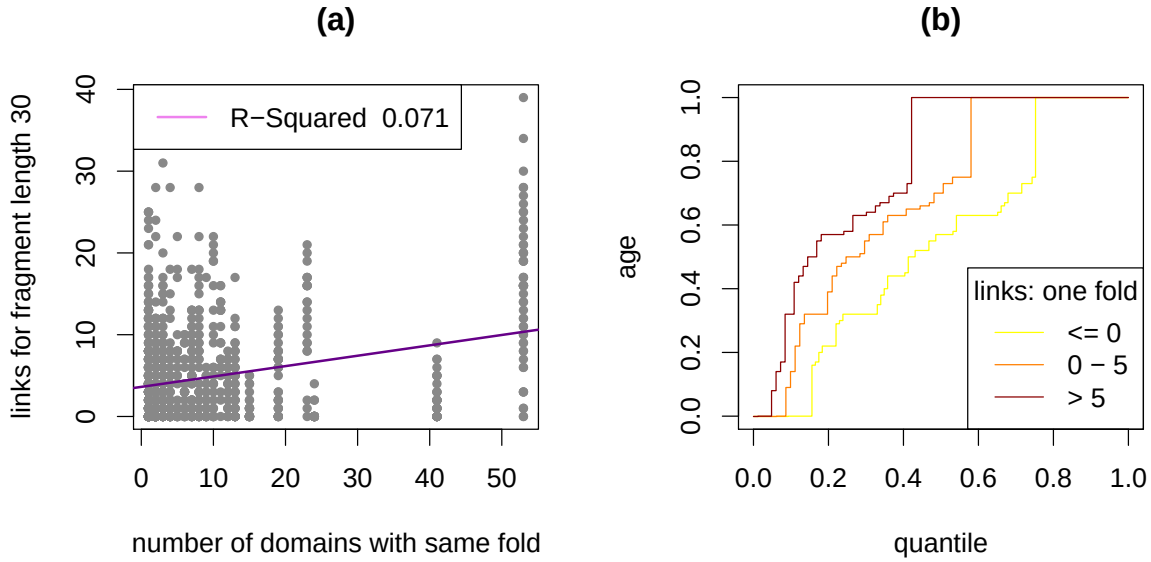


Figure 6.12: Fold size, fragment links and age
(a) The number of domains in the set with the same fold versus the number of links per domain for a fragment length of 30. (b) Superfamily age against domain quantile. The distributions are split into roughly three similar sized groups based on the numbers of links per domain. A link between two domains was established when the linkage score was greater than the threshold as given in Figure 6.8. Domains that are the only example of their fold within the dataset are included. The separation in age distribution between domains with no links and more than five is significant with a p-value of 4.1e-7.

perfamily size or the prefilter used to generate the fragments. In addition, the correlation becomes significantly stronger for longer fragment lengths or when sequence similarity is taken into account.

6.3.7 Self similarity

So far pairwise fragments within the same structural domain have been ignored. It is clearly not very interesting to look at fragments within the same region of the domain: a linkage score would then merely reflect the domain length of the protein. However, the structural similarity between different regions within a domain may be measured, while filtering out fragments that overlap.

Figure 6.13 shows the age distributions for domains split on different internal linkage scores (no cutoffs have been used). The trends seen here seem completely different from those in Figure 6.10. At longer fragment lengths, a higher number of internal fragments is observed for young superfamilies (Figure 6.13 (c) and (d)). This may indicate that a

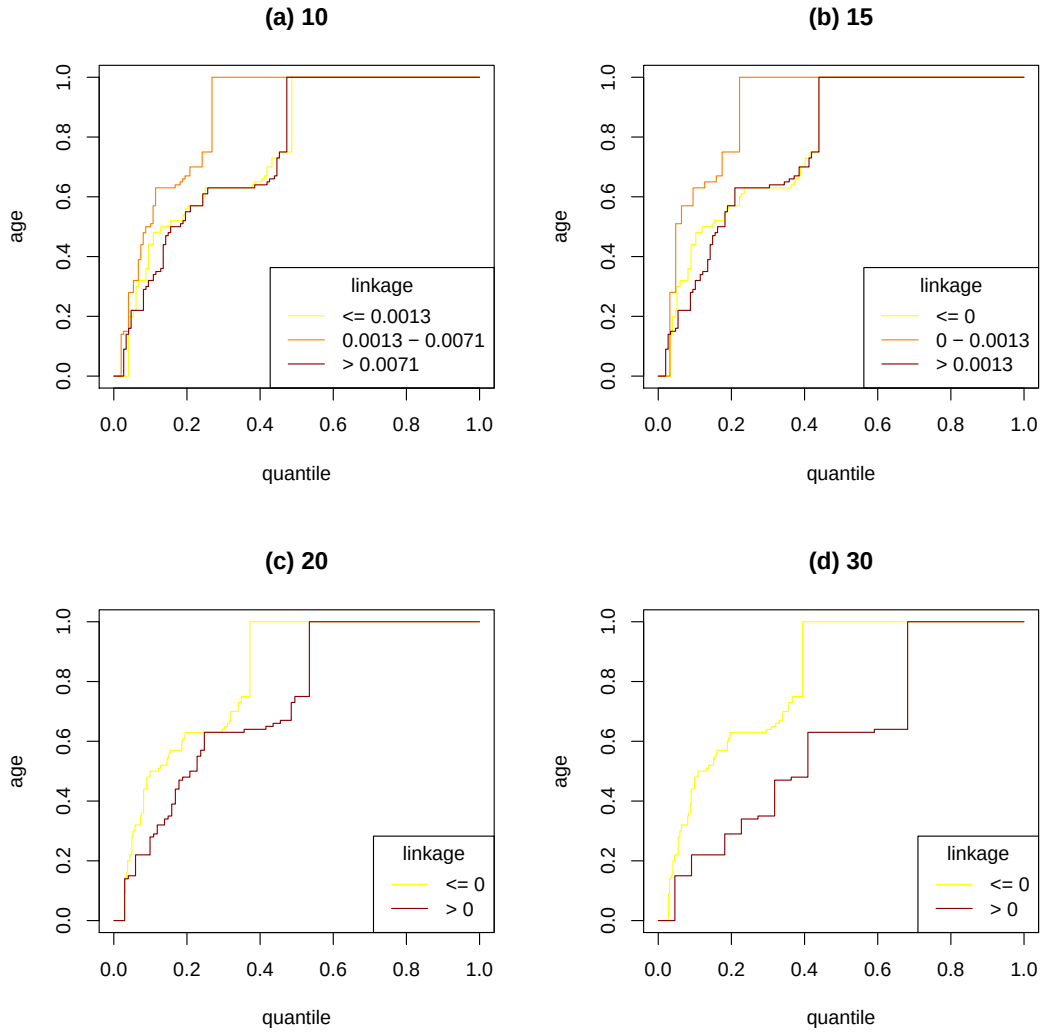


Figure 6.13: Self similarity and age
Superfamily age against domain quantile for internal linkage scores. The fragment lengths used to obtain the links between domains were fixed at 10 for Figure (a), 15 for Figure (b) , 20 for Figure (c) and 30 for Figure (d). In Figure (c) and (d) domains with a higher amount of self-similarity are significantly shifted towards younger folds; the Wilcoxon rank-sum test gives p-values of 3.0e-3 (c) and 7.0e-4 (d). In Figure (a) and (b) the domains in the middle range of self-similarity are significantly shifted towards older folds; the Wilcoxon rank-sum test gives p-values of 5.3e-5 (a) and 5.4e-4 (b).

way to create a new superfamily is to duplicate parts of the existing structure.

For shorter fragments, there is no difference between the age distributions of domains with a very high amount of self similarity and a very low amount of self similarity. However, the middle range domains tend to be older. Such a correlation is typical for a variable that does not correlate with fold age and has been observed before: older folds converge to more average properties and avoid extremes.

6.4 Discussion and conclusion

New proteins are thought to be created through point mutations, duplication and recombination of structural domains. Here we show evidence that this might also occur on a scale below the domain level: fragments are shared more often with older superfamilies, which is expected in a model where new topologies can be built through an assembly of, or multiple insertions of, fragments from existing proteins.

The separation between age distributions of folds with many links and few links is stronger in the Fragnostic dataset (Figure 6.5) than in our structure based fragment dataset (Figure 6.10). This is not surprising, as Fragnostic uses a prefilter that strongly selects fragments with a similar sequence profile, which may make the fragments more evolutionarily relevant.

The aim of creating the structure-based fragments was not to create the best separator between fold ages, but to investigate whether the separation in age distributions by Fragnostic's links was purely due to an effect of sequence similarity. Since the sequence filter in Fragnostic's method is based on similar principles as the sequence profile methods used to obtain the fold ages, the results may be biased. Figure 6.10 shows that links based on structural fragments alone still give a significant separation of age distributions for long fragments (> 20 residues). Using our set of fragments we have also shown that this separation is not due to other variables known to be correlated with fold age.

When considering what the results signify, a little care has to be taken as domains sharing a common structural motif may be caused by either divergent or convergent evolution. In a convergent scenario, the similarity between the regions in the domain would be due to a favourable confirmation which has evolved more than once during evolution through different paths. In a divergent scenario, the similarity is the result of a shared common ancestor. Note that this does not necessarily mean that the entire domains are evolutionarily related, but one of the two domain may have 'picked up' the structural fragment through an insertion.

In this work the most common cases of convergent evolution (e.g. secondary structure

elements) have been penalised by a normalisation procedure (see Section 6.3.4). Furthermore, the differences between age groups become more pronounced with increased fragment length. When increasing the fragment length the probability of convergence should decrease, favouring a divergent scenario. In addition links based on pairwise fragments that indicate structural as well a sequence similarity (Fragnostic) show a stronger correlation with superfamily ages than those based on structure alone. Sequence similarity between fragment pairs makes it more likely that such fragments are indeed evolutionarily related.

Creating a measure that is able to link above the fold level with evolutionary significance would provide valuable guidance when extending existing protein structure classifications. It would also have important implications for our understanding of fold evolution. Moreover, the ability to recognise evolutionarily related subdomains could push forward the field of protein structure prediction, bridging the gap between homology modelling and 'ab initio' modelling.

This study does not aim to show that fragment-based links are the best of the proposed measures to separate young and old folds, or the best measure to indicate evolutionary relatedness of a protein region. However, it does suggest that fragments are likely to have evolutionary relevance and that there is a possibility of finding evolutionary links above fold level. Further work is needed to see if stronger links can be established. Many previously proposed measures, e.g. GRATH (Harrison et al., 2002) or DALI (while including lower confidence scores) (Holm and Sander, 1993), could now also be investigated for their evolutionary relevance.

Chapter 7

Conclusions

We have shown that fold usage patterns on completed genomes can give us an idea behind the mechanisms of protein fold evolution.

We searched for SCOP and CATH domains on the genes of 150 completed genomes from all three kingdoms of life (Bacteria, Eukaryotes, Archaea) using the fold recognition methods, PSI-BLAST, SUPERFAMILY and Gene3D. The domain assignments provide information if (occurrence) and how many times (copies) a superfamily or a fold appears on a given genome.

A power-law like distribution was observed for the number of copies of a fold per genome and for the number of families per superfamily and superfamilies per fold. All these distributions may be explained by a process where the rich get richer. This could reflect the duplication process during evolution, or the fold classification process. Comparing fold usage across classification systems, we show that SCOP and CATH follow similar trends, with a few important exceptions. CATH's architecture classification level shows distinctly different distributions and correlations, perhaps reflecting its lack of evolutionary relevance.

The genomic occurrence distribution is not power-law-like, and is more variable with respect to kingdom, structural class, classification level or classification system than the other usage measures. The observed distribution can be explained by a simple evolutionary model. This model also suggests that genomic occurrence patterns of folds can be

used to create a phylogeny of species, predict function or estimate ages for a fold, but that it may be difficult to obtain an evolutionary fold tree based on such data.

Occurrence patterns for folds were used to estimate their relative ages. We developed a parsimony based method that models evolution by assigning gains and losses for a fold on a precalculated phylogenetic tree of genomes. The highest gain in this species tree estimates the relative age of the fold.

The estimated fold ages show correlations with measures that have previously been associated with protein age (copies and interactions). These measures are both expected to become more abundant when a fold has had more time to evolve. The correlation between the age of a fold and the number of copies (or interactions) per genome appears indeed restrictive: folds with many copies (or interactions) are usually old, but old folds do not necessarily have many copies (or interactions).

Structural classes, as defined by SCOP, have different fold usage patterns and age distributions: folds of SCOP's alpha/beta class tend to be older than folds of SCOP's other classes; folds in the 'small protein' class tend to be younger. In addition, the number of different folds seen for longer genomes is relatively small for alpha/beta folds and large for 'small protein' folds. These results indicate that alpha/beta folds have arisen relatively early in evolution.

When domain properties are correlated with age in a systematic way we find that: longer domains are generally older; domains with a higher proportion of alpha helical residues or residues in parallel beta bridges are generally older; whereas domains with a larger proportion of anti-parallel beta bridges are generally younger. In addition hydrophobic amino acids occur more often in old domains, whereas amino acids that are polar, but not charged, occur more often in younger domains. All these properties indicate that older folds are more stable.

Occurrence patterns were also used to assess and improve the reliability of the fold recognition methods. Occurrences that lie isolated within the tree can be identified by a parsimony algorithm as a 'gain' at leaf level. This, together with information about the e-values of the assignment is used to detect possible false positives. Independent tests

confirm the predictive power of this method to detect false positives. The method may be used to estimate and improve the reliability of any set of fold recognition data on completed genomes.

Pairs of structurally similar fragments between domains can suggest links between domains above the fold level. We developed a method to determine such fragment pairs using a fast structural prefilter. Domains that share significant structural fragments with many other domains (of different folds) tend to be older than domains that have few or no links. This correlation becomes stronger when the fragments become longer, or when the fragments also share sequence similarity. The correlation may be explained in a similar fashion as the number of copies of a fold: the older the fold, the more time it has had to spread structural fragments. This suggests that shared fragments may provide links between domains above the fold level that are evolutionary relevant.

In summary, we show that domain assignments on completed genomes enable us to assess and compare structural classification systems, assess and compare fold recognition methods, estimate the evolutionary ages of folds and investigate new methods that can relate protein domains above the topology level.

Bibliography

Abeln, S., Deane, C., Jul 2005. Fold usage on genomes and protein fold evolution. Proteins 60 (4), 690–700.

Abeln, S., Teubner, C., Deane, C. M., Jan 2007. Using phylogeny to improve genome-wide distant homology recognition. PLoS Comput Biol 3 (1).

Adamic, L. A., Huberman, B. A., 2002. Zipf's law and the internet. Glottometrics 3, 143–150.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., Oct 1990. Basic local alignment search tool. J Mol Biol 215 (3), 403–410.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., Sep 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17), 3389–3402.

Andreeva, A., Howorth, D., Brenner, S., Hubbard, T., Chothia, C., Murzin, A., Jan 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue, 226–229.

Andreeva, A., Prlic, A., Hubbard, T. J., Murzin, A. G., Jan 2007. Sisyphus–structural alignments for proteins with non-trivial relationships. Nucleic Acids Res 35 (Database issue), 253–259.

Apic, G., Gough, J., Teichmann, S. A., Jul 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310 (2), 311–325.

Baker, D., Sali, A., Oct 2001. Protein structure prediction and structural genomics. Science 294 (5540), 93–96.

Barabási, A., Oltvai, Z., Feb 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5 (2), 101–113.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P., Jan 2000. The Protein Data Bank. Nucleic Acids Res 28 (1), 235–242.

Bernhardt, G., Ldemann, H. D., Jaenicke, R., Knig, H., O, S. K., Nov 1984. Biomolecules are unstable under 'black smoker' conditions. Naturwissenschaften V71 (11), 583–586.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M., May 1977. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112 (3), 535–542.

Bhattacharyya, R., Pal, D., Chakrabarti, P., Oct 2002. Secondary structures at polypeptide-chain termini and their features. Acta Crystallogr D Biol Crystallogr 58 (Pt 10 Pt 2), 1793–1802.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., Jan 2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Res 31 (1), 365–370.

Bradley, P., Misura, K. M., Baker, D., Sep 2005. Toward high-resolution de novo structure prediction for small proteins. Science 309 (5742), 1868–1871.

Branden, C., Tooze, J., 1999. Introduction to protein structure. Garland.

Brenner, S., Koehl, P., Levitt, M., Jan 2000. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 28 (1), 254–256.

Brenner, S. E., Nov 2000. Target selection for structural genomics. Nat Struct Biol 7 Suppl, 967–969.

Brooks, D. J., Fresco, J. R., Lesk, A. M., Singh, M., Oct 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol Biol Evol 19 (10), 1645–1655.

Caetano-Anollés, G., Caetano-Anollés, D., Jul 2003. An evolutionarily structured universe of protein architecture. Genome Res 13 (7), 1563–1571.

Casbon, J. A., Saqi, M. A., Dec 2004. Analysis of superfamily specific profile-profile recognition accuracy. BMC Bioinformatics 5, 200–200.

Cheek, S., Krishna, S. S., Grishin, N. V., May 2006. Structural classification of small, disulfide-rich protein domains. J Mol Biol 359 (1), 215–237.

Cherkasov, A., Jones, S., Apr 2004. Structural characterization of genomes by large scale sequence-structure threading. BMC Bioinformatics 5 (1), 37–37.

Chew, L. P., Huttenlocher, D. P., Kedem, K., Kleinberg, J. M., 1999. Fast detection of common geometric substructure in proteins. Journal of Computational Biology 6 (3/4).

Chothia, C., Lesk, A. M., Apr 1986. The relation between the divergence of sequence and structure in proteins. EMBO J 5 (4), 823–826.

Chou, K. C., Némethy, G., Scheraga, H. A., Aug 1983. Role of interchain interactions in the stabilization of the right-handed twist of beta-sheets. J Mol Biol 168 (2), 389–407.

Chou, P. Y., Fasman, G. D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47, 45–148.

Christopher, J. A., Baldwin, T. O., Mar 1996. Implications of n and c-terminal proximity for protein folding. J Mol Biol 257 (1), 175–187.

Congreve, M., Murray, C., Blundell, T., Jul 2005. Structural biology and drug discovery. Drug Discov Today 10 (13), 895–907.

Coulson, A., Moult, J., Jan 2002. A unifold, mesofold, and superfold model of protein fold use. Proteins 46 (1), 61–71.

Crippen, G. M., Maiorov, V. N., Sep 1995. How many protein folding motifs are there? J Mol Biol 252 (1), 144–151.

Danson, M. J., Hough, D. W., Aug 1998. Structure, function and stability of enzymes from the archaea. Trends Microbiol 6 (8), 307–314.

Day, R., Beck, D. A., Armen, R. S., Daggett, V., Oct 2003. A consensus view of fold space: combining scop, cath, and the dali domain dictionary. Protein Sci 12 (10), 2150–2160.

Deeds, E. J., Shakhnovich, B., Shakhnovich, E. I., Feb 2004. Proteomic traces of speciation. J Mol Biol 336 (3), 695–706.

Di Giulio, M., Apr 2000. The universal ancestor lived in a thermophilic or hyperthermophilic environment. J Theor Biol 203 (3), 203–213.

Du, P., Andrec, M., Levy, R. M., Jun 2003. Have we seen all structures corresponding to short protein fragments in the protein data bank? an update. Protein Eng 16 (6), 407–414.

Efimov, A. V., Jan 1995. Structural similarity between two-layer alpha/beta and beta-proteins. J Mol Biol 245 (4), 402–415.

Efimov, A. V., Jun 1997. Structural trees for protein superfamilies. Proteins 28 (2), 241–260.

Eisen, J. A., Mar 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8 (3), 163–167.

Eisenberg, E., Levanon, E., Sep 2003. Preferential attachment in the protein network evolution. Phys Rev Lett 91 (13), 138701–138701.

Engelhardt, B. E., Jordan, M. I., Muratore, K. E., Brenner, S. E., Oct 2005. Protein molecular function prediction by bayesian phylogenomics. PLoS Comput Biol 1 (5).

England, J. L., Shakhnovich, E. I., May 2003. Structural determinant of protein designability. Phys Rev Lett 90 (21), 218101–218101.

Feng, D., Doolittle, R., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25 (4), 351–360.

Finkelstein, A. V., Badretdinov, A., Gutin, A. M., Oct 1995. Why do protein architectures have boltzmann-like statistics? Proteins 23 (2), 142–150.

Finkelstein, A. V., Gutun, A. M., AYa, B., Jun 1993. Why are the same protein folds used to perform different functions? FEBS Lett 325 (1-2), 23–28.

Fitch, W. M., Margoliash, E., Jan 1967. Construction of phylogenetic trees. Science 155 (760), 279–284.

Flory, P., 1953. Principles of Polymer Chemistry. Cornell University.

Friedberg, I., Godzik, A., Jul 2005. Fragnostic: walking through protein structure space. Nucleic Acids Res 33 (Web Server issue), 249–251.

Ganesh, C., Eswar, N., Srivastava, S., Ramakrishnan, C., Varadarajan, R., Jul 1999. Prediction of the maximal stability temperature of monomeric globular proteins solely from amino acid sequence. FEBS Lett 454 (1-2), 31–36.

Gerstein, M., 1998. How representative are the known structures of the proteins in a complete genome? a comprehensive structural census. Fold Des 3 (6), 497–512.

Ginalski, K., Pas, J., Wyrwicz, L. S., von Grotthuss, M., Bujnicki, J. M., Rychlewski, L., Jul 2003. Orfeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31 (13), 3804–3807.

Goldstein, R. A., Pollock, D. D., Jul 2006. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. Mol Biol Evol 23 (7), 1444–1449.

Gotoh, O., Dec 1982. An improved algorithm for matching biological sequences. J Mol Biol 162 (3), 705–708.

Gough, J., Chothia, C., Jan 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 30 (1), 268–272.

Gough, J., Karplus, K., Hughey, R., Chothia, C., Nov 2001. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 313 (4), 903–919.

Govindarajan, S., Goldstein, R. A., Apr 1996. Why are some proteins structures so common? Proc Natl Acad Sci U S A 93 (8), 3341–3345.

Govindarajan, S., Recabarren, R., Goldstein, R. A., Jun 1999. Estimating the total number of protein folds. Proteins 35 (4), 408–414.

Grishin, N. V., May-Jun 2001. Fold change in evolution of protein structures. J Struct Biol 134 (2-3), 167–185.

Grosberg, A. Y., Khokhlov, A. R., 1994. Statistical Physics of Macromolecules (translated by Atanov YA). AIP Press.

Hadley, C., Jones, D. T., Sep 1999. A systematic comparison of protein structure classifications: Scop, cath and fssp. Structure 7 (9), 1099–1112.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White,

R., Jan 2004. The gene ontology (go) database and informatics resource. Nucleic Acids Res 32 (Database issue), 258–261.

Harrison, A., Pearl, F., Mott, R., Thornton, J., Orengo, C., Nov 2002. Quantifying the similarities within fold space. J Mol Biol 323 (5), 909–926.

Hegyi, H., Lin, J., Greenbaum, D., Gerstein, M., May 2002. Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. Proteins 47 (2), 126–141.

Herning, T., Yutani, K., Inaka, K., Kuroki, R., Matsushima, M., Kikuchi, M., Aug 1992. Role of proline residues in human lysozyme stability: a scanning calorimetric study combined with x-ray structure analysis of proline mutants. Biochemistry 31 (31), 7077–7085.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., Vriend, G., Dec 1992. A database of protein structure families with common folding motifs. Protein Sci 1 (12), 1691–1698.

Holm, L., Sander, C., Sep 1993. Protein structure comparison by alignment of distance matrices. J Mol Biol 233 (1), 123–138.

Holm, L., Sander, C., Jan 1998. Touring protein fold space with dali/fssp. Nucleic Acids Res 26 (1), 316–319.

Hou, J., Sims, G. E., Zhang, C., Kim, S. H., Mar 2003. A global representation of the protein fold space. Proc Natl Acad Sci U S A 100 (5), 2386–2390.

Hubbard, T. J., Blundell, T. L., Jun 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng 1 (3), 159–171.

Hurst, L. D., Feil, E. J., Rocha, E. P., Aug 2006. Protein evolution: causes of trends in amino-acid gain and loss. Nature 442 (7105), 11–12.

Huynen, M. A., van Nimwegen, E., May 1998. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15 (5), 583–589.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., Godzik, A., Jul 2005. Ffas03: a server for profile–profile sequence alignments. Nucleic Acids Res 33 (Web Server issue), 284–288.

Jones, D., Taylor, W., Thornton, J., Jul 1992. A new approach to protein fold recognition. Nature 358 (6381), 86–89.

Jones, T. A., Thirup, S., Apr 1986. Using known substructures in protein model building and crystallography. EMBO J 5 (4), 819–822.

Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf, Y. I., Koonin, E. V., Kondrashov, A. S., Sunyaev, S., Feb 2005. A universal trend of amino acid gain and loss in protein evolution. Nature 433 (7026), 633–638.

Kabsch, W., Sander, C., Dec 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637.

Karev, G., Wolf, Y., Rzhetsky, A., Berezovskaya, F., Koonin, E., Oct 2002. Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evol Biol 2 (1), 18.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14 (10), 846–856.

Kearsley, S. K., 1989. On the orthogonal transformation used for structural comparisons. Acta Cryst. A45 (2), 208–210.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., Philips, D. C., Mar 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181 (4610), 662–666.

Khorana, H. G., 1966. Polynucleotide synthesis and the genetic code. Harvey Lect 62, 79–105.

Kinch, L. N., Grishin, N. V., Jun 2002. Evolution of protein structures and functions. Curr Opin Struct Biol 12 (3), 400–408.

Koehl, P., Levitt, M., Oct 2002. Sequence variations within protein families are linearly related to structural variations. J Mol Biol 323 (3), 551–562.

Konstantinidis, K. T., Tiedje, J. M., Mar 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101 (9), 3160–3165.

Koonin, E., Wolf, Y., Karev, G., Nov 2002. The structure of the protein universe and genome evolution. Nature 420 (6912), 218–223.

Krishna, S. S., Grishin, N. V., Jul 2004. Structurally analogous proteins do exist! Structure 12 (7), 1125–1127.

Krishna, S. S., Grishin, N. V., Apr 2005. Structural drift: a possible path to protein fold change. Bioinformatics 21 (8), 1308–1310.

Kumar, S., Nussinov, R., Aug 2001. How do thermophilic proteins deal with heat? Cell Mol Life Sci 58 (9), 1216–1233.

Lee, D., Grant, A., Buchan, D., Orengo, C., Jun 2003. A structural perspective on genome evolution. Curr Opin Struct Biol 13 (3), 359–369.

Lee, D., Grant, A., Marsden, R. L., Orengo, C., May 2005. Identification and distribution of protein families in 120 completed genomes using gene3d. Proteins 59 (3), 603–615.

Levitt, M., Jul 1992. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226 (2), 507–533.

Levitt, M., Chothia, C., Jun 1976. Structural patterns in globular proteins. Nature 261 (5561), 552–558.

Lieph, R., Veloso, F. A., Holmes, D. S., Oct 2006. Thermophiles like hot t. Trends Microbiol 14 (10), 423–426.

Lin, J., Gerstein, M., Jun 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res 10 (6), 808–818.

Lindahl, E., Elofsson, A., Jan 2000. Identification of related proteins on family, superfamily and fold level. J Mol Biol 295 (3), 613–625.

Liu, J., Rost, B., Oct 2001. Comparing function and structure between entire proteomes. Protein Sci 10 (10), 1970–1979.

Liu, X., Fan, K., Wang, W., Feb 2004. The number of protein folds and their distribution over families in nature. Proteins 54 (3), 491–499.

Lupas, A., Ponting, C., Russell, R., May-Jun 2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134 (2-3), 191–203.

Madera, M., Gough, J., Oct 2002. A comparison of profile hidden markov model procedures for remote homology detection. Nucleic Acids Res 30 (19), 4321–4328.

Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., Eisenberg, D., Jul 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285 (5428), 751–753.

Marsden, R. L., Ranea, J. A., Sillero, A., Redfern, O., Yeats, C., Maibaum, M., Lee, D., Addou, S., Reeves, G. A., Dallman, T. J., Orengo, C. A., Mar 2006. Exploiting protein structure data to explore the evolution of protein function and biological complexity. Philos Trans R Soc Lond B Biol Sci 361 (1467), 425–440.

Matsuda, K., Nishioka, T., Kinoshita, K., Kawabata, T., Go, N., Oct 2003. Finding evolutionary relations beyond superfamilies: fold-based superfamilies. Protein Sci 12 (10), 2239–2251.

McCloskey, M., Poo, M. M., 1984. Protein diffusion in cell membranes: some biological implications. Int Rev Cytol 87, 19–81.

Michie, A. D., Orengo, C. A., Thornton, J. M., Sep 1996. Analysis of domain structural class using an automated class assignment protocol. J Mol Biol 262 (2), 168–185.

Miller, S. L., May 1953. A production of amino acids under possible primitive earth conditions. Science 117 (3046), 528–529.

Miller, S. L., 1987. Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harb Symp Quant Biol 52, 17–27.

Mirkin, B., Fenner, T., Galperin, M., Koonin, E., Jan 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3 (1), 2.

Moult, J., Jun 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15 (3), 285–289.

Moult, J., Pedersen, J. T., Judson, R., Fidelis, K., Nov 1995. A large-scale experiment to assess protein structure prediction methods. Proteins 23 (3).

Murzin, A., Jun 1998. How far divergent evolution goes in proteins. Curr Opin Struct Biol 8 (3), 380–387.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., Apr 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247 (4), 536–540.

Needleman, S. B., Wunsch, C. D., Mar 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (3), 443–453.

Nilsson, J., Persson, B., von Heijne, G., Sep 2005. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. Proteins 60 (4), 606–616.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., Thornton, J., Aug 1997. CATH–a hierarchic classification of protein domain structures. Structure 5 (8), 1093–1108.

Orengo, C. A., Taylor, W. R., 1996. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 266, 617–635.

Ortiz, A. R., Strauss, C. E., Olmea, O., Nov 2002. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11 (11), 2606–2621.

Pál, C., Papp, B., Lercher, M. J., Dec 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37 (12), 1372–1375.

Panchenko, A. R., Wolf, Y. I., Panchenko, L. A., Madej, T., Nov 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 61 (3), 535–544.

Park, J., Holm, L., Chothia, C., Feb 2000. Sequence search algorithm assessment and testing toolkit (sat). Bioinformatics 16 (2), 104–110.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C., Dec 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284 (4), 1201–1210.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., Orengo, C., Jan 2005. The cath domain structure database and related resources gene3d and dhs provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33 (Database issue), 247–251.

Pearl, F. M., Lee, D., Bray, J. E., Buchan, D. W., Shepherd, A. J., Orengo, C. A., Feb 2002. The cath extended protein-family database: providing structural annotations for genome sequences. Protein Sci 11 (2), 233–244.

Pearl, F. M. G., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., Orengo, C. A., Jan 2003. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res 31 (1), 452–455.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., Yeates, T. O., Apr 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96 (8), 4285–4288.

Peng, K., Obradovic, Z., Vucetic, S., 2004. Exploring bias in the protein data bank using contrast classifiers. Pac Symp Biocomput, 435–446.

Perl, D., Mueller, U., Heinemann, U., Schmid, F. X., May 2000. Two exposed amino acid residues confer thermostability on a cold shock protein. Nat Struct Biol 7 (5), 380–383.

Perutz, M., Kendrew, J., Watson, H., 1965. Structure and function of haemoglobin ii. some relations between polypeptide chain configuration and amino acid sequence. J. Mol. Biol. 13, 669–678.

Perutz, M., Rossmann, M., Cullis, A., Muirhead, H., Will, G., A.C.T., N., 1960. Structure of hmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by x-ray analysis. Nature 185, 416–422.

Plaxco, K. W., Simons, K. T., Baker, D., Apr 1998. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277 (4), 985–994.

Ponting, C. P., Russell, R. B., May 1995. Swaposins: circular permutations within genes encoding saposin homologues. Trends Biochem Sci 20 (5), 179–180.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., 1992. Numerical Recipes in C, 2nd Edition. Cambridge University Press, Ch. 11.

Qian, J., Luscombe, N., Gerstein, M., Nov 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 313 (4), 673–681.

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286.

Ranea, J. A., Buchan, D. W., Thornton, J. M., Orengo, C. A., Feb 2004. Evolution of protein superfamilies and bacterial genome size. J Mol Biol 336 (4), 871–887.

Ranea, J. A., Sillero, A., Thornton, J. M., Orengo, C. A., Oct 2006. Protein superfamily evolution and the last universal common ancestor (luca). J Mol Evol 63 (4), 513–525.

Richardson, J. S., Aug 1977. beta-sheet topology and the relatedness of proteins. Nature 268 (5620), 495–500.

Richardson, J. S., 1981. The anatomy and taxonomy of protein structure. Adv Protein Chem 34, 167–339.

Rossmann, M. G., Argos, P., Jul 1976. Exploring structural homology of proteins. J Mol Biol 105 (1), 75–95.

Rost, B., Feb 1999. Twilight zone of protein sequence alignments. Protein Eng 12 (2), 85–94.

Rufino, S. D., Donate, L. E., Canard, L., Blundell, T. L., 1996. Analysis, clustering and prediction of the conformation of short and medium size loops connecting regular secondary structures. Pac Symp Biocomput, 570–589.

Russell, R. J., Ferguson, J. M., Hough, D. W., Danson, M. J., Taylor, G. L., Aug 1997. The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 a resolution,. Biochemistry 36 (33), 9983–9994.

Rychlewski, L., Fischer, D., Elofsson, A., 2003. Livebench-6: large-scale automated evaluation of protein structure prediction servers. Proteins 53 Suppl 6, 542–547.

Sadreyev, R. I., Grishin, N. V., Apr 2004. Quality of alignment comparison by compass improves with inclusion of diverse confident homologs. Bioinformatics 20 (6), 818–828.

Sadreyev, R. I., Grishin, N. V., 2006. Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. BMC Struct Biol 6, 6–6.

Sanchez-Ruiz, J. M., 1995. Differential scanning calorimetry of proteins. Subcell Biochem 24, 133–176.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., Altschul, S. F., Jul 2001. Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29 (14), 2994–3005.

Schein, C. H., Apr 1990. Solubility as a function of protein structure and solvent components. Biotechnology (N Y) 8 (4), 308–317.

Serrano, L., Day, A. G., Fersht, A. R., Sep 1993. Step-wise mutation of barnase to binase. a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. J Mol Biol 233 (2), 305–312.

Shakhnovich, B. E., Deeds, E., Delisi, C., Shakhnovich, E., Mar 2005. Protein structure and evolutionary history determine sequence space topology. Genome Res 15 (3), 385–392.

Shi, J., Blundell, T. L., Mizuguchi, K., Jun 2001. Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310 (1), 243–257.

Shindyalov, I. N., Bourne, P. E., Sep 1998. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Eng 11 (9), 739–747.

Shindyalov, I. N., Bourne, P. E., Feb 2000. An alternative view of protein fold space. Proteins 38 (3), 247–260.

Siew, N., Elofsson, A., Rychlewski, L., Fischer, D., Sep 2000. Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16 (9), 776–785.

Simons, K. T., Bonneau, R., Ruczinski, I., Baker, D., 1999. Ab initio protein structure prediction of casp iii targets using rosetta. Proteins Suppl 3, 171–176.

Sjölander, K., Jan 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20 (2), 170–179.

Smith, T. F., Waterman, M. S., Mar 1981. Identification of common molecular subsequences. J Mol Biol 147 (1), 195–197.

Snel, B., Bork, P., Huynen, M., Jan 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12 (1), 17–25.

Sobolevsky, Y., Trifonov, E. N., Nov 2006. Protein modules conserved since luca. J Mol Evol 63 (5), 622–634.

Struyf, A., Hubert, M., Rousseeuw, P., 1996. Clustering in an object-oriented environment. Journal of Statistical Software 1.

Summers, N. L., Karplus, M., Dec 1990. Modeling of globular proteins. a distance-based data search procedure for the construction of insertion/deletion regions and pro—-non-pro mutations. J Mol Biol 216 (4), 991–1016.

Swindells, M. B., MacArthur, M. W., Thornton, J. M., Jul 1995. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 2 (7), 596–603.

Taverna, D. M., Goldstein, R. A., Jan 2000. The distribution of structures in evolving protein populations. Biopolymers 53 (1), 1–8.

Taylor, W., Heringa, J., Baud, F., Flores, T., Feb 2002. A Fourier analysis of symmetry in protein structure. Protein Eng 15 (2), 79–89.

Taylor, W. R., Oct 2006. Topological accessibility shows a distinct asymmetry in the folds of betaalpha proteins. FEBS Lett 580 (22), 5263–5267.

Tekaia, F., Yeramian, E., Dujon, B., Sep 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297 (1-2), 51–60.

Thornton, J. M., Sibanda, B. L., Jun 1983. Amino and carboxy-terminal regions in globular proteins. J Mol Biol 167 (2), 443–460.

Todd, A. E., Marsden, R. L., Thornton, J. M., Orengo, C. A., May 2005. Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 348 (5), 1235–1260.

Trifonov, E. N., Aug 2004. The triplet code from first principles. J Biomol Struct Dyn 22 (1), 1–11.

Trifonov, E. N., Berezovsky, I. N., Feb 2003. Evolutionary aspects of protein structure and folding. Curr Opin Struct Biol 13 (1), 110–114.

Troll, G., beim Graben, P., 1998. Zipf s law is not a consequence of the central limit theorem. Physical Review E 57, 1347–1355.

van Vlijmen, H. W., Karplus, M., Apr 1997. Pdb-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 267 (4), 975–1001.

Venclovas, C., Margelevicius, M., 2005. Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. Proteins 61 Suppl 7, 99–105.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., Teichmann, S. A., Feb 2004. Supradomains: evolutionary units larger than single protein domains. J Mol Biol 336 (3), 809–823.

Vogel, C., Chothia, C., May 2006. Protein family expansions and biological complexity. PLoS Comput Biol 2 (5).

Vogel, C., Teichmann, S. A., Pereira-Leal, J., Feb 2005. The relationship between domain duplication and recombination. J Mol Biol 346 (1), 355–365.

Wallin, E., von Heijne, G., Apr 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 7 (4), 1029–1038.

Wilson, D., Madera, M., Vogel, C., Chothia, C., Gough, J., Jan 2007. The superfamily database in 2007: families and functions. Nucleic Acids Res 35 (Database issue), 308–313.

Winstanley, H. F., Abeln, S., Deane, C. M., Jun 2005. How old is your fold? Bioinformatics 21 Suppl 1, 449–458.

Wolf, Y., Brenner, S., Bash, P., Koonin, E., Jan 1999. Distribution of protein folds in the three superkingdoms of life. Genome Res 9 (1), 17–26.

Wolf, Y. I., Grishin, N. V., Koonin, E. V., Jun 2000. Estimating the number of protein folds and families from complete genome data. J Mol Biol 299 (4), 897–905.

Wood, T. C., Pearson, W. R., Aug 1999. Evolution of protein sequences and structures. J Mol Biol 291 (4), 977–995.

Wu, J., Kasif, S., DeLisi, C., Aug 2003. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19 (12), 1524–1530.

Xenarios, I., Salwínski, L., Duan, X., Higney, P., Kim, S., Eisenberg, D., Jan 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30 (1), 303–305.

Yan, Y., Moult, J., Oct 2005. Protein family clustering for structural genomics. J Mol Biol 353 (3), 744–759.

Yang, A. S., Honig, B., Aug 2000. An integrated approach to the analysis and modeling of protein sequences and structures. ii on the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 301 (3), 679–689.

Yang, S., Doolittle, R., Bourne, P., Jan 2005. Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A 102 (2), 373–378.

Zhang, C., DeLisi, C., Dec 1998. Estimating the number of protein folds. J Mol Biol 284 (5), 1301–1305.

Appendix A

Assignments on genomes

key	genes	hits PB	%	hits SF	%	hits G3	%	name
A	40214	20274	50.42	23197	57.68	18050	44.88	All Archaea
ACrTh_Aper	1841	832	45.19	952	51.71	739	40.14	Aeropyrum_pernix
ACrTh_Paer	2605	1086	41.69	1270	48.75	981	37.66	Pyrobaculum_aerophilum
ACrTh_Ssol	2977	1393	46.79	1655	55.59	1217	40.88	Sulfolobus_solfataricus
ACrTh_Stok	2825	1278	45.24	1515	53.63	1144	40.50	Sulfolobus_tokodaii
AEuAr_Aful	2420	1283	53.02	1493	61.69	1153	47.64	Archaeoglobus_fulgidus
AEuHa_Hsp.	2075	1076	51.86	1200	57.83	1064	51.28	Halobacterium
AEuMe_Mace	4540	2171	47.82	2465	54.30	1929	42.49	Methanosarcina_acetivorans
AEuMe_Mjan	1786	924	51.74	1049	58.73	826	46.25	Methanocaldococcus_jannaschii
AEuMe_Mkan	1687	825	48.90	932	55.25	740	43.86	Methanopyrus_kandleri
AEuMe_Mmar	1722	983	57.08	1082	62.83	869	50.46	Methanococcus_maripaludis
AEuMe_Mmaz	3371	1657	49.15	1903	56.45	1473	43.70	Methanosarcina_mazei_Go
AEuMe_Mthe	1873	1027	54.83	1141	60.92	927	49.49	Methanothermobacter_thermautotrophicus
AEuTh_Paby	1896	986	52.00	1149	60.60	886	46.73	Pyrococcus_abyssi
AEuTh_Pfur	2125	1092	51.39	1263	59.44	958	45.08	Pyrococcus_furiosus
AEuTh_Phor	1955	930	47.57	1079	55.19	823	42.10	Pyrococcus_horikoshii
AEuTh_Ptor	1535	930	60.59	1034	67.36	788	51.34	Picrophilus_torridus
AEuTh_Taci	1482	904	61.00	1012	68.29	770	51.96	Thermoplasma_acidophilum
AEuTh_Tvol	1499	897	59.84	1003	66.91	763	50.90	Thermoplasma_volcanium
B	340113	181773	53.44	201017	59.10	148799	45.64	All Bacteria
BAcAc_Blon	1727	1028	59.53	1121	64.91	916	53.04	Bifidobacterium_longum
BAcAc_Cdip	1259	664	52.74	1315	104.45	1093	86.81	Corynebacterium_diphtheriae
BAcAc_Ceff	2950	1539	52.17	1724	58.44	1407	47.69	Corynebacterium_efficiens
BAcAc_Cglu	3057	1641	53.68	1790	58.55	1040	34.02	Corynebacterium_glutamicum
BAcAc_Lxyl	2030	1153	56.80	1276	62.86	1053	51.87	Leifsonia_xyli
BAcAc_Mavi	4350	2647	60.85	2893	66.51	2436	56.00	Mycobacterium_avium
BAcAc_Mbov	3920	2108	53.78	2344	59.80	1060	27.04	Mycobacterium_bovis
BAcAc_Mlep	1605	944	58.82	1014	63.18	882	54.95	Mycobacterium_leprae
BAcAc_Mtub	3991	2137	53.55	2375	59.51	853	21.37	Mycobacterium_tuberculosis_H37Rv
BAcAc_Pacn	2297	1329	57.86	1449	63.08	1190	51.81	Propionibacterium_acnes
BAcAc_Save	7577	4160	54.90	4702	62.06	3887	51.30	Streptomyces_avermitilis
BAcAc_Scoe	7769	4320	55.61	4854	62.48	4094	52.70	Streptomyces_coelicolor
BAqAq_Aaeo	1529	989	64.68	1082	70.77	856	55.98	Aquifex_aeolicus
BBaBa_Bthe	4778	2459	51.47	2778	58.14	2177	45.56	Bacteroides_thetaiotaomicron
BBaBa_Pgin	1909	969	50.76	1097	57.46	886	46.41	Porphyromonas_gingivalis
BChCh_Ctep	2252	1210	53.73	1325	58.84	1081	48.00	Chlorobium_tepidum
BCyCh_Gvio	4430	2236	50.47	2543	57.40	2024	45.69	Gloeobacter_violaceus
BCyCh_Ssp*	3167	1682	53.11	1885	59.52	1624	51.28	Synechocystis
BCyCh_Ssp.	2517	1264	50.22	1435	57.01	1168	46.40	Synechococcus
BCyCh_Telo	2475	1358	54.87	1493	60.32	-	-	Thermosynechococcus_elongatus
BCyNo_Nsp.	5366	2621	48.84	2946	54.90	2587	48.21	Nostoc
BCyPr_Pmar	1882	954	50.69	1064	56.54	877	46.60	Prochlorococcus_marinus
BDeDe_Drad	2997	1635	54.55	1818	60.66	1506	50.25	Deinococcus_radiodurans
BDeDe_Tthe	1982	1239	62.51	1361	68.67	1038	52.37	Thermus_thermophilus
BFiBa_Bant	5311	2567	48.33	2820	53.10	650	12.24	Bacillus_anthracis
BFiBa_Bcer	5234	2627	50.19	2914	55.67	2254	43.06	Bacillus_cereus
BFiBa_Bhal	4066	2178	53.57	2402	59.08	1949	47.93	Bacillus_halodurans
BFiBa_Bsub	4105	2252	54.86	2466	60.07	1987	48.40	Bacillus_subtilis
BFiBa_Bthu	5117	2652	51.83	2926	57.18	1775	34.69	Bacillus_thuringiensis_serovar_konkukian
BFiBa_Linn	2968	1657	55.83	1807	60.88	1469	49.49	Listeria_innocua_Clip
BFiBa_Lmon	2846	1695	59.56	1842	64.72	1406	49.40	Listeria_monocytogenes_EGD-e
BFiBa_Oihe	3500	1966	56.17	2126	60.74	1764	50.40	Oceanobacillus_iheyensis
BFiBa_Saur	2697	1505	55.80	1628	60.36	556	20.62	Staphylococcus_aureus
BFiBa_Sepi	2419	1333	55.11	1428	59.03	893	36.92	Staphylococcus_epidermidis
BFiCl_Cace	3672	1974	53.76	2183	59.45	1852	50.44	Clostridium_acetobutylicum
BFiCl_Cper	2660	1499	56.35	1647	61.92	1361	51.17	Clostridium_perfringens
BFiCl_Ctet	2373	1323	55.75	1443	60.81	1190	50.15	Clostridium_tetani
BFiCl_Tten	2588	1441	55.68	1619	62.56	1284	49.61	Thermoanaerobacter_tengcongensis
BFiLa_Efae	3113	1543	49.57	1691	54.32	1462	46.96	Enterococcus_faecalis
BFiLa_Ljoh	1821	1025	56.29	1135	62.33	943	51.78	Lactobacillus_johnsonii
BFiLa_Llac	2321	1267	54.59	1407	60.62	1125	48.47	Lactococcus_lactis
BFiLa_Lpla	3009	1718	57.10	1873	62.25	1546	51.38	Lactobacillus_plantarum
BFiLa_Saga	2124	1180	55.56	1273	59.93	803	37.81	Streptococcus_agalactiae
BFiLa_Smut	1960	1155	58.93	1257	64.13	1054	53.78	Streptococcus_mutans
BFiLa_Spne	2094	1123	53.63	1232	58.83	879	41.98	Streptococcus_pneumoniae
BFiLa_Spy*	1697	976	57.51	-	0.00	665	39.19	Streptococcus_pyogenes_M1GAS
BFiLa_Spyo	1886	1017	53.92	1111	58.91	731	38.76	Streptococcus_pyogenes_MGAS10394
BFuFu_Fnuc	2067	1077	52.10	1219	58.97	929	44.94	Fusobacterium_nucleatum
BPrAl_Bjap	8317	4561	54.84	5068	60.94	4105	49.36	Bradyrhizobium_japonicum

key	genes	hits PB	%	hits SF	%	hits G3	%	name
BPrAl_Bmel	3198	1883	58.88	2037	63.70	1517	47.44	Brucella_melitensis
BPrAl_Bsui	3271	1770	54.11	1902	58.15	1427	43.63	Brucella_suis
BPrAl_Ccre	3737	2131	57.02	2357	63.07	1926	51.54	Caulobacter_crescentus
BPrAl_Mlot	6743	3724	55.23	4100	60.80	3574	53.00	Mesorhizobium_loti
BPrAl_Rpal	4813	2808	58.34	3095	64.31	2517	52.30	Rhodopseudomonas_palustris
BPrAl_Rsph	11582	2162	18.67	2394	20.67	-	-	Rhodobacter_sphaeroides
BPrAl_Smel	3341	1991	59.59	2195	65.70	3224	96.50	Sinorhizobium_meliloti
BPrBe_Bbro	4994	3079	61.65	3432	68.72	2058	41.21	Bordetella_bronchiseptica
BPrBe_Bpar	4185	2652	63.37	2943	70.32	1767	42.22	Bordetella_parapertussis
BPrBe_Bper	3436	2273	66.15	2476	72.06	1526	44.41	Bordetella_pertussis_Tohama
BPrBe_Cvio	4407	2419	54.89	2654	60.22	2138	48.51	Chromobacterium_violaceum
BPrBe_Neur	2461	1400	56.89	1561	63.43	1224	49.74	Nitrosomonas_europaea
BPrBe_Nmen	2079	1084	52.14	1185	57.00	838	40.31	Neisseria_meningitidis
BPrBe_Rsol	3440	1937	56.31	2158	62.73	2468	71.74	Ralstonia_solanacearum
BPrDe_Bbac	3587	1661	46.31	1875	52.27	1527	42.57	Bdellovibrio_bacteriovorus
BPrDe_Dpsy	3116	1623	52.09	1782	57.19	1481	47.53	Desulfotalea_psychrophila_LSv
BPrDe_Dvul	3379	1677	49.63	1831	54.19	1544	45.69	Desulfovibrio_vulgaris
BPrDe_Gsul	3446	1942	56.36	2182	63.32	1728	50.15	Geobacter_sulfurreducens
BPrEp_Cjej	1629	946	58.07	1041	63.90	738	45.30	Campylobacter_jejuni
BPrEp_Hhep	1875	957	51.04	1050	56.00	821	43.79	Helicobacter_hepaticus
BPrEp_Hpyl	1576	770	48.86	848	53.81	672	42.64	Helicobacter_pylori
BPrEp_Wsuc	2043	1205	58.98	1316	64.42	1054	51.59	Wolinella_succinogenes
BPrGa_Asp.	3325	1908	57.38	2097	63.07	1674	50.35	Acinetobacter
BPrGa_Cbur	2010	949	47.21	1053	52.39	854	42.49	Coxiella_burnetii
BPrGa_Ecar	4472	2542	56.84	2762	61.76	2222	49.69	Erwinia_carotovora
BPrGa_Ecol	4237	2508	59.19	2721	64.22	1607	37.93	Escherichia_coli
BPrGa_Hduc	1717	889	51.78	975	56.79	793	46.19	Haemophilus_ducreyi
BPrGa_Hinf	1657	1077	65.00	1153	69.58	955	57.63	Haemophilus_influenzae_Rd
BPrGa_Paer	5567	3374	60.61	3675	66.01	2960	53.17	Pseudomonas_aeruginosa
BPrGa_Plum	4683	2267	48.41	2581	55.11	2021	43.16	Photorhabdus_luminescens
BPrGa_Pmul	2015	1284	63.72	1372	68.09	1130	56.08	Pasteurella_multocida
BPrGa_Pput	5350	3095	57.85	3351	62.64	2733	51.08	Pseudomonas_putida
BPrGa_Psyr	3971	2174	54.75	3248	81.79	2639	66.46	Pseudomonas_syringae_pv_tomato
BPrGa_Sent	4395	2384	54.24	2602	59.20	871	19.82	Salmonella_enterica
BPrGa_Sfle	4182	2293	54.83	2513	60.09	948	22.67	Shigella_flexneri_2a
BPrGa_Sone	4324	2233	51.64	2545	58.86	2017	46.65	Shewanella_oneidensis
BPrGa_Styp	4425	2545	57.51	2764	62.46	1768	39.95	Salmonella_typhimurium
BPrGa_Vcho	3835	1983	51.71	2191	57.13	1757	45.81	Vibrio_cholerae_O1_biovar_eltor
BPrGa_Vpar	4756	2482	52.19	2789	58.64	2255	47.41	Vibrio_parahaemolyticus
BPrGa_Vvu*	4488	2459	54.79	2728	60.78	1958	43.63	Vibrio_vulnificus_CMCP6
BPrGa_Vvul	4955	2502	50.49	2770	55.90	2002	40.40	Vibrio_vulnificus_YJ016
BPrGa_Xaxo	4312	2404	55.75	2677	62.08	2129	49.37	Xanthomonas_axonopodis_pv_citri
BPrGa_Xcam	4181	2368	56.64	2625	62.78	2100	50.23	Xanthomonas_campestris_pv_campestris
BPrGa_Xfas	2766	1186	42.88	1309	47.32	1047	37.85	Xylella_fastidiosa_9a5c
BPrGa_Ypes	3885	2122	54.62	2346	60.39	787	20.26	Yersinia_pestis
BPrGa_Ypse	3901	2190	56.14	2376	60.91	1264	32.40	Yersinia_pseudotuberculosis
BSpSp_Lint	4727	1737	36.75	1946	41.17	1220	25.81	Leptospira_interrogans_serovar_Lai
BThTh_Tmar	1858	1127	60.66	1233	66.36	952	51.24	Thermotoga_maritima
E	651775	311201	47.75	323078	49.57	101035	42.40	All Eukaryotes
EAlAp_Pfal	5417	1959	36.16	2234	41.24	1869	34.50	Plasmodium_falciparum
EAlAp_Pyoe	7861	1873	23.83	2176	27.68	-	-	Plasmodium_yoelii
EFuAs_Agos	4718	2474	52.44	2765	58.61	2239	47.46	Ashbya_gossypii
EFuAs_Anid	9541	5113	53.59	5646	59.18	-	-	Aspergillus_nidulans
EFuAs_Calb	12017	5722	47.62	3427	28.52	-	-	Candida_albicans
EFuAs_Cgla	5181	2705	52.21	2983	57.58	2439	47.08	Candida_glabrata
EFuAs_Dhan	6317	3109	49.22	3667	58.05	2725	43.14	Debaromyces_hansenii
EFuAs_Fgra	11640	5690	48.88	6327	54.36	-	-	Fusarium_graminearum
EFuAs_Klac	5327	2664	50.01	2977	55.89	2398	45.02	Kluyveromyces_lactis
EFuAs_Kwal	5214	2626	50.36	2932	56.23	-	-	Kluyveromyces_waltii
EFuAs_Mgri	11109	4707	42.37	5274	47.48	-	-	Magnaporthe_grisea
EFuAs_Ncra	10620	3906	36.78	4398	41.41	-	-	Neurospora_crassa
EFuAs_Sbay	18770	3044	16.22	3375	17.98	-	-	Saccharomyces_bayanus
EFuAs_Scer	11686	3017	25.82	3346	28.63	2685	22.98	Saccharomyces_cerevisiae
EFuAs_Smik	18032	3074	17.05	3411	18.92	-	-	Saccharomyces_mikatae
EFuAs_Spar	17878	3011	16.84	3301	18.46	-	-	Saccharomyces_paradoxus
EFuAs_Spom	5000	2748	54.96	3039	60.78	2487	49.74	Schizosaccharomyces_pombe
EFuAs_Ylip	6521	3259	49.98	3682	56.46	2895	44.40	Yarrowia_lipolytica
EFuBa_Umay	6522	3260	49.98	3626	55.60	-	-	Ustilago_maydis
EMeAr_Agam	15212	7697	50.60	7758	51.00	-	-	Anopheles_gambiae
EMeAr_Dmel	10538	5581	52.96	11317	107.39	8972	85.14	Drosophila_melanogaster
EMeCh_Cint	21574	12426	57.60	9013	41.78	-	-	Ciona_intestinalis
EMeCh_Drer	29663	18154	61.20	22321	75.25	6147	20.72	Danio_rerio
EMeCh_Frub	33003	20443	61.94	14528	44.02	-	-	Fugu_rubripes
EMeCh_Ggal	28416	16097	56.65	13433	47.27	-	-	Gallus_gallus
EMeCh_Hsap	33869	19095	56.38	21427	63.26	20374	60.16	Homo_sapiens
EMeCh_Mmus	31535	17873	56.68	23208	73.59	17866	56.65	Mus_musculus
EMeCh_Ptro	39648	18421	46.46	20986	52.93	-	-	Pan_troglodytes
EMeCh_Rnor	32543	19644	60.36	22160	68.09	6720	20.65	Rattus_norvegicus
EMeCh_Xtro	52786	33092	62.69	17024	32.25	-	-	Xenopus_tropicalis
EMeNe_Cbri	19334	7260	37.55	9000	46.55	-	-	Caenorhabditis_briggsae
EMeNe_Cele	21124	8279	39.19	11177	52.91	8646	40.93	Caenorhabditis_elegans
EMyDi_Ddis	13049	5079	38.92	6410	49.12	-	-	Dictyostelium_discoideum
EViSt_Atha	28860	14784	51.23	18019	62.44	12573	43.57	Arabidopsis_thaliana
EViSt_Osat	61250	23315	38.07	26711	43.61	-	-	Oryza_sativa
all	1032102	513248	49.73	547292	53.03	276209	45.69	All Genomes