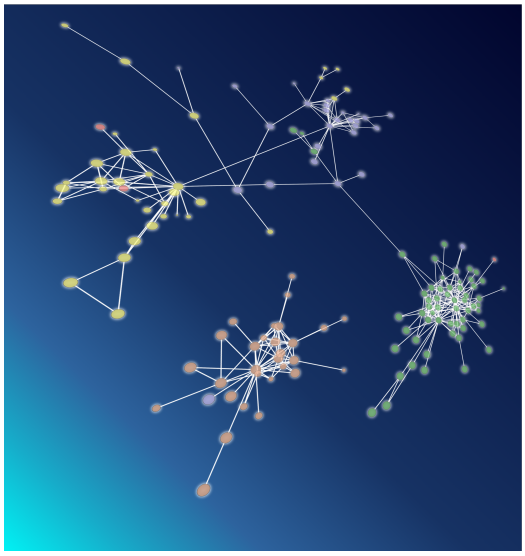
Department of Statistics
University of Oxford

Exploring the fold space preferences of ancient and newborn protein superfamilies

Hannah Edwards



a thesis submitted for the degree of
Doctor of Philosophy

Trinity term 2014

This thesis is dedicated to my family.

In memory of my father Ray, and with love to my mother Ann, and sister Ruth.

For the love, the laughter, the tears, and the constant support.

I love you all.

Acknowledgements

This thesis is the culmination of three wonderful years in the presence of many people without whom it would not have been possible. My first and most significant thanks go to my supervisor Charlotte Deane, who has been a constant supportive, creative and lively influence. I am constantly in awe of your seemingly endless motivation and your infectious enthusiasm has left me, I suspect, with a lifelong passion for both proteins and bioinformatics. I'd also like to thank Sanne Abeln, whose PhD thesis helped inspire much of the work in mine, and with whom collaborations and conversations have greatly contributed to my understanding of its themes.

I've been ridiculously lucky to be surrounded by an array of such incredible personalities in my family and friends. To Mireille, Tiago, Jamie, JP, James, Henry, Rey, Claire and all the rest of the OPIGlets, thanks for all the pool, punting, pints, parties (and proteins). To Anna, a true friend, so much more than a housemate, and emergency chauffeur, thanks for everything... especially the giggles. To Mole, Nick, Jo, Robin and the rest of my London family, thanks for being my home away from home and being there through it all. To Anna Jones, scouser, poet and all round happy person, thanks for the good times, the wire and the ode!

Finally I'd like to thank my family. Super mum and phlegmy. It isn't possible to fully express how grateful I am for your love, understanding and support over the last three years (and the twenty four before that).

Abstract

Protein evolution is a complex and diverse process, yielding an incredible assortment of biological functions and pathways occurring in the cells of living organisms. The way in which a protein's structure is constrained by its functional role and its notable conservation across even distant evolutionary relationships highlight structure as an important unit when considering the evolutionary dynamics of proteins. This thesis attempts to place the structural landscape of the protein universe within an evolutionary framework.

We investigate potential evolutionary histories of protein superfamilies by introducing an age, which estimates when the ancestor of that superfamily first evolved. The range of ages of known protein superfamilies goes right back to those which evolved before the diversification of life into three major superkingdoms. The structures of these proteins are varied but those which have evolved more recently tend to be shorter and have a less elaborate globular packing.

Protein structures sit within a complex global landscape of three-dimensional folds and we attempt to model the dynamics of this space using networks of folds. These networks consist of a structurally diverse core of folds with older ages, and neighbouring folds tend to be of similar ages. Moreover, there are a few pivotal folds which appear repeatedly as central in the landscapes, connecting together otherwise disparate portions of the space.

Sequence profiles which capture patterns of conservation and variation amongst naturally occurring proteins within a superfamily can be compared to identify distant evolutionary relationships. The power of these profiles to detect such relationships is improved by seeding them with structural alignments. A landscape of evolutionary links crossing between different protein folds is presented.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

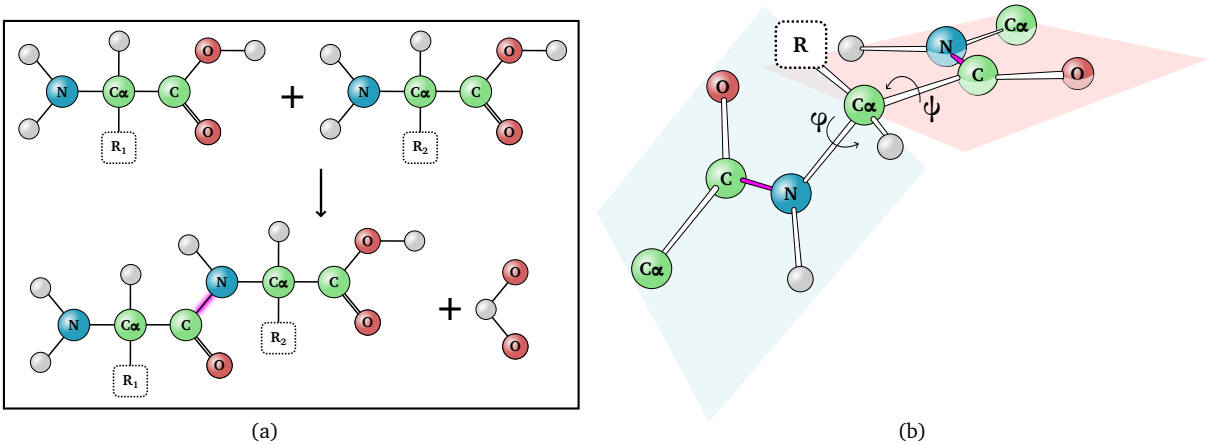
Proteins are the molecular machinery of the cell. The evolution of their mechanistic diversity is one of the main reasons for the incredible complexity and diversity of life we see around ourselves today. With an ever increasing wealth of data from protein sequences and structures there is great potential to examine the evolutionary behaviours, histories and relationships of different protein progenies. While the evolutionary forces which act on proteins are many and complex there is evidence that they are, at least in part, guided by structural constraints. The aim of this thesis is to place the structural landscape of proteins, or fold space, within an evolutionary framework.

1.1 Outline

In this introductory chapter, several themes which are relevant to the overall aim of the thesis are introduced. An initial section introduces the basics of protein anatomy and what we mean by the different levels of protein structure. Protein evolution is discussed in the next three sections. Initially, the biological mechanisms behind evolutionary change are examined. In two further sections we look at the concept of structure as an evolutionary unit: both exploring structure as a conserved property as well as examining evidence for structural variation between related proteins. Following this discussion of evolutionary mechanisms, different methods of determining homology are summarised. The methods looked at in this section can be divided into two categories: automatic alignment based methods which attempt to capture similarities between two or more chains, and classification schemes, many of which also use alignment based methods along with a host of other tools to establish homology. The chapter then introduces the concept of fold space: the global structure space within which naturally occurring proteins sit. Finally a thesis outline is given, briefly summarising the work in the following five chapters.

1.2 Protein anatomy and folding

Proteins are macromolecules which are found inside, on or near the cells of all life forms. They carry out an incredible diversity of different roles and are key players in almost every biological process. They are made up of chains of amino acids. These amino acids are chemical compounds with a variety of different physical and chemical properties. Each amino acid contains an amine (NH_2) and a carboxylic acid (COOH) group, as well as a side chain group which is unique to each amino acid. Amino acids are joined by a condensation reaction where a hydrogen atom is lost from the amine group and the hydroxide (OH) is lost from the carboxylic acid to form water. This reaction results in the residues of the amino acids being linked by a peptide bond. This reaction is shown in Figure 1.1a. Joined by peptide bonds, the chain of nitrogen and carbon atoms belonging to each amino acid are known as the backbone of the polypeptide. To



(a) (b)

Figure 1.1: The polymerisation of amino acids to form peptides. Carbon atoms are shown in green, nitrogen in blue, oxygen in red and hydrogen atoms in grey. Peptide bonds are shown in magenta. (a) Two amino acids join through a condensation reaction. Hydroxide from the carboxylic acid group of the first amino acid and hydrogen from the amine group of the second amino acid combine to form a water molecule, leaving the residues of these amino acids linked by a bond known as the peptide bond. (b) Peptide bonds are restricted in conformation but the other two bonds between backbone atoms of the polypeptide are not. Two angles, ϕ and ψ, capture their conformation. ϕ gives the angle between successive C atoms, from the reference point of the N-C_α bond. Similarly, ψ gives the angle between successive N atoms, from the reference of the C_α-C bond.

distinguish it from the other carbon atom along the backbone chain the carbon atom which is bound to the side chain R group is known as C_α. The peptide bond is restricted in conformation due to the partial double bond which it shares with the carbonyl carbon atom. This bond has only two possible conformations, *cis* and *trans*, with the *trans* conformation 1000 times more stable than the *cis* [Kessel and Ben-Tal, 2011]. The other bonds along the backbone are unconstrained however and can be characterised by two angles, ϕ and ψ, known as dihedral angles. These angles are illustrated in Figure 1.1b.

The overall conformation of these dihedral angles along the polypeptide chain is driven by several forces, particularly the interactions between the different side chains with each other and their surroundings. There are twenty different side chains which occur in basic amino acids. These twenty amino acids are shown in a table in Figure 1.2. Each amino acid's unique side chain causes it to exhibit specific properties, however there are several categories which can be used to group amino acids together. Most notable is the distinction between hydrophobic and

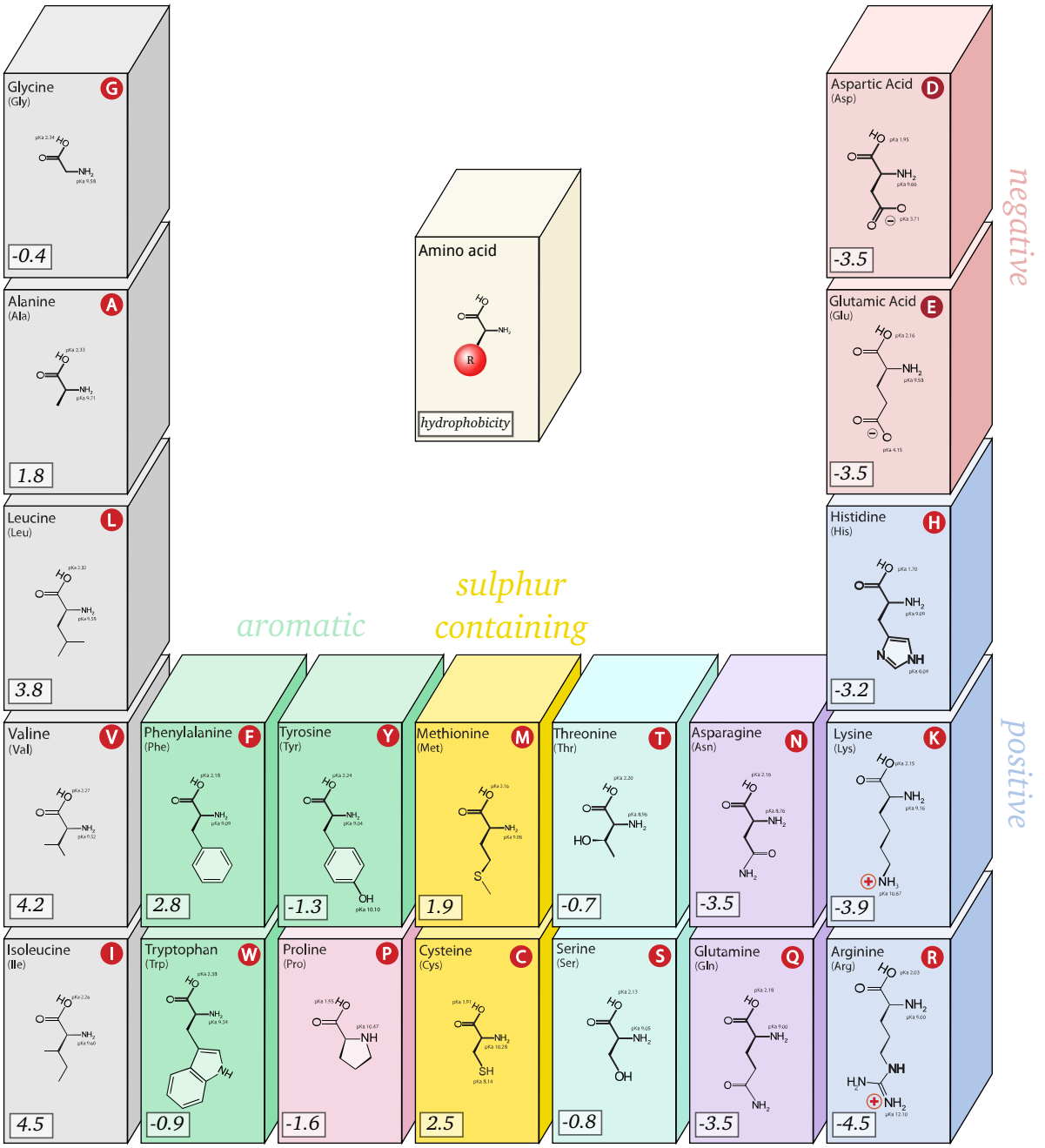


Figure 1.2: A periodic table of amino acids. The chemical structures of each of the twenty amino acids are shown, as well as their name and three letter and single letter abbreviations. Their hydrophobicities according to the scale of Kyte and Doolittle [1982] are shown in boxes. Amino acids are grouped according to their properties. Broadly speaking the hydrophobic residues are shown in grey, polar, charged amino acids are shown in red (negative) and blue (positive), aromatic residues in green, and sulphur-containing residues in yellow. This figure has been adapted with permission from Guzzetta [2000].

polar side chains. Hydrophobic amino acids, roughly speaking the left half of the table in Figure 1.2, tend to avoid water and thus prefer to be protected from the aqueous solution which most proteins are found in. Polar amino acids are either partially or fully charged, and thus interact preferably with water. Polar amino acids are found on the right of the table in Figure 1.2 and charged residues are the rightmost column.

1.2.1 Levels of protein structure

The folding of protein chains, driven by the interactions between their constituent residues, yields compact three-dimensional structures. A protein's structure acts as a scaffold for its function, and also prevents the aggregation of different proteins which can be detrimental to the cell. Moreover, examination of these structures reveals a hierarchy of different levels of organisation. An illustration of how these different levels contribute to the protein fold is shown in Figure 1.3. This figure shows the cAMP receptor protein (CRP), a transcriptional activator, which binds to DNA and promotes transcription of that section of the genome into RNA through an interaction with RNA polymerase. Typical of globular proteins, CRP also exhibits a structural organisation at multiple levels.

At the lowest level, the structures of each amino acid constituting the chain is known as the primary structure of the protein. There are also regular substructures known as secondary structures. The formation of these substructures is driven by the formation of hydrogen bonds between backbone atoms of the chain. The formation of such bonds helps stabilise the polar backbone amide (NH) and carbonyl (CO) groups, particularly in the hydrophobic core of the protein [Kessel and Ben-Tal, 2011]. Two particular substructures maximise this hydrogen bonding potential as well as providing particularly compact conformations of the atoms making up the chain. They are the α-helix and the β-strand.

α-helices Helices are the most common secondary structure in globular proteins. While the helical content of different chains differ widely, around 26% of residues in globular proteins are

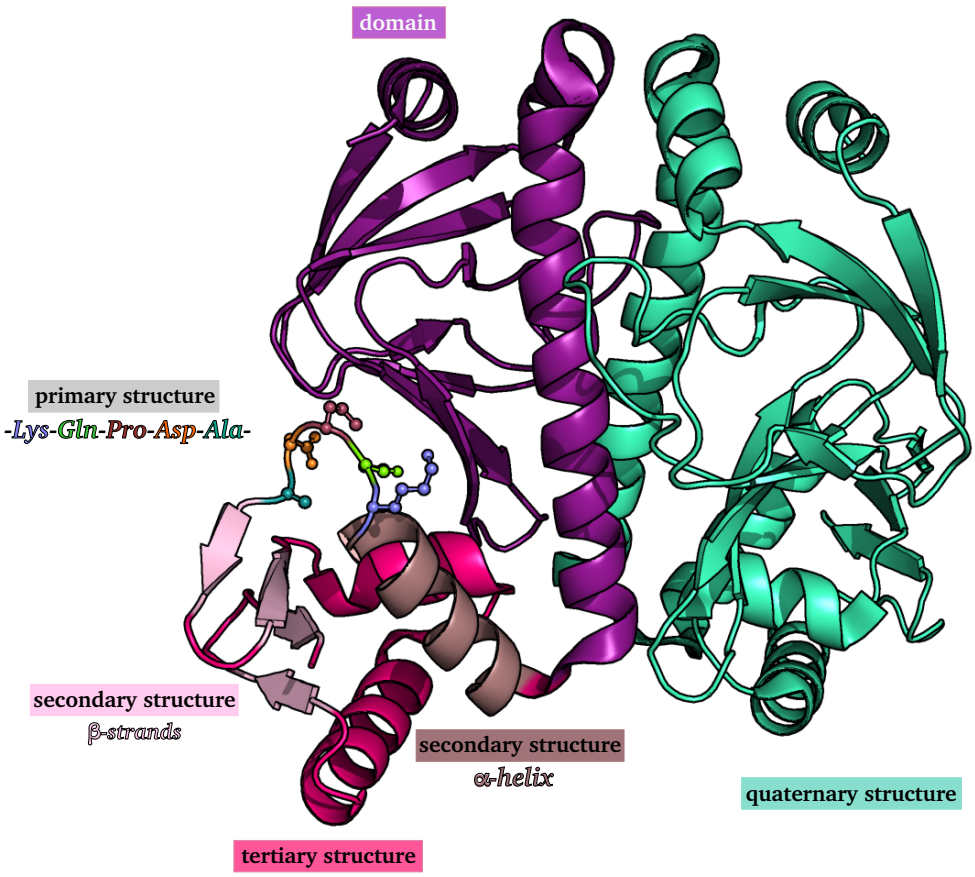


Figure 1.3: Levels of protein structure within the cAMP receptor protein (PDB entry 1G6N). The polypeptide chains which make up a protein form regular structures at multiple different levels. At the primary level, the carbonyl and amide groups of successive amino acids form peptide bonds resulting in a backbone chain. Each amino acid also consists of a unique side chain group. Here, five amino acids are shown along part of the protein. The backbone chain is shown as a smooth loop, and the side chain atoms are shown as balls with sticks representing their covalent bonds. Secondary structure refers to local substructures within the protein. There are two popular types of secondary structure: the α-helix shown here in maroon, and the β-strand shown in light pink. These structures form through hydrogen bonding between backbone carbonyl and amide groups: locally in the case of the helix, and between different β-strands which align to form sheets or barrels. The overall three-dimensional organisation of the protein chain and its secondary structure units is called the tertiary structure, and independent folding units within a chain are called domains. The CRP has two chains, and within each chain is two domains, shown here in pink and purple. In the N-terminal domain (purple), the tertiary structure consists of a long *alpha*-helix, and an unusual β-sheet formation, known as a β-helix. This structure forms the basis of the binding site of the protein with cAMP. The C-terminal domain's (pink) tertiary structure forms a three α-helical bundle next to a four stranded β sheet. The helical bundle has a specific orientation known as a 'winged' bundle, and binds to DNA. The quaternary level of structure is the orientation of the different chains to form the protein unit. In the CRP the relative orientation of the two chains is mitigated by the N-terminal domain and differs when the protein binds to cAMP [Passner *et al.*, 2000].

found within α-helical substructures [Kessel and Ben-Tal, 2011]. α-helices are local substructures resembling a spiral and are formed by hydrogen bonding between the carbonyl group of a residue, at position i within the sequence, and the amide group of the residue at position $i+4$.

β-strands β-strands are the next most common secondary structure elements. Successive residues are found in a pleated or zigzag conformation to form a strand. Hydrogen bonds form between residues in neighbouring strands, so β-strands are considered a non-local secondary structure. Strands are rarely seen in isolation, but are arranged in successively hydrogen bonded sheets. In some cases, sheets curve round on themselves and the first and last strand also share hydrogen bonds and become a barrel. β-strands are slightly less common than α-helices but still account for around 19% of residues in globular proteins [Kessel and Ben-Tal, 2011]. The way that β-strands align next to each other within the sheet or barrel can also differ. If successive strands run (in the N-C direction) in the same direction, they are known as parallel strands. On the other hand if they run in opposite directions they are known as anti-parallel. Anti-parallel strands can be connected by short strings of only a few residues, known as β-turns. In general, parallel strands require much longer linking sections between successive strands. Where successive parallel strands are separated by an α-helix, it is known as an β-α-β motif. In Figure 1.3, the highlighted sheet is an anti-parallel sheet of four β-strands. There are two β-turns connecting the first two and the last two strands. The sequence ordering of strands within the sheet is 1243, and two helices are found between the second and third strands.

There are other regular secondary substructures, although these are much less common than the α-helices and β-strands. They include the 3_{10} and π helices. Sections of the chain without regular hydrogen bonded secondary structure are known as loops.

Non-covalent interactions, particular between side chain groups, drive the next level of structural organisation: the tertiary structure of a chain. Different chains may have identical secondary structure content and ordering, yet look completely different as three-dimensional objects. By burying hydrophobic side chains in the core of the protein away from the aqueous

environment within the cell, and maintaining polar residues at its surface, the tertiary structure of a protein chain ensures its stability and solubility. Moreover, these three-dimensional conformations also drive the functional specialisation of the protein, for example by allowing the chain to fold and produce binding sites which recognise and bind to specific molecules. Often a single protein chain contains separate three-dimensional assemblages, all of which are independently folding units and often perform specific functions [Kessel and Ben-Tal, 2011]. These units are known as domains, and are considered the structural, functional and evolutionary building blocks of the protein universe. The tertiary structures of different domains are often referred to as their folds.

The final level of structure evident within a protein is in the interactions between separate chains. Proteins with multiple chains in conformation moderated by non-covalent interactions are known as oligomers. In Figure 1.3, the CRP contains two chains which form a dimer. This quaternary organisation of protein structures often results in some form of symmetry [Goodsell and Olson, 2000].

1.2.2 Determining protein structure

Proteins are sub-cellular macromolecules and are typically less than 100 nanometres in size, orders of magnitude smaller than the wavelength of the visible light spectrum. Because of this it is impossible to determine their structures visually. The first high resolution image of a protein structure wasn't seen until 1958, when [Kendrew *et al.*, 1958] used a method called X-ray crystallography to solve the structure of the protein myoglobin.

1.2.2.1 The PDB

Since this first structure was determined, over 100,000 different structures have been solved and deposited in the Protein Data Bank (PDB) [Berman *et al.*, 2000]. The PDB is a worldwide online archive dedicated to storing records of experimentally determined structures of macromolecules, including proteins, nucleic acids and complex assemblies. Each entry in the PDB contains a list

of the coordinates of the atoms comprising the macromolecule, as well information about the data acquisition and processing.

1.2.2.2 X-ray crystallography for protein structure determination

X-ray crystallography bombards crystallised proteins with X-rays. The electrons of the crystalline atoms cause the X-rays to diffract and form a pattern, from which the structure of the atoms can be deduced [Yaffe, 2005].

X-ray diffraction remains the most popular method for determining high quality protein structures, with around 87% of PDB structures solved using X-ray crystallography. However, the method is not without problems. Firstly, crystallisation of a protein requires a very specific set of conditions, and the ideal conditions differ depending on the protein. Some proteins such as membrane proteins are very hard to crystallise, due to the high hydrophobicity of their surfaces. Secondly, the phases of the X-rays need to be inferred in order to determine the structure [Hauptman, 1991]. Thirdly, the protein needs to be in crystal form, which is not its natural environment. Therefore the resultant structure is not guaranteed to resemble the protein as it is in the cell. In particular, any dynamic behaviour in the protein's conformation is largely lost.

1.2.2.3 NMR spectroscopy for protein structure determination

An alternative to crystallography is nuclear magnetic resonance (NMR) spectroscopy. NMR spectroscopy determines the structure of a protein based on the magnetic fields of its constituent atoms and their intramolecular environment. The method exposes a protein chain in solution to a strong magnetic field and then measures the radio waves re-emitted by the polypeptide. At the time of writing, structures solved using NMR make up about 10% of the PDB. Unlike crystallography NMR can be used to determine protein structures in solution and detect their natural dynamics, and even binding processes.

1.2.3 Visualising protein structure

High resolution protein structures at the atomic level can be visualised using PyMOL, a molecular visualisation system [Schrödinger, LLC, 2010]. In Figure 1.4 three such representations are shown. Stick representations, such as Figure 1.4a, show covalently bonds between atoms in the protein as sticks. Cartoon representations, as in Figure 1.4b, which are used frequently throughout this thesis, are derived from the backbone chains of the protein. The chain is smoothed for ease of vision, and secondary structure elements are highlighted: helices as spirals and strands as arrows. Using this type of representation, it is usually easy to examine the fold of a protein: in this case a barrel with 8 strands surrounded by helices. The final representation, shown here in Figure 1.4c, is the surface representation. Here, atoms are visualised as spheres, and the surface of the protein is shown as the accessible surface area of these spheres. Using this representation, notable features relating to the realistic surface of the protein can be examined: for example, any binding pockets or clefts.

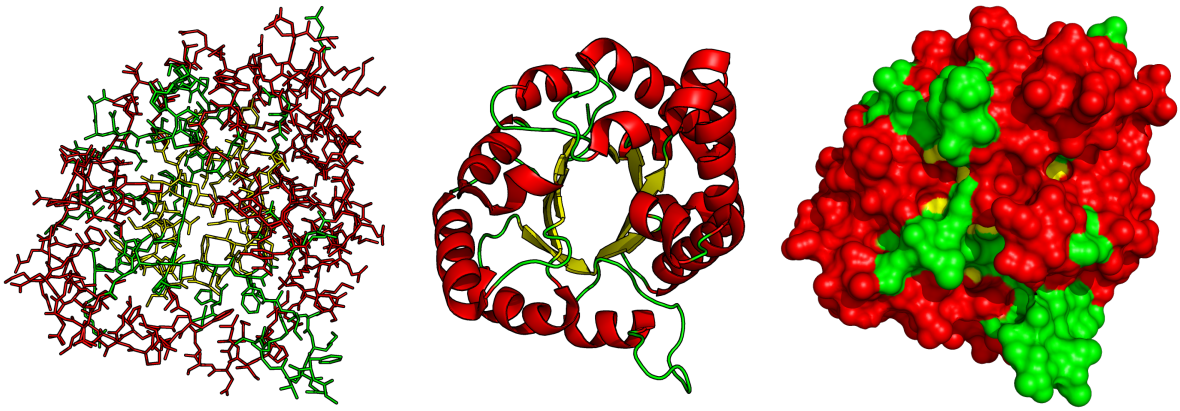


Figure 1.4: Graphical representations of protein structures produced with PyMOL [Schrödinger, LLC, 2010]. Three representations of the TIM barrel protein triosephosphate isomerase (PDB entry 1N55). Residues are coloured by their secondary structure types: with helices in red, strands in yellow, and loops in green. **(a)** Stick representation. Bonds between atoms are drawn as sticks. **(b)** Cartoon representation of the protein backbone. The backbone is smoothed and helical and strand sections are highlighted. Side chains are invisible. **(c)** Surface representation. The accessible surface area of the protein is shown by treating each atom as a sphere.

1.3 Evolution of protein sequence and structure: biological mechanisms

Protein sequences are encoded by genetic material such as DNA and RNA. As such they are susceptible to evolutionary change as the genes which encode them mutate. In very simple terms evolutionary change is the consequence of two separate forces. The first is stochastic change, or mutation, to the genetic material which can cause modification to a protein. Mutations can leave a protein sequence unchanged, due to codon degeneracy, or involve large scale changes which can disrupt the very core of its structure. The second force is that of selection. Mutational changes which result in deleterious effects on the protein, for example, distorting the active site or causing misfolding and aggregation of the protein, will in turn have negative effects on the cell, or organism, it is found in. These effects induce a diminished capacity to reproduce and gradually the deleterious mutation is lost [Ross and Poirier, 2004; Worth *et al.*, 2009]. Similarly, a beneficial or neutral mutation may remain in the population.

1.3.1 Mutational mechanisms

The processes through which proteins develop mutational change are complex and beyond the scope of this thesis. However, a non-exhaustive list of some of the most common types of mutations is given in this section.

Substitutions The most common mutation is a change to a single character of a protein's sequence, known as a substitution or point mutation. This occurs when one nucleotide is replaced by another, often as a result of errors in DNA replication. Substitutions to nucleotides can have no effect on a protein's amino acid, if the altered codon still codes for the same amino acid. On the other hand, substitutions which alter stop or start codons can have a dramatic effect on a protein, producing phenomena such as gene fusion or premature truncation of a protein.

Indels Insertions and deletions, collectively termed indels, are the second most common type of mutation. They occur about an order of magnitude less frequently than substitutions [Benner *et al.*, 1993], although as with other types of mutation their rate varies both within and between different genomes [Cooper *et al.*, 2004; Pál *et al.*, 2006; Hahn *et al.*, 2007; Chen *et al.*, 2009]. Indels occur when sections of one or more nucleotides are either inserted or deleted from the genetic sequence. Insertions are often the result of transposable elements: mobile genetic elements which can change their position on the genome [Fedoroff, 2012].

Gene duplication Gene duplication occurs when a mutation gives rise to multiple copies of a single section of genetic material. It can happen as the result of unequal crossovers in the recombination of DNA, or through retroposition of RNA, where mRNA is transcribed back into DNA and then inserted into the genome [Zhang, 2003]. Duplication is one of the major forces behind the evolution of novelty in the protein universe. This is because duplication of part or the whole of a protein creates functional redundancy in the copy [Zhang, 2003]. The redundant protein can therefore accommodate changes which may alter its structure and function.

Exon shuffling Exon shuffling is the phenomenon where the nucleotides of a genetic sequence remain unchanged, yet occur in a different order. It can occur due to recombination, alternative splicing or duplication events followed by partial deletion [Patthy, 1999]. In proteins, exon shuffling events can result in relationships such as circular permutations [Jeltsch, 1999].

Lateral gene transfer Lateral gene transfer refers to mutation events where foreign genetic material is transferred to an organism by methods other than vertical descent and then incorporated into its genome. An example of a mechanism which results in this type of mutation is the transduction of DNA from one bacteria to another through infection by a virus. Lateral gene transfer has a powerful potential for evolutionary innovation, as completely novel yet functionally mature sequences can appear in a genome as the result of a single evolutionary event [Gogarten and Townsend, 2005]. It is often hard to incorporate lateral gene transfer in

traditional models of evolution, which focus on descent by reproduction [Doolittle and Bapteste, 2007]. Moreover, a majority of the mechanisms which result in lateral gene transfer are unique to Prokaryotes [Gogarten and Townsend, 2005]. Thus, while it is recognised as a prominent process amongst Prokaryotic genomes, its relative frequency amongst Eukaryotes is still debated [Andersson, 2005].

1.3.2 Natural selection and the protein record

The array of proteins which we see in nature represent a record of such mutational changes, but moderated by selective pressure. The changes listed above can affect proteins in a variety of different ways. Those which are largely deleterious will not be visible in this record. Early studies of the protein record and observed mutations suggested that the majority of selected change is neutral [Kimura, 1968; Ohta, 1973]. This phenomenon is known as genetic drift.

In this thesis we focus on evolution at the domain level of protein structure. We examine this force using the protein record so are concerned with how the combined forces of mutation and selection affect these units.

1.4 Structure as a conserved evolutionary unit

Ever since the first protein structures were solved it became evident how often the tertiary structures of seemingly distinct proteins resembled each other. Work by Chothia and Lesk [1986] later demonstrated that, under evolutionary processes, protein structures were highly conserved relative to their sequences. Since then protein structures have been viewed as important units within the evolutionary process.

This view is motivated by the importance of maintaining function as a constraint on evolutionary selection pressures [Worth *et al.*, 2009]. Since the tertiary structure of a protein is inherently linked to its function and the role it plays in the cell, evolutionary pressure to maintain the function of a protein will be evident in the conservation of its structure. This motivation

is further supported by observing that conservation under evolutionary change appears stronger in functionally sensitive areas, such as the active site. Figure 1.5 demonstrates this observation in the case of the protein triose phosphate isomerase. On examination of 494 closely related homologous proteins, positions in the sequences which correspond to a close proximity in tertiary space to the binding pocket contain fewer mutations than positions further away from the active site.

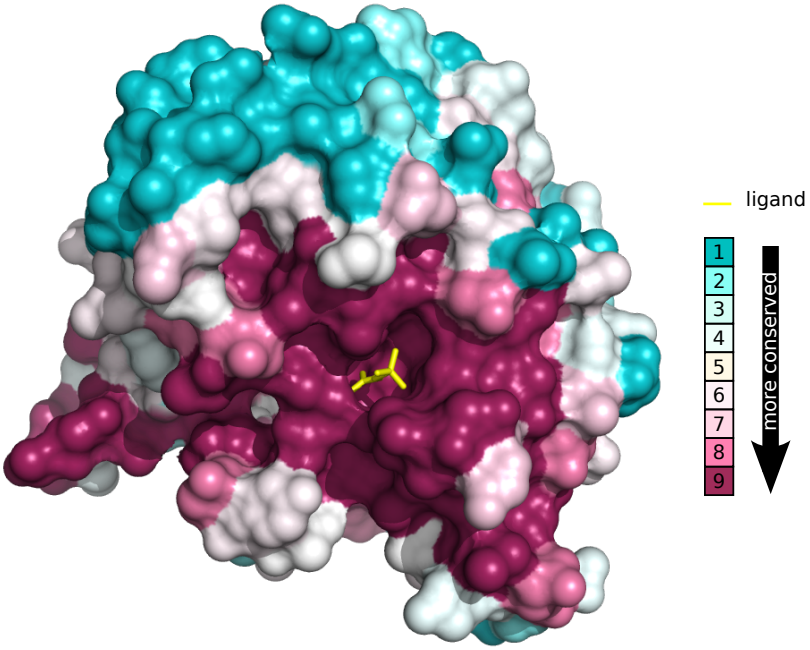


Figure 1.5: Evolutionary conservation of residues in triose phosphate isomerase. A surface representation of PDB entry 1AMK is shown where residues are coloured according to their conservation calculated using the ConSurf server [Glaser *et al.*, 2003]. The method calculates a position-specific conservation score based on a multiple sequence alignment of the seed sequence (1AMK) and 494 homologous sequences detected by a PSI-BLAST search with an E-value ≤ 0.0001. The score is then normalised between 1 (variable position) and 9 (conserved position). The enzyme's substrate is shown in yellow and binds at the centre of the most conserved region of the structure. This example illustrates the importance of maintenance of function as a constraint on the evolution of proteins.

Furthermore, even non-conserved positions tend to disrupt the overall structure very little. The pure scale of this observation has been used recently to great effect in the area of protein structure prediction. Huge sequence alignments of related proteins can be studied to identify

residues were mutations appear to be dependent on one another. These so-called correlated mutations are signatures of the common fold belonging to all the sequences and tend to represent residues in contact within that fold [Marks *et al.*, 2011].

1.5 Structural variation in related proteins

While the conservation of structure is a powerful force in evolutionary studies it is not universal. There are several examples where related proteins differ in their overall fold. Figure 1.6 shows five examples of such changes. In all these cases, significant similarity at the sequence and functional levels suggests a homologous link. However, the overall structures of these related domains differ.

Figure 1.6a shows an example of fold change in the SasA N-terminal domain and KaiB protein. These domains share a strong sequence similarity and have identical binding partners [Andreeva and Murzin, 2006]. The structures of their N terminals are also very similar, comprising a common β-α-β motif. However, their C terminal sections differ from each other, with the β-β-α motif in SasA substituted for a α-α-β motif in KaiB.

In Figure 1.6b, an example of homologous β-barrel and β-sandwich domains are shown. Here, structures from two Rieske proteins are shown, one taken from the mitochondria, the other from chloroplasts [Murzin, 1998]. These proteins are once again sequence and functionally similar. However, the arrangement of their β strands differ with the mitochondrial protein folding into a sandwich, and the chloroplast Rieske forming a barrel. Such examples have suggested the possibility of a continuous spectrum of β-sheet organisation, with no firm distinction between barrel and sandwich substructures [Richardson, 1981].

Figure 1.6c shows an example of circular permutation, thought to occur as a result of duplication followed by partial deletion [Grishin, 2001]. These two domains share high sequence identity (when aligned accordingly) and also share a functional similarity. Their overall secondary structure placement is very similar but the N and C terminals are found at different

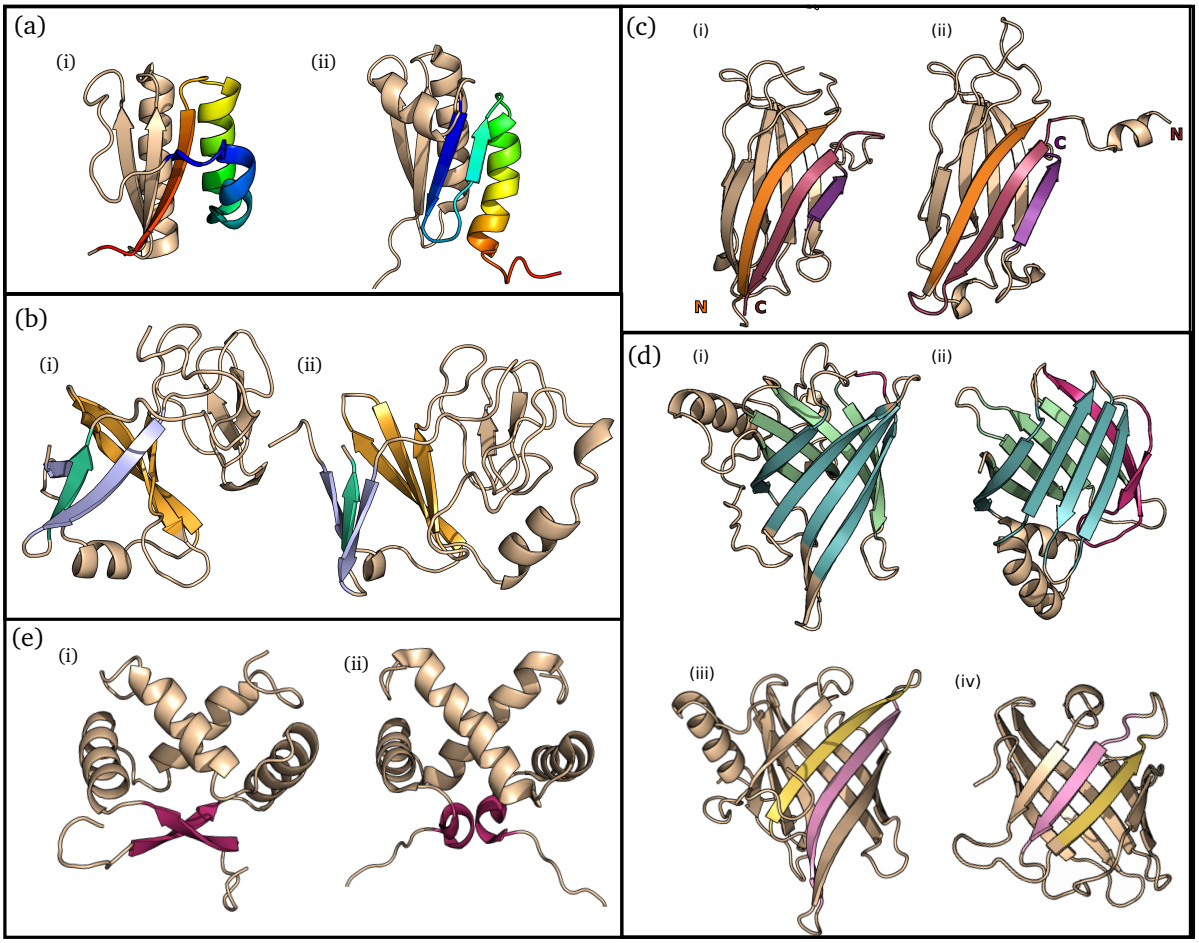


Figure 1.6: Five examples of structural variation between homologous proteins. Each pair of domains share significant sequence similarity to indicate homology yet their structures differ. Figures are produced using PyMOL using examples discussed in the cited publications. **(a)** Fold change in (i) SasA N terminal domain (1T4Y) and (ii) KaiB (1R5P) [Andreeva and Murzin, 2006]. The domains share a common β-α-β motif at their N terminals. The C terminals are coloured as a rainbow from blue to red to illustrate the C terminal structural variation from a $\beta\beta\alpha$ unit in SasA to a $\alpha\alpha\beta$unit in KaiB. **(b)** Evolutionary links between a β-sandwich and a β-barrel in the N terminal domains of Rieske proteins (i) a barrel from a chloroplast protein (1RFS) and (ii) a sandwich from a mitochondrial protein (1RIE) [Murzin, 1998]. **(c)** Circular permutation between (i) Phospholipase C C2 domain (1QAS) and (ii) Synaptotagmin C2 domain (1RSY) [Grishin, 2001]. The overall arrangement of secondary structures remain conserved, however the amino and carboxyl termini occur at different locations and thus the connectivity of the domains differ. For clarity, the three strands which occur at a terminus in either domain are coloured by their aligned regions. **(d)** strand invasion or withdrawal between (i) Retinol binding protein (1HBQ) and (ii) Retinoic acid binding protein (1CBS) [Grishin, 2001]. Two strands have been inserted into the β-barrel in 1CBS or deleted from 1HBQ. For clarity this region has been shown in magenta and equivalent strands at the front and back of the barrel have been coloured by their aligned regions (blue and green respectively). Also shown is the β-hairpin swap between (iii) Retinol-binding protein (a different view of (i)) and (iv) Thrombin inhibitor triabin (1AVG) [Grishin, 2001]. The domains have similar structures except that two β-strands have swapped places in the β-barrel. **(e)** Secondary structure swap in ARC repressor proteins between (i) the classic fold (1ARG) and (ii) a mutant (1NLA) [Cordes *et al.*, 2000].

places, altering the connectivity of the secondary structure elements.

Figure 1.6d shows examples of structural variability within Lipocalin proteins. In the two examples shown here three homologous proteins differ by a β-strand invasion/withdrawal in one pair of domains and a β-strand swap in another pair [Grishin, 2001]. The two domains which differ structurally through a β-strand invasion/withdrawal are shown first. The structural changes are dramatic, requiring the swapped elements to completely alter their hydrogen bonding patterns. The extra strands in the retinoic acid binding protein have been coloured pink. These strands alter the shape of the binding cavity leading to an adaption of function [Grishin, 2001].

The second example of structural variation within Lipocalins shows a β-strand swap. Like the previous example the hydrogen bonding patterns of strands within these protein differ due to the structural divergence while both sequences share significant sequence identity. Again the sequences of the two proteins share similarity indicative of homology, although the swapped strands are less conserved than other areas of the protein. The retinol-binding protein exhibits the typical Lipocalin fold consisting of an up and down meander, while the triabin consists of a unique topology derived from the switch. Despite their sequence similarities, the structural divergence has functional implications. While most lipocalins, including the retinol-binding protein, bind ligands inside the barrel, the new loops between the two central strands of triabin block the open end of the barrel. Triabin functions instead as a protease inhibitor.

Finally, Figure 1.6e shows an example of secondary structure substitution in ARC repressor proteins [Cordes *et al.*, 2000], where two helices in one domain are substituted for two β-strand ribbons in the other. The proteins are almost identical in sequence and function.

1.5.1 Other evidence for structural plasticity

While there are only relatively few examples of structural variation amongst naturally-occurring homologous proteins, much more dramatic changes have been seen in designed proteins. For example, a recent study found a series of three amino acid substitutions at varied positions

throughout a designed protein sequence, each of which corresponded, independently, to fold switching between a 3α-helical and a $4\beta + \alpha$ sandwich [He *et al.*, 2012].

Moreover, structural transitions have been known to occur in the same protein, mediated by changes in environment. Whilst transitions are common amongst proteins, very few of these involve drastic structural rearrangements. However, recent studies suggest fold changes can occur in these circumstances [Bryan and Orban, 2010; Burmann *et al.*, 2012].

1.5.2 Structural analogy and homology

In general, although not absolutely as the above examples have shown, naturally occurring proteins related by evolution tend to be structurally similar. However, the converse is not necessarily true: that structurally similar proteins will always be homologous. In fact, there is an important distinction to be made between homology and structural analogy [Cheng *et al.*, 2008]. Determining homology requires a diverse set of tools, and while structural similarity can strengthen the case for an evolutionary link, it cannot be used in the absence of any other similarity. This is one of the reasons why homology detection is such a challenging field as will be discussed in the following sections.

1.6 Detecting homology

The evolutionary trajectories discussed in Section 1.3 cover a range of events from single residue mutations to larger scale structural changes which are the result of more dramatic modification. One of the most fundamental challenges in computational biology, and a subject on which much of the work presented in this thesis rests, is the identification of such evolutionary links between proteins. For closely related proteins, which have diverged through modest numbers of amino acid substitutions and small insertions and deletions, homology can be detected relatively simply. However, for more distantly related proteins, in particular those for whom a series of different evolutionary events have eliminated any detectable sequence similarity, this task

becomes harder. In general, establishing homology using protein sequences alone, gets markedly harder in the 'twilight zone' of $< 30\%$ sequence identity [Holm and Sander, 1996]. While structural similarity can suggest the possibility of homology beyond this zone, as we have already seen, this relationship remains in no way strict. It has been suggested that both similarity at the structural and functional levels are required for homology detection in the absence of sequence similarity [Murzin *et al.*, 1995].

In this section, two different approaches which can be used to assess homology are discussed: classification and alignment. Classification schemes use a full repertoire of sequence-based, structural, and functional analyses to establish evolutionary relationships between protein domains. Classification schemes can adapt to accommodate significant structural variation. For example, almost all the examples in Figure 1.6 are recognised as homologous under the most commonly used classification schemes. While powerful tools, these schemes are often limited by the lengthy, and often manual, process of establishing distant relationships, as well as the requirement that every member of the scheme have known structure. As such, the portion of the protein universe classified under such schemes remains low. Alignment methods meanwhile tend to involve algorithms which identify a correspondence between the characters of one protein and those of another. This correspondence can then be quantified and statistically evaluated to determine its significance. The algorithms tend to be much quicker than the classification process as they aim to maximise similarities of only one type: be it sequence or structure similarities, and as such are limited to finding only close homologues in the case of sequence alignment, and requiring additional analysis to prove homology in the case of a structural alignment.

1.6.1 Classification schemes

Classification schemes collate evolutionary information using a variety of sources. They aim to establish homology between proteins of known structure using sequence-based, structural and functional analyses. Protein domains are organised first into families, with significant sequence similarity, and then into superfamilies which represent groups of families with a probable

common evolutionary origin supported by similar structure and function. Often they include further categories relating to the fold or topology of a domain: its overall structure, with no reference to evolutionary histories. While classification schemes can organise protein domains using a variety of methods, often they resort to manual analysis to determine homology. Two such schemes are SCOP [Murzin *et al.*, 1995] and CATH [Orengo *et al.*, 1997]. These schemes both aim to describe structural and evolutionary relationships between protein domains, but differ in the methodologies used for this task. While there is a broad consensus between the two classification schemes, there are still important differences, which derive largely from differences in domain annotation [Day *et al.*, 2003]. Further differences can be seen in the structure of the clustering, with SCOP agreeing more with an average-linkage clustering of automated alignments on consensus domains and CATH agreeing more with the single-linkage clustering of the same data [Pascual-García *et al.*, 2009].

1.6.1.1 SCOP

SCOP stands for the **S**tructural **C**lassification **O**f **P**roteins [Murzin *et al.*, 1995]. Figure 1.7 shows the main categories in its hierarchical organisation. Protein domains are assigned first to a family, where domains typically share at least 30% sequence identity. Superfamilies are the scheme's homologous units, with families that share significant structural and functional similarities being clustered together in a single superfamily. Beyond this level the scheme focuses on structural, rather than evolutionary, organisation of proteins. Folds, the subsequent level in the hierarchy, need not share an evolutionary history (although it is possible that many do) but contain superfamilies with a common structural topology at their core. The final level of the hierarchy is the class level consisting of seven different categories describing the fold type. Four of these categories describe the majority secondary structure content of their domains. **All-$\boldsymbol{\alpha}$** domains consist of mainly α-helical substructures, **all-$\boldsymbol{\beta}$** similarly consist of a majority β-strand structure, $\boldsymbol{\alpha/\beta}$ consist of helices and strands arranged into β-α-β units, and $\boldsymbol{\alpha+\beta}$ domains contain segregated helices and strands. The three remaining classes cover additional categories:

multi-domain proteins, which contain two or more domains belonging to different classes but for which no distinct homologues have been identified, **membrane proteins**, which fold in a different environment (lipid bilayer) to the aqueous cell interior (mostly water) occupied by most soluble proteins, and **small proteins**, which often require additional stabilising features such as disulfide bridges and metal ion cofactors. There are further categories, such as coiled coil proteins, low resolution structures, peptides and designed proteins, although these are not counted as true classes.

As can be seen in Figure 1.7, each unit within the SCOP hierarchy is given a classification string: like c.23.10.3, which refers to the Acetylhydrase family, within the SGNH hydrolase superfamily (c.23.10), and the Flavodoxin-like fold (c.23), which is part of the α/β class (c). These strings are used throughout this thesis to refer to SCOP superfamilies and folds of interest. While these strings are unique identifiers, they sometimes vary between different SCOP releases. Lists of superfamily and fold classification strings along with their full names can be found in Appendix Tables A1 and A2 respectively.

A new version of SCOP, called SCOP2, has recently been released [Andreeva *et al.*, 2014]. This new classification scheme significantly alters the hierarchical structure typical of SCOP as shown in Figure 1.7. Traditional SCOP places the structural classifications of fold and class as parents in the hierarchy to evolutionary clusters of families and superfamilies. In SCOP2, the topology of the classification structure becomes a directed acyclic graph, rather than a tree, and separates evolutionary and structural relationships. In other words, simply because two domains can be placed within an evolutionary unit (superfamily) does not necessarily mean they share the same structure (fold). However, at the time of writing, SCOP2 remains a prototype with only 139 structural superfamilies defined. In this thesis SCOP 1.75, the latest full release of traditional SCOP, is used exclusively. This version classifies 110,800 domains (10,569 at $< 40\%$ identity) into 1962 superfamilies and 1195 folds. Table 1.1 gives a more detailed breakdown of the coverage of this version.

Sequence information and atomic structure files for domains classified under SCOP are

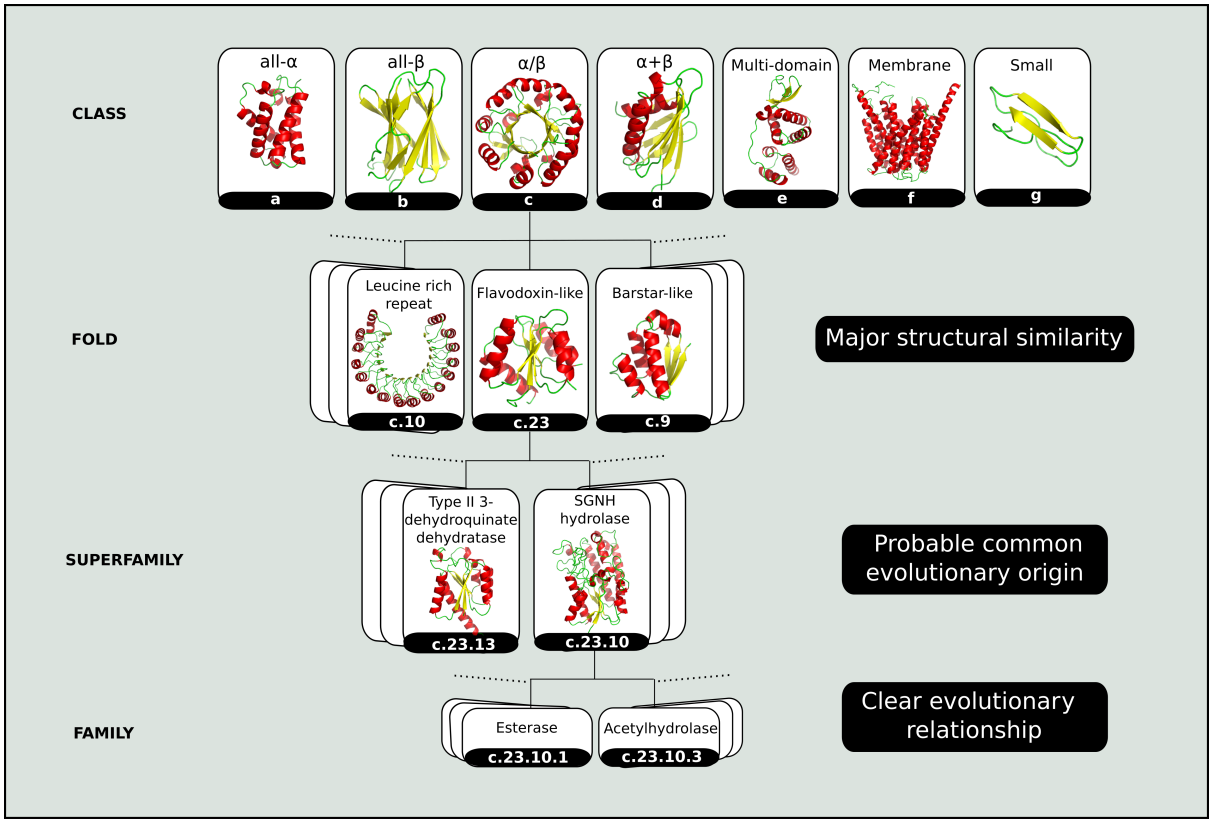


Figure 1.7: Schematic of the SCOP hierarchy. Protein domains are first clustered together into families representing a clear evolutionary link. In general this is equivalent to sharing at least 30% sequence identity. However, in some cases domains with lower identities are placed in the same family due to exceptionally strong structural and functional similarities. Families are then grouped into superfamilies when they share a common evolutionary origin based on similarities in structure and function. Superfamilies are further clustered into folds, with similar structures but which may not share an evolutionary origin. Folds are defined predominantly by the structure at their core, but may differ in their periphery. Finally, all folds are classified into one of seven classes, describing the majority secondary structure type or an additional category for those domains (see text for details). Classified domains are referenced by a letter, representing their class, and then a series of numbers, representing their fold, superfamily and family respectively.

Class	Folds	Superfamilies	Families	Sequences
All-α	284	507	871	1990
All-β	174	354	742	2218
α/β	147	244	803	2679
$\alpha+\beta$	376	552	1055	2667
Multi-domain	66	66	89	191
Membrane	58	110	123	198
Small	90	129	219	626
Total	**1195**	**1962**	**3902**	**10569**

Table 1.1: SCOP 1.75 statistics. Numbers of folds, superfamilies, families and sequences at $< 40\%$ identity in SCOP 1.75

available as part of the ASTRAL compendium [Brenner *et al.*, 2000]. Atomic coordinates for residues from classified domains are curated from the appropriate PDB files. Structures are then assessed using an AEROSPACI score, which combines a SPACI score with a manual penalty for aberrant or atypical domains of a superfamily [Chandonia *et al.*, 2004]. SPACI scores assess a structure based on three criteria: the resolution of the original data, how well the model fits the data, and a stereochemical check which indicates how well the structure complies with standard molecular geometry [Brenner *et al.*, 2000].

1.6.1.2 CATH

The CATH database is similar to SCOP in its hierarchical organisation of the structural and evolutionary relationships of protein domains. CATH stands for **C**lass, **A**rchitecture, **T**opology and **H**omologous superfamily, which are the structural categories within its hierarchy [Orengo *et al.*, 1997]. Additional categories are Sequence families, Orthologous families, Like domains, Identical domains and Domains. There are more categories in CATH than there are in SCOP, and it tends to focus more on structural relationships, while SCOP's emphasis is on evolutionary connections. Furthermore, it tends to rely more on automatic, as opposed to manual classification.

In particular, for assignment into the same homologous superfamily, CATH uses a variety of different automatic methods, including sequence identity, overlap of equivalence, a structural

alignment score SSAP [Orengo and Taylor, 1993], and alignments of hidden Markov models using a variety of different methods. However, this process has not been entirely automated and in many cases establishing homology still requires an analysis of the literature for functional similarities between two domains of suspected homology. Like SCOP, this homologous superfamily classification is the hardest to define.

CATH's superfamilies are further clustered into topologies which are structural units similar to SCOP's folds. The architecture of a topology captures the overall organisation of a domains secondary structure elements, regardless of their connectivity. Examples of such architectures are the up-and-down α-helical bundle, and α-β-α layers. Topologies belonging to these architectures may differ drastically in the number of secondary structure elements they contain, as well as in the connectivities between these elements. The final level of the CATH hierarchy is a domain's class which summarises the majority secondary structure content of a domain. However, the organisation of this level is much simpler in CATH than SCOP, symptomatic of its structural rather than evolutionary focus. CATH classes are mainly α, mainly β, α and β, and few secondary structure.

The latest version of CATH classifies 173,536 different domains (11,926 at $< 35\%$ identity) into 2,626 superfamilies and 1,313 topologies. Table 1.2 gives a more detailed breakdown of the classified domains.

Class	Architectures	Topologies	Superfamilies	Sequences
Mainly α	5	386	875	2917
Mainly β	20	229	520	2618
α and β	14	594	1113	6183
Few	1	104	118	208
Total	**40**	**1313**	**2626**	**11926**

Table 1.2: CATH 4.0 statistics. Numbers of architectures, topologies, superfamilies and sequences at $< 35\%$ identity in CATH 4.0

1.6.2 Alignment

Despite their limitations, automatic methods, which produce alignments between domains, form a vital part of the homology detection toolbox. As mentioned above, the breadth of classification schemes tend to be limited by two things. Firstly, as structural schemes, they require domains to have known structure. While a necessary stipulation, this requirement is a significant limiting factor for the portion of the protein universe defined under these classifications. For example, of the 14,831 well-defined sequence families in the Pfam database (as of March 2013), less than half (6,411) have at least one member whose structure has been solved [Punta *et al.*, 2012]. Secondly, while the procedure for establishing homology under a classification scheme is superior to current alignment methods it is also a far longer process and often requires manual analysis. Over 100,000 structures currently reside in the PDB, but of these only 69,058 have at least one domain been classified under CATH, and 38,221 under SCOP.

Alignment methods have the potential to address these limitations and can often be used as a complementary tool to classification schemes. For example, to identify further members of an existing classification. This concept is expanded on in the following sections and in particular through the description of the SUPERFAMILY database, a powerful tool used to predict members of SCOP superfamilies using sequence alignment methods [Gough *et al.*, 2001]. Primarily though, the following section covers a general introduction to the principles underlying all types of alignment: the identification of a one-to-one correspondence between the characters of one protein and those of another.

1.6.2.1 General principles of alignment

In general, for any two chains of characters $C = (C_1, C_2, \ldots, C_n)$ and $D = (D_1, D_2, \ldots, D_m)$ we can construct an array of pairwise correspondences between the characters in C and those in D. An *alignment* between the two chains is a path through this array (see Figure 1.8 for an example).

In order to measure the strength and significance of such an alignment two constructs are

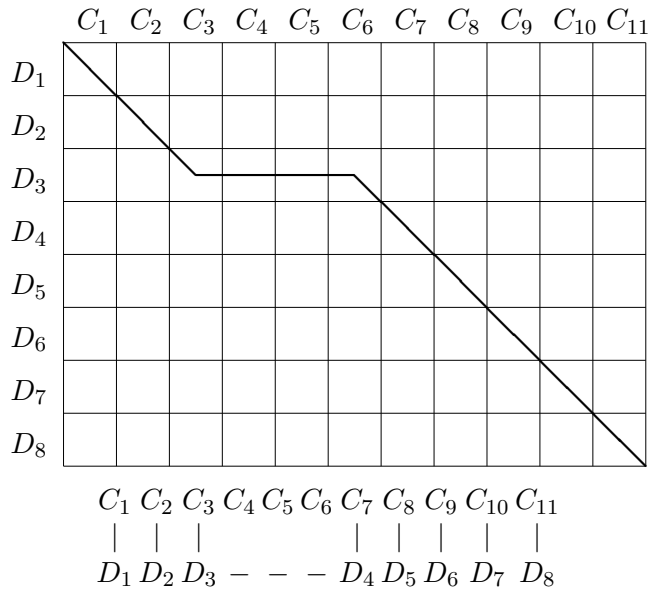


Figure 1.8: An example alignment between chains $C = (C_1, \ldots, C_{11})$ and $D = (D_1, \ldots, D_8)$. The alignment is a path through the pairwise array of characters in C and those in D from the top left to the bottom right. The character correspondences for this path are shown beneath the array.

needed:

1. A *similarity* measure which can quantify the distance between two characters and assign a score to their pairwise alignment.

2. A set of rules governing the set of legitimate steps through the array and a quantification of how each step affects the alignment score.

For any such pairwise comparison of chains, the *optimal alignment* is the path through the array which maximises the alignment score outlined by these constructs.

For example, aligning two protein sequences involves constructing a pairwise array of amino acids. Similarity scores between these characters can be calculated using substitution tables, which measure the frequency of mutations from one amino acid to another. Moreover, within a traditional sequence alignment, legitimate steps must increase along C and D by at most one amino acid. Step types include diagonal steps (match), horizontal steps (insertion) and vertical

steps (deletion). Diagonal steps contribute the similarity score between the matched amino acids to the overall alignment score. Horizontal and vertical steps can penalise the alignment score by a gap penalty.

1.6.2.2 Aligning sequences

As mentioned above, aligning two sequences can be viewed as identifying the optimal path through a pairwise array of their amino acids. In other words, the optimal threading, in terms of maximising amino acid similarities and minimising gapped regions, is sought for one sequence onto another. This is however a somewhat simplistic and naïve approach which detects only very close homologues. Dramatic improvements can be made to the power of such tools by including information which captures the evolutionary constraints acting on these sequences.

For example, one method is to consider the local environment of positions within the sequence. Different substitution tables can be used which are specific to different environments, for example, to the secondary structure type of a residue [Hill *et al.*, 2011]. In the above scheme the quantification of the similarity between two characters would be measured differently depending on each character's local structural configuration. If a residue is found in a helix, for example, it may be able to tolerate different mutations than if it occurred in a strand or surface loop.

Another tactic is to infer at least some of the evolutionary constraints on a protein through examining other members of its family. For example, there will likely be several positions of the protein, mitigated by its structural and functional landscape, which are critical to maintaining that landscape and thus remain highly conserved across naturally occurring members of the family. Similarly, there will be positions which tolerate amino acid substitutions even between 'dissimilar' residues. The advantage of this method would be that the similarity score between a pair of residues at these positions would be treated very differently. In other words, the alignment remains analogous to the scheme outlined above but finds the optimal threading of a sequence, not against another sequence, but against a 'profile' capturing the family-specific

constraints on evolution.

Hidden Markov models Hidden Markov models (HMMs) are a perfect example of how such profiles can be used to align sequences. HMMs can be interpreted as finite state machines capturing the statistical profile of a set of protein sequences [Eddy, 1998]. Figure 1.9 gives an example of such a machine. At each position i of the model, the machine can be in either a match state, an insert state or a delete state. In a match or insert state the machine emits an amino acid. The particular amino acid α emitted from a state X_i depends on a position and state specific set of emission probabilities, $f(\alpha \mid X_i)$. From a certain state X, the machine can transition to another state Y according to another set of probabilities, transition probabilities $p(X \rightarrow Y)$. The allowed transitions for the machine depend on its topology. For example, if the machine described in Figure 1.9 is in state M_2, it can transition to states I_2, M_3 or D_3. The position specific transition and emission probabilities can be inferred from the sequence

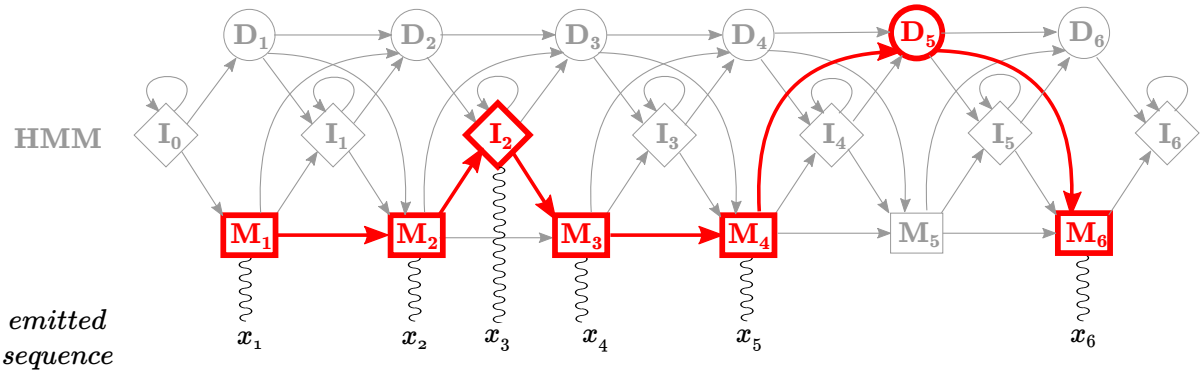


Figure 1.9: The structure of a hidden Markov model as a finite state machine.

profile. For example, highly conserved positions will have a high match emission probability for the conserved residue, while more variable regions will have a more uniform set of emission probabilities.

Aligning a sequence to an HMM is equivalent to determining the most likely path through the model which could have emitted the sequence. For example, in Figure 1.9 a path is shown

in red, which emits a sequence $x_1 \cdots x_6$. The log-likelihood of a path emitting a sequence is calculated as the sum of the logs of the transition and emission probabilities along that path.

Alignments can be assessed by comparing this log likelihood score to a null model which incorporates the background distributions of amino acids and the length of the model [Eddy, 1998].

1.6.2.3 The SUPERFAMILY database

The SUPERFAMILY database is a library of hidden Markov models specifically designed to recognise sequences belonging to SCOP superfamilies [Gough and Chothia, 2002]. There are multiple models in the library representing each superfamily, each seeded by a different domain in the SCOP database. The models are constructed using the SAM-T2K method [Karplus *et al.*, 1998], which iteratively refines a model describing the seed sequence by gradually adding more homologues to the profile. We also use this method to construct HMMs from structure alignments in Chapter 5.

SAM-T2K uses PSI-BLAST to identify very close homologues of the seed sequence [Altschul *et al.*, 1997]. These homologues are then aligned to the original sequence or alignment. By training the hidden Markov model on these sequences the model can learn the best alignment of the homologues. This alignment can then be translated into a model using the background distributions described above. The process is then repeated using this final model to search for a new set of homologues. Each iteration uses a more relaxed E-value in the homologue search, and attaches a lower relative weight to the observed alignment frequencies. The process terminates after the fourth iteration, which uses an E-value of 0.005 and a weight of 0.5 bits per column for the alignment information relative to the background distribution. A schematic of the procedure is presented in Figure 1.10.

SUPERFAMILY assignments on completely sequenced genomes The SUPERFAMILY database also contains assignments, using their library of HMMs, of protein sequences from completely

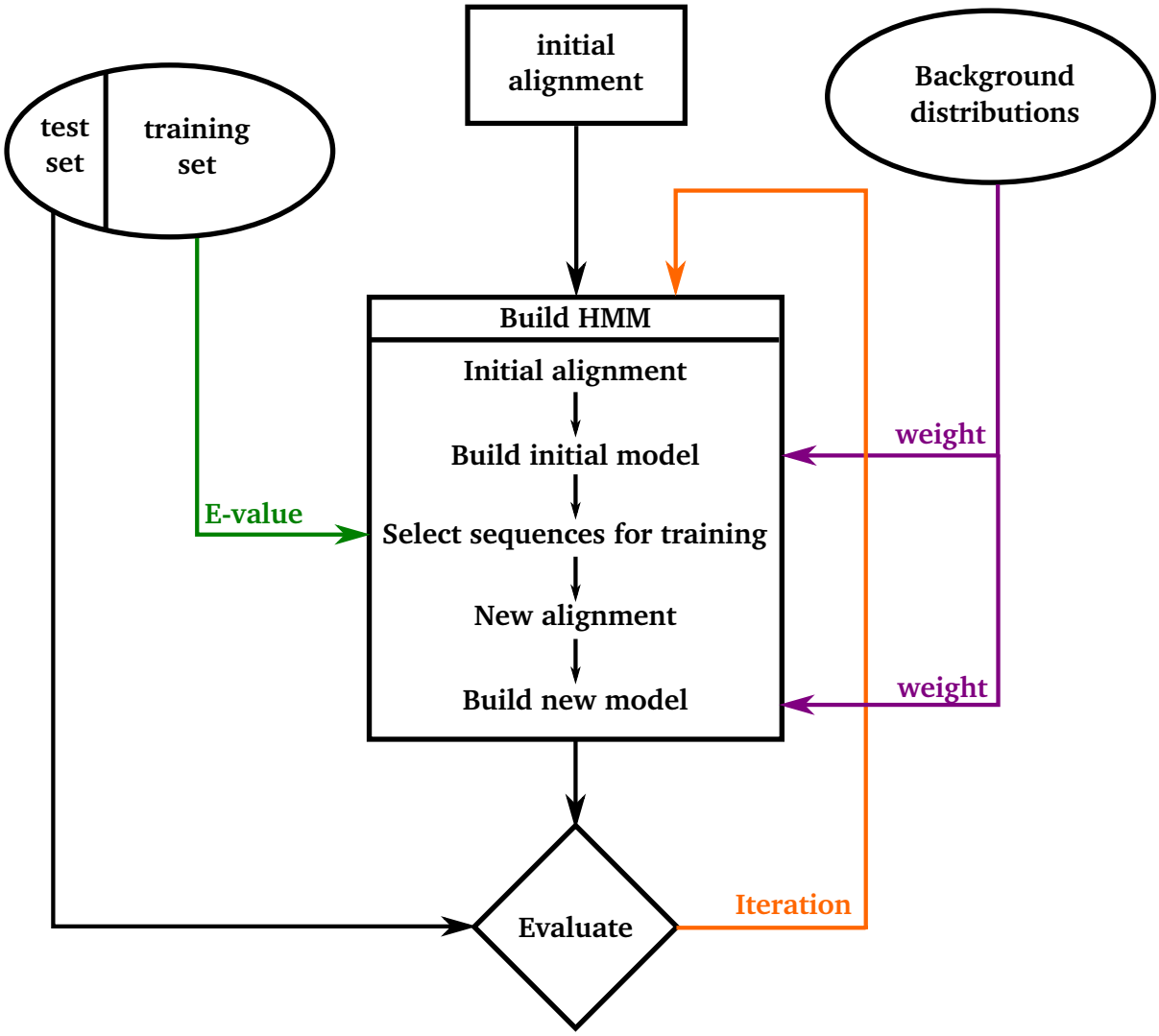


Figure 1.10: Schematic of the SAM-T2K pipeline. SAM's iterative procedure starts with an initial sequence or alignment. By searching for homologues of the seed sequence(s) the program learns a hidden Markov model by gradually adding homologues to the multiple alignment, then using the subsequent model to search for more distant homologues. The procedure incorporates four iterations, shown above in orange. Alignments are converted into models by combining the observed frequencies with background distributions. This process is modulated by a weighting factor, shown in purple. In SAM-T2K this weight starts at 0.8 bits per column for the alignment frequencies in the first iteration. In subsequent iterations the weight is lowered to 0.7, 0.6 and then 0.5 in the final iteration. Homologues are selected by aligning each sequence to the current model. At each iteration, the E-value of this alignment required for selection increases. Initially, it is 0.0001, then 0.0002, 0.001, and finally 0.005. This figure has been reproduced and adapted from the information found at [Hughey *et al.*, 2003]

sequenced genomes to SCOP superfamilies [Gough *et al.*, 2001]. SUPERFAMILY assigns a sequence to a superfamily if the alignment of that sequence to any of the models seeded by constituent members of that superfamily, produces an E-value, when compared to the null model, below 0.0001.

1.6.2.4 Aligning structures

Aligning two protein structures follows a similar process to that outlined above, although there are several variations amongst the available algorithms. Specific examples of such algorithms are illustrated in more depth in Chapter 4 of this thesis so only a brief introduction is given here. Instead of quantifying similarities between amino acids based on their mutational preference for one another, structural alignment methods derive their similarity scores from the local structures surrounding each residue. Side chains are typically ignored in this process and similarities are calculated from the superposition of a backbone fragment centred on a residue in one structure to a fragment from the other. Although not always the case, the optimal threading is found using similar constructions to those for sequence threading, although gap penalties tend to be more lenient.

A very simple example of a structure alignment would be to calculate the similarity between different residues based on their secondary structure type. A similarity score assigning, for example, 1 to a pair of residues with the same secondary structure annotation (i.e. they both appear in a helix) and 0 if they are found in different secondary structure environments.

1.7 The structure of the protein universe

Earlier in this chapter, classification schemes such as CATH and SCOP were introduced in the context of determining evolutionary relationships in the protein universe. An additional feature of these schemes is their structural organisation of domains. Categories such as folds, architectures and classes place naturally occurring proteins into a 'structure space'. The landscape of

this structure space is hard to define and, because of the elusive relationships between protein sequence and structure, its affiliation with the corresponding sequence space is complex.

It has long been observed that the occupation of this space by naturally occurring proteins follows a skewed distribution: with a small number of highly populated 'structures' and large numbers of more sparsely populated folds. However, the reasons for such a distribution have yet to be resolved. High population of popular folds could be due to the existence of highly designable structures which have been visited multiple times by convergent evolution: 'structural attractors'. Alternatively, it is possible that common biochemical activities may remain in demand in the cellular environment, prompting preferential proliferation of domains involved in such activity. A final alternative is that this distribution is simply a relic of a stochastic evolutionary process [Koonin *et al.*, 2002]. In any of these cases the landscape of structures, and the limited repertoire of popular motifs is a fascinating subject. For example, Figure 1.11 shows three highly populated motifs: the singly-wound barrel, the doubly-wound sheet and the Greek key. These motifs describe β-sheet topologies which appear not only across a large number of domains, but also across several superfamilies and even different folds.

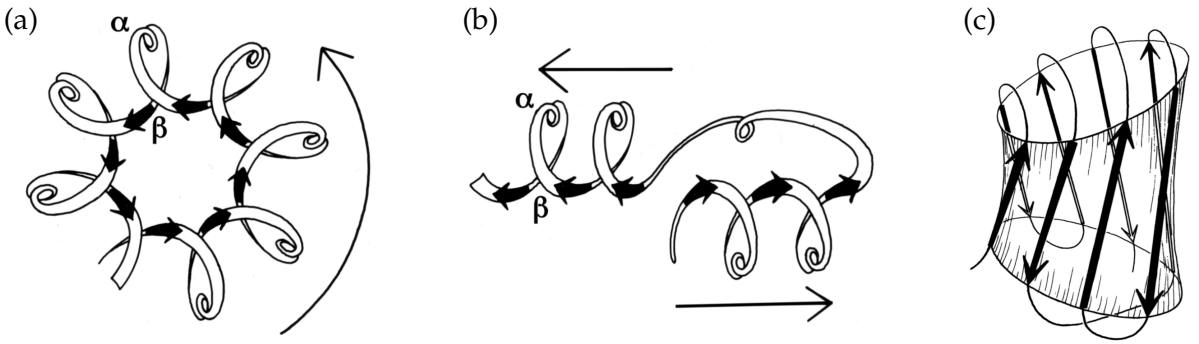


Figure 1.11: Common β-sheet topologies. **(a)** The singly-wound parallel α/β-barrel. **(b)** The doubly-wound parallel α/β-sheet. **(c)** The anti-parallel Greek key β-barrel. These figures have been reproduced with permission from Richardson [1981] © Jane Richardson.

There are many questions which centre on the nature of protein fold space. One of the most fundamental such questions is if it makes sense to consider folds as units within this space. The

presence of structural variation within members of a single fold as we saw in section 1.5, and the presence of highly populated motifs uniting a variety of folds has raised the possibility of fold space being represented by a continuum, rather than a discrete space [Ptitsyn and Finkelstein, 1980; Sadowski and Taylor, 2010]. In this thesis we largely consider fold space as discrete, focusing on folds as a single entity. This is largely because in SCOP, the classification scheme we use, evolutionary information is still used in the partitioning of different folds [Murzin *et al.*, 1995].

1.8 Thesis outline

The remainder of this thesis is contained in five further chapters, each of which are summarised briefly below.

Chapter 2: Superfamilies on genomes and estimates of their ages The second chapter of this thesis looks at a method of estimating the age of homologous protein superfamilies. The method uses the predicted occurrences of superfamilies across a large species tree of completely sequenced genomes. We examine the current coverage of these predictions on the genomes. This gives an estimate of the portion of the protein universe with known fold. We show that, for many species, this is high, but it is in predominantly Eukaryotic species where the largest portions of unannotated sequences sit. A simple evolutionary model is then applied to the occurrence profiles of superfamilies in order to predict a likely scenario for the evolution of each superfamily, and in particular, the internal node at which the superfamily's ancestor first evolved: its age. In this chapter we examine the robustness of age estimates to variability in the evolutionary model, in particular to changes in the topology of the phylogenetic tree and in the relative weighting of gene gain and loss events in each evolutionary scenario. We show that as a whole, the estimates remain largely unperturbed by even quite significant changes to the underlying model.

Chapter 3: Preferences of ancient and new-born superfamilies The third chapter of this thesis uses the age estimates from the previous chapter to partition the protein universe into populations of ancient and new-born superfamilies. Ancient superfamilies are thought to have an ancestor which can be traced back right to the very root of the tree of life, whereas new-born superfamilies are estimated to have evolved much more recently, near the leaves of the tree. In this chapter these populations are compared according to a variety of different characteristics, including features of their primary, secondary and tertiary structures, as well as their functional annotations. We postulate that, in general, the structures of new-born superfamilies appear to be less elaborate, shorter, with fewer buried residues and long range contacts, than their ancient counterparts.

Chapter 4: Bridges through fold space In the fourth chapter of this thesis we present several networks which attempt to represent the global structure space within which different protein folds sit. Each network representation is the result of a different method of comparing structures through aligning their backbone chains, and the methods are compared to each other with the introduction of a posterior probability attached to their alignment score. This probability captures the likelihood of a score defining a fold level relationship as defined by the structural classification of proteins database (SCOP). At different probability thresholds, ranging from 0.5 to 0.9, networks are presented which capture a set of inter-fold similarities, which we call structural bridges. Folds are nodes in these networks and are annotated with the age estimates which are calculated in Chapter 2. We examine the differences between these networks, looking in particular, at the consensus of the different alignment methods. We show that, while there are certainly areas of disagreement, there is still a well-defined consensus between these methods. This consensus space comprises a community structure resembling the SCOP class divisions, but with an additional discrimination between all-β barrels and sandwiches. The core of these networks, as calculated using different centrality measures, is dominated by ancient folds, whereas new-born folds occur more often at the periphery of the landscapes. Moreover struc-

tural bridges appear to occur between folds of similar ages, and even more so when considering bridges which form the consensus of the different methods.

Chapter 5: Tunnels through sequence space The fifth chapter of this thesis looks at constructing sequence profiles for superfamilies from alignments of homologous sequences. These profiles are constrained using multiple structural alignments of their superfamily constituents. Each sequence profile is built iteratively and is used to construct a hidden Markov model. 1,728 different models representing superfamilies across fold space are aligned together in a pairwise fashion. Significant hits between models representing different superfamilies are then examined. We visualise the resultant landscape of relationships both within and between different fold units. We examine two well known and strongly linked clusters: the Rossmannoid folds and the β-propeller folds, as well as looking at less prominent relationships.

Chapter 6: Conclusions, context and future directions In the final chapter of this thesis, the previous chapters are discussed and considered within a global context. Improvements to the methodologies used are proposed and avenues for further investigation are highlighted.

CHAPTER 2

Superfamilies on genomes and estimates of their ages

The methods from this chapter as well as Figures 2.5 and 2.9 have been previously published by the author in Edwards et al. [2013]. They are included here with permission under the Creative Commons Attribution (CC BY) license.

In this chapter we begin by examining the evolution of protein superfamilies. As discussed in Chapter 1, superfamilies are thought to represent homologous progenies consisting of several different protein domains which have evolved, most likely divergently, from a single ancestor.

The recent explosion in the sequencing of genomes offers the potential for examining this history in more detail. For example, by considering what species' genomes include proteins belonging to a superfamily, an assessment of that superfamily's evolutionary history can be made. In this chapter these principles are used to generate estimates for the structural ancestors of SCOP superfamilies.

The methods and results we present here are partially reproduced from previously published work [Edwards *et al.*, 2013].

2.1 Motivation

The current wealth of freely available genetic sequences offers the potential for uncovering the evolutionary history of genes and their products, proteins. This approach is particularly well suited to homologous structural superfamilies. This is primarily due to the fact that, as has already been discussed, structures remain far more conserved during evolutionary drift than their corresponding sequences [Ponting and Russell, 2002]. They thus preserve a deep phylogenetic signal.

A particular example of this can be found in a study by Lin and Gerstein [2000]. Using the protein repertoires of eight well annotated and completely sequenced genomes, phylogenetic trees showing the relationships between these genomes can be constructed. When comparing trees built using families detected by sequence similarity alone against trees built using fold occurrences, Lin and Gerstein [2000] found that fold occurrences produced tree topologies which agreed in general with the traditional phylogeny. On the other hand they found that sequence family units failed even to resolve the trees into the three different superkingdoms. As such, their study supports the use of structure as a fundamental molecular unit in phylogenetic analysis.

2.1.1 Genome evolution

The evolution of proteins described in the previous chapter, through processes such as substitutions, insertions, deletions and duplications, produces variation between a parent and its offspring. However, the vast majority of this change, while altering a protein's sequence, will leave it occupying the same structure and functional role within the cell. In other words, it leaves the superfamily intact. As an alternative to considering lists of genes and their protein sequences, completely sequenced genomes can instead be viewed as repertoires of superfamilies. Comparing two different genomes then becomes a matter of comparing their superfamily repertoires. In this context evolutionary change between parent and offspring genomes is comprised of superfamily duplication, deletion or acquisition events.

Superfamily acquisition can occur through the innovation of a new structural and functional unit or by lateral gene transfer from another species. Superfamily duplication occurs due to gene duplication and affects the number of copies of a certain superfamily on each genome. Superfamily deletion results in a loss of a particular superfamily in the repertoire of the offspring. The relative frequencies of each of these events has not been fully explored. However, lateral gene transfer is thought to be a considerable evolutionary force particularly in Prokaryotic evolution [Mirkin *et al.*, 2003]. Lateral gene transfer, where genetic material is transmitted through interference from another species rather than through a common ancestry, is not thought to be so prominent in Eukaryotic species however [Rogozin *et al.*, 2005]. Purely from comparing the repertoires of superfamilies in a parent and offspring, lateral gene transfer and *de novo* innovation are indistinguishable from each other.

2.2 Outline

In this chapter a method for estimating superfamily and fold ages is presented. In order to establish the methodology for this estimation and explore its underlying assumptions a four step process is described. These four steps include superfamily classification, superfamily prediction,

phylogenetic trees of completely sequenced genomes and models for superfamily evolution. Each of these steps is discussed and placed within context in the literature.

Superfamily and fold ages are presented based on an analysis of 1,014 completely sequenced genomes from across the tree of life. The estimates are also calculated across different possible topologies for this tree and modifying the model of evolution used to identify structural ancestors. The assignments of superfamilies to genomes underlying this estimation are explored in detail. For example, the coverage of these genomes is shown to be substantial but not complete, and comparatively lower in Eukaryote species. The set of superfamilies which are estimated to be ancient with an ancestor at the root of the tree of life are shown to be similar to extant Archaeal genomes. Finally the fold age estimates are compared to a dissimilar method which also attempts to calculate the age of structures.

2.3 Age estimation

Age estimates for families of proteins have previously been calculated using a variety of different methods but there is no established standard for the technique. Figure 2.1 shows a simple schematic for age estimation consisting of four fundamental steps for which different interpretations have been implemented:

1. The classification of homologous protein progenies, usually based on either sequence or structural features.

2. The identification of proteins from a set of completely sequenced genomes as members of these progenies.

3. The construction of a phylogeny used to describe the evolutionary history and relatedness of these genomes.

4. A model for the evolution of that protein progeny across the set of genomes, resulting in the identification of the ancestral node.

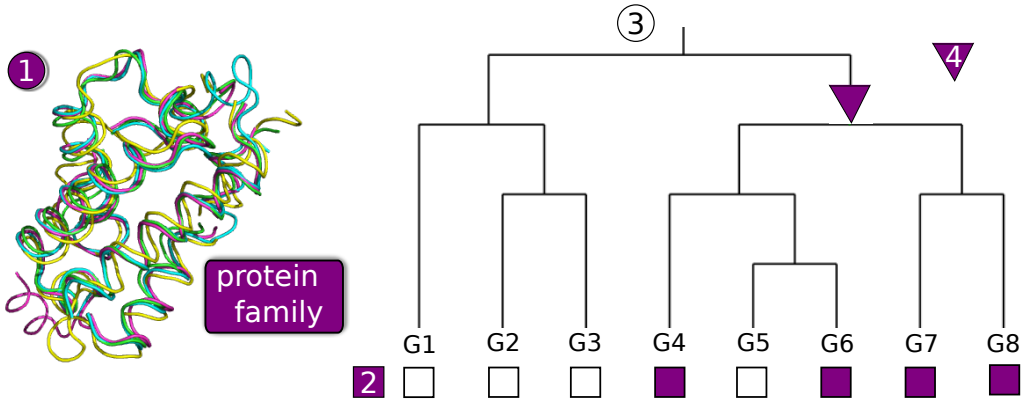


Figure 2.1: Simple schematic for age estimation. Protein families, grouped by either sequence or structural similarities, can be identified on different species' genomes across the tree of life. Based on the underlying phylogeny of these species the internal node representing the emergence of the ancestral structure can be identified.

In this section the different approaches which have been applied at each step will be explored in order to place the method and work outlined in this Chapter within an appropriate context.

2.3.1 Protein progenies

The first component of this method is the classification of proteins into homologous progenies. Correct classification is important as progenies defined either too stringently or leniently will result in incorrect age estimates. As has been discussed both here and in Chapter 1 there is an increasing body of evidence that more distant homologous relationships are apparent beyond the detection of current sequence similarity methods [Chothia and Lesk, 1986; Gough, 2005; Valas *et al.*, 2009].

In this work both structural superfamilies and folds are considered to be the most appropriate units for this method, a choice which has largely been adopted in previous work of this kind [Caetano-Anollés and Caetano-Anollés, 2003; Winstanley *et al.*, 2005]. While a few studies have adapted the technique to calculate ages of sequence families [Toll-Riera *et al.*, 2012; Capra *et al.*, 2012; Choi and Kim, 2006] these methods tend to involve a much reduced phylogenetic tree and evolutionary scale. In particular they have been used to look at specific proteins of interest on

Eukaryotic subtrees [Toll-Riera *et al.*, 2012; Capra *et al.*, 2012].

One advantage of protein sequence families over structural superfamilies and folds is that, in general, these units result in a more complete coverage of genomes after assignment. This is because of the relative simplicity of classifying sequence families over structural classification, and as such a greater portion of the protein universe are members of sequence families than belong to structural superfamilies or folds. However, it is important to note that even with this lack of coverage, as Lin and Gerstein [2000] show, structural repertoires can reconstruct long term evolutionary relationships more faithfully than sequence families.

In this chapter we primarily consider SCOP superfamilies as evolutionary units, but also apply the method at the fold level. While folds remain a structural classification and do not necessarily represent evolutionary relationships it is possible that a common fold uniting two different superfamilies is a symptom of a shared ancestry [Ponting and Russell, 2002; Gough, 2005]. Previous studies have also been applied using the CATH classification system [Abeln, 2007].

2.3.2 Predictions on genomes

The second step is assigning each protein sequence appearing on a genome to one or more of the protein progenies. As proteins often contain multiple domains, several assignments can be made to a single sequence. Since there is no direct structural information for the majority of the proteins on these genomes predictions must be made using just sequence information.

Predicting membership of sequence families is relatively simple. However, predicting the superfamily or the fold of an unknown sequence is much harder. The most effective methods combine structural classification data with powerful sequence alignment techniques. For example, profiles seeded by a member of a superfamily can be used to express the sequence constraints and signatures representing that superfamily. This is true even if the other sequences making up the profile are not classified. Predicting membership of a superfamily does not have to be the responsibility of a single profile, rather several different profiles can capture the sequence

fingerprints of the different families making up that superfamily.

Examples of methodologies which can be used to construct such profiles or HMMs are PSI-BLAST, SAM and HMMer [Altschul *et al.*, 1990; Karplus *et al.*, 1998; Eddy, 1998]. Several databases exist which build their own profile HMMs and apply them to completely sequenced genomes. Gene 3D [Buchan *et al.*, 2002] and SUPERFAMILY [Gough *et al.*, 2001] both use the SAM protocol to construct profile HMMs of CATH and SCOP superfamilies respectively. These existing databases are popular choices for the prediction stage although alternatives such as using PSI-BLAST to predict superfamilies have also been used [Winstanley *et al.*, 2005; Toll-Riera *et al.*, 2012]

In this chapter the SUPERFAMILY database is used as a prediction tool. SUPERFAMILY, as discussed in Section 1.6.2.3, is a library of profile HMMs specifically fine-tuned to assign SCOP superfamilies with high accuracy and coverage [Gough *et al.*, 2001]. Previous results show that SUPERFAMILY assignments increase the coverage of assignments on genomes over a PSIBLAST protocol [Abeln, 2007; Winstanley *et al.*, 2005]. We also collapse the full repertoire of predicted superfamilies and folds to a simple binary occurrence for each progeny. This means the subsequent evolutionary model need only take into account superfamily innovation, death and vertical descent, rather than including every duplication and deletion event.

2.3.3 The Tree of Life

The third step is to construct the underlying phylogenetic relationships between the species from which the completely sequenced genomes have been taken. There is much debate regarding the accuracy of a single Tree of Life [Doolittle and Bapteste, 2007]. While there is a general consensus concerning several well-studied branches of interest, trees of thousands of species across the three superkingdoms, as we use here, remain unresolved. For this reason it is preferred to retain several possible phylogenetic trees to contribute to the analysis [Winstanley *et al.*, 2005].

Trees of life consist of three features. The first is the topology of the tree, illustrating the

divergence of species as branching events. The second is the rooting of the tree, which places more recent branching events ahead of ancestral speciation events. Finally branch lengths capture the amount of time (relatively) between different speciation or branching events. These are often calculated separately. For the trees in this chapter, we impose a root at the trifurcation of the three superkingdoms, branch lengths are calculated separately to the topology, and we further normalise branch lengths so that the extant species are the same vertical distance from the root. The exact methods used to construct the trees are described in the Methods.

For the age estimates presented here we construct several different tree topologies but essentially use three underlying methodologies: the NCBI common taxonomy, Neighbour-joining on distances between genomes, and Wagner parsimony using the presence or absence of a superfamily as a character state on the evolving genomes. Each methodology is introduced here briefly.

2.3.3.1 NCBI common taxonomy

The NCBI taxonomy uses both nucleotide and protein sequences, as well as collating the consensus from scientific literature, to construct a tree of sequenced life [Federhen, 2012]. A tree of life is downloadable from their website using organisms' taxonomy IDs and is in the form of a tree rooted at the trifurcation of Bacteria, Archaea and Eukarya. The NCBI taxonomy is not considered a phylogeny so branch lengths are not calculated.

2.3.3.2 Neighbour-joining

Neighbour-joining is an algorithm for constructing trees from distances between taxa [Saitou and Nei, 1987]. In our case we consider two different distance metrics which capture the overlap of superfamilies appearing on any pair of genomes (see Methods). For any given distance matrix between taxa, neighbour-joining aims to construct a tree of minimum *length* which describes the relationships defined by the matrix (see Figure 2.2a). The tree length is simply the sum of all branch lengths along the tree. Neighbour-joining is a greedy algorithm and,

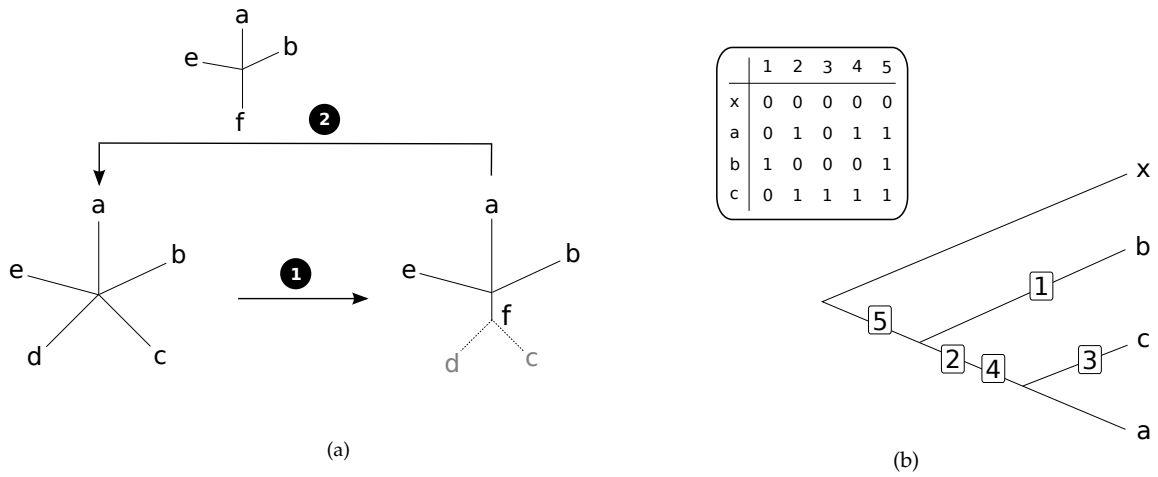


Figure 2.2: Constructing phylogenetic trees. (a) Schematic showing the steps of the neighbour-joining algorithm. Beginning with an unresolved tree (a star graph), the algorithm joins that pair of nodes associated with the greatest decrease in tree length. To do this a modified distance matrix is consulted, where distances are normalised by a node's average divergence from all other nodes. This pair is then replaced by a new node representing their ancestral node. New distances are recalculated and the algorithm is repeated until all pairs have been joined. (b) Wagner parsimony construction of phylogenetic trees. Most parsimonious trees are those involving the fewest state changes along their branches.

while not guaranteed to produce the tree of minimal length it is generally quite accurate and a computationally efficient method [Saitou and Nei, 1987]. It produces an unrooted tree with branch lengths.

2.3.3.3 Wagner parsimony

Trees constructed using Wagner parsimony attempt to select the tree topology associated with the minimum number of character state changes [Kluge, 1969]. Each taxon is represented by a set of characters (in our case the set of protein superfamilies or folds), each of which can have multiple states (presence/absence). When two taxa are joined the cost of this mergence is the number of differences in their character states. The algorithm produces an unrooted tree with branch lengths. Figure 2.2b shows a schematic of this principle.

The above methods are not intended to encompass all available tree-building approaches. Instead, we have chosen two methodologies expected to produce realistic topologies without

incurring unreasonable computational cost. There are alternative algorithms for constructing topologies from data such as ours, such as UPGMA [Sneath and Sokal, 1973], and maximum likelihood [Felsenstein, 1981].

2.3.4 Evolutionary model

The final assumption of the method is the model underlying the identification of the ancestral node. In this case the model is equivalent to outlining the most likely sequence of gain and loss events at internal nodes of the tree which explain the occurrence profile at its leaves (see Figure 2.3). The likelihood of these events is based on parsimonious assumptions relating to the evolution of protein domains. The principle underlying all types of parsimony is that the scenario of events involving the least evolutionary change is most likely. Gain events can represent *de novo* superfamily innovation, lateral gene transfer of a superfamily between genomes, but can also represent a false positive assignment of a superfamily to a genome. Loss events can represent the deletion of a superfamily and also false negative assignments to a genome.

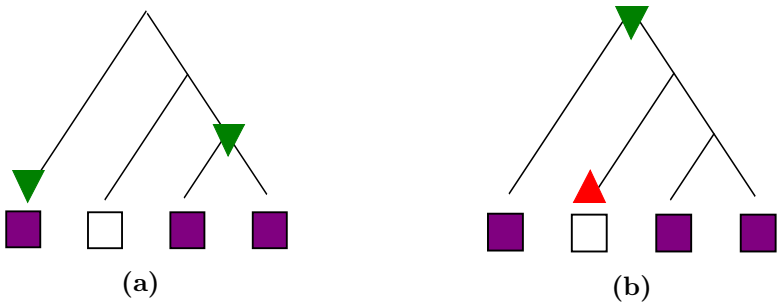


Figure 2.3: Two evolutionary scenarios resulting in the same occurrence pattern. This figure has been adapted using an example given in Mirkin *et al.* [2003]

A maximum parsimony model for superfamily evolution has been largely adopted for this step [Winstanley *et al.*, 2005; Yang and Bourne, 2009], although alternatives include DOLLO parsimony: taking the most recent common ancestor (MRCA) [Winstanley *et al.*, 2005; Choi and Kim, 2006]. Moreover, it is possible to consider scenarios in between these possibilities. Maximum parsimony will select the smallest number of events which explain an occurrence

pattern, while Dollo parsimony minimises gain events at a cost of introducing more loss events. In fact, we can consider any relative weighting (g) of the likelihood of loss events over gain events, by defining a score:

$$S = \lambda + g\gamma \tag{2.1}$$

where λ and γ are the number of loss and gain events respectively. For a given gain weight g, the most parsimonious model of events is the scenario which minimises this score.

Mirkin *et al.* [2003] developed an efficient algorithm for identifying these events, which can be generated from a single sweep of the tree, from leaves to root. At each node, most parsimonious events are retained for the scenario where a superfamily is inherited from its parent node and the case where it is not. These can be calculated directly from the most parsimonious events of its children (see Figure 2.4). At the final node (the root) the algorithm terminates after selecting the most parsimonious scenario. If the superfamily is inherited at the root a final gain event must be introduced.

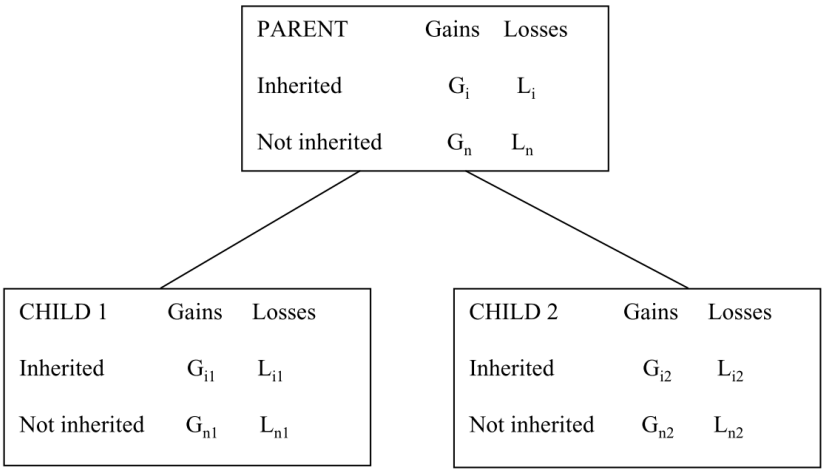


Figure 2.4: Patterns of events in a parent-children triple according to a parsimonious scenario. At any parent node, the most parsimonious series of events for both scenarios where a superfamily is inherited or not can be directly inferred from the events retained for its children. This figure has been reproduced with the permission of the publishers BioMed Central from Mirkin *et al.* [2003].

It is also possible, at this stage of the method, to treat evolution on particular genomes

differently. For example, the rates of horizontal gene transfer are known to vary significantly between Prokaryotes and Eukaryotes [Rogozin *et al.*, 2005], with very few transfer events occurring on Eukaryotic genomes. The parsimony model outlined above can thus be applied using different gain weights on the Eukaryotic and Prokaryotic subtrees.

2.3.5 Alternative methodologies

Other studies have been published which produce a fold timeline generated from a phylogenetic tree of topologies [Caetano-Anollés and Caetano-Anollés, 2003; Wang *et al.*, 2006; Kim and Caetano-Anollés, 2011]. This approach differs from that presented here though it is based on the same data: the predictions of superfamilies on completely sequenced genomes. In these studies, however, genomic abundance (the copy count of a superfamily on a genome) is used instead of its binary occurrence. This method is based on further assumptions that this set of superfamilies can be represented cladistically, with current protein diversity originating from a limited set of architectural designs. Trees are constructed using a directed parsimony algorithm on the genomic abundance of each superfamily across the completely sequenced genomes. Directed parsimony assumes that a particular character state is ancestral. In this case high genomic abundance is assumed to be ancestral. Node ages are then assigned to superfamilies based on the number of nodes between a superfamily and the root of the tree.

For the work presented here we do not explore an in depth comparison between this method and our own. However some simple analysis is included in Section 2.5.5.3.

2.3.6 What do we mean by the 'age' of a superfamily?

Superfamily ages calculated from the above scheme are relative, and a course-grained measure which depend on both the identification of sequences homologous to known superfamilies and on the phylogenetic construction of the tree of life.

It is important to remember that this analysis is undertaken on a set of extant protein structures and that an *old* superfamily is not the same as an old protein. In fact, our current

set of superfamilies could be described as a single time slice of structural units at different stages of evolution [Choi and Kim, 2006]. As such, the current structure space can be understood to contain superfamilies with long evolutionary histories and those with newly evolved ones, as well as those in between.

It must also be emphasised that the ancient set of superfamilies do not by any means represent the first protein repertoire of early life. The ageing method presented here uses only a subset of extant species in order to infer the evolutionary histories of a set of superfamilies. As such the phylogenetic signal goes only as far as the last universal common ancestor (LUCA) of these modern species. While LUCA existed before the diversification of superkingdoms it is likely to have existed long after the origin of life. Moreover this predicted set of ancient superfamilies is only the subset of LUCA's proteome which have survived and thrived till the present day.

2.4 Methods

Superfamily ages are estimated from the phylogenetic profiles of SCOP superfamilies across completely sequenced genomes. The method follows the general principles outlined in Figure 2.1. For a more specific schematic describing the principal steps detailed below see Figure 2.5.

2.4.1 Superfamily predictions

The data we use in this study were taken from the SUPERFAMILY database (v1.75) [Gough *et al.*, 2001]. SUPERFAMILY uses families of HMMs to identify homologues of SCOP superfamilies. The database comprises protein sequences taken from completely sequenced and annotated genomes and assignments of these sequences to SCOP superfamilies.

We downloaded predictions of $2,019$ superfamilies on all $1,496$ species available in the SUPERFAMILY database on September 11th 2012. This set was then filtered as follows:

- 407 species annotated as pathogens in the GOLD (v.4) database [Kyrpides, 1999] were

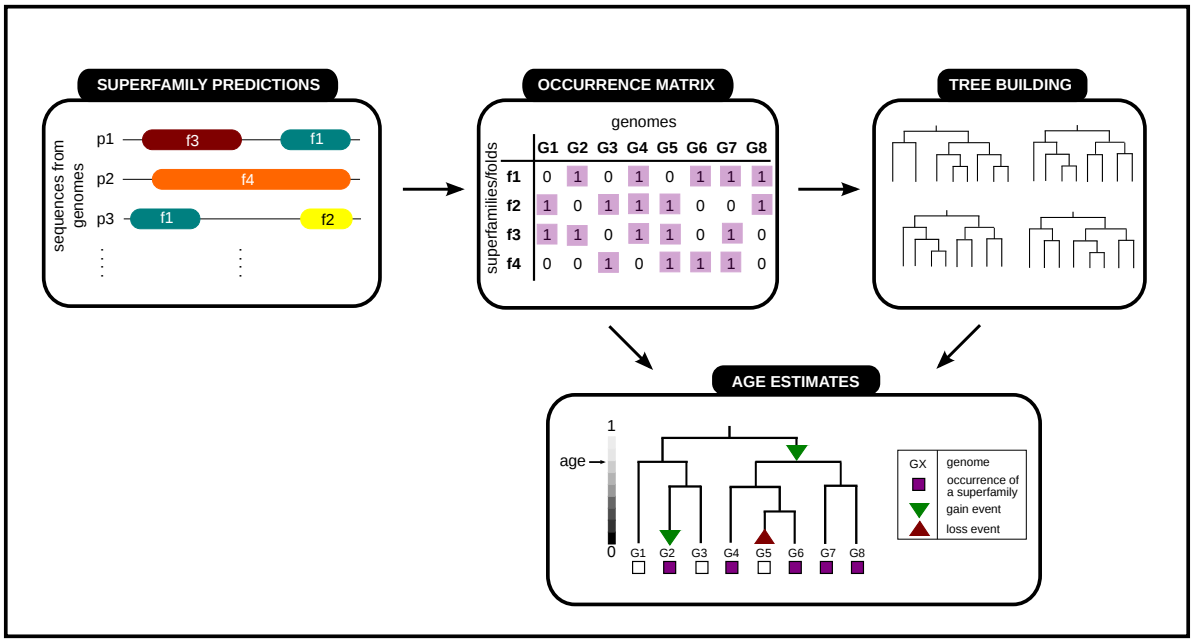


Figure 2.5: Schematic diagram of the method for predicting ages for superfamilies or folds. Predictions of superfamily/fold domains on genome sequences are calculated using SUPERFAMILY's families of HMMs [Gough *et al.*, 2001]. These predictions are then collapsed to a binary occurrence matrix where each entry represents the presence (1) or absence (0) of a superfamily or fold on a genome. Multiple phylogenetic trees are constructed: some using the NCBI taxonomy, and others inferred from the column vectors of the occurrence matrices, either using distance matrices or treating each superfamily as a discrete character and calculating the most parsimonious topology. Age estimates are calculated using these trees (after branch length estimation and normalisation) and considering the rows of the occurrence matrix as phylogenetic profiles on the leaves of each tree. Gain and loss events are then predicted at internal nodes of the tree based on a model for the evolution of superfamilies. Different models used here included Dollo parsimony, maximum parsimony, using a range of different gain weights, and a fusion parsimony model, determining events using Dollo parsimony on the Eukaryotic subtree and maximum parsimony elsewhere. See Methods for further details. This figure has been reproduced from Edwards *et al.* [2013].

removed as pathogens are often associated with incomplete genomes and with increased rates of lateral gene transfer.

- 31 species which were classified in the category *candidatus*, a provisional status for putative taxa were also removed [Murray and Schleifer, 1994].

- 44 species found, during the later stages of the method, to lead to poor resolution on the phylogenetic tree were manually identified and removed. These species were largely characterised by having small genomes or were pathogens with annotations missing in the GOLD database and are listed in Supplementary Table B1.

This left 649 Bacteria, 265 Eukaryotes and 100 Archaea. We called this set the ALLgenomes and it was intended to represent the diversity in the currently known tree of life as accurately as possible. A second set, MULTIgenomes, was created that contained 210 multi-cellular Eukaryotes, a subset of ALLgenomes. Ages from MULTIgenomes are used in further chapters of this thesis where we analyse superfamily dynamics within the context of multi-cellular evolution. For example, we look at the preferences of superfamilies for disulphide bonds, which are found particularly in extra-cellular proteins, in Chapter 3. These species are listed in full in the Supplementary material of Edwards *et al.* [2013] and those which have been manually removed are included in Supplementary Table B1.

2.4.2 Occurrence Matrix

These superfamily predictions were collapsed to a binary occurrence matrix representing the presence or absence of a superfamily on a genome and superfamilies with no predictions across the remaining genomes were removed. This left a table of 1,896 superfamilies on ALLgenomes and 1,584 superfamilies across MULTIgenomes. Matrices at the fold level were constructed in the same way, with 1,119 folds on ALLgenomes and 959 on MULTIgenomes.

2.4.3 Tree Building

Multiple species trees were considered as the underlying phylogeny for the completely sequenced genomes. Using numerous trees helps to ensure that the results presented here are robust to inaccuracies in estimating this Tree of Life. We considered both the NCBI common taxonomy tree [Federhen, 2012] as well as phylogenies constructed using the superfamily and fold occurrence profiles calculated above. For completeness the constructed trees were estimated using both parsimony and distance-based algorithms. All the trees were inferred using the PHYLIP package [Felsenstein, 1989]. The name of the specific algorithms used during the tree building process are given in brackets throughout the sections below. A total of 8 different trees were constructed for each of the genome sets (ALLgenomes and MULTIgenomes).

2.4.3.1 NCBI trees

The NCBI common taxonomy tree for ALLgenomes and MULTIgenomes were downloaded from the NCBI website. Branch lengths were added using the presence-absence of superfamilies or folds as unweighted, symmetric states using the Wagner parsimony algorithm (PARS) which averages the number of state transitions over all sites and over all possible most parsimonious placements of the state transitions among branches.

2.4.3.2 Distance trees

A neighbour-joining algorithm (NEIGHBOR) was used to construct trees from pairwise distance matrices. The distance metrics used were calculated using a comparison of the numbers of folds or superfamilies on two different genomes. A contingency table was constructed comparing any two genomes G_i and G_j. This table counts the number of folds or superfamilies occurring on both genomes (a), those occurring only on G_j (b), and those occurring just on G_i (c):

The distance $D_{i,j}$ between genomes G_i and G_j was then calculated using two different dissimilarity metrics defined as follows:

		G_i	
		1	0
G_j	1	*a*	*b*
	0	*c*	*d*

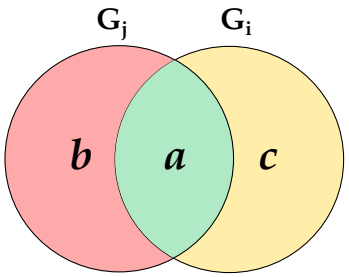


Figure 2.6: Comparing two genomes' superfamily/fold content. Given the occurrences of a set of superfamilies or folds on two genomes, G_i and G_j, count the number of superfamilies/folds on both genomes (a) and compare it to the number on G_j only (b) and G_i only (c).

- Jaccard distance: $D_{i,j} = (b + c)/(a + b + c)$
- Bray-Curtis distance: $D_{i,j} = (b + c)/(2a + b + c)$

Matrices were composed of the distances between every pairwise combination of species in a set and used as input to the tree building algorithm. For each genome set four distance matrices were calculated: using the Jaccard and the Bray-Curtis distances on superfamily and fold occurrence data.

In all these cases, an extended majority rule consensus tree (CONSENSE) was calculated from individual trees constructed using neighbour-joining on 100 delete-half jackknife samples of the original occurrence data. Branch lengths were added to this consensus topology using the Fitch-Margoliash algorithm (FITCH) using the complete distance matrix.

2.4.3.3 Parsimony trees

Trees were also built using Wagner parsimony (PARS) and treating the presence-absence data of folds or superfamilies as unweighted, symmetric character states. Extended majority-rule consensus trees (CONSENSE) were summarised from trees built from 100 delete-half jackknife samples of the occurrence data where up to 10 trees tied for the best parsimony score were retained per sample. Branch lengths were added to the consensus trees using a final implementation of the Wagner parsimony algorithm (PARS).

2.4.3.4 Tree transformations

The trees for ALLgenomes were rooted at the trifurcation of the three superkingdoms and the trees for MULTIgenomes were rooted by including the archaeal species *Acidianus hospitalis* and using this as an outgroup. Branch lengths were normalised to lie between 0 and 1, with the leaves at 0 and the root at 1. An example tree is shown in Figure 2.7

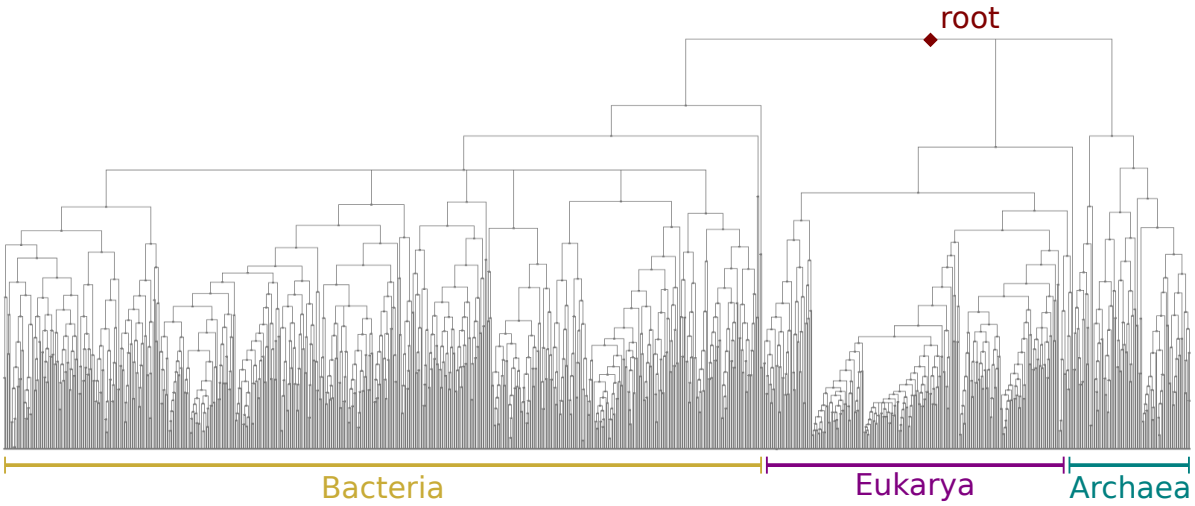


Figure 2.7: Example of one of the phylogenetic trees constructed for age estimation. This tree was built using Wagner parsimony using superfamily occurrence data as described in the main text. Species names are not shown for clarity.

2.4.3.5 Tree comparison

Tree topologies were compared in a pairwise manner using the symmetric distance of Robinson and Foulds [1981]. This measures the number of partitions existing on one tree and not the other. It is not affected by the rooting of the tree or by branch length estimates. This distance is illustrated in Figure 2.8.

2.4.4 Age Estimation

For each tree, the age of a superfamily is the result of a parsimony analysis on potential gain and loss events of the superfamily. Relative ages are quantified as the height of the node of

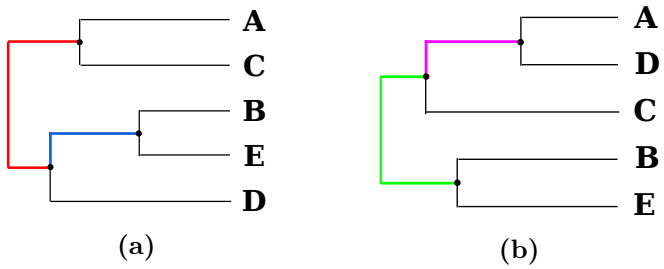


Figure 2.8: Symmetric distance between two trees. The symmetric distance compares the number of partitions on one tree but not the other. Partitions are defined as the partitioning of taxa induced by an internal branch. For example, tree (a) contains two internal branches. The red branch induces the partition $\{A, C|B, D, E\}$, and the blue branch induces $\{A, C, D|B, E\}$. The green branch of tree (b) induces the partition $\{B, E|A, C, D\}$, and its pink branch $\{A, D|B, C, E\}$. All other branches are leaves on the trees and induce a singular partition of the form $\{A|B, C, D, E\}$ and are common across both trees. The symmetric distance between the two trees is thus 2, counting the partitions (induced by the red and pink branches) which appear only on one of the trees.

the earliest event and as such are a number between 0 and 1, where an age of 0 refers to a superfamily whose structural ancestor first appeared on one or more leaves of the tree and an age of 1 refers to a superfamily whose structural ancestor first appeared before the trifurcation of the superkingdoms.

2.4.4.1 Maximum Parsimony

The maximum parsimony analysis was undertaken as implemented by Mirkin *et al.* [2003]. Given the occurrence profile of a superfamily across the genomes, several scenarios of gain and loss events at internal and external nodes of the tree can be proposed which explain the profile. Maximum parsimony attempts to find the scenario which minimises the score $S = \lambda + g\gamma$, where λ and γ are the numbers of loss and gain events respectively and g is the gain weight.

By minimising this score the algorithm considers vertical descent of superfamilies to be by far the most common evolutionary scenario at any speciation event on the tree. Both lateral gene transfer and *de novo* gene gain are considered as gain events and the likelihood of these events occurring, relative to gene loss, is parametrised as the gain weight g. We primarily used a gain weight of $g = 1$, maintaining an equal penalty for both loss and gain events and supported

by Mirkin *et al.* [2003]. Further analysis was also carried out using values of g ranging from $0.1 - 10$ incorporating up to a 10-fold penalty on either loss events or gain events relative to each other.

2.4.4.2 Dollo Parsimony

On the trees of MULTIgenomes species Dollo parsimony was adopted as the default model for age estimation. Dollo parsimony allows at most a single gain event in the evolution of a superfamily and aims to minimise the number of subsequent loss events.

2.4.4.3 Fusion Parsimony

The above parsimony models were combined to define a fusion method, which allowed at most one gain event to occur on the Eukaryotic subtree but multiple gain events to occur across other parts of the tree. As such, fusion parsimony assumes Dollo parsimony on Eukaryotic genomes and maximum parsimony elsewhere as the most likely evolutionary model for domain evolution.

2.5 Results

Superfamily and fold ages were calculated using the methods outlined above and are available for download at `http://www.stats.ox.ac.uk/~edwards/Resources.html`. They are also included as Supplementary Table B2. We include here further analysis of these ages and the data used to estimate them.

2.5.1 Superfamily ages

Superfamily ages were estimated for 1,896 different superfamilies. A histogram of the distribution of ages calculated using maximum parsimony and a gain weight of $g = 1$ is shown in Figure 2.9. Perhaps surprisingly, an age of 1.0 is the most popular estimate, associated with roughly a third of the superfamilies. An age of 1.0 indicates a predicted structural ancestor

for a superfamily at the root of the tree of life, and its popularity illustrates that a significant number of superfamilies, as structural progenies, have a strong presence right across the tree of life. There is a reduction in the number of superfamilies with an age estimate between 0.8 and 1.0. This is due to the long branch lengths between each superkingdom and the root of the tree, which differ between different trees.

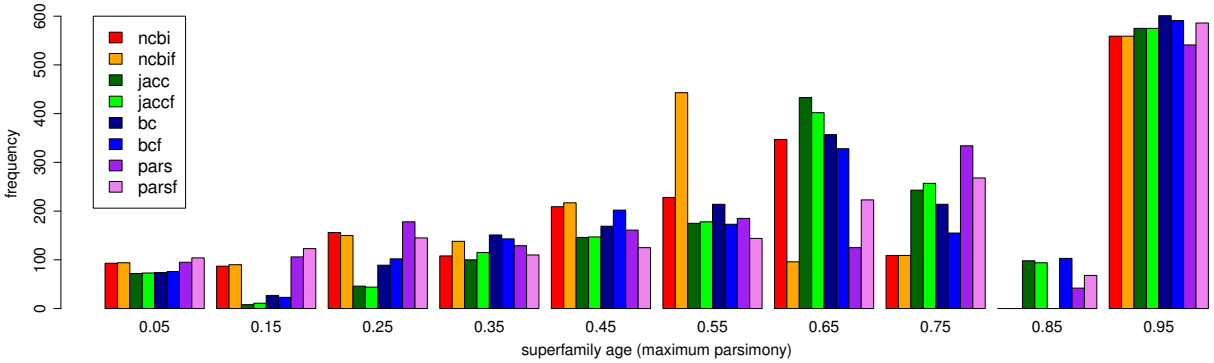


Figure 2.9: Histogram of superfamily ages. Ages are estimated using maximum parsimony and a gain weight of one. Ages are displayed using all eight phylogenetic trees of ALLgenomes. The ncbi and ncbif trees are the NCBI topology with branch lengths estimated using superfamily and fold occurrences respectively. The jacc and jaccf trees use a neighbour-joining algorithm on the Jaccard distances between genomes, where jacc is constructed using the distance between superfamily occurrences and jaccf from fold occurrences.. The bc and bcf trees are similarly built using Bray-Curtis distances between superfamily and fold occurrences respectively. Finally, pars and parsf are trees built using Wagner parsimony, treating the occurrence data (superfamily and fold respectively) as discrete, binary characters. This figure has been reproduced from Edwards *et al.* [2013].

2.5.2 Genome coverage

Looking at the coverage of the SUPERFAMILY assignments across genomes is an important part of assessing the relevance of the age estimates we present here, but is also an interesting question in its own right. In particular, this coverage can indicate the proportion of the protein universe for which either structural data exists or for which a structural model can be postulated (that is, an identifiable homologue exists which has structural data).

Genome coverage can be measured in two ways: either as the proportion of protein sequences on that genome with at least one domain assignment, or the proportion of amino acid residues

with at least one assignment. Figure 2.10 shows the coverage in terms of the number of proteins with assignments, both at a significant ($E \leq 10^{-4}$), and a non-significant ($E \leq 0.1$) level.

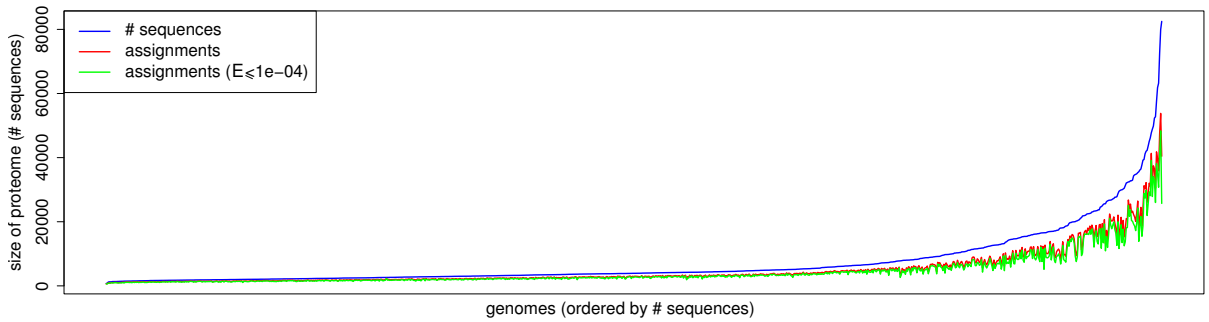
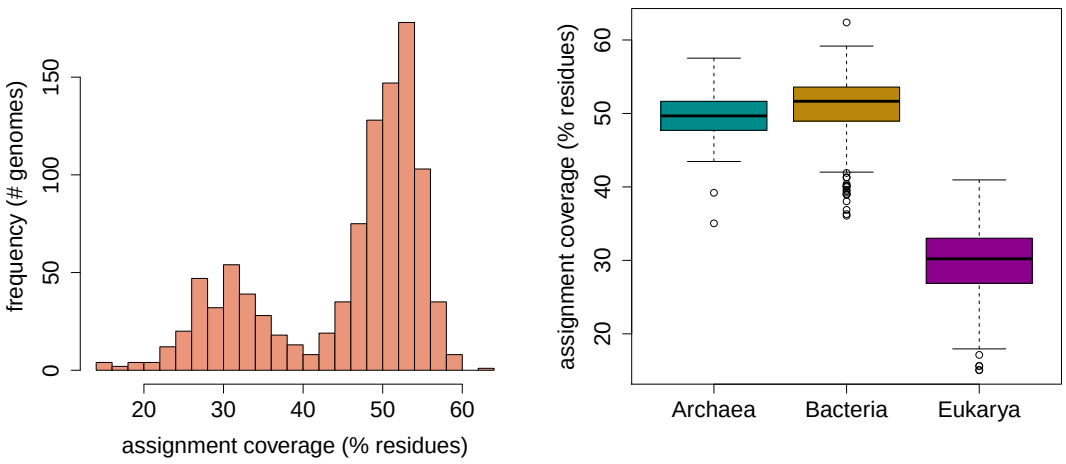


Figure 2.10: SUPERFAMILY coverage on completely sequenced genomes by the number of proteins with assignments. The 1014 genomes were ordered by genome size (# sequences). The number of these sequences with at least one assignment are shown in red ($E \leq 0.1$) and green ($E \leq 10^{-4}$)

Interestingly there is only a small difference between the proportion of significant assignments and those which are non-significant. The cutoff $E \leq 10^{-4}$ was chosen to minimise the number of false positives in the assignments, as suggested by Gough *et al.* [2001]. Also interesting is the fact that genome coverage is, for the majority of genomes, relatively high. On average, 65% of the sequences on a genome are assigned at least one superfamily. The assignment percentage ranges from 31.1% on the Eukaryote *Melampsora laricis-populina* to 87.2% on the Bacteria *Thermatoga*. In general, it is on the larger genomes (which tend to be Eukaryotes) where coverage is lost.

Figure 2.11 summarises genome coverage calculated as the proportion of amino acid residues within the sequences with at least one significant assignment. In general this is lower than the percentage of sequences with assignments. This is because it counts unassigned residues within a sequence even if that sequence has a domain assignment elsewhere. However, a similar picture is presented here in that, on average, a lower proportion of the Eukaryote genomes receive assignments. In fact, the difference between assignments on the different superkingdoms produces a bimodal distribution, with Eukaryote coverage peaking at around 30%, while Prokaryotic coverage tends to be $\sim 50\%$.

The coverage on the genomes of certain organisms may also be informative, particularly in regards to assessing the proportion of the protein universe which exist without structural resolution. For example, the genome of *Homo sapiens* has significant assignments on 60.5% of its protein sequences, covering 29.7% of the residues in its proteome.



(a) **(b)**

Figure 2.11: SUPERFAMILY coverage by the proportion of residues with at least one assignment. (a) A histogram of the frequency of genomes with a particular proportion of residues assigned. The distribution is bimodal with an initial peak at $\sim$ 30% followed by another at $\sim$ 50%. (b) Splitting the distribution by the superkingdom of the genome shows that the low peak captures the coverage on Eukaryote genomes and the higher peak applies to Bacteria and Archaea.

2.5.3 Superfamily predictions

As we have just seen, predictions of superfamilies on the genome sequences are not complete. There are still several sections of the completely sequenced genomes where we cannot suggest with good authority what superfamily or fold is represented. For these remaining sequences there are two possible scenarios. Firstly, a sequence may be a member of a superfamily or fold which has already been classified, but for whom its sequence is not close enough to a domain with existing structural annotation and classification. Secondly, the sequence represents a novel

structural unit which has not yet been observed.

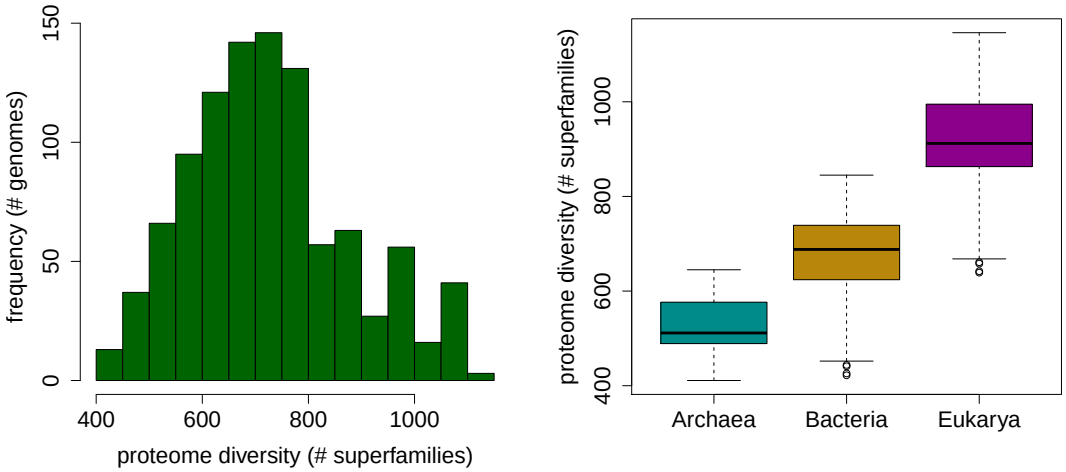
Previous sections have shown that these areas of missing coverage are more common on Eukaryotic genomes but it is of interest if there is a bias towards missing a particular type of superfamily. While it is not possible to fully examine this possibility, the E-values of existing assignments can be analysed to test for a bias in prediction significance. The Kendall's tau correlation coefficient between the average length of domains within a superfamily and the average E-value of that superfamily's assignment is -0.10 which indicates a weak, but still significant, relationship between longer domains and lower E-values.

2.5.4 Superfamily repertoires on proteomes

As well as looking at the coverage of assignments on genomes it is also interesting to look at the diversity of superfamilies assigned to a particular genome. This diversity can be viewed as a measurement of the structural breadth on a genome. Figure 2.12 shows the distribution of proteome diversity, measured as the number of distinct superfamilies assigned to a genome. In general Eukaryotic species have a higher structural breadth than Bacteria, who, in turn, contain a larger number of distinct superfamilies than Archaea. A further dimension to this analysis is available by splitting this superfamily diversity by SCOP class (see Figure 2.13). Eukaryotes tend to contain a greater diversity of superfamilies from all classes except α/β superfamilies, where the structural breadth of the different superkingdoms is more balanced. In fact, on average, Bacterial species contain more α/β superfamilies than either Eukaryotes or Archaea. Furthermore, all genomes tend to have smaller repertoires of all-β superfamilies and larger repertoires of α/β superfamilies.

2.5.4.1 Ancient superfamilies

We compared the set of ancient superfamilies (superfamilies with an age estimate of 1.0) to the superfamily repertoires on the 1,014 genomes we built occurrence data for. There were 485 superfamilies annotated with an age of 1.0 using maximum parsimony ($g = 1$) on all 8 trees



(a) **(b)**

Figure 2.12: Superfamily repertoires on genomes by SUPERFAMILY assignment. (a) A histogram of the frequency of genomes with a particular structural breadth measured by the number of distinct superfamilies assigned to its sequences. (b) Boxplots of this distribution when the genomes are split by superkingdom.

over the ALLgenomes set. We compared this set to the structural repertoires appearing on the completely sequenced genomes by calculating the Jaccard distances between them. Figure 2.14 shows how this set of ancient superfamilies compared to the profiles of superfamilies found on the completely sequenced genomes. The ancient superfamilies, as a structural repertoire, were more similar to the Archaeal genomes than either Bacteria or Eukaryotes. The tree of species was recalculated adding the ancestral repertoire as a separate profile, and this set was place alongside Archaea, as an outgroup to this subtree (see Figure 2.14b).

2.5.5 Robustness of age estimates

In order to assess the relevance of these estimates an investigation is presented here into aspects of their robustness. In particular, two of the four steps within the age estimation process were perturbed. Age estimates were compared when altering the tree of life or the age estimation algorithm to demonstrate how sensitive these estimates are to these underlying assumptions.

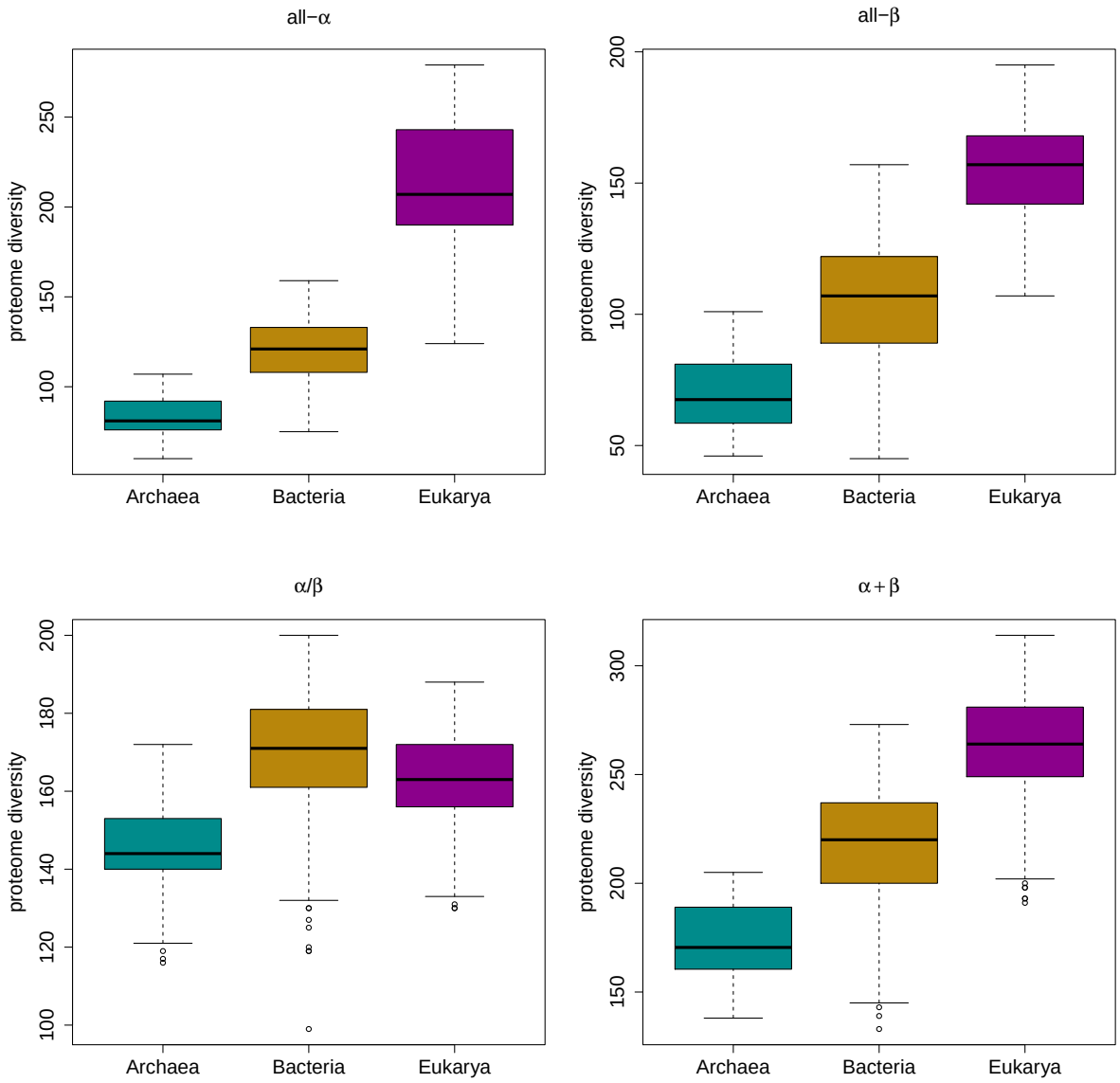


Figure 2.13: Superfamily repertoires split by SCOP class. Boxplots of the number of distinct superfamilies of the four main SCOP classes (all-α, all-β, α/β, $\alpha+\beta$) assigned to each genome.

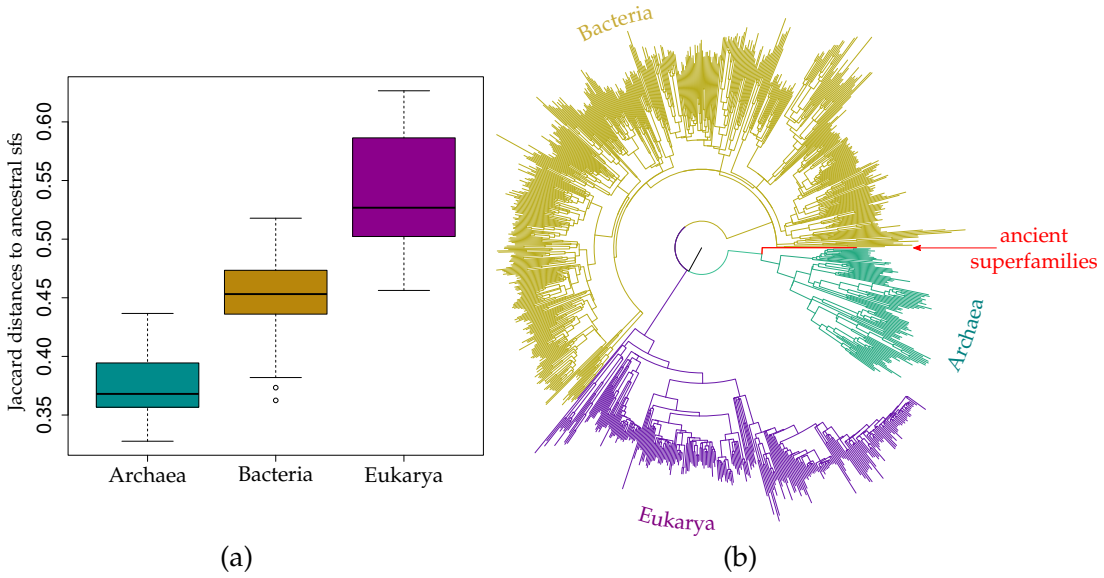


Figure 2.14: Comparing the set of ancestral superfamilies with the structural repertoires on completely sequenced genomes. (a) Boxplot showing the Jaccard distances between each genome, split by superkingdom, to the set of ancient superfamilies. (b) Neighbour-joining tree built using Jaccard distances between the 1014 proteomes in ALLgenomes the set of ancient superfamilies.

Moreover, the method outlined here was compared to another dissimilar methodology for age estimation.

2.5.5.1 Phylogenetic trees

Several possible trees of life were generated for this analysis. All trees successfully segregated the three superkingdoms. All topologies were retained to contribute to analysis in further chapters. Figure 2.15a shows a heatmap and dendogram of the symmetric distances, as defined by Robinson and Foulds [1981]. This symmetric distance measures the number of partitions on one tree but not the other (see Methods 2.4.3.5). Based on this distance the NCBI tree topology is most dissimilar to the trees constructed using the occurrence profiles. This is to be expected as the constructed trees all use the same occurrence data, while NCBI uses a wider range of taxonomic information. However, it is reassuring to note that the distances from the NCBI trees to the other topologies remain of a similar magnitude as the differences between the

constructed topologies. Trees built using Neighbour-joining with Bray-Curtis distances, and those built by Wagner parsimony are clustered together. Within this group trees built using superfamily occurrences are clustered together, as more similar than the trees built from fold profiles. However, trees built using Jaccard distances on superfamily and fold profiles are more similar to each other than they are to the other methods.

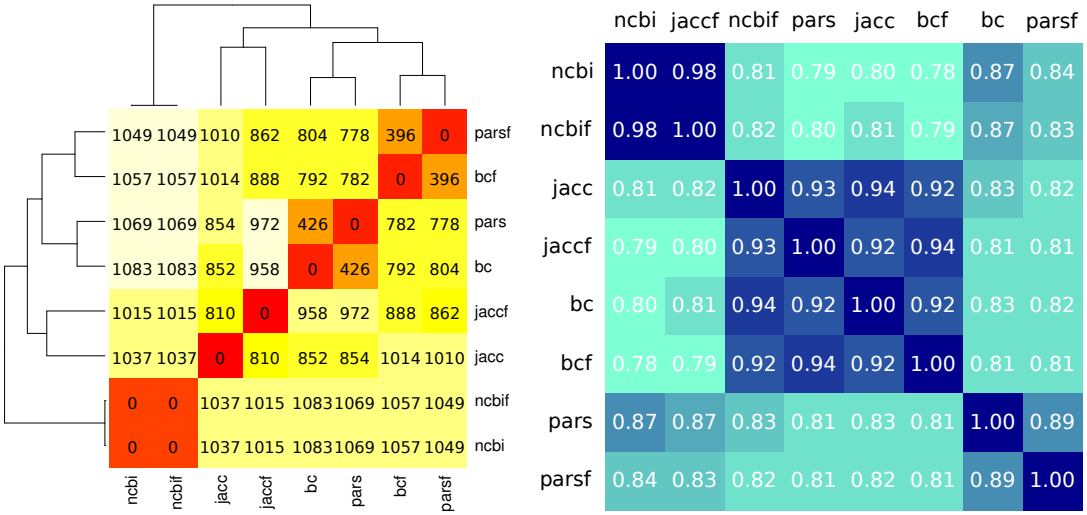


Figure 2.15: Phylogenetic tree comparison. (a) Heatmap and dendogram of the symmetric distances between tree topologies. (b) Pairwise Kendall-τ correlation coefficients between superfamily ages generated by maximum parsimony on the different trees.

More important than the overall similarity of the trees is the level of sensitivity of the age estimates to the differences in topology. Figure 2.15b shows a table of pairwise correlation coefficients comparing the distribution of ages estimated on the different trees. Correlation is calculated using Kendall's tau coefficient [Kendall, 1938]. These ages are all calculated using maximum parsimony and a gain weight of $g = 1$. Ages across these trees are highly correlated ($\tau \geq 0.78$), with the biggest differences being between the NCBI trees and the distance based trees. The parsimony tree ages correlate strongly with both the NCBI ages and the distance ages. Despite being topologically more similar to the parsimony trees, ages from trees built using Bray-

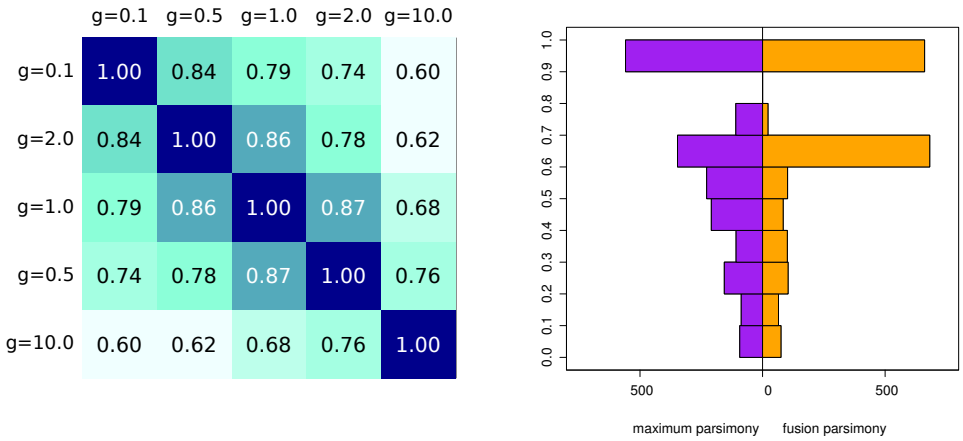
Curtis distances are more highly correlated to those generated using Jaccard distances. In fact patterns of correlations between age estimates do not appear to match the similarities between topologies. This is most likely due to the fact that the symmetric distances vary substantially with the different topological arrangements at the leaves of the tree, whereas differences closer to the root of the tree and those involving differences in branch lengths are more likely to affect the age estimates. However, as we saw in Figure 2.9, the majority of superfamilies have an age estimate at either the root of the tree or at the root of a superkingdom. Thus the age estimates are affected more by the organisation of the tree at its root, which is more sensitive to the algorithm used to construct it. Moreover, we can see how strongly the age estimates correlate, despite their topological differences.

2.5.5.2 Parsimony model

We varied the age estimation algorithm by either a substantial alteration to the gain weight parameter, or by considering a different model of evolution, prohibiting multiple gain events on the Eukaryotic subtree.

Gain weights In the maximum parsimony algorithm the gain weight g represents the ratio of the probability of a loss event relative to a gain event. One of the most significant assumptions within the maximum parsimony model is this parameter [Omland, 1999]. There is, to our knowledge, no comprehensive assessment of the biological relevance for different values of this parameter for structural superfamily evolution. The majority of the results presented in this thesis follow previous studies which support the hypothesis that these two types of event are equally likely to occur at any internal node [Mirkin *et al.*, 2003]. However we also calculated age estimates using a range of values for this parameter, up to a ten-fold asymmetry in the relative likelihood of both gain and loss events. As expected, the age estimates were sensitive to the change in this parameter, although they still maintained a strong correlation to ages calculated with a relative gain weight of 1 ($\tau \geq 0.68$), where correlation is calculated using Kendall's tau

coefficient. Figure 2.16a shows a full table of pairwise correlations between the age distributions generated on the NCBI taxonomy with branch lengths estimated using superfamily occurrences for this range of gain weights.



(a) (b)

Figure 2.16: Comparing age distributions generated under different parsimony algorithms. (a) Exploring the effects of different gain weights. Pairwise Kendall-τ correlation coefficients of age distributions generated using a range of different gain weights from g = 0.1 (gains 10 times more likely than loss events) to g = 10.0 (loss events 10 times more likely than gain events) (b) Altering the parsimony method. Back to back histogram comparing the age distributions generated using maximum parsimony ($g = 1$) to those from the fusion parsimony algorithm. These ages used the NCBI tree with branch lengths estimated using superfamily occurrences.

Fusion parsimony As we mentioned in the Introduction, gain events in the tree represent gene gain but also false positives in the occurrence data as well as lateral gene transfer events. Since lateral gene transfer rarely occurs among Eukaryotic genomes it is perhaps more biologically relevant to consider the weights placed on gain events differently when considering the Eukaryotic tree of life [Rogozin *et al.*, 2005]. We therefore also calculated ages using a fusion parsimony method: assigning events based on Dollo parsimony within the Eukaryotic subtree and according to maximum parsimony at the root and within the Bacterial and Archaeal subtrees (see Methods). These fusion ages are strongly correlated to those estimated using the maximum parsimony model on the entire tree ($\tau \geq 0.85$ over equivalent phylogenetic trees).

Figure 2.16b shows a back to back histogram comparing ages generated by this fusion method when compared to maximum parsimony ages with a gain weight $g = 1$. For simplicity, and because these are the most dissimilar set of equivalent ages, only ages using the NCBI tree with branch lengths estimated using superfamily occurrences are shown. The principal contributions to the differences between these age distributions were the ages of superfamilies appearing on a smaller number of diverse Eukaryote genomes being increased to the root of the Eukaryote superkingdom (0.7).

2.5.5.3 Comparing to alternative methodologies

We compared age estimates calculated using our method to those calculated using the construction outlined in Section 2.3.5, where the same data (predictions of folds on genomes) are used to construct a tree of folds. We were unable to access more recently calculated age estimates so this comparison is performed using previously calculated estimates of SCOP (v1.67) folds on a reduced set of genomes. Figure 2.17b shows a density scatter plot of ages calculated using our method (fold ages) against those calculated from the fold tree (fold node age). Our ages are based on the relative height of the ancestral node in a tree of species where 1.0 represents an ancestral node at the root of the tree of life. Fold node ages are based on the number of nodes between a fold and the ancestral node where 0.0 represents a fold at the root. Fold node ages have been projected onto a bifurcating species tree in Kim and Caetano-Anollés [2011] and are shown in Figure 2.17a. On this tree the Archaeal branch diverges from the rest of the tree at a node age of 0.210. The remaining branch bifurcates into Eukaryotes and Bacteria at a node age of 0.415. Our assumption of a trifurcation at the tree's root would fall somewhere between these two values. To compare age estimates between the methods we collapsed any fold with a fold node age estimate smaller than the last common ancestor to the same fold node age as the ancestor. While there is a negative correlation as expected (Kendall-$\tau = -0.47$ where folds with $n_d < 0.415$ were collapsed to 0.415, and $\tau = -0.45$ where folds with $n_d < 0.210$ were collapsed to 0.210), there are several important areas of disagreement between the two methods.

In particular, there are folds which we annotate with an age of 1.0 that are the furthest away from the root fold according to Kim *et al.* [2006]. These tend to be folds with a strong occurrence but relatively low abundance across Eukaryote species, and with a more complex profile involving low copy counts across Bacteria and/or Archaea. Examples of such folds include the Concanavalin A-like lectins/glucanases (b.29), Leucine-rich repeat right-handed β-α superhelix (c.10) and the histone fold (a.22).

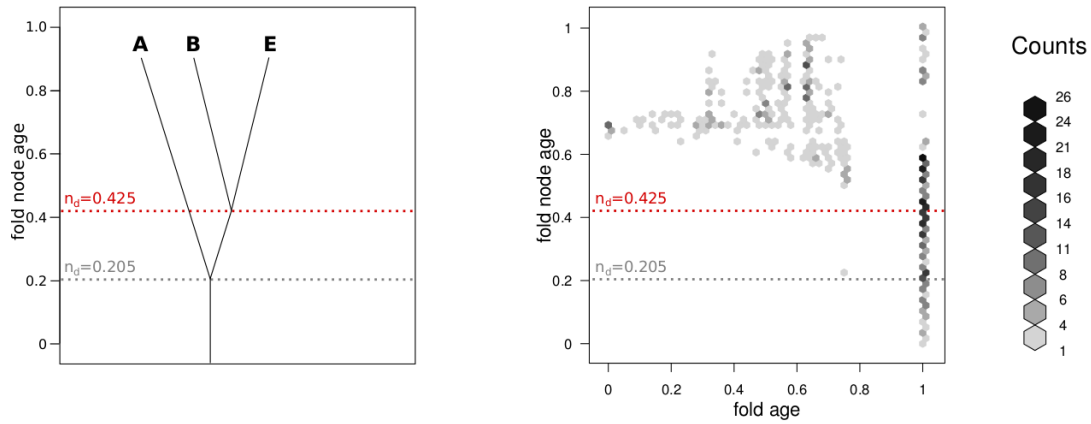


Figure 2.17: Comparison of our ages (fold age) with the node ages of those calculated from a phylogenomic tree of folds (fold node age). Fold node ages for 784 folds were downloaded from the MANET database [Kim *et al.*, 2006] using SUPERFAMILY assignments at the fold level of SCOP v1.67 and as such are out of date. However, a comparison is still worthwhile. Fold ages for SUPERFAMILY 1.67 were taken from previously published work [Winstanley *et al.*, 2005] and were estimated using SUPERFAMILY predictions on a tree constructed from Jaccard distances on superfamily occurrences. As fold ages are only calculated up to the last common ancestor of the genomes a comparison between the two estimates is only meaningful under that timeframe. (a) Kim and Caetano-Anollés [2011] projected the fold trees onto a bifurcating species tree. The last universal common ancestor of all genomes occurred at a fold node age of 0.210 and the three main superkingdoms were estimated to have split by a fold node age of 0.415. (b) Hex-binned heat plot of the fold ages and the fold node ages. Each bin is coloured by the frequency of folds found at that point.

In order to further explore the relationship between the ages calculated from gain and loss events on a tree of genomes and those derived from a tree of fold structures, we estimated our own fold tree ages. We constructed trees of folds from the binary occurrence data we used to construct the trees of species. Trees were constructed using the Camin-Sokal parsimony

algorithm which produces a rooted tree topology from the assumption that ancestral character states are known and using a maximum parsimony criteria to identify the optimal topology Camin and Sokal [1965]. The algorithm was performed using the MIX program from PHYLIP Felsenstein [1989]. We considered the 1,112 folds as taxa, each associated with 1,014 binary characters representing their presence (1) or absence (0) on one of the 1,014 completely sequenced genomes. We further assumed an ancestral state of 1 for each character in this set. 100 delete-half jackknifed samples of 507 characters per fold were considered and 10 most parsimonious trees were stored for each sample. Fold tree ages were calculated from the resultant topologies by counting the number of nodes between each fold and the root of the tree, and then normalised to lie between 0 and 1 where 0 was at the root of the tree and 1 was the fold furthest away from this ancestral set of characters. Finally the fold tree ages were calculated as the median age across each of the 1000 trees. These fold tree ages were negatively correlated to those from species trees (Kendall's-tau = -0.73). Figure 2.18 shows the relationship between these fold tree ages and the age estimates of folds on an NCBI species tree with branch lengths estimated using superfamily occurrences. Folds which were estimated high fold tree ages (were far from the root) but were estimated to have an age of 1.0 on the species trees were almost all folds which occurred strongly across Eukaryote and Archaeal species but weakly across Bacterial species. Examples of such folds are the Ribosomal protein L31e (d.29) and the Heme-binding protein HasA (d.35).

2.6 Conclusions

In this Chapter we have generated age estimates for protein superfamilies and folds representing the currently determined structural universe. These estimates summarise probable evolutionary scenarios resulting in the emergence of protein superfamilies or folds. They are based on predictions of these structural units on completely sequenced genomes across the tree of life. On average 65% of proteins on these genomes received at least one superfamily assignment.

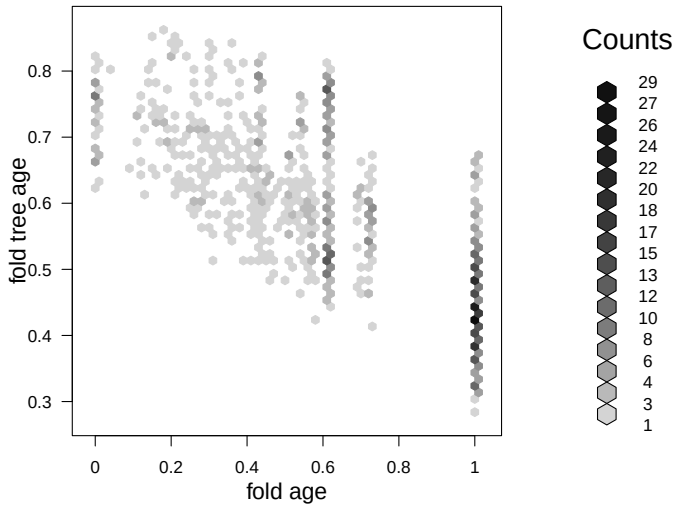


Figure 2.18: Comparison of the species tree ages (fold age) with the fold tree ages of those calculated from a phylogenomic tree of folds (fold tree age). A hex-binned heat plots of fold tree ages and fold ages. Each hex shows the frequency of folds found at that point.

Moreover, despite this incomplete structural coverage, both superfamily and fold occurrence data were used to construct phylogenetic trees of life which successfully partitioned the three main superkingdoms and were comparable to the NCBI taxonomy. Compared to structural superfamilies and folds, sequence family annotation, though at a higher coverage, fails to hold the required phylogenetic signal to resolve such trees.

In general, Eukaryotic genomes tend to be much larger than their Bacterial and Archaeal counterparts, both in the number of protein sequences and in the structural breadth of distinct superfamilies and folds appearing in their proteomes. They also suffer from the lowest percentage coverage of assignments, suggesting that this endowment in terms of structural breadth may be even greater. We compared the set of ancient structures to the extant proteomes and found that as a repertoire of superfamilies, they were clustered with Archaeal genomes, over Bacteria and Eukaryotes.

We investigated the robustness of these ages, showing that they agree fairly well with other

methodologies, and also remain highly robust to changes within our method: to different phylogenetic trees, as well as changing the underlying age estimation algorithm. Adjustments to the parsimony algorithm, such as allowing at most one gain event on the entire Eukaryotic subtree, and even tolerating a 10-fold asymmetry in the likelihood of loss events to gain events across the whole tree, failed to significantly perturb the estimates.

Robustness has been investigated in terms of two of the four stages in the age estimation process, namely the phylogenetic tree of genomes and the model for superfamily evolution. The primary two stages, superfamily classification and prediction, remain uninvestigated. In particular it would be valuable to apply this methodology to domains classified under CATH and investigate how this change in classification affects the estimates. However, systematic comparisons between the two schemes have shown a majority agreement between classifications [Hadley and Jones, 1999]. While we have not included in this chapter a thorough examination of the accuracy of the superfamily prediction pipeline, a brief analysis has shown that the SUPERFAMILY assignments do not appear to be biased towards domains of any length.

CHAPTER 3

Preferences of Ancient and New-born Superfamilies

The majority of the material found in this chapter as well as Figures 3.4, 3.5, 3.6, 3.8, 3.10 and 3.11 have been previously published by the author in Edwards et al. [2013]. They are included here with permission under the Creative Commons Attribution (CC BY) license.

This chapter examines the structures, sequences and functions of superfamilies and relates these properties to the age estimates of that superfamily. Using the ages calculated in Chapter 2 we can partition the protein universe into different populations based on their esti-

mated age. In particular, this chapter examines in detail superfamilies which have evolved at the root of the tree of life (ancients) and those with a recent evolutionary history (new-borns). We examine whether either of these populations show a significant preference for certain properties pertaining to their primary, secondary and tertiary structures, their sequence profiles and functional annotations. The majority of the methods and results in this chapter have been reproduced from the publication Edwards *et al.* [2013].

3.1 Motivation

The evolution of proteins is one of the fundamental processes that has delivered the diversity and complexity of life we see around ourselves today. Through this process nature has explored a vast number of different shapes and configurations. These structures, as we have already discussed, are highly conserved during evolutionary drift because of their optimisation for certain functions. This leads us to question the driving force behind the evolution of new superfamilies and folds.

In Chapter 2 of this report we estimated ages for superfamilies: the evolutionary units which make up the protein universe. The varying age estimates for these superfamilies support the idea, proposed by Choi and Kim [2006], that ancestors of these superfamilies emerged at different points during evolution (see Figure 3.1). Under this model, superfamilies are proposed to have emerged under completely different evolutionary pressures. The range in the age estimates presented in the previous chapter represent substantial evolutionary timescales. For example, superfamilies originating right at the root of the tree have a structural ancestor that was present before the diversification of the three different superkingdoms, whereas those appearing on just a single species' genome will have likely evolved far more recently. It is known that the structural space explored by proteins over evolution is far from homogeneous [Brenner *et al.*, 1997; Choi and Kim, 2006; Goldstein, 2008]. It is possible that the different environments which each superfamily has been subjected to, as a result of the evolutionary age of its ancestor, have affected some of the features making up its structural landscape.

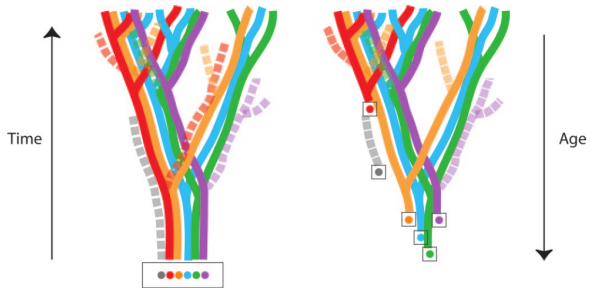


Figure 3.1: Different models of protein evolution illustrated by Choi and Kim [2006]. Structural ancestors of protein families are shown as solid coloured circles and evolutionary paths of that family are shown as coloured lines (solid lines represent extant families whereas dotted lines refer to families which have died out). The single birth model on the left implies that all extant proteins have evolved from proteins that were present in the last common ancestor. The multiple birth model on the right was proposed by Choi and Kim [2006] and suggests that the ancestors of modern day proteins emerged at different points during evolution. This figure has been reproduced with permission from Choi and Kim [2006] (Copyright (2006) National Academy of Sciences, U.S.A.)

For example, while the lengths of protein sequences appear to differ greatly between those present in Eukaryotes and those in Bacteria and Archaea, these differences seem to result from longer non-domain linker regions in Eukaryote proteins while the length distributions of the domains themselves vary much less [Wang *et al.*, 2011]. This result suggests that different selective pressures apply to domains than to their linking regions. While reductive pressure from genome size restrictions can be compensated for by shorter linkers, domain evolution may be driven by alternative pressures. This has prompted investigations into the lengths of protein domains. Several such studies have shown a significant positive correlation between the age of a domain's structure and its length [Choi and Kim, 2006; Abeln, 2007; Capra *et al.*, 2012]. These results remain pronounced over different methods for calculating the age of a superfamily or protein sequence.

This seemingly fundamental relationship between the age of a structure and its length has supported the idea that the primitive protein universe was populated mainly by small folds [Choi and Kim, 2006]. In fact, the recent success in using structural fragments to predict protein structures (see, for example, Kolodny *et al.* [2002]) has further stimulated debate as to

whether the evolutionary origins of the current fold space are in fact short peptide fragments that have combined to form larger folds [Friedberg and Godzik, 2005].

On the other hand however, while considering fragments as a unit of protein structure has indeed been a useful tool in comparing and predicting protein structures, these fragments are never found in isolation in nature. Particularly for globular proteins, the necessity of a hydrophobic core in maintaining stable structure also imposes a minimum length restriction. Even then, the shortest globular structures often require additional stability derived from disulphide bonds or amino acids with aromatic side chains [Greenwald and Riek, 2012]. Both cysteine, which forms disulphide bonds and the aromatic amino acids (phenylalanine, tryptophan, histidine and tyrosine) are thought to have been among the most recent amino acids to evolve [Trifonov, 2004] so are unlikely to have been found in the very first proteins.

It has also been noted that domains in the α/β class domains tend to be significantly older than superfamilies belonging to other classes [Winstanley *et al.*, 2005]. α/β domains also tend to be significantly longer than other classes but they are also distinguishable in several other respects [Hou *et al.*, 2005]. They are unique among the classes in containing a majority of parallel β-strands as opposed to the anti-parallel structure which characterise all-β and $\alpha + \beta$ classes. α/β folds also contain a large number of the so-called 'superfolds' [Orengo *et al.*, 1994]: folds containing large numbers of different superfamilies and a high proportion of all determined structures. Such α/β superfolds include P-loop NTPases, Rossmann folds and TIM barrels [Koonin *et al.*, 2002].

3.2 Outline

In this chapter superfamilies are summarised in terms of several properties relating to their structures, sequences and functions. The purpose of this is to show whether superfamilies which have evolved at different times differ in terms of any of these features. The following sections lay down how superfamilies are divided into two populations: the first, ancients, have a predicted

ancestor at the root of the tree of life, and the second, new-borns, have a relative age estimate of at most 0.4, and have evolved closer to the leaves of the tree. These two populations are then compared using properties of several representative and non-redundant crystal structures from each superfamily.

Initially, secondary structure features are compared, supporting previous work which recognised differences in the age estimates between SCOP classes. We look at, in addition to the class of a superfamily, the strand topologies of their structures, both in terms of their direction and in terms of four-stranded motifs seen in their β-sheets. Additional tertiary packing features are also examined. For example, the proportion of long-range contacts between residues, the proportion of residues which are buried within the structure, and the hydrophobicity of this buried core are investigated. For all these features, dependence on the length of the domain is eliminated by normalising appropriately. In terms of all these features, ancient and new-born superfamilies are found to differ significantly. Newly evolving structures are less elaborate in terms of their strand architecture, have fewer long-range contacts, buried residues and a less hydrophobic core.

Further investigation was also carried out into the sequence and structural properties of these populations. Firstly, propensities for different amino acids within the sequences of ancient and new-born domains are examined. Residues more often found in ancient domains tend to be hydrophobic while polar uncharged and aromatic residues are overrepresented in new-born proteins.

Functional properties for superfamilies are taken from their domains' GO annotations and the populations of superfamilies are tested for propensities of these terms. Fundamental functional roles are found to be enriched in ancient proteins but no specific terms are overrepresented in the newly evolving set. These annotations as they stand do not account for the structural patterns we observe, which is illustrated in the case of preference for strand direction.

Finally, a specific case study is presented of two motifs found in parallel all-β proteins. The Greek key and jelly roll motifs are found in a large number of different folds. Geometrically

they are very similar, with the jelly roll containing a Greek key at its core but extending the topology. Despite this similarity there is a clear difference in the age distributions of superfamilies annotated with these different motifs, with Greek key carrying structures tending to be older than those with jelly rolls.

3.3 Methods

Summary properties of superfamilies were obtained using domains from the ASTRAL (1.75) database [Brenner *et al.*, 2000]. This database collates the relevant atomic coordinates corresponding to domains annotated under the SCOP classification. As suggested by Brenner *et al.* [2000] we used only structures with an aerospaci score > 0.4 and, in order to remove bias by over-representation of certain sequence families we filtered the set of representative structures to $< 40\%$ sequence identity. The remaining set of $5,488$ domains will be referred to as the ASTRAL40 set. This included representative domains from all superfamilies within the 7 major classes in SCOP. They are all-α, all-β, α/β, $\alpha+\beta$, multi-domain, membrane and small proteins.

3.3.1 Superfamily ages

Superfamily ages were calculated as described in Chapter 2 on both ALLgenomes and MULTIgenomes. For simplicity, ages on ALLgenomes were calculated using maximum parsimony on the NCBI tree, although results were also supported using ages from other trees and also those calculated using Fusion parsimony. Ages on MULTIgenomes were calculated using Dollo parsimony on the NCBI tree.

Ancient superfamilies were those superfamilies assigned a relative age of 1.0, and new-born superfamilies had an age of ≤ 0.4. This cutoff for new-born superfamilies was chosen to allow for a large enough dataset to be compared to the set of ancient superfamilies. Where applicable, middle-aged superfamilies were any superfamily not counted as ancient or new-born.

Significance tests

Comparisons between the properties of new-born and ancient superfamilies were carried out using the Mann-Whitney U test [Mann and Whitney, 1947]. Since multiple superfamilies shared the same age and therefore tied in rank the standard deviation of the distribution for the test statistic was appropriately adjusted (see [Sheskin, 2007]).

The Mann-Whitney U test assigns ranks to a list of observations. The test compares the sums of the ranks of two populations and states whether the expected rank of an observation from one population is higher or lower than the expected rank of an observation from the other. Explicitly, for N observations, n_1 from population 1 and n_2 from population 2, the U statistic is calculated:

$$U = \min\left(n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2\right)$$

where R_1 and R_2 are the sums of the ranks of observations belonging to populations 1 and 2 respectively. Mann and Whitney [1947] showed that for $N > 20$ independent observations U could be approximated by a normal distribution and thus the significance of a U statistic could be assessed.

This test was used primarily because it makes few assumptions about the distributions the observations are drawn from. Because the test uses the ranks of each observation rather than their actual value there is no assumption of normality [Mann and Whitney, 1947] in their underlying distribution. However, as stated above an important assumption is that each observation, from either population, is independent. For the structural properties and ages of superfamilies and folds it is clearly unreasonable to assume independence. Superfamilies in the same fold will have similar structural properties and evolutionary relationships between different superfamilies will also affect their age estimates.

3.3.2 Length

Lengths of superfamilies were defined as the mean of the lengths of domains representing that superfamily in the ASTRAL40 set. The 66 superfamilies classified as multi-domain proteins in SCOP were omitted from this analysis.

3.3.3 Secondary structure

Secondary structure was assigned using PROMOTIF [Hutchinson and Thornton, 1996].

3.3.3.1 Strand direction

Strand direction was calculated using PROMOTIF [Hutchinson and Thornton, 1996]. Only domains in the ASTRAL40 set with $> 10\%$ strand content were considered. Each domain was then annotated as parallel if $> 75\%$ of its strand residues were in parallel strands, anti-parallel if $> 75\%$ of its strand residues were in anti-parallel strands and mixed otherwise. The label for a superfamily was summarised as the majority label for its representative domains.

3.3.3.2 Sheet topology

Sheet topologies were assigned using PROMOTIF [Hutchinson and Thornton, 1996]. Four-stranded motifs were extracted from the topologies by considering any four consecutive strands that were also consecutive in the β-sheet. Figure 3.2 gives an example of PROMOTIF's topology strings and their interpretation in terms of consecutive strands. Topology strings such as [3,-1X,-1] were selected as four-stranded motifs, but strings such as [3X,-2,3] were omitted as this captures 4 strands not consecutive in the sheet. A superfamily was annotated as containing a motif if it was found on any of the domains representing that superfamily.

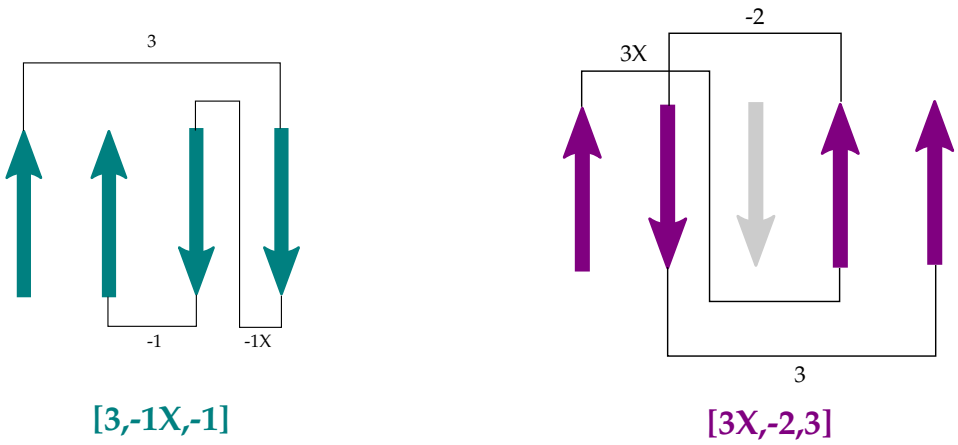


Figure 3.2: Topology strings used in PROMOTIF. Each β-sheet can be assigned a topology string specifying a list of its connections. A meander is given the string 1, and a crossover is given the string 1X. (a) The four strand motif denoted by the sheet topology string [3X,-1X,-1]. The string captures the topology of a sheet or motif within a sheet uniquely up to a change of sign. [-3X,1X,1] also captures this topology. (b) The topology string [3X,-2,3] does not denote four consecutive strands and is omitted from our search for four stranded motifs.

3.3.4 Radius of gyration

The centre of mass (R_c) and the radius of gyration (R_g) of a domain were calculated from the coordinates of the C_α atoms (r_i for $i = 1..N$)

$$R_g = \sum_{i=1}^{N}(r_i - R_c)^2 \text{ , where } R_c = \frac{\sum_{i=1}^{N} r_i}{N}. \tag{3.1}$$

The radius of gyration was then normalised by the radius of a spherical compact globule of N residues [Rawiso, 1999]:

$$R_g^* = \frac{R_g}{N^{1/3}}. \tag{3.2}$$

The radius of gyration for a superfamily was taken as the mean of the radius of gyration values for its representative domains in the ASTRAL40 set.

3.3.5 Non-local contacts

Two residues were said to be in contact if their C_α atoms are $\leq$ 6Å apart (see, for example Plaxco *et al.* [1998]). Contacts are defined as non-local if they occur between atoms in amino acids $\geq$ 10 residues apart. The number of non-local contacts for a domain is normalised by dividing by its radius of gyration:

$$C^* = \frac{\#\text{ non-local contacts}}{R_g}. \tag{3.3}$$

Non-local contacts were summarised for a superfamily as the mean value of C^* on its representative domains.

3.3.6 Buried residues

The solvent accessibility of a residue was assigned using JOY [Mizuguchi *et al.*, 1998]. A residue was assigned as buried if $< 7\%$ of its surface area is exposed to water. The proportion of buried residues in a domain was normalised by the radius of gyration, an estimate of the volume of the structure. The proportion of buried residues for a superfamily was taken as the mean of the values across its representative domains.

3.3.7 Hydrophobicity

The hydrophobicity of a residue was measured using two scales for comparison: the OMH scale [Sweet and Eisenberg, 1983] and also the Kyte-Doolittle scale [Kyte and Doolittle, 1982]. The hydrophobicity of a sequence of amino acids is the sum of hydrophobicities of each residue divided by the length of the sequence. Summary values for the hydrophobicity of a superfamily were calculated by averaging over the hydrophobicities of its representative domains in the ASTRAL40 set.

3.3.8 Disulphide bonds

Disulphide bonds were annotated with JOY [Mizuguchi *et al.*, 1998]. Each domain in the ASTRAL40 set was annotated as to whether it contained disulphide bonds or not. If more than half of the representative domains for a particular superfamily contained at least one disulphide bond it was counted as a superfamily with disulphide bonds. A superfamily was considered to contain no disulphide bonds only if all its domains in the ASTRAL40 set contained no disulphide bonds.

3.3.9 Amino acid content

The Propensities of an amino acid aa for ancient and new-born domains were calculated as:

$$P(aa)_g = \frac{N(aa)_g/N(aa)}{N(total)_g/N(total)} \tag{3.4}$$

where $N(aa)$ is the number of aa residues across all domains in the ASTRAL40 set, $N(total)$ is the total number of amino acids in these domains and $N(*)_g$ is the number of amino acids in domains representing superfamilies predicted to belong to an age group $g \in \{\text{ancient}, \text{new-born}\}$.

Propensities have an expected value of 1, with values > 1 indicating over-representation of that amino acid in a particular age group compared to the background distribution and values < 1 indicating under-representation. We calculated the significance of these propensities using a χ^2-test with a single degree of freedom on the observed occurrences of that amino acid in that age group ($N(aa)_g$). To account for multiple testing the Bonferroni correction was used and only propensities with $P < 0.01/40 = 2.5 \times 10^{-4}$ were considered significant.

3.3.10 Function

GO functional annotations [Ashburner *et al.*, 2000] for SCOP superfamilies were downloaded from the SUPERFAMILY website [Gough *et al.*, 2001]. These functional annotations were assembled using GO terms on Uniprot proteins [Bairoch *et al.*, 2005] with known SCOP clas-

sifications. Annotations for a superfamily were derived from all Uniprot proteins assigned to a superfamily, including multi-domain Uniprots. Annotations supported by just the single domain with a superfamily's classification were also retained as truly domain-centric functional annotations but consisted of a poor coverage across the superfamilies.

Functional enrichment analysis was performed on this set, assuming the number of superfamilies annotated with a particular GO term followed a hypergeometric distribution [Rivals *et al.*, 2007], and significance calculated with a one-sided test for the enrichment of a term in a particular age group $g \in \{\text{ancient}, \text{middle-aged}, \text{new-born}\}$. As above, the Bonferroni correction was used to account for multiple testing. A total of $7,394$ GO terms were investigated so terms with a P-value $< 0.01/22182 = 4.5 \times 10^{-7}$ were considered significant.

3.3.11 Greek key and jelly roll motifs

Greek key motifs were extracted from ASTRAL40 domains using the method outlined in Hutchinson and Thornton [1993]. Strand hydrogen bond partners were assigned using PROMOTIF [Hutchinson and Thornton, 1996]. As the jelly roll motif is formed by adding two extra strands to a Greek key motif, these were then identified from the Greek key set. Superfamilies with any representative domain containing a jelly roll motif contributed to the jelly roll set. All other superfamilies containing domains annotated with a Greek key motif were counted as the Greek key set.

3.4 Results

We examined several properties of the ancient and new-born populations of superfamilies. In the following sections preferences which resulted in a statistically significant difference under the Mann-Whitney U test are reported.

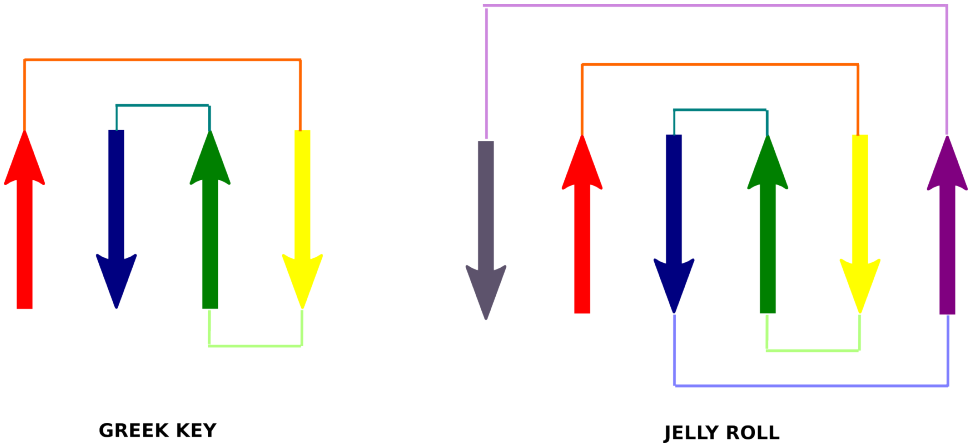


Figure 3.3: Example topologies of the Greek key and jelly roll motifs. Alternative configurations can be obtained by reversing the chain direction.

3.4.1 Structural Preferences

We examined the structural preferences of ancient and new-born superfamilies by looking at properties of their representative domains that related to their lengths, secondary structure and tertiary packing arrangements.

3.4.1.1 Secondary Structure: SCOP class and strand direction

Most globular proteins are classified by their majority secondary structure content in one of the four main SCOP classes (all-α, all-β, α/β and $\alpha+\beta$). This distinction, while potentially arbitrary from an evolutionary perspective, appears to characterise a large part of the structural variation within fold space [Hou *et al.*, 2003]. We observe, in consensus with previous work [Winstanley *et al.*, 2005; Choi and Kim, 2006], that the age distributions of these classes differ substantially. Figure 3.4a gives a percentile plot for the age distributions of the SCOP classes. Each line represents the percentiles of an age distributions for a class from one of the eight different trees. Most notably, α/β superfamilies were significantly older than all other SCOP classes ($p \leq 8.29 \times 10^{-7}$). α/β domains tend to be longer than other classes (Figure 3.5b) and they also contain a large number of the so-called 'superfolds': folds containing large numbers

of different superfamilies [Orengo *et al.*, 1994].

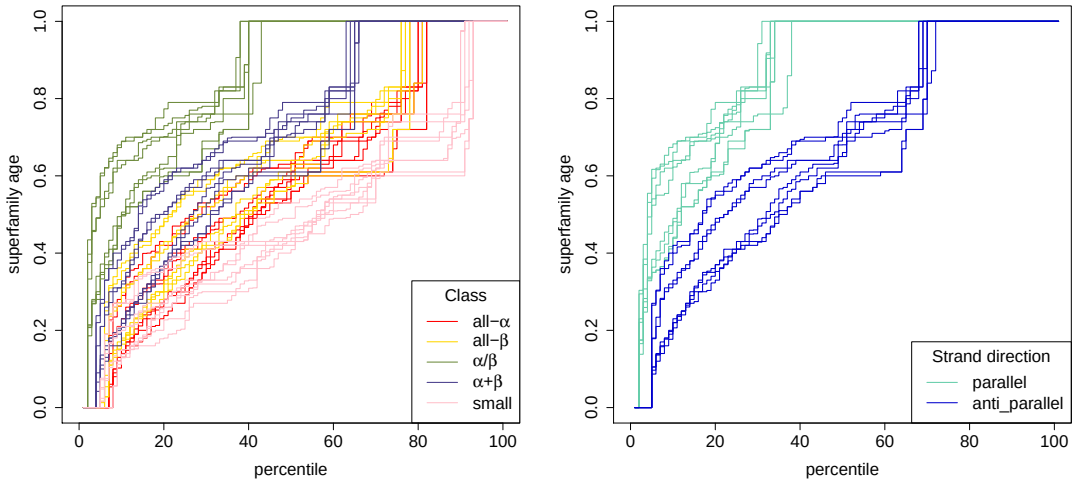
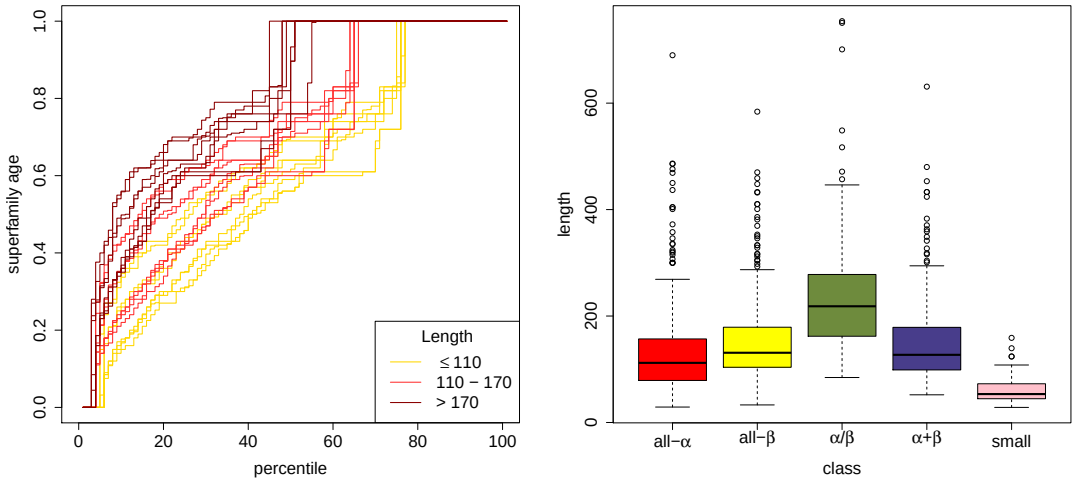


Figure 3.4: Class and strand direction by age (a) Percentile plot of the age distributions of four main SCOP classes as well as the small protein class. Age distributions of multi-domain and membrane proteins have been omitted for clarity. (b) Percentile plot of the age distributions for superfamilies dominated by either parallel or anti-parallel strand structure. These figures have been reproduced from Edwards *et al.* [2013].

α/β domains are also unique among the classes in containing a majority of parallel β-strands as opposed to the anti-parallel structure which characterise all-β and $\alpha + \beta$ classes. We found that, when looking just at domains with primarily either parallel or anti-parallel sheet structure there was a strong, significant preference for superfamilies containing parallel strands to be older than those with anti-parallel strands ($p = 5.20 \times 10^{-11}$, Figure 3.4b). Parallel sheets are rarely seen containing less than five strands so seem to require the cooperation of a more elaborate hydrogen-bonded network than anti-parallel sheets. Parallel strands also tend to have tighter restrictions to the torsion angles of their backbone conformation and tend to be buried by other main chain structures [Richardson, 1981].

3.4.1.2 Domain length

Previous studies have demonstrated a significant positive correlation between the length of a domain and its age [Choi and Kim, 2006; Capra *et al.*, 2012]. The fact that new-born structures appear to be shorter has supported the hypothesis that the primitive protein universe was populated mainly by small folds [Choi and Kim, 2006]. We find that ancient superfamilies are significantly longer than new-born superfamilies ($p = 1.74 \times 10^{-16}$, Figure 3.5a). We also observe that the SCOP class of small proteins is significantly younger than the other classes (for all-α: $p = 8.17 \times 10^{-4}$, all-β: $p = 1.02 \times 10^{-4}$, α/β: $p = 7.57 \times 10^{-28}$, $\alpha + \beta$: $p = 6.51 \times 10^{-11}$, Multi-domain: $p = 1.66 \times 10^{-4}$, Membrane: $p = 1.93 \times 10^{-2}$).



(a) **(b)**

Figure 3.5: Domain lengths and their relationship to superfamily age. (a) Percentile plot of the age distributions for superfamilies partitioned by their domain length. (b) Domain length distributions for superfamilies of five SCOP classes. These figures have been reproduced from Edwards *et al.* [2013].

The observation that α/β superfamilies are both older and longer than other domains raises the question of whether there are other properties unique to these folds which drive their difference in ages and result in a residual correlation between the length of a domain and its age. In order to investigate this we studied the relationship between domain length and

superfamily age stratified by SCOP class.

The relationship between length and age within different classes showed a much weaker correlation than that seen overall. Ancient superfamilies within the all-α and $\alpha + \beta$ classes were still statistically significantly longer than new-born superfamilies within the same classes but other classes failed to show a significant preference (see Figure 3.6). However, this lack of significance could be due to insufficient numbers of superfamilies in both age groups within these classes. It seems that the relationship between the length of a domain and its age is not purely a residual effect of the age distributions of different SCOP classes.

3.4.1.3 β-sheet topologies

In order to look more closely at the β-sheet structure of our superfamilies, we also investigated the distributions on the ancient and new-born populations of different four-strand β-sheet motifs. Figure 3.7 shows two frequency plots for each motif in the ancient and new-born populations. While the simple up-and-down β meander (with topology string [1,1,1]) is the dominant motif in both populations, there is a marked preference in the ancient superfamilies for more elaborate topologies. This not only includes parallel motifs (top row), as we expect from the previous sections, but also along the more interwoven anti-parallel topologies (second row).

3.4.1.4 Non-local contacts

We compared the number of non-local contacts with superfamily age and found that ancient superfamilies had significantly more non-local contacts, normalised by radius of gyration, than new-born superfamilies ($p = 4.38 \times 10^{-11}$). We found no significant difference between the numbers of overall contacts, including local contacts, of ancient and new-born superfamilies. Thus, newly evolved superfamilies appear by this measure to be, on average, simpler and less elaborate structures, with fewer long-range contacts.

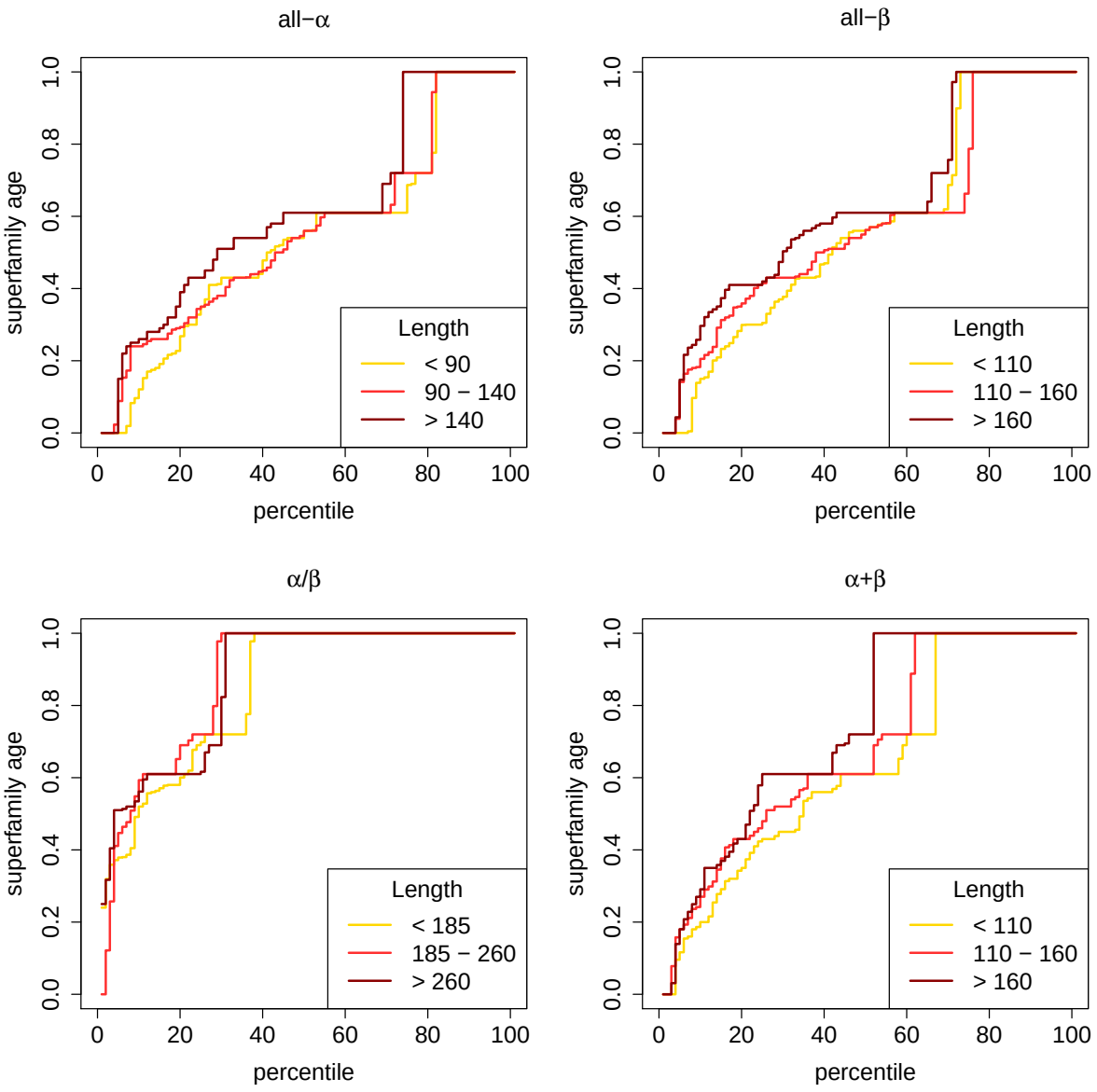
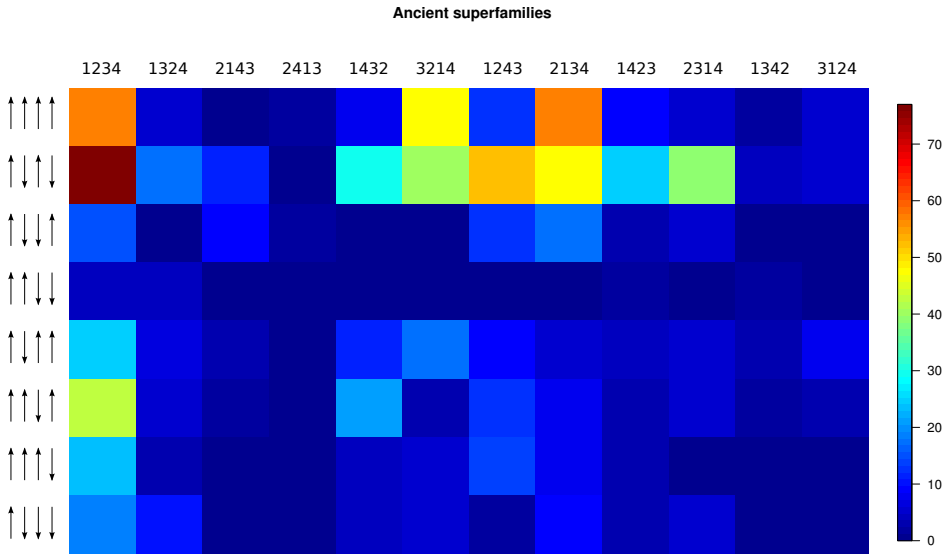
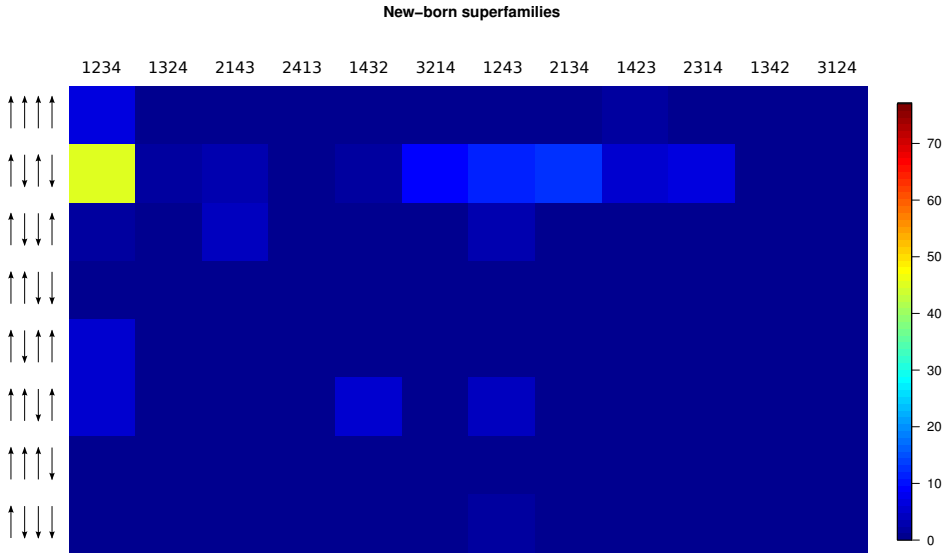


Figure 3.6: Domain lengths and their relationship to superfamily age when stratified by their class. Percentile plots of the ages for different domain lengths within the four main SCOP classes. These figures have been reproduced from Edwards *et al.* [2013].



(a)



(b)

Figure 3.7: Four-strand β-sheet motifs on ancient and new-born superfamilies. Each square shows the number of (a) ancient or (b) new-born superfamilies which have representative domain with a particular motif. The columns of each grid correspond to the ordering of the strands, while the rows correspond to their direction. For example the top left square corresponds to the topology string [1X,1X,1X].

3.4.1.5 Buried Residues

The residues in the core of a protein structure are key to maintaining the overall architecture of the domain, and its structural stability. There are also more evolutionary constraints on these residues than on surface residues [Overington *et al.*, 1992].

Here we studied whether there was a correlation between the ages of our superfamilies and the proportion of their residues that were buried. We found that amongst all domains ancient superfamilies contained a significantly higher proportion of buried residues, normalised by the radius of gyration of the structure, than new-born superfamilies ($p = 3.67 \times 10^{-7}$). This normalised value for the proportion of buried residues indicates the buried portion of the domain relative to its size. New-born superfamilies therefore tend to have a higher surface area to volume ratio than superfamilies in other age groups.

3.4.1.6 Hydrophobicity

The hydrophobic collapse of a globular polypeptide is thought to be one of the primary forces behind protein folding [Sadowski and Taylor, 2010]. The hydrophobicity of the core of a protein structure is thus an important indication of its thermostability and of its folding rate. Given that new-born superfamilies have a higher surface area to volume ratio and there is a marked difference in the hydrophobicities of the core and surface residues in a domain, we investigated whether the age of a domain modulated the hydrophobicity of either its core or its surface.

There was no indication that any age group preferred a highly hydrophilic surface. However, using two different hydrophobicity scales, ancient superfamilies tended to contain a more hydrophobic core than new-born superfamilies ($p = 1.10 \times 10^{-3}$ using the OMH scale and $p = 3.10 \times 10^{-4}$ with the Kyte-Doolittle scale).

3.4.1.7 Disulphide bonds

Another feature that stabilises particular protein structures is the presence of disulphide bonds. These are formed between the thiol groups of two cysteine residues. They are particularly im-

portant for the stability of some small proteins and those secreted in the extra-cellular medium [Wong *et al.*, 2011]. Here we looked at the age distributions of superfamilies containing disulphide bonds compared to those containing none.

Due to the enrichment of disulphides in extra-cellular proteins we carried out the analysis using ages estimated by Dollo parsimony from their occurrences in multi-cellular Eukaryotes only (for details of this see Chapter 2). Even with this constraint superfamilies containing disulphide bonds appear to be significantly younger than those containing none ($p = 1.00 \times 10^{-3}$ and Figure 3.8). The set of superfamilies containing disulphides contained, as expected, a greater

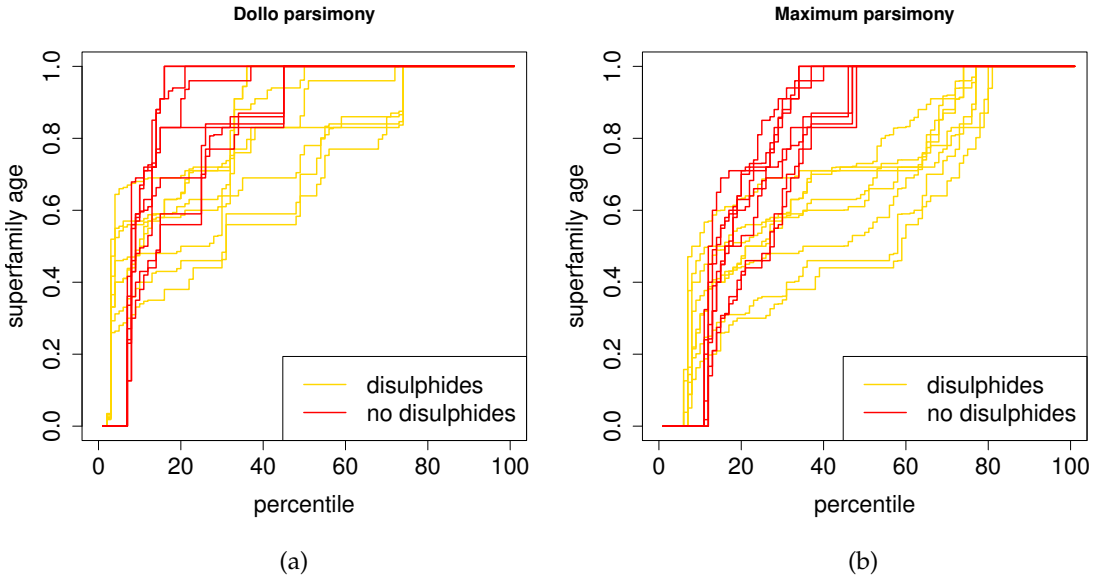


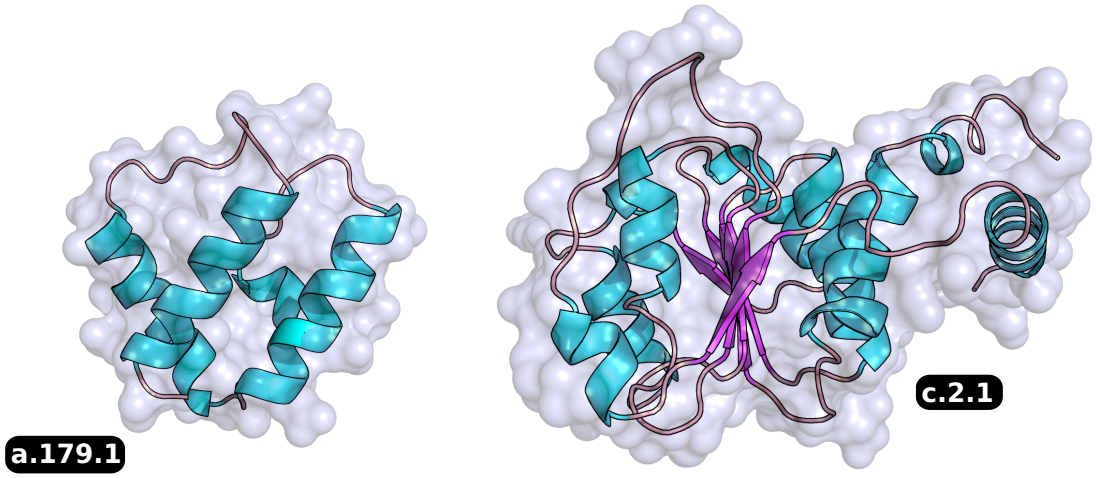
Figure 3.8: Disulphide bonds and their relationship to superfamily ages on multicellular Eukaryotes. Percentile plots for (a) Dollo parsimony and (b) maximum parsimony age estimates on the MULTIgenomes set. Superfamilies with disulphide bonds appearing in more than half their representative domains were compared to those with no bonds found on any representative domain. These figures have been reproduced from Edwards *et al.* [2013].

proportion of the small protein class. However, there was no significant difference in the length distributions of superfamilies with disulphide bonds and those containing no disulphide bonds. It is possible that, in new-born superfamilies, disulphide bonds provide extra stability for more simple, less globular structures.

3.4.2 Visualising the structural preferences of superfamilies

Figure 3.9 shows summary properties of ancient and new-born superfamilies and their tertiary structures. This figure helps to illustrate the magnitude of the preferences identified in the previous sections.

(a)



(b)

	new born	ancient
length	125	186
non-local contacts	14%	17%
buried residues	18%	26%
hydrophobic core	1.85	1.97

Figure 3.9: Structural preferences of new-born and ancient superfamilies. (a) Example domains from the ancient and new-born populations. The superfamily a.179.1 is an all-α domain known as the Replisome organiser. On our tree it is only found in the species *Halobacillus halophilus*. c.2.1 is the populous NAD(P)-binding Rossmann fold consisting of three layers with a doubly-wound β-sheet in the centre. (b) Summary properties for the whole ancient and new-born populations. Properties were calculated as the mean of the values for each superfamily, which were in turn calculated as the mean of the values for each representative domain in the ASTRAL40 set.

Ancient superfamilies are on average 61 residues longer than new-born superfamilies. Approximately 17% of their intra-domain contacts are long range as opposed to only 14% of the contacts in new-born superfamilies. On average 26% of ancient superfamilies' residues are buried whereas only 18% of new-born superfamilies residues are buried. The average hydrophobicity of a buried residue according to Kyte and Doolittle [1982] is 1.97 in an ancient superfamily, and

1.85 in a new-born superfamily.

3.4.3 Sequence level preferences

The enrichment of disulphide bonds among new-born superfamilies indicated a potential over-representation of cysteine residues among these superfamilies. We investigated whether there were further relationships with other amino acids.

Very little is known about the evolution of early life but it is a common theory that the twenty amino acids we see today did not appear simultaneously. It is likely therefore that the earliest peptides consisted of only a subset of these amino acids: the first to evolve. Trifonov suggests a chronological order for the evolution of these amino acids: Gly, Ala, Asp, Val, Pro, Ser, Glu, Leu, Thr, Arg, Ile, Gln, Asn, His, Lys, Cys, Phe, Tyr, Met, Trp [Trifonov, 2004].

We looked here at the sequence composition of different domains and the propensity for different amino acids for ancient or new-born superfamilies. Since sequence change is rapid compared to structural change it is unlikely that the composition of the earliest peptides could be detected from their extant descendants. However, the propensities calculated here may still hold some signal of preference for certain amino acids.

Propensities were calculated for all 20 amino acids across the two age groups and are shown in Table 3.1. While amino acids predicted by Trifonov to occur early during protein evolution were more likely to be enriched in ancient superfamilies this relationship was by no means strict. Amino acids significantly over-represented in ancient superfamilies are Arg, Gly, and Val. Gly and Val are hydrophobic, non-polar residues, and Arg, is polar and positively charged. Four other amino acids which were non-significantly preferred in ancient superfamilies (Ala, Ile, Leu, Met) are hydrophobic, and Pro and His (positively charged) are also non-significantly over-represented. Residues over-represented in new-born superfamilies are Asn, Cys, Gln, Ser, Thr, Trp and Tyr. These residues are mostly polar and uncharged. Trp and Tyr also contain large, aromatic side chains. Non-significantly preferred amino acids amongst new-borns are Asp (polar, negatively charged), Lys (polar, positively charged) and Phe (hydrophobic, aromatic).

The propensities in new-born superfamilies for polar residues further supports our previous observation that newly evolving structures may have a larger surface area to volume ratio.

amino acid	ancient propensity	p-value	new-born propensity	p-value
Ala	1.03	2.93e-03	**0.94**	4.50e-05
Arg	**1.06**	2.14e-05	**0.89**	5.59e-09
Asn	**0.91**	2.13e-09	**1.17**	< 2.2e-16
Asp	0.97	1.24e-02	1.06	6.03e-04
Cys	**0.84**	4.59e-09	**1.31**	8.88e-16
Gln	**0.92**	3.09e-06	**1.14**	1.57e-10
Glu	1.00	7.46e-01	0.99	6.57e-01
Gly	**1.07**	1.23e-08	**0.88**	5.66e-15
His	1.03	1.18e-01	0.94	3.21e-02
Ile	1.04	2.50e-03	**0.93**	3.37e-05
Leu	1.03	1.34e-02	0.95	6.90e-04
Lys	0.97	1.14e-02	1.06	5.19e-04
Met	1.03	1.95e-01	0.95	7.56e-02
Phe	0.99	4.89e-01	1.02	3.43e-01
Pro	1.03	4.92e-02	0.95	6.97e-03
Ser	**0.93**	1.92e-07	**1.13**	9.15e-13
Thr	0.96	2.50e-03	**1.08**	3.38e-05
Trp	0.91	9.01e-04	**1.18**	5.27e-06
Tyr	0.94	3.69e-04	**1.11**	1.03e-06
Val	**1.05**	2.98e-06	**0.90**	1.46e-10

Table 3.1: Preferences of different amino acids for new-born or ancient superfamilies. Propensities for amino acids for a particular age group were calculated using representative domains from the ASTRAL database. P-values were based on a χ^2-test on the proportions of that amino acid observed in each age group. Values were considered significant and given in bold if the adjusted value (using the Bonferroni correction) was less than 0.01. That is, if $p < 2.5 \times 10^{-4}$.

3.4.4 Functional Preferences

So far we have primarily focused on the structural properties characterising superfamilies rather than on their functional roles.

We performed enrichment analysis of GO functions for populations of superfamilies in the different age groups. We compared three different age groups: new-born, ancient and middle-aged superfamilies (those superfamilies in neither the new-born or ancient groups). A list of all

terms which were significantly enriched can be found in Supplementary Table C1.

It has been observed in a study of the protein interaction network of yeast that older proteins tend to have more interaction partners than either middle-aged or young proteins [Rito *et al.*, 2012]. This would appear to indicate that older superfamilies will tend to have more enriched functional terms than younger superfamilies, since partners in the interaction network will tend to share functional annotations. Indeed we find this to be the case. Of 189 GO terms found to be enriched in any one of the three age groups (ancient, middle-aged or new-born), none were enriched in new-born superfamilies, 8 in middle-aged superfamilies and the remaining 181 were enriched in ancient superfamilies.

The terms enriched in middle-aged superfamilies refer mostly to the regulation of developmental growth unique to Eukaryotes. The majority of terms enriched in ancient superfamilies correspond to fundamental cellular processes common to the vast majority of the tree of life.For full details of the functional terms enriched in our age groups see Supplementary Table C1.

3.4.4.1 Does structure or function drive the structural preferences?

We considered the possibility that the structural biases of ancient and new-born superfamilies we report here might be a residual effect of a more fundamental relationship with function. For example, we observe a strong relationship between ancient superfamilies and parallel strands. But, as mentioned before, α/β folds are often superfolds, and are known to be associated with a large repertoire of fundamental functions. Perhaps it is the enrichment of these functions in the α/β class that drives the preference for ancient superfamilies to have parallel strands.

We compared our structural ages (Figure 3.4c) with ages for populations of superfamilies annotated with functional terms enriched in either parallel or anti-parallel superfamilies. In order to do this we constructed lists of parallel/anti-parallel functions: GO terms significantly enriched in the subset of parallel/anti-parallel superfamilies. We then compared the ages of the superfamilies annotated with these terms. The results of this comparison are shown in Figure 3.10. We found that the structural partition resulted in a much more dramatic age

difference than the functional groupings. In particular, the functional annotations failed to divide the space efficiently, with many superfamilies annotated with both 'parallel' terms and 'anti-parallel' terms. Even when considering superfamilies unique to a directional functional annotation, there was a less marked distinction than seen in superfamilies distinguished by structural features alone.

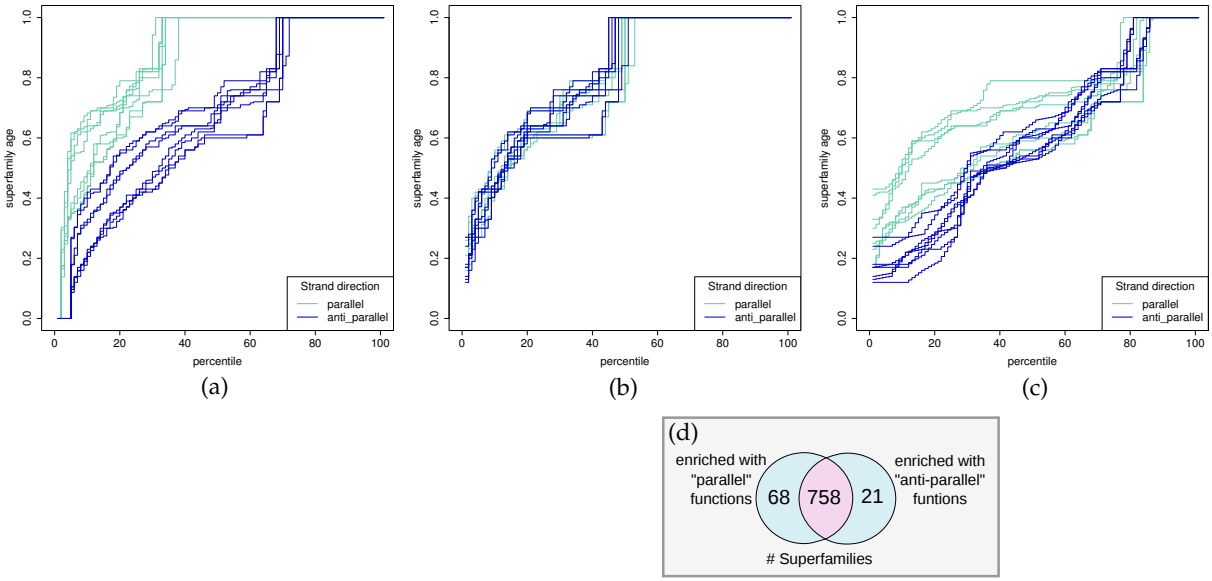


Figure 3.10: Structure vs. functional annotations on fold space preferences. (a) A percentile plot showing the age distributions of superfamilies annotated with a majority parallel/anti-parallel strand structure. It is a reproduction of Figure 3.4b. (b) A percentile plot of the age distributions of superfamilies annotated with parallel/anti-parallel functions: that is functional terms significantly enriched in the parallel or anti-parallel set of superfamilies. (c) A percentile plot of unique superfamilies annotated with parallel/anti-parallel functions. This plot differs from that in panel b in that superfamilies annotated with at functions associated with both parallel and anti-parallel structures are omitted. (d) Venn diagram showing the partitioning of superfamilies into parallel and anti-parallel functional groups. The blue area represents the 89 superfamilies annotated exclusively with either parallel (68) or anti-parallel (21) functional terms. These superfamilies were the ones whose age distributions are plotted in panel c. These figures have been reproduced from Edwards *et al.* [2013].

3.4.5 Case study: Common β-sheet motifs

Not only can these ages be related to general properties of proteins but they also provide a framework for examining more specific questions. For example, we present here a case study

for analysing the evolutionary dynamics of certain structural motifs common in domains in a number of different folds.

As was discussed earlier, anti-parallel β-sheet structures appear to be significantly younger than parallel sheets. Anti-parallel topologies are, however, as we saw in Figure 3.7, more common and more varied than parallel motifs. The most common topology in anti-parallel sheets is the hairpin meander where neighbouring strands in a sheet are consecutive in the amino acid sequence. Apart from the simple meander the next two most common topological motifs are the Greek key and the jelly roll. Around 30% of all-β folds in SCOP are annotated as containing either a Greek key or a jelly roll and these motifs form a considerable role in their classification. Proteins containing these motifs rarely share either sequence similarity or a common function [Hutchinson and Thornton, 1993]. The topological architecture of these two common motifs is very similar, with the jelly roll containing a Greek key at its core. While some papers treat the jelly roll motif as a special case of the Greek key [Stirk *et al.*, 1992], others argue that they occupy a unique portion of fold space [Cheng and Brooks, 2013].

In this study the age distributions of superfamilies classified as containing a Greek key or a jelly roll were compared. Greek keys were significantly older than jelly rolls ($p = 0.01$, Figure 3.11). Moreover, we could find no other disparity (for example, in the lengths of these populations) that helped explain this difference.

3.5 Conclusions

In this Chapter we have explored the preferences of newly evolved superfamilies when compared to their ancient counterparts. In general, new-born superfamilies are structurally less elaborate than ancient structures. They appear, on average, to have a less hydrophobic core and a greater surface area to volume ratio even when these properties are normalised for any dependence on length. They differ from ancient superfamilies in terms of their amino acid composition, containing more polar residues, and tend to contain more additional stabilising features such as

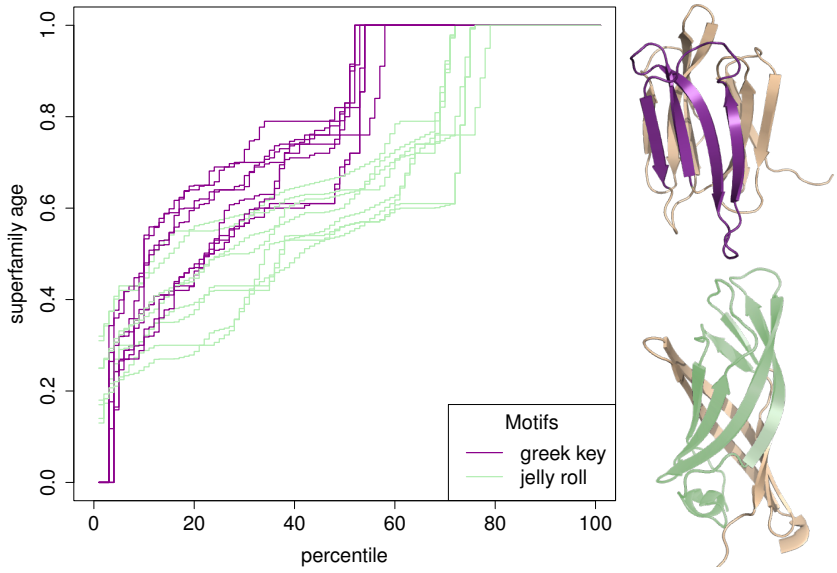


Figure 3.11: Superfamily ages of Greek key and jelly roll motifs. Percentile plots for the age distributions of superfamilies containing a Greek key or a jelly roll motif within their beta-sheet topologies. Domains annotated as containing at least one Greek key motif are significantly older than those containing the jelly roll motif. This figure has been reproduced from Edwards *et al.* [2013].

disulphide bonds and aromatic residues. Ancient superfamilies on the other hand are dominated by α/β superfamilies and are enriched for many fundamental cellular functions.

The age of a superfamily could also be described as the depth at which it can be traced back through evolution. As such, there are several interpretations of our results, in particular in the case of what we have termed new-born superfamilies. Firstly, it could be that an entirely new domain was formed at some point in evolution. This could indicate that the evolution of a new superfamily as a transition from an already existing structure is a rare event, or that evolutionary transitions through fold space, when they occur, are more often reductive. It could also suggest that, through evolutionary drift, there is a tendency towards an increasingly elaborate structure.

Secondly, a superfamily with a low age estimate might have originated earlier in evolution but the family recognition profiles have failed to identify homologues in distantly related species. In this case, such a superfamily may lack a representative deposition of solved struc-

tures, or be rapidly evolving and highly divergent. Certainly, characteristics such as a high solvent accessibility are correlated with the rate of sequence evolution [Toll-Riera *et al.*, 2012]. Nevertheless, by using multiple profiles to build their Hidden Markov Models, SUPERFAMILY improves detection of sequence-divergent families compared to pairwise comparison and single profile searches [Gough *et al.*, 2001]. As a greater coverage of proteins in such superfamilies are solved structurally, the likelihood of an incorrect low age estimate will decrease.

Thirdly, a young superfamily may be the result of an unidentified evolutionary link between superfamilies. As such the structural ancestor of these superfamilies may be earlier than their given age estimates. In order to address this possibility we have shown that the preferences are preserved at both the superfamily and fold level of the SCOP hierarchy.

Finally, what appears to be a young superfamily may actually be ancient but has been lost at several more internal nodes than a parsimonious scenario suggests. This could be the result of functional specialisation within a superfamily. At present our understanding of the evolutionary history of individual superfamilies is not advanced enough to alter the evolutionary model behind age estimation for each superfamily. Our work concerning the robustness of the dataset of ages (see Chapter 2) suggests that the results will be upheld within a moderate level of variation between different superfamilies.

In this study we considered the structural universe of proteins and showed that the age preferences of structural characteristics are not a residual effect derived from functional annotations. This result alone justifies the use of protein structures as a fundamental evolutionary unit.

Using our age estimates we examined the specific case of Greek key and jelly roll motifs, and identified a significant difference between their ages of origin. Given their similarity in topology it is possible that some superfamilies containing these motifs were involved in evolutionary transitions, where a Greek key acted as a scaffold during the innovation of a jelly roll topology.

This example demonstrates that these ages can be used to examine specific properties or motifs of interest, as well as explore more general fold space preferences for proteins at different

stages in their evolution.

The structural preferences for superfamilies exposed in this chapter prompt a consideration of the landscape of different protein structures and how evolution has explored this space. The next chapter looks at this idea. In particular, network constructions of the global structure space are presented and examined.

CHAPTER 4

Bridges through fold space

In this chapter several dynamic network representations of protein structure space are presented where each fold is a node and different folds are connected to each other when they share significant structural similarity. Each network is representative of a different algorithm evaluating structural similarity, and dynamically alters as the threshold for determining the significance of this similarity is changed. The edges in these networks represent landscapes of structural bridges: relationships between different folds. These landscapes are examined, both in terms of their general properties as well as specific nodes and edges of interest.

In the previous chapter we examined how superfamilies of different ages appear to occupy different structural neighbourhoods, in terms of their lengths, secondary structure, non-local

contacts and hydrophobic cores. The work presented in this chapter formalises this concept of structural neighbourhoods, within a global space, and examines how folds of different ages sit within the space.

4.1 Motivation

The general view of the protein structure universe is well represented by two classification schemes SCOP and CATH, where discrete and distinct structural folds or topologies exist segregated from each other, and with sharp boundaries between them [Murzin *et al.*, 1995; Orengo *et al.*, 1997]. Increasingly however, evidence is being presented for similarities *between* different folds [Yang and Honig, 2000; Shindyalov and Bourne, 2000; Friedberg and Godzik, 2005; Sadowski and Taylor, 2010]. At a structural level, these relationships could exist for a variety of reasons. It is possible they are the result of a misannotation of fold boundaries, or indeed that fold space is wrongly assumed to be discrete [Sadowski and Taylor, 2010]. They may also be the result of convergent evolution to a particularly favourable confirmation. They could also represent the structural relic of an evolutionary transition from one fold to another. Whatever their cause, such inter-fold similarities are deserving of further study, to illuminate the overall structure and dynamic of naturally occurring fold space.

While several different strategies have been previously employed in order to discern the global organisation of fold space, there is little consensus on the best way to visualise the space. Figure 4.1 shows seven examples taken from the literature of different visualisations for fold space. Each method is summarised briefly in the figure legend. While this is not an exhaustive review of different visualisation strategies it is clear there is a wide variety amongst the techniques presented here. For example, several methods consider folds as discrete points in their space (see Figure 4.1 (a), (e) and (f)), while others consider each protein domain separately (Figure 4.1 (b), (c) and (g)). Furthermore, there is little consensus on how to measure distances between points in the spaces. A variety of different comparison techniques are used, including

analyses of protein structures' fragment content (Figure 4.1 (e) and (g)), their topological connectivity (Figure 4.1 (d)), and even the results of sequence similarity searches (Figure 4.1 (b) and (c)). Finally, different approaches are taken to consider the overall organisation of the spaces. Performing Principal Component Analysis (PCA) or Multidimensional Scaling (MDS) in order to reduce the dimensionality of the highly complex landscape of similarities is a common tactic (Figure 4.1 (a), (f) and (g)), as are network representations (Figure 4.1 (b), (c) and (e)).

4.2 Outline

In this chapter representations of fold space as dynamic networks of discrete folds are presented. Edges in the networks represent structural bridges between folds and are weighted by the magnitude of their structural similarity. Changing the threshold at which the structural similarity score determines a bridge yields a dynamic landscape of structural relationships across fold space.

Four dissimilar methods are used for comparing protein structures and different networks are generated using each of these methods. The consensus network constructed using bridges identified by all four methods is also presented. This consensus space is used to examine and identify high confidence regions, as well as bridges which may be simply artefacts of a particular method. Significant portions of each space are identified as possible artefacts of each method, with about half of the bridges in each landscape unsupported by any other methods. However, we expose a common organisation to these spaces centred on their division into five communities: all-α, all-β barrels, all-β sandwiches, α/β and $\alpha + \beta$.

In Chapter 2 of this thesis the concept of the age of a fold was introduced. Assuming a common ancestry for domains of a particular fold, the age gives an estimate for a fold's structural ancestor based on its abundance across the tree of life. Fold ages are also used here to annotate each node in the dynamic networks. In particular, we find that ancient folds sit centrally in the networks while new-born folds occur more often at the periphery of the spaces. Highly central

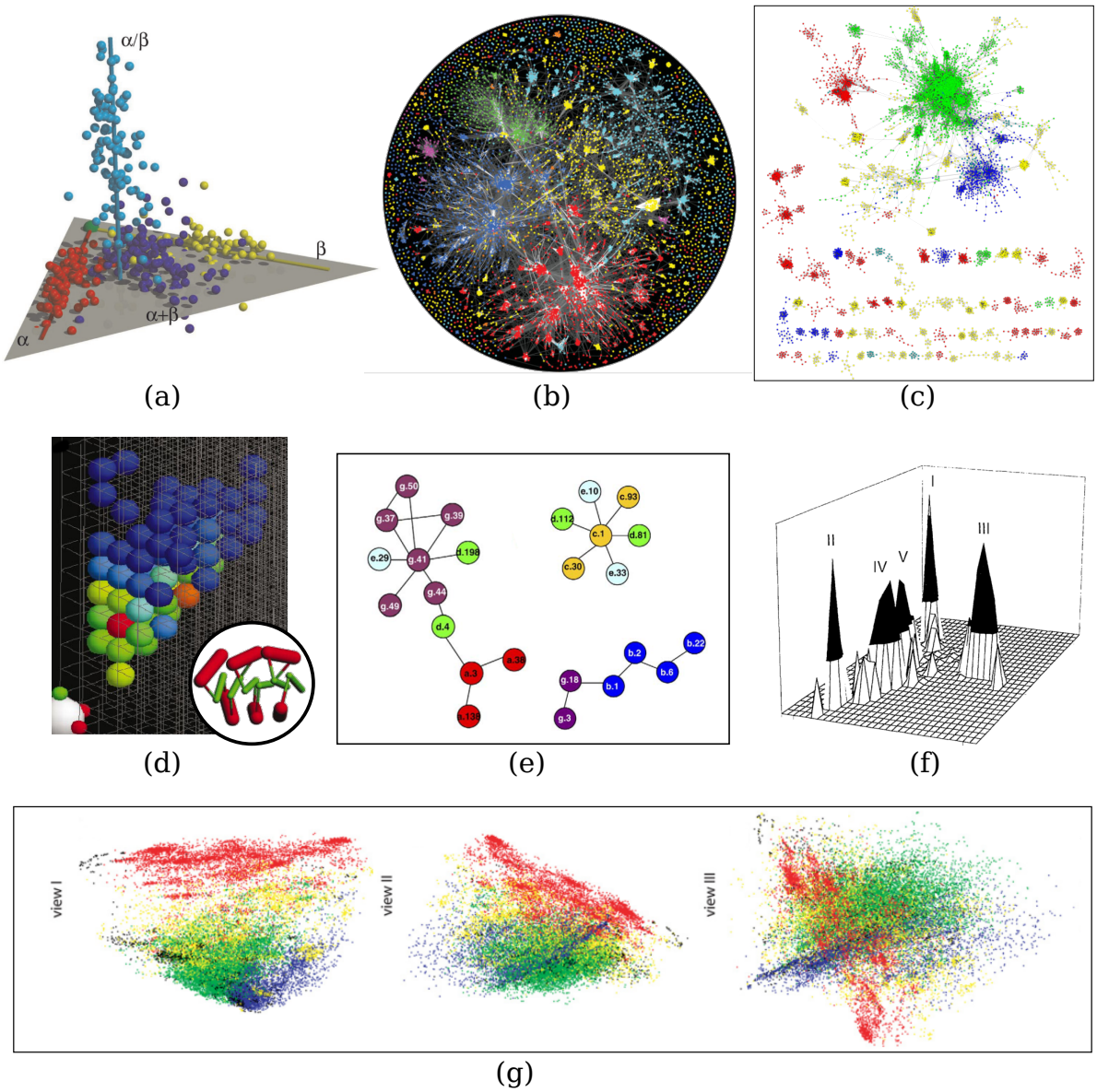


Figure 4.1: Visualisations of fold space. **(a)** A three-dimensional map generated using Multidimensional Scaling (MDS) on a pairwise matrix of structural similarity scores produced using DALI [Holm and Sander, 1993]. Each sphere represents a SCOP fold and are coloured by class (all-α (red), all-β (yellow), α/β (cyan) and $\alpha+\beta$ (blue)). The three eigenvectors identified during MDS are labelled the α, β and α/β axes. Reproduced with permission from Hou *et al.* [2003] © 2003 National Academy of Sciences, U.S.A. **(b)** A Galaxy of Folds from Alva *et al.* [2010]. Fold space is interpreted as a network where each node represents a different protein domain. Domains are coloured by SCOP class (all-α (blue), all-β (cyan), α/β (red), $\alpha+\beta$ (yellow), small proteins (green), multi-domain (orange) and membrane proteins (magenta)) Edges represent the p-value of a sequence comparison using HHsearch. The lighter the edge the smaller the p-value. Reproduced with permission from Alva *et al.* [2010] © 2009 The Protein Society. **(c)** Domain network demonstrating shared motifs. Each node represents a domain and domains are coloured by their SCOP class (all-α (navy), all-β (red), α/β (green), $\alpha+\beta$ (yellow), other (light blue)). Edges represent a shared motif: a region of at least 75 residues with similar sequence and structure. Reproduced with permission from Nepomnyachiy *et al.* [2014] © 2014 National Academy of Sciences, U.S.A. (Continues on next page)

Figure 4.1: (Continued from previous page.) **(d)** A subsection of fold space corresponding to three layer (α-β-α) proteins (left) based on the Forms of Taylor [2002]. Fold space is measured as a grid with the origin (0-0-0) marked as a large white sphere. The directions of increasing numbers of β strands and α helices in each layer are indicated by attached red and green spheres. Grid cells are populated with a sphere if a protein domain has been matched to that Form, and spheres are coloured by the number of matched proteins (red, most; blue, least). The most populated cell has the Form 0-4-2. Reprinted by permission from Macmillan Publishers Ltd. from Taylor [2002] © 2002. **(e)** Network visualisation of fold space. Each node represents a fold and nodes are connected based on the proportion of fragments they have in common. Common fragments are identified using both sequence and structural similarities. Reprinted from Friedberg and Godzik [2005] © 2005 with permission from Elsevier. **(f)** 2-dimensional MDS on a pairwise matrix of structural similarity scores. The third dimension (vertical axis) measures the popularity of each unit area in terms of the number of protein domains. The top five peaks are labelled and termed *attractors* of fold space. They correspond to: (I) parallel β (II) β-meander (III) α-helical (IV) β-zigzag and (V) $\alpha\beta$-meander. Reprinted from Holm and Sander [1996] with permission from AAAS. **(g)** Three perspectives of a map of fold space from Osadchy and Kolodny [2011]. Axes are generated from a 3-dimensional Principal Component Analysis on FragBag vector descriptors of protein domains. Each point represents a protein domain and are coloured by SCOP class (all-α (blue), all-β (red), α/β (green), $\alpha+\beta$ (yellow) and other (black)). Reprinted with permission from Osadchy and Kolodny [2011] © 2011 National Academy of Sciences, U.S.A.

folds are found across the four main SCOP classes and the higher age estimates of central nodes cannot be attributed simply to a bias towards α/β folds.

Furthermore, the structural bridges that define these landscapes can be attributed an age difference: the absolute difference between the ages of the endpoint folds. These age differences are shown to be lower on bridges within the networks than the age differences between two unconnected folds. Moreover, age differences are even lower on bridges which are identified by more than one method.

By examining the network spaces and their consensus in more detail, four pivotal folds of interest are identified. Each fold is from a different SCOP class and occupies a prominent position in each of the networks under consideration. These folds are all ancient and contain popular structural motifs connecting otherwise disparate communities in the landscapes. Finally two pairs of structural siblings connected by bridges in all networks are presented in more detail.

This chapter consists of five further sections. Initially, the four different structural alignment algorithms used to generate the networks are described. Secondly, terminology specific to network analysis is introduced. The Methods section describes the network construction process.

In particular, it covers a method by which the different alignment scores can be compared in order to establish the cut-offs which will define comparable network landscapes. Finally, the Results and Conclusions sections present the results of the analysis in terms of both general and specific properties of these spaces, and place the results in the context of other work.

4.3 Structure alignment algorithms

Structurally aligning between two or more protein chains is currently an active area of research [Hasegawa and Holm, 2009; Hollup *et al.*, 2011; Sadowski and Taylor, 2012]. Several aspects of this problem remain open questions. In particular, the absence of a consistent measure of distance in structure space means there is no consensus on an appropriate alignment score [Hasegawa and Holm, 2009]. Measures such as the RMSD of aligned atoms scale with protein length so, for longer domains, comparatively insignificant local structural divergence can result in disproportionately large RMSDs [Irving *et al.*, 2001]. Where length corrections are made to the RMSD, for example as with the TM-score [Zhang and Skolnick, 2004], the biological interpretation of this score is uncertain. Where scores are left unnormalised, often statistical measures such as Z-scores or p-values are used to assess an alignment's quality [Wrabl and Grishin, 2008]. However, these scores tend to be calibrated by a method's ability to replicate the existing classification schemes (SCOP and CATH), which are not purely structurally defined. As such, these measures do not provide a meaningful sense of the distance between two structures. A recent publication introduced the use of elastic shape analysis to determine the level of distortion between two protein structures as a well defined mathematical metric [Liu *et al.*, 2011]. This method, while not yet specialised for protein domain comparison, represents a novel approach to the complex problem of structure alignment and as such we have included it in our analysis.

The dynamic networks constructed in this chapter are derived from an all vs. all structural comparison amongst a set of representative domains from the four primary structural classes of

the SCOP database. Given the complexity of the structural alignment problem several different algorithms are considered to compute these comparisons. Specifically, four different methods are used: MAMMOTH [Ortiz *et al.*, 2002], FATCAT [Ye and Godzik, 2003], TM-align [Zhang and Skolnick, 2005] and elastic shape analysis (ESA) [Liu *et al.*, 2011]. These different algorithms are briefly outlined below.

4.3.1 MAMMOTH

The MAMMOTH (**Ma**tching **m**olecular **m**odels **o**btained from **th**eory) algorithm constructs a pairwise array of heptapeptide fragments from each protein structure [Ortiz *et al.*, 2002]. Such fragments are then compared using the unit-vector root mean square (URMS) distance between their C_α atoms.

This distance involves placing unit vectors, each representing the direction between successive C_α atoms of a fragment, at the origin of a unit sphere. After the optimal superposition of unit vectors representing one heptapeptide onto the set of vectors representing another, the URMS distance between the two is calculated as the root of the squared distances between corresponding unit vectors [Kedem *et al.*, 1999].

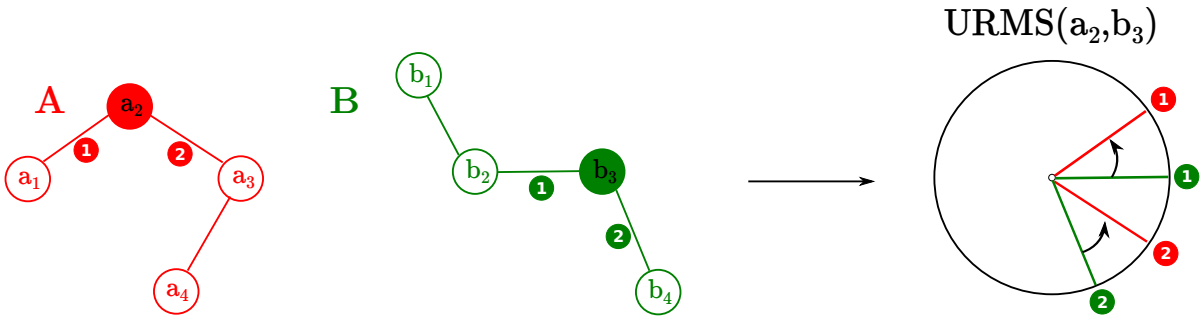


Figure 4.2: A simplified diagram illustrating the URMS distance between structural fragments. Here two-dimensional tripeptides are considered as opposed to MAMMOTH's three-dimensional heptapeptides. Each n-peptide can be considered as an ordered set of (n-1) unit vectors capturing the direction between consecutive C_α atoms. For example, the first tripeptide of chain A above is made up of the 2 vectors: $\overrightarrow{\mathbf{a_1 a_2}}$ and $\overrightarrow{\mathbf{a_2 a_3}}$. By transposing each vector to begin at the origin of a unit sphere (2D circle in the above example) two fragments can be compared simply by rotating one set of vectors to lie over the other.

A similarity score between any two heptapeptide structures A and B can be calculated by comparing their URMS distance to the expected distance between two random sets of vectors (URMS^R). Explicitly:

$$S_{AB} = \begin{cases} \dfrac{10(URMS^R - URMS^{AB})}{URMS^R} & \text{if } URMS^R > URMS^{AB} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where URMS^R for two random n-peptides has been shown by Kedem *et al.* [1999] to be:

$$URMS^R = \sqrt{2.0 - \frac{2.84}{\sqrt{n-1}}}. \tag{4.2}$$

For the heptapeptides (7-peptides) considered by MAMMOTH $URMS^R \approx 0.92$.

An optimal alignment between two chains of heptapeptides can be identified by finding the path through the pairwise array of the heptapeptides comprising each chain which maximises the alignment score. Gap opening and extension penalties of 7.0 and 0.45 are used. That is, any n horizontal or vertical steps through the pairwise array carry a penalty to the alignment score of $7 + 0.45(n-1)$. In other words, this alignment will pair together fragments with low URMSDs, while attempting to minimise any gaps in this correspondence.

Given this alignment, which represents a one-to-one correspondence between heptapeptides in one protein with the heptapeptides of the other, a superposition of chains can be calculated which minimises the distance between corresponding fragments. Finally, MAMMOTH calculates the percentage structural identity (PSI): the maximal subset of aligned fragments that have their C_α coordinates $\leq$ 4.0Å after superposition, as a percentage of the length of the shortest structure. These last steps are completed using the iterative, heuristic algorithm MaxSub [Siew *et al.*, 2000].

The MAMMOTH score is in the form of a Z-score based on the likelihood of a better PSI score occurring by chance given the length of the shortest protein. It is the result of fitting an extreme value distribution to the distribution of PSI scores. This fit also results in a p-value.

4.3.2 FATCAT

Similar to MAMMOTH, FATCAT (**F**lexible structure **a**lignmen**t** by **c**haining **A**FPs with **t**wists) performs an alignment between two chains of structural fragments [Ye and Godzik, 2003]. In the case of FATCAT, fragments of 8 amino acids in length are considered. The similarity score for a given pair of fragments, A and B, is:

$$S_{AB} = \begin{cases} 24F(d_{AB}) & \text{if } d_{AB} < 3\text{Å} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where $F(d_{AB}) = 1 - d_{AB}/4$ rewards a low RMS distance d_{AB} between the C_α atoms of A and B. For $d_{AB} < 3$Å, A and B are said to be an aligned fragment pair (AFP).

In MAMMOTH, an initial alignment is sought from local similarities before a global superposition is carried out. FATCAT however, attempts to combine these steps by calculating the global compatibility of fragments and using that to affect their alignment score. In principle, any two AFPs can be chained together provided the first strictly precedes the second (that is, the last residue of the first fragment must be before the first residue of the second fragment in both protein sequences). The alignment score for connecting two AFPs is a function of their global compatibility as well as penalties for any gaps or mismatched regions involved in their chaining.

Given two AFPs, m and k, between proteins X and Y, their *compatibility* is defined as:

$$D_{mk} = \sqrt{\sum_{i=1}^{8} \left(d_{X_i^m, X_i^k} - d_{Y_i^m, Y_i^k} \right)^2} \tag{4.4}$$

the root mean square deviation of the RMSDs between corresponding residues in the two fragment pairs for each protein (see Figure 4.3).

If $D_{mk} \leq 5$Å, then m and k are deemed globally compatible and can appear in the same alignment, although their similarity scores will be penalised if $D_{mk} > 1$Å. FATCAT alignments

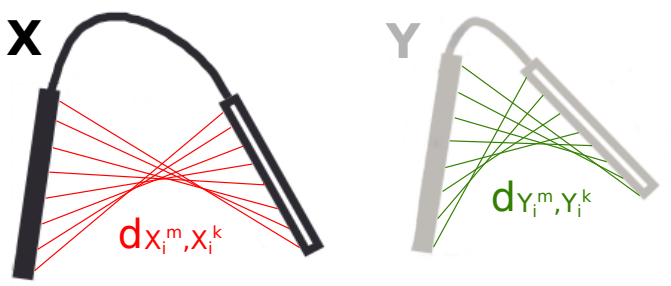


Figure 4.3: Compatibility in FATCAT alignments. To quantify the compatibility of two AFPs the distances between corresponding residues of each fragment are compared between the different chains. Reproduced from Ye and Godzik [2003] by permission of Oxford University Press.

can still result from chaining together incompatible AFPs but require the introduction of a twist in one of the structures (see Figure 4.4). Twists are penalised in the alignment score and the number of twists in a full alignment is limited to 5.

The explicit cost (penalty to the alignment score) for a step from AFP m to k is given as:

$$c(m \to k) = 25W(D_{mk}) + 0.5(p+q) \tag{4.5}$$

where p and q are the gaps and the mismatched regions introduced by the chaining and $W(D_{mk})$ weighs the cost of the compatibility between m and k:

$$W(D_{mk}) = \begin{cases} 1 & \text{if } D_{mk} > 5\text{Å} \\ \left(\dfrac{D_{mk}-1}{4}\right)^2 & \text{if } 1\text{Å} < D_{mk} \leq 5\text{Å} \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

By combining local fragment similarities with the concept of global compatibility FATCAT's algorithm generates a single alignment which maximises structural correspondence between residues but which is also legitimate in a global flexible superposition. Finally, FATCAT's alignment is post-processed by removing twists which do not affect the overall RMSD and iteratively refined. The FATCAT algorithm produces a FATCAT-score (FS) for each optimal

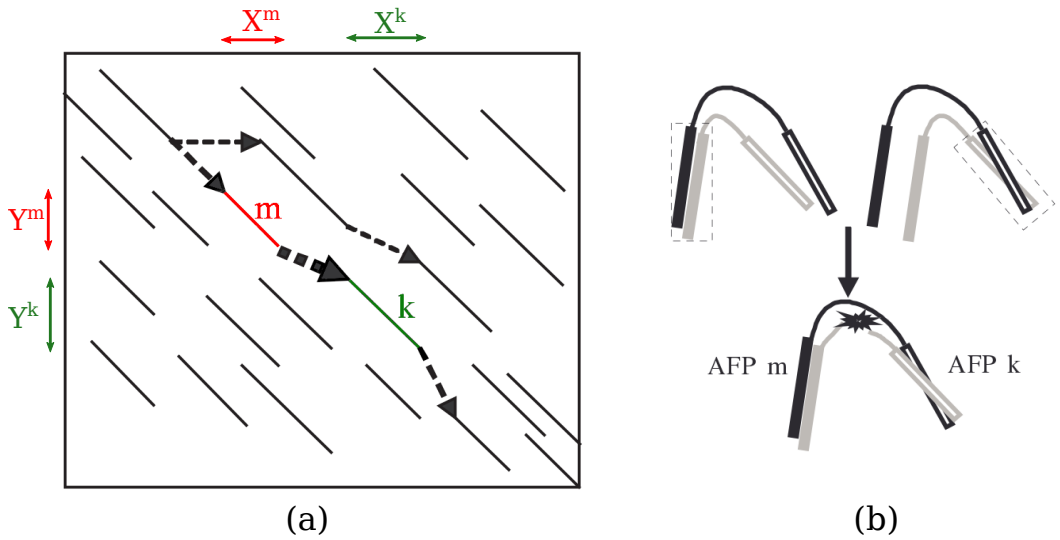


Figure 4.4: Flexibility in the FATCAT algorithm. An aligned fragment pair (AFP) m is defined where, after superposition, 8-peptides X^m and Y^m have an RMSD < 3Å. (a) Two AFPs, m and k, can be aligned if they are globally compatible: that is, they require similar superposition transformations. (b) Incompatible AFPs can still be aligned by introducing a *twist* in one of the structures. Reproduced from Ye and Godzik [2003] by permission of Oxford University Press.

alignment, calculated as:

$$FS = AS \times \sqrt{\frac{L}{RMSD \times N}} \tag{4.7}$$

where L is the length of the alignment, $RMSD$ is the RMS distance between the aligned atoms after twists have been introduced, and N is the number of *blocks* in the alignment (i.e. the number of twists + 1). The alignment score, AS, is the sum of similarities between AFPs matched under the alignment, minus the costs of chaining these AFPs together, as described above. It also produces a p-value which is the result of fitting an extreme value distribution to the FATCAT-scores.

4.3.3 TM-align

TM-align is an alignment algorithm constructed from the pairwise comparison of amino acid residues. It is designed to maximise the TM-score between the two structures [Zhang and

Skolnick, 2005]. The TM-score is defined to be:

$$\text{TM-score} = \frac{1}{L_{mean}} \sum_{i}^{L_{ali}} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \tag{4.8}$$

where L_{mean} is the average length of the native protein structures, L_{ali} is the alignment length, d_i is the distance between the ith pair of aligned residues and $d_0 = 1.24\sqrt[3]{(L_{min} - 15)} - 1.8$ where L_{min} is the length of the shortest structure. The TM-score is designed through this equation to be independent of protein length.

The algorithm begins with three computationally efficient initial alignments and then performs successive iterative refinement to these alignments in order to maximise the TM-score.

The first alignment is between the secondary structure states of each residue. Three different secondary structure states (α,β and coil) can be assigned to a C_α trace using the method outlined in Zhang and Skolnick [2005], which, for α and β states, consider the coordinates of 5 neighbouring residues around the position of interest. Similarities between characters $A, B \in \{\alpha, \beta, \text{coil}\}$ are simply:

$$S_{AB} = \begin{cases} 1 & \text{if A = B} \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

The alignment is calculated using a gap opening penalty of -1 but no further penalty for gap extension.

The second initial alignment is in the form of a gapless structure alignment. Like a global alignment this is the result of iterative refinement on an initial guess. However, the absence of gaps in this alignment make it computationally efficient as an initial step. The similarity between two residues and a given superposition of one protein on the other is:

$$S(i,j) = \frac{1}{1 + d_{ij}^2/d_0^2}. \tag{4.10}$$

The final alignment constructs similarity scores between residues by combining the similarities from the previous alignments in equal measures and uses a gap opening penalty of -1. At this stage no gap extension penalty is considered.

The TM-align algorithm then performs iterative refinement to these alignments. Given an alignment, the two structures can be superposed in such a way that the TM-score between the aligned residues is maximal. The corresponding translation and rotation can then be applied to the entire protein chains and similarity scores calculated between any two residues. The next alignment in the iterative path can then be identified as the optimal path through this pairwise array. At this refinement stage the gap opening penalty is set to -0.6 and, like the initial alignments, there is no extension penalty.

The TM-align algorithm outputs this final alignment between the protein structures as well as the TM-score corresponding to that alignment.

4.3.4 ESA

Similar to TM-align, ESA (**E**lastic **s**hape **a**nalysis), aims to shift the concept of *distance* in protein space away from the RMSD. Whereas TM-align focuses on TM-score, in ESA the theory of Riemannian geometry has been applied to define an elastic metric between two curves in 3D space [Liu *et al.*, 2011].

Explicitly, given a 3D curve $\beta(t)$ (where $\beta(t_i) = [\beta_x(t_i), \beta_y(t_i), \beta_z(t_i)]$) consider its representation $\beta(t) = r(t)\Theta(t)$, where $r(t)$ is the magnitude of $\beta(t)$ and $\Theta(t)$ is its direction. The elastic distance of a small perturbation $\delta\beta(t)$ is defined as:

$$\|(\delta r, \delta\Theta)\| = \sqrt{a\int_0^1 |\delta r(t)|^2 \frac{1}{r(t)}dt + b\int_0^1 |\delta\Theta(t)|^2 r(t)dt} \tag{4.11}$$

where the first term, governed by the parameter a, measures the amount of stretching in the deformation, while the second term, governed by b is a measurement of the bending. This norm can be used to find the distance between two protein structures by modelling their backbones

as continuous curves. Figure 4.5 shows an example visualisation of this distance in terms of a path of intermediate structures representing the distance travelled, in terms of stretching and bending, between two curves.

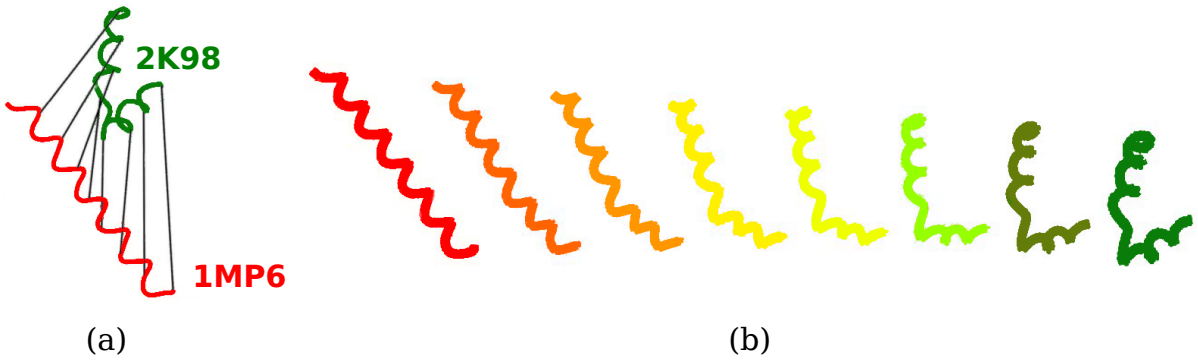


Figure 4.5: Structural similarity using elastic curves. Computing the geodesic curve between PDB structures 1MP6 and 2K98. (a) The alignment between the two protein structures considered as 3D curves. (b) The geodesic between the two curves, representing the path involving the least stretching and bending in the underlying curve. The distance between these two curves, as measured in this method, is 0.895. These figure have been reproduced with permission from Liu *et al.* [2011] © 2011.

Moreover, when the curve $\beta(t)$ is represented by its square root velocity function (SRVF):

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \tag{4.12}$$

this norm is computationally tractable. Furthermore, $q(t)$ is unique for all 3D curves $\beta(t)$ up to translation, and the SRVF of an arbitrary rotation of $\beta(t)$, $O\beta(t)$ where $O \in SO(3)$, is $Oq(t)$. Explicitly, the elastic metric (θ) between two SRVFs, $q_1(t)$ and $q_2(t)$ is defined as:

$$\theta = \cos^{-1}\left(\int_0^1 \langle q_1(t), q_2^*(t)\rangle dt\right) \tag{4.13}$$

where q_2^* is the SRVF q_2 after optimising the rotation and parametrisation of the curve to minimise this distance.

4.3.5 Four diverse methods for structure alignment

The above four methods provide a varied range of different approaches to structure alignment. Probably the least established, ESA is highly dissimilar to the other algorithms presented here. In particular, it has not been fully developed as software for the comparison of protein structures. Rather, the mathematical framework has been established from the theoretical problem of shape analysis, prompting the application to protein structures to be considered.

However, there are certainly similarities between the different methods. FATCAT and MAMMOTH both initiate an alignment based on fragment comparison, and TM-align uses an initial secondary structure alignment, following similar principles. ESA and TM-align both produce spatial quantifications with which to measure the similarity between two aligned structures: the elastic metric and the TM-score respectively. On the other hand, FATCAT's RMSD and MAMMOTH's PSI are compared between different alignments by fitting extreme value distributions to these scores, thus generating statistical measures.

While these methods do not represent the vast scope of available alignment methodologies, they provide a reasonable cross-section of the current field.

4.4 Network analysis

In this chapter we consider the landscape of structure space as undirected, weighted graphs with folds as nodes (or vertices) and significant structural similarities as weighted edges. Concepts relating to the general theory of graphs in the context of our analysis will now be defined.

A *graph* G is defined as an ordered pair $G = (V, E)$, consisting of a set of *nodes* $V(G)$ and a set of *edges* $E(G)$, where each element of $E(G)$ is specified by a pair of nodes $(i, j) \in V(G) \times V(G)$. These nodes are called the *endpoints* of an edge. In a *weighted graph*, each edge is also associated with a *weight* w capturing the strength of an edge. In our case these weights refer to the degree of structural similarity between two folds. In general, an unweighted graph can be considered as a special case of a weighted graph where the weights are either 0

or 1. A *path* through a graph is a sequence of edges connecting a sequence of nodes. A graph is *connected* if a path exists between any two vertices. A graph is *disconnected* if it is not connected. A *connected component* of a graph is a connected subgraph which is not connected to any other vertex in the rest of the graph.

4.4.1 Shortest paths

A *shortest path* between two nodes in an unweighted graph is simply a path between them involving the least number of edges.

In a weighted graph, this concept is slightly harder to define. Dijkstra [1959] proposed that the inverse of each weight as the *cost* of a connection and calculating the path of least resistance between two nodes. Explicitly, the shortest path length between nodes i and j is

$$d(i,j) = \min\left(\frac{1}{w_{ih}} + \ldots + \frac{1}{w_{hj}}\right) \tag{4.14}$$

where w_{ih} is the weight associated with the edge between nodes i and h and $ih, \ldots, hj$ is a shortest path between nodes i and j. Conceptually, shortest paths give a sense of the *distance* between any two nodes in the graph.

4.4.2 Community detection

Community structures were detected using the Louvain method for non-overlapping partitions in weighted networks [Blondel *et al.*, 2008]. This method is a greedy algorithm that attempts to maximise the *modularity* of partitions within the graph. Given a partition of the nodes within a graph the modularity of this segregation measures the density of the graph's edges which fall within a single partition compared to those which fall between two different partitions. The partition which optimises this modularity will have the majority of edges connecting nodes within the same partition and only sparse connections between different partitions.

4.4.3 Node centrality

Several measures exist to quantify how *central* a node is to a graph. In general, a node is thought to be central if it is the endpoint to several different edges (and these edges are of relatively high weight), if its average shortest path length to other nodes across the graph is low, and if it bridges together otherwise isolated components of the graph. Explicitly, these three properties can be quantified as the degree, closeness and betweenness centralities of nodes in the graph.

4.4.3.1 Degree centrality

The degree centrality of a node i is the strength of the edges it is attached to. In a weighted graph this can be defined as:

$$C_D(i) = \sum_j w_{ij} \tag{4.15}$$

the sum of edge weights (w_{ij}) along all the edges in the graph with an endpoint at node i. The sum is calculated over nodes j, which are connected to i by a single edge.

4.4.3.2 Closeness centrality

The closeness centrality of a node is how close it is to the other nodes in the graph. Here, closeness is defined as the inverse of farness, that is the inverse of the shortest path length defined in Equation 4.14.

$$C_C(i) = \sum_j \frac{1}{d(i,j)}. \tag{4.16}$$

The sum is calculated over all nodes j in the same connected component as i

4.4.3.3 Betweenness centrality

The betweenness centrality of a node measures how often it appears on shortest paths between other nodes within the graph. In general, the betweenness centrality measure will favour nodes

which connect communities over nodes which lie within a community.

$$C_B(i) = \sum_{j,k \neq i} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{4.17}$$

where $\sigma_{jk}(i)$ is the number of shortest paths, using the distance $d(j,k)$, between nodes j and k which pass through i, and σ_{jk} is the total number of shortest paths between j and k. The sum is taken over all pairs of nodes (j and k) which do not include i, but are in the same connected component as i.

4.5 Methods

4.5.1 Domain dataset

Domain coordinate files for structures from the four main SCOP classes (all-α, all-β, α/β and $\alpha+\beta$) were taken from the ASTRAL database (version 1.75) [Brenner *et al.*, 2000]. To ensure these structures were of sufficient accuracy we removed any file with an assigned aerospaci score of < 0.4, as suggested by Brenner *et al.* [2000]. The set of domains was further filtered to $< 40\%$ sequence identity in order to avoid over-representation from particular sequence families. Due to the requirements of the structural alignment algorithms the dataset was further refined by omitting structures with only backbone C_α coordinates, and those which contained one or more chain breaks. Chain breaks were assigned using the `Bio.PDB` module in BioPython [Cock *et al.*, 2009]. This program diagnoses chain breaks where successive C_α atoms are further than 4.3Å apart. This resulted in a dataset of 4,098 domains, comprising 793 from the all-α class, 948 classified as all-β, 1,215 α/β domains and 1,142 from $\alpha+\beta$. These domains represent a total of 1025 different SCOP superfamilies and 631 folds.

4.5.2 Pairwise comparison

For each of the four structural alignment methods outlined above, $8,394,753 = \binom{4098}{2}$ pairwise comparisons were computed. Some of the algorithms employed heuristic computations, resulting in alignments sensitive to the input order of domains. However, cases which generated different alignments were rare (for example, using MAMMOTH, $< 2\%$ of comparisons resulted in an alternative alignment being found when the input order was reversed). Furthermore, in tests of $\sim 100,000$ reversed alignments, none of these cases resulted in a significant difference to the scores. As a result, for computational efficiency, we performed a comparison just once for each pair of domains in the dataset.

Each method was run using the default parameters, unless otherwise stated. ESA initially characterised each domain backbone as a curve of N points, where N is the average length of each pair of domains. FATCAT was run in flexible mode and TM-align used a TM-score normalised by the average length of the two domains as described in the Methods. Computational times for each method, in CPU seconds, are given for a range of domain lengths in Figure 4.6.

4.5.3 Edge detection

Out of the 8,394,753 domain comparisons, 78,563 represent comparisons between domains in the same fold. Using this set of domain pairs as a gold standard we examined each method's ability to identify SCOP fold level relationships.

For several of the alignment methods, different scores were generated. These scores assess the quality of the structural alignment, its significance or the distance between the aligned domains. We initially used receiver operating characteristic (ROC) analysis to identify the most appropriate discriminating score.

In order to standardise these scores and compare alignment significance across different methods we introduce a posterior probability attached to each method's discriminating score. This Bayesian approach allowed us to set comparable thresholds at different levels. It also meant consensus networks could be constructed at varying degrees of structural similarity.

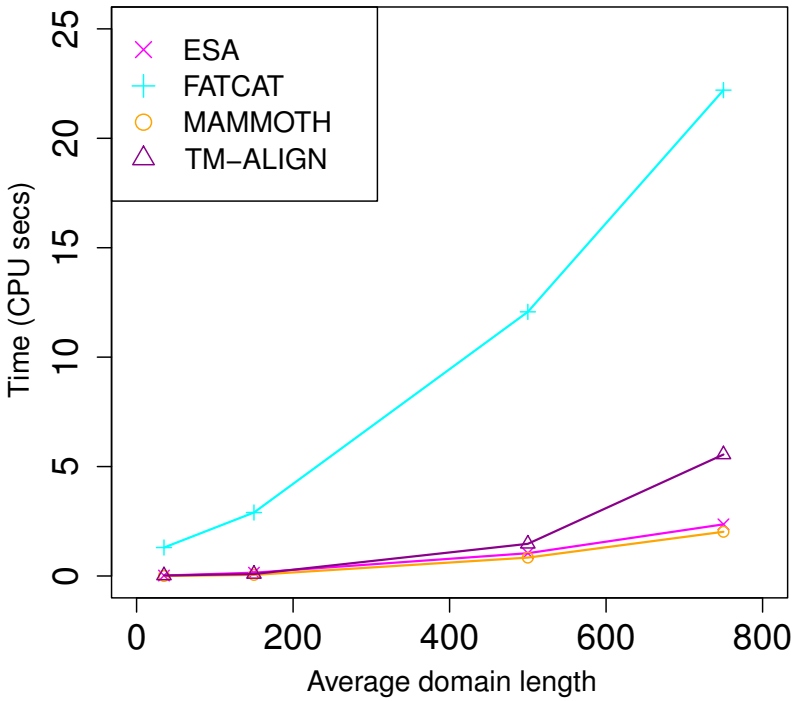


Figure 4.6: Run-times for the structural alignment algorithms. To illustrate the performance of each algorithm, four pairwise comparisons were used, where domains within a pair were unrelated under SCOP and had similar lengths. The average domain length for each pair was 34, 150, 500 and 753, representing comparisons between short, average, long and very long chains. Run-times were processed using all four alignment methods for each pair, and were calculated on the same computer (Intel Core 2 Duo E8500 (3.16GHz))

4.5.3.1 Discriminating scores

The quality of each structural alignment can be evaluated in several ways. In the case where a method generates multiple scores for each alignment we performed a ROC analysis on the distribution of scores and used the area under the ROC curve (AUC) as a measure of that score's ability to discriminate between different folds. Scores generated can be in the form of a similarity (high values indicate closer relationships), or represent a distance (low values imply closer relationships).

Explicitly, for a candidate threshold similarity score $\bar{s}$, we define true and false positives and negatives depending on whether the similarity agrees with SCOP as in Figure 4.7.

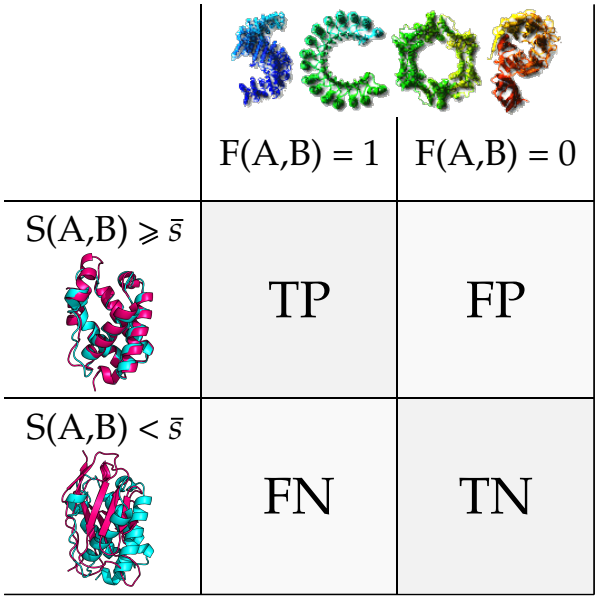


Figure 4.7: Contingency table for comparing similarity scores to SCOP. True and false positive and negatives are the numbers of pairwise comparisons (A,B) in each cell. $F(A, B)$ is a binary classifier, dependent on whether domains A and B are in the same fold in SCOP, and $S(A, B)$ is the similarity score generated by the alignment method. If the score generated captures the distance between the two domains these definitions are similar. In these cases, the positive sets (TP and FP) count pairs (A, B) where $S(A, B) \geq \bar{s}$ and the negative sets (TN and FN) represent pairs where $S(A, B) < \bar{s}$.

True and false positive rates can then be calculated as:

$$TPR(\bar{s}) = \frac{TP}{TP + FN} \quad \text{and} \quad FPR(\bar{s}) = \frac{FP}{FP + TN} \tag{4.18}$$

where TPR and FPR lie in the interval $[0, 1]$. Plotting the pairs $(TPR(\bar{s}), FPR(\bar{s}))$ for values of $\bar{s}$ spanning the range of possible values for $S(A, B)$ gives a ROC curve and the AUC can be estimated using the trapezium rule. For a particular alignment method, the score maximising the AUC was chosen as the most discriminating score.

Table 4.1 shows the different scores we considered as potential discriminating scores for each method. In particular, four different scores were considered from both the MAMMOTH and FATCAT algorithms. TM-align and ESA both produce only one scoring measure.

MAMMOTH	**FATCAT**	**TM-align**	**ESA**
Z-score	FATCAT-score	TM-score	elastic metric
PSI	RMSD		
TM-score	TM-score		
p-value (log scale)	p-value (log scale)		

Table 4.1: Table of scores per alignment method. Lists cover each score we considered as a potential measure for the quality of a structural alignment. For methods which generate multiple scores (MAMMOTH and FATCAT) we compared the ability of these scores to identify fold siblings using a ROC analysis, as described in the main text.

4.5.3.2 Posterior probabilities

In order to make these different scores comparable we performed a Bayesian analysis similar to that outlined in Xu and Zhang [2010]. We calculated the posterior probability of two domains being in the same fold, given a similarity score above some threshold. Explicitly:

$$P(F = 1 \mid S > \bar{s}) = \frac{P(S > \bar{s} \mid F = 1)P(F = 1)}{P(S > \bar{s} \mid F = 1)P(F = 1) + P(S > \bar{s} \mid F = 0)P(F = 0)} \tag{4.19}$$

where $F = F(A, B)$ and $S = S(A, B)$ are as above. In the case of distance scores these inequalities are inverted, to calculate the posterior probability given a score *below* the threshold.

The conditional probabilities $P(S > \bar{s} \mid F = 1)$ and $P(S > \bar{s} \mid F = 0)$ are $TPR(\bar{s})$ and $FPR(\bar{s})$ respectively, and defined above. The prior probabilities, $P(F = 1)$ and $P(F = 0)$ are the probabilities that two domains will either be fold siblings or unrelated. They are taken from the observed frequencies of SCOP fold siblings and unrelated domains in the dataset. Specifically:

$$\begin{aligned} P(F = 1) &= \sum_{A,B} F(A, B)/T \\ &= 78,563/8,394,753 \\ &= 0.0094 \end{aligned} \tag{4.20}$$

and

$$\begin{aligned} P(F = 0) &= 1 - P(F = 1) \\ &= 0.9906 \end{aligned} \tag{4.21}$$

where T is the total number of pairs (A, B).

Plotting these posterior probabilities along a series of values for $\bar{s}$ results in a rapid phase transition from $P(F = 1 \mid S > \bar{s}) \approx 0$ to $P(F = 1 \mid S > \bar{s}) \approx 1$ (see Figure 4.8). At each point along this transition, the value of $\bar{s}$ as a score threshold can define a set of bridges between folds and thus construct a fold space network. In this work we consider networks where the posterior probability $P(F = 1 \mid S > \bar{s}) \geq 0.5$. We construct static networks at probability thresholds of 0.5, 0.6, 0.7, 0.8 and 0.9.

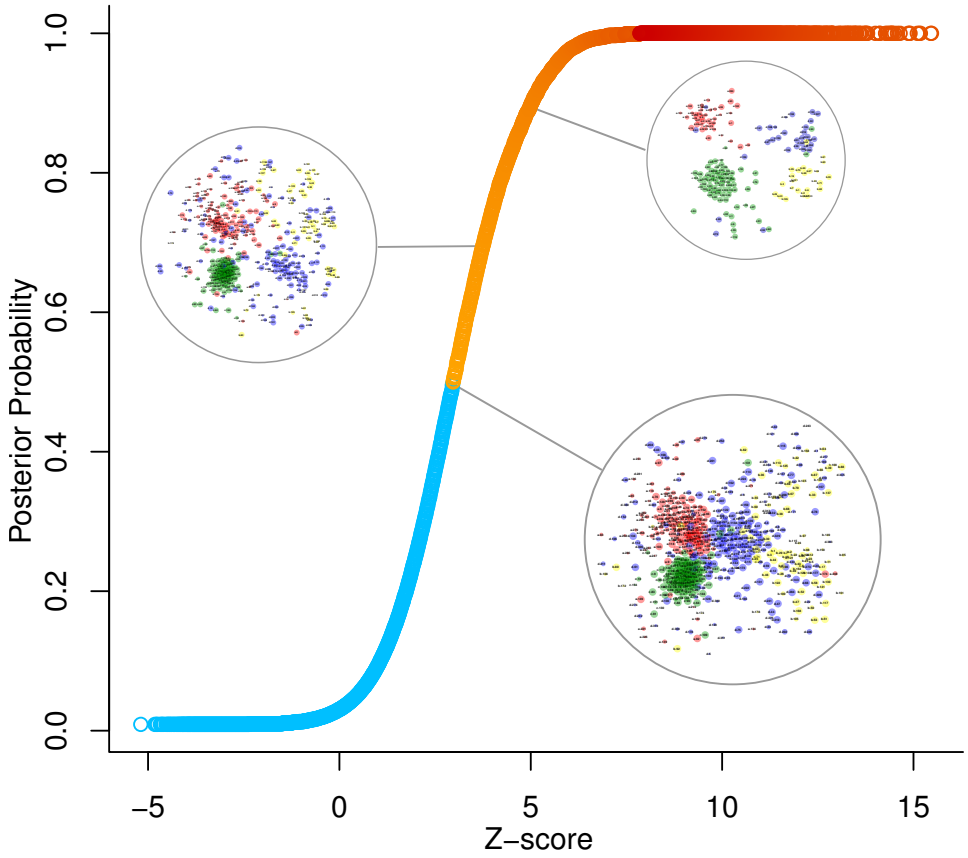


Figure 4.8: Bayesian analysis for different values of a cutoff for the Z-score produced by MAMMOTH. The figure plots the posterior probability of two domains being in the same fold given their Z-score is greater than a threshold. At each probability the corresponding score threshold determines a different landscape of structural bridges. Because each score is linked to a posterior probability in this way, different methods can be compared through translating their scores to the corresponding probabilities.

4.5.4 Network construction

Once appropriate cutoffs were identified, the lists of significant pairwise structural relationships between domains were collapsed to networks of folds. In order to do this we let each node represent a fold with at least one representative domain in the dataset. We placed an edge between two folds if there was a significant similarity between any of their constituent domains. Figure 4.9 is a schematic illustration of this concept.

4.5.4.1 Edge weights

Several of the scores considered here are statistical measures derived from fitting extreme value distributions to the method's scores (e.g p-values, Z-score etc). While statistical measures may be applicable as discriminating scores, we have chosen to use measures of the geometric distance between two structures as weights on the edges of each network. For this reason, in cases where the selected discriminating score was a statistical rather than a distance measure, we estimated the TM-score as a weight for edges in these networks as defined in Equation 4.8, derived from the RMSD and the lengths of the aligned domains.

In the case of ESA the discriminating score is in the form of a distance rather than a similarity. To convert these distances into edge weights we calculated the inverse of the elastic metric.

Edges between folds are annotated with the weight of the strongest similarity between any two of their representative domains. As such, relationships between folds are characterised by their most similar domains. This is motivated by the belief that it is strong similarities which represent significance, rather than a multiplicity of weaker connections.

4.5.4.2 Consensus network

Consensus networks can be generated by taking the intersection of the edges in each of the networks considered. Consensus edge weights are generated by combining the weights attached to each of the edges within their own network. Weights corresponding to an individual alignment

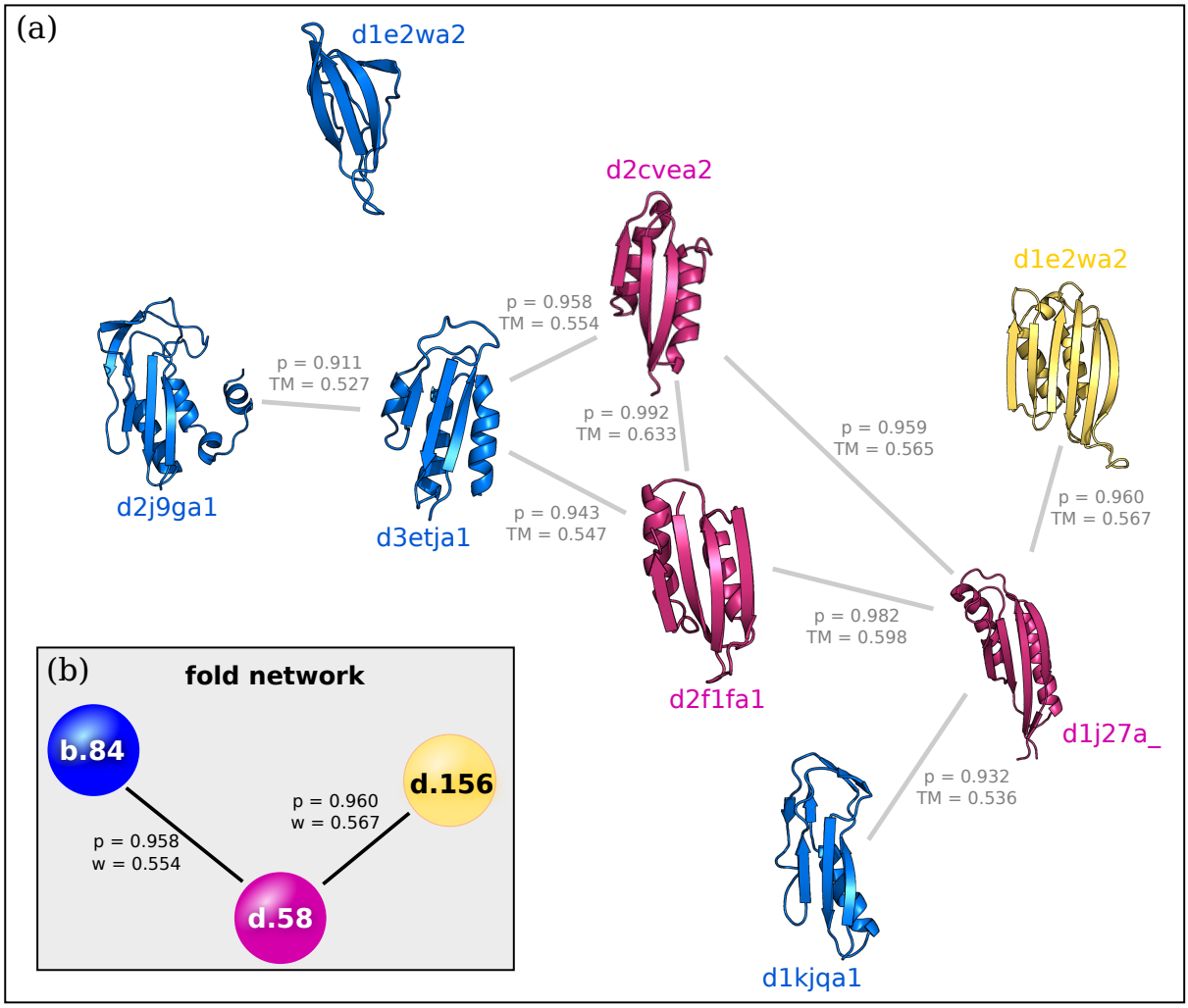


Figure 4.9: Collapsing domain comparisons to a fold network. (a) Relationships between protein domains from the ASTRAL database can be attributed a TM-score and a posterior probability. Given a probability threshold (here 0.9) these relationships can define a domain network. Domains are coloured according to their SCOP fold. (b) This domain network is collapsed to a fold network by representing all domains from the same fold as a single node. Edges are placed between two folds if there is an edge between any of their representative domains. The edges are weighted using the TM-score of the most similar pair of representative domains.

method are first centred by dividing them by their mean. Consensus weights can then be calculated by averaging the respective normalised weights.

4.5.5 Network analysis

We annotated the nodes in the networks with an estimate of fold age, as calculated in Chapter 2 of this thesis, and looked at several properties of the nodes and edges in these networks. We considered the age difference on each edge in the networks, calculated as the absolute difference between the ages of the endpoint nodes. We calculated degree, betweenness and closeness centralities as defined in Section 4.4. Betweenness and closeness centralities were calculated for nodes in different connected components separately. Degree centralities were calculated for each fold, regardless of its connectivity in the network. Central and peripheral sets of folds were identified as follows:

1. Central folds were the top 30% of nodes ranked by their centrality measures.

2. Peripheral folds were assigned differently depending on the centrality measure:

    (a) As defined by degree they were connected by at most one bridge in their network.

    (b) As defined by closeness they were the bottom 30% of nodes ranked by this centrality.

    (c) As defined by betweenness they were nodes with a betweenness value of zero.

Peripheral folds were defined differently according to the various measures because of the skewed distributions of both degree and betweenness, with a large portion of folds being assigned a betweenness or degree close to zero. Using each of these three distinctions between central and peripheral folds we compared the age distributions of these two node types using the Mann-Whitney U test [Mann and Whitney, 1947]. Furthermore, we also identified a set of pivotal nodes in the networks, where folds were found to be highly central according to all three measures and in all networks.

4.6 Results

4.6.1 Structural alignment

For each structural alignment method 8,394,753 pairwise comparisons between different protein domains were computed. As described in the Methods these scores were subject to analyses, first using the AUC of a ROC curve to determine the most appropriate discriminating score for each method, and then to a Bayesian analysis to identify the posterior probability associated with each score.

4.6.1.1 Discriminating scores

In order to identify each score's ability to discriminate between folds we calculated the area under the ROC curve. A summary table for these scores and their corresponding AUCs is shown in Table 4.2.

MAMMOTH		**FATCAT**		**TM-align**		**ESA**	
Z-score	0.93	FATCAT-score	0.85	TM-score	0.98	elastic metric	0.72
PSI	0.90	RMSD	0.57				
TM-score	0.92	TM-score	0.91				
p-value	0.91	p-value	0.94				

Table 4.2: Table of AUCs per score. For each score the area under the ROC curve is calculated as a measure of that score's ability to identify SCOP fold siblings.

In general, where it is calculated, the TM-score is a successful discriminator, with AUC scores over 0.9. Statistical measures and their p-values, which have been optimised against the SCOP classification, are also highly effective. Despite its popularity as a measure of similarity between protein structures, RMSD is relatively ineffective in identifying fold level siblings. The elastic metric is, as expected, not as efficient in identifying fold relationships as scores from the other methods, which have been optimised for this purpose. However, it is markedly more effective than the RMSD.

We thus chose the discriminating scores to be the Z-score for MAMMOTH, p-value for FATCAT, TM-score for TM-align and the elastic metric for ESA.

4.6.1.2 Posterior probabilities

Using the Bayesian analysis outlined in the Methods we calculated appropriate cutoffs for each score based on the posterior probability of a pair of domains being SCOP fold siblings given a particular score. The FATCAT team recommend using a more stringent p-value for alignments involving an all-α domain [Ye and Godzik, 2003]. As a result we calculated separate probability values for FATCAT alignment scores involving all-α domains and those involving any other class. We found that without this consideration the FATCAT network was dominated by bridges between all-α folds. While the class dynamics did vary in the other alignment methods, there was no indication that any of these methods generated unfairly skewed landscapes. The cutoffs we found using this method are summarised in Table 4.3 for probabilities ranging from 0.5 to 0.9.

	MAMMOTH	**FATCAT**		**TM-align**	**ESA**
		all-α	**other**		
Posterior Probability	Z-score	$-\ln(p)$	$-\ln(p)$	TM-score	Elastic metric
0.5	2.97	10.58	4.24	0.44	0.814
0.6	3.35	11.64	4.77	0.45	0.796
0.7	3.77	12.80	5.42	0.47	0.776
0.8	4.28	14.26	6.34	0.50	0.751
0.9	5.02	16.39	7.99	0.53	0.715

Table 4.3: Table of threshold scores for each method and their corresponding probability. Thresholds indicate the value of a score where the posterior probability of two domains sharing a fold is either 0.5, 0.6, 0.7, 0.8 or 0.9. Separate thresholds were found for FATCAT comparisons involving an all-α domain and those involving only domains from other classes. The elastic metric is a distance rather than a similarity and in this case thresholds represent an upper limit below which alignments are accepted. All other scores represent lower limits.

4.6.2 Structural bridges and fold space landscapes

For each method the structural bridges at each probability threshold collectively determine a landscape for the global organisation of fold space. Figure 4.10 shows visualisations of the four networks of structure space, as the thresholds determined by the posterior probabilities transition from 0.5 to 0.9, representing increasingly stringent thresholds for determining similarity.

Some general network statistics relating to each construction can be found in Figure 4.11. In general, as the probability threshold increases, network sizes shrink, the density of edges decreases, and the average shortest path length between any two folds increases. The number of edges in the static landscapes vary from 5,571 in the MAMMOTH network at a 0.5 threshold to 250 in the ESA network at a threshold of 0.9. In none of these networks are all 631 folds connected by structural bridges. However, the number of connected folds remains a sizeable portion, with at least 450 connected nodes in networks constructed at a probability threshold of 0.5.

There are however significant differences between the alignment algorithms, as well as similarities. In general, in the networks for ESA MAMMOTH and FATCAT, about 50% of the bridges identified are only identified by that method. For the TM-align networks, this proportion is lower at about 30%. Importantly, this proportion does not seem to be affected by increasing the stringency of the cutoff, but remains relatively constant even as the posterior probability approaches 0.9 (see Figure 4.12). In other words, networks constructed using different alignments remain the same distance apart regardless of similarity threshold. In isolation, a proportion of edges in these networks will always be an artefact of the alignment method.

Despite these differences, several properties remain conserved across every landscape. In general, and in concert with previous observations, the networks partition fold space into the four secondary structure classes. In the majority of these landscapes the all-α and α/β folds form densely packed clusters, whereas folds with anti-parallel β sheets, belonging to the all-β and $\alpha + \beta$ classes are more dissipated throughout the space. Applying a community detection

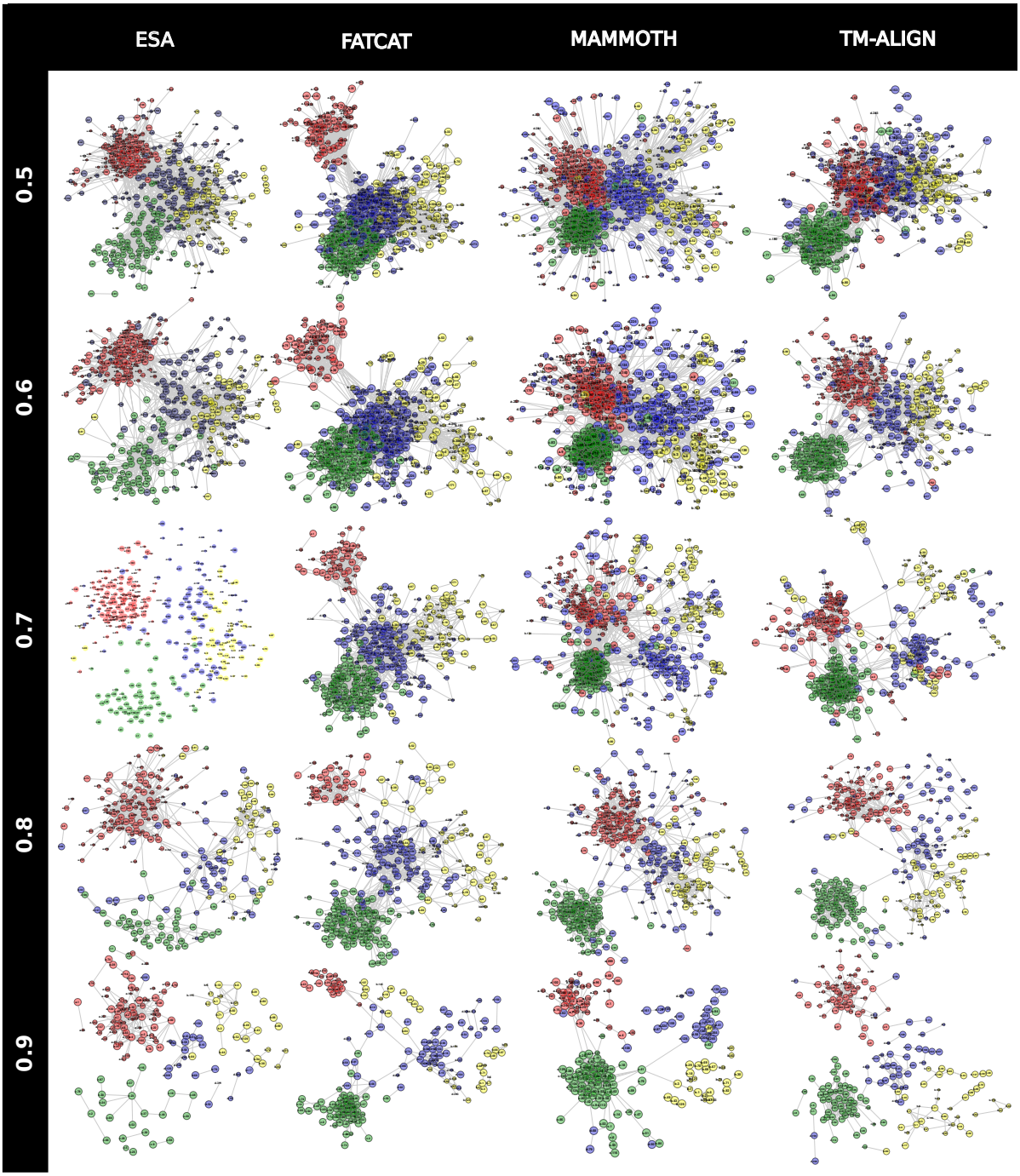


Figure 4.10: Fold space networks for the four structural alignment methods at increasingly stringent probability cutoffs, ranging from 0.5 to 0.9. Nodes are coloured by their SCOP class (all-α: red, all-β: yellow, α/β: green, $\alpha+\beta$: blue) and have a size proportional to their evolutionary age estimate (an age of 0.0 (younger folds) corresponds to a smaller node size, while folds with an age of 1.0 (ancient folds) are larger). Only connected nodes are shown in these representations. Visualisations were produced using Cytoscape [Shannon *et al.*, 2003]

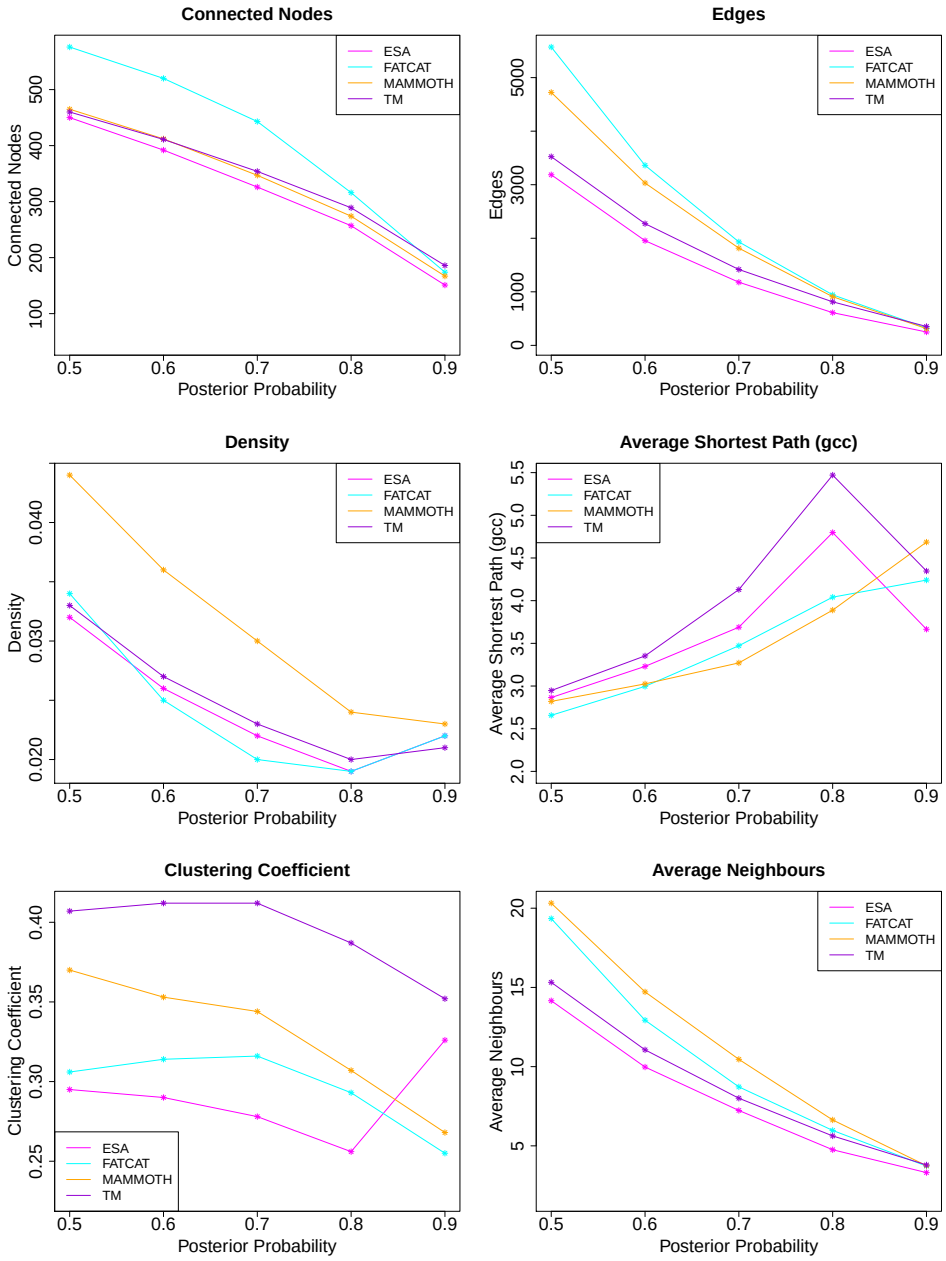


Figure 4.11: Network statistics for the four networks at different probability thresholds. The statistics are calculated using the tnet package in R. The number of nodes refers to the number of folds connected by at least one bridge in each network. The number of edges is the number of bridges determined as significant in each network. The density is the proportion of all possible edges between this set of connected nodes which are bridges. The shortest path between any two nodes is calculated as the smallest sum of weights along bridges forming a path between those folds. The average shortest path for a network is the average of these path lengths across all pairs of nodes in its largest connected component. The clustering coefficient is the proportion of triplets of connected folds which have bridges connecting all three. Average neighbours is the average number of bridges connecting a fold in the network.

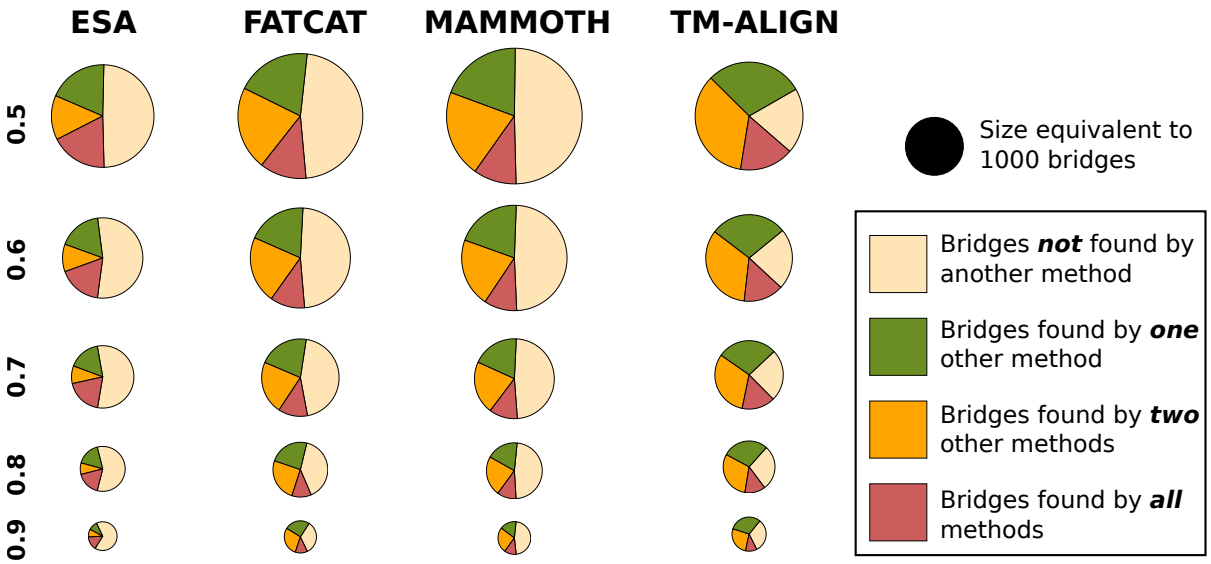


Figure 4.12: Disagreement between structural alignments. Pie charts representing the portion of structural bridges in each method which are also found in other networks at the same probability threshold. Pie charts are scaled according to the number of bridges in the network and are divided into segments depending on how much agreement there is as to each bridge. Segment portions remain relatively constant as the probability threshold becomes more stringent.

algorithm to these landscapes identifies five predominant communities with a high density of structural bridges within each group, but sparsely connected externally. These communities can be generally defined as all-α, α/β, $\alpha + \beta$, all-β sandwiches and all-β barrels by the prevailing population of folds within these clusters. Figure 4.13 shows the communities in the consensus network at a probability threshold of 0.5. The all-β sandwiches and barrels tend to remain detached from each other even at the least stringent probability threshold of 0.5 and are often closer in the landscapes to the $\alpha + \beta$ community than they are to each other.

4.6.3 Structural bridges by averaged similarities

The structural bridges we have presented in this chapter define relationships between folds by the similarities of their closest domains. This is due to the belief that different evolutionary and structural pressures will apply to the different domains within a fold. As a result we postulated that a strong similarity between two representative structures would be more indicative of an

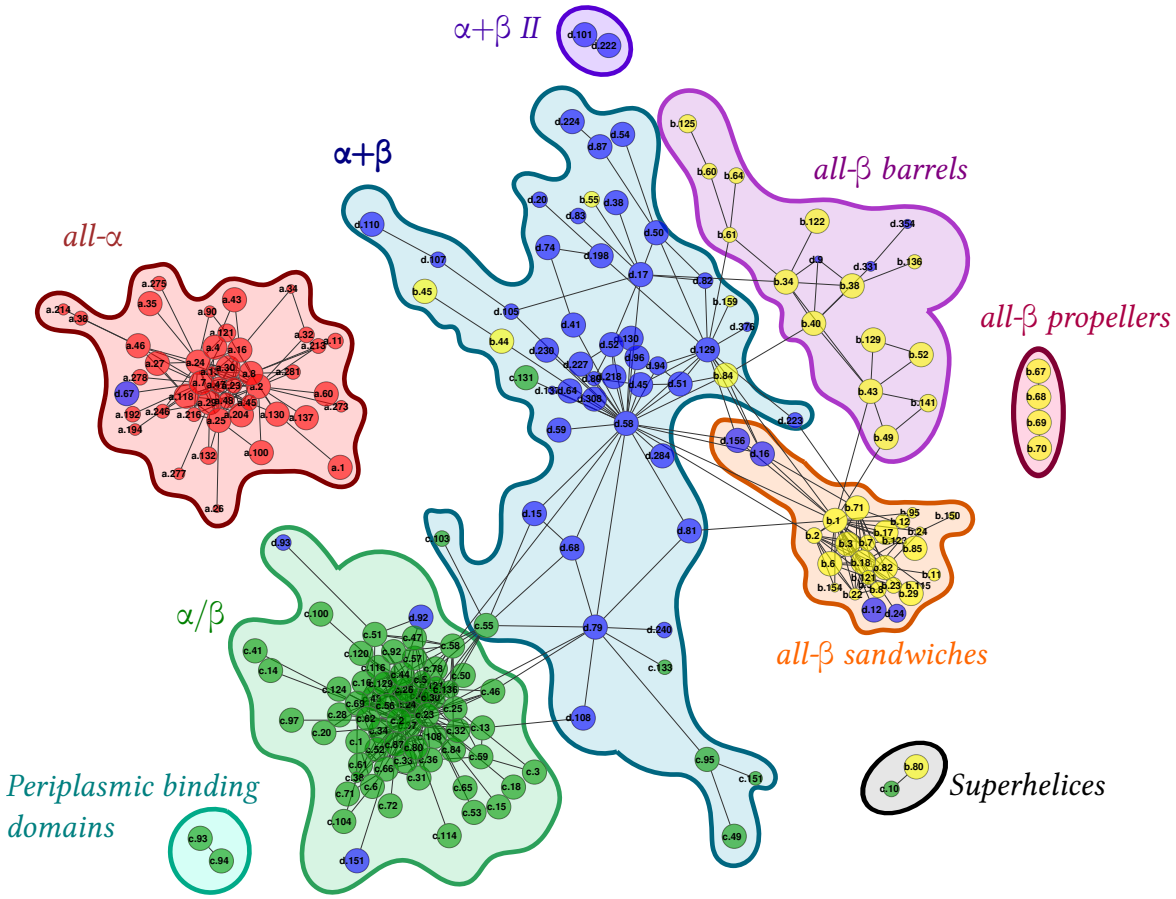


Figure 4.13: The community structure of the consensus network at a probability threshold of 0.5 as calculated using the Louvain method for weighted networks. Nodes have a size proportional to their age and are coloured by their SCOP class (all-α: red, all-β: yellow, α/β: green, $\alpha+\beta$: blue). Clusters are circled and manually assigned labels are included based on each community's fold content.

evolutionary relationship than the collective behaviour of all domains within each fold.

Nevertheless we did examine how the network spaces would vary if edges were defined according to the average posterior probability of all possible scores between the different domains. Networks of both mean and median scores were examined and were found to be very similar. For simplicity, the results presented here are for the median networks.

As expected these networks were far smaller than those constructed using the maximum similarities. At a threshold of 0.5 the number of edges found in the median networks ranged from 218 in the ESA network to 584 in the TM-Align network. Significantly, there was little consensus between the four methods. The maximum similarity networks we present in the rest of this chapter typically contained 50% edges identified by a single method alone and 15% edges found by all four methods. On the other hand the averaged networks at a threshold of 0.5 consisted of up to 75% edges unsupported by any other method with only about 2% edges identified by all four algorithms.

We believe this effect is due in part to the nature of the structure alignment algorithms. An insignificant alignment will still generate a score under all of the alignment algorithms but the quantitative interpretation of this score is debatable. The reason for this is that it is derived from a meaningless superposition of the two structures. Allowing insignificant alignments between any two representative structures of different folds to contribute to their pairwise score has the potential to skew this score towards a meaningless value.

Another argument is that if relationships exist between different folds they will likely span long evolutionary time periods. As the two folds diverge it is unlikely that the similarities between different families within the two folds will remain constant. We believe that if significant similarities exist between just a few families representing these folds this is more indicative of their relationship than the average divergence between all representative families. Nevertheless it is important to indicate that the following results are dependent on this assumption.

4.6.4 Fold ages and the fold networks

This section shows the result of a more in depth analysis of these networks. In particular the fold ages calculated in Chapter 2 are used as an annotation to each network and this extra dimension is explored in terms of conserved properties across each landscape. Firstly age differences across each bridge in the network are defined as the age difference between folds at the endpoint of each bridge. These age differences are compared to the differences between two unconnected folds. Prominent nodes are identified using network centrality analysis and the relationship between their centrality and their age estimate is examined.

4.6.4.1 Age differences on bridges

Edges in these networks represent, not simply the phenomenon of structural similarity between proteins, but structural bridges between folds: distinct and separate structural, and possibly evolutionary, units. We annotated each node in the networks with a measure of its fold age, as calculated in Chapter 2. These ages estimate the emergence, on a tree of sequenced life, of a fold's structural ancestor. Each age estimate falls between zero and one, where an age of one represents an ancestral fold emerging at the root of the tree, and an age of zero signifies an ancestor at its leaves. We were thus able to consider the difference in age attributed to each of the bridges in our networks. As edges were undirected we considered the absolute difference in age of the endpoint folds to each edge (bridge). We investigated the distribution of age differences, comparing those of structural bridges to a background distribution of random pairs of unconnected folds. As described above, a large number of these bridges were identified by just a single alignment method so we examined separately the distribution of age differences on edges found on one, two, three and four networks to those found on none. Figure 4.14 shows a boxplot of these age differences on the set of networks built at a probability cutoff of 0.6. Distributions for the other thresholds are similar and show the same trend. Not only are the pairings of unconnected folds associated with higher age differences than bridged folds, but this difference increases with the number of methods supporting an edge. Bridges identified by at least two

different methods had a median age difference of zero as opposed to the 0.25 of unconnected folds. In other words, folds which share structural features tend to have a similar estimated

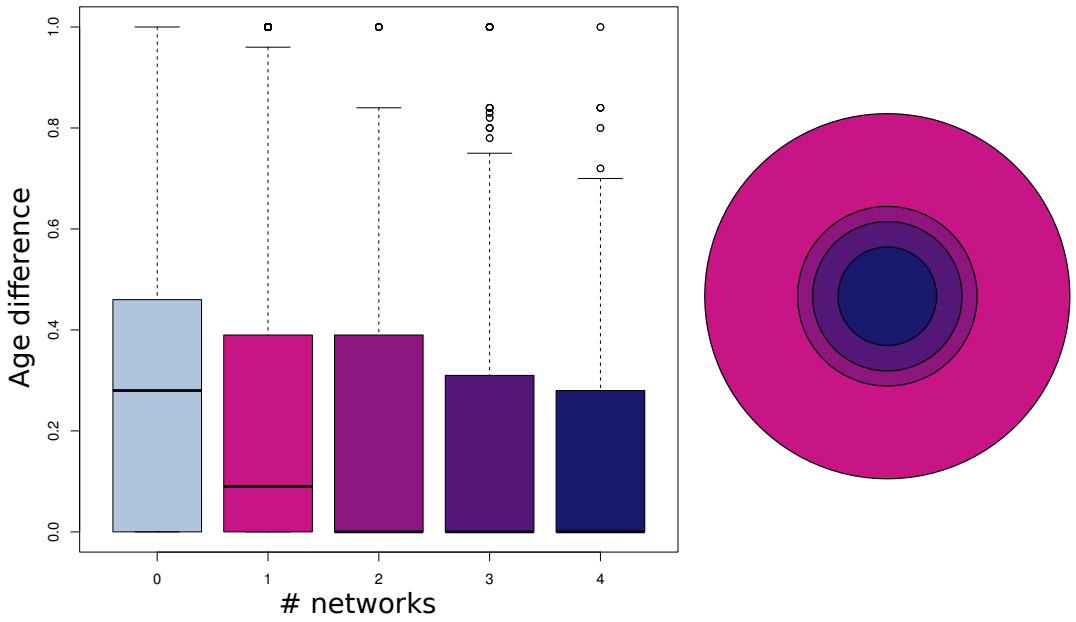


Figure 4.14: Edges and their age differences in our fold networks at a probability cutoff of 0.6. The median age difference for pairs of unconnected folds is 0.25. This falls to 0 for pairs of folds connected in at least two networks. Concentric circles with an area proportional to the number of edges in each set. There are 9336 edges appearing in just one network, 2250 in just two, 1566 in three, and 678 in all four.

age, with fewer significant relationships between folds with disparate evolutionary histories. In Chapter 3 of this report we found that several general structural attributes were correlated with the age estimates of both superfamilies and folds. The observation here that structurally similar folds tend to be closer in age than dissimilar folds offers an interesting perspective that more specific structural attributes may be connected to age as well. In particular, these structural features appear to have been explored by different folds at similar evolutionary timescales.

4.6.4.2 Node centrality

The prominence of each fold within these landscape was calculated using three different centrality measures: degree, closeness and betweenness. For each measure, two populations were

identified: central folds and peripheral folds, and the age distributions of these two populations were compared. Figure 4.15 shows the mean fold age for both central and peripheral nodes according to the three measures in each of the four networks as their posterior probabilities vary from 0.5 to 0.8. Consensus networks and those at a threshold of 0.9 were omitted for simplicity as they contain more than one large connected component. However, similar results were seen when folds within each of the connected components of these networks were considered. In all these cases, fold ages of nodes with strong centralities have a higher distribution than those of nodes in less central positions. In fact by all three of these measures central nodes tend to be older than more peripheral nodes in every network ($p \leq 7.23 \times 10^{-3}$).

While the different measures all attempt to classify each fold in terms of its prominence to the architecture of the network, the three measures used here all have different interpretations. To illustrate how the measures partition the spaces Figure 4.16 visualises the centralities on the TM-align network at a probability threshold of 0.6. For reference the original network is shown with the folds labelled and coloured according to their SCOP classification. Three further versions of the same network are shown with highly central nodes coloured yellow and peripheral nodes in dark blue. Highly central folds are seen from all four classes and, overall, the measures do not appear to be biased towards a certain class. Percentile plots are also shown, illustrating the distribution of ages for the central and peripheral folds identified for this network. Further exploration of each of the different measures and their interpretation in the context of a fold space network is given below.

Degree centrality Conceptually, degree centrality is the easiest of these measures to appreciate. A fold of high degree is one with several structural neighbours. It contains popular structural features found across several separate protein evolutionary and structural units. Degree centralities were calculated for all folds, regardless of their connectivity in the networks. That is, degrees were calculated as outlined above, with a degree of zero assigned to disconnected nodes. The distributions of degrees were similar across all the networks, and were skewed, with the

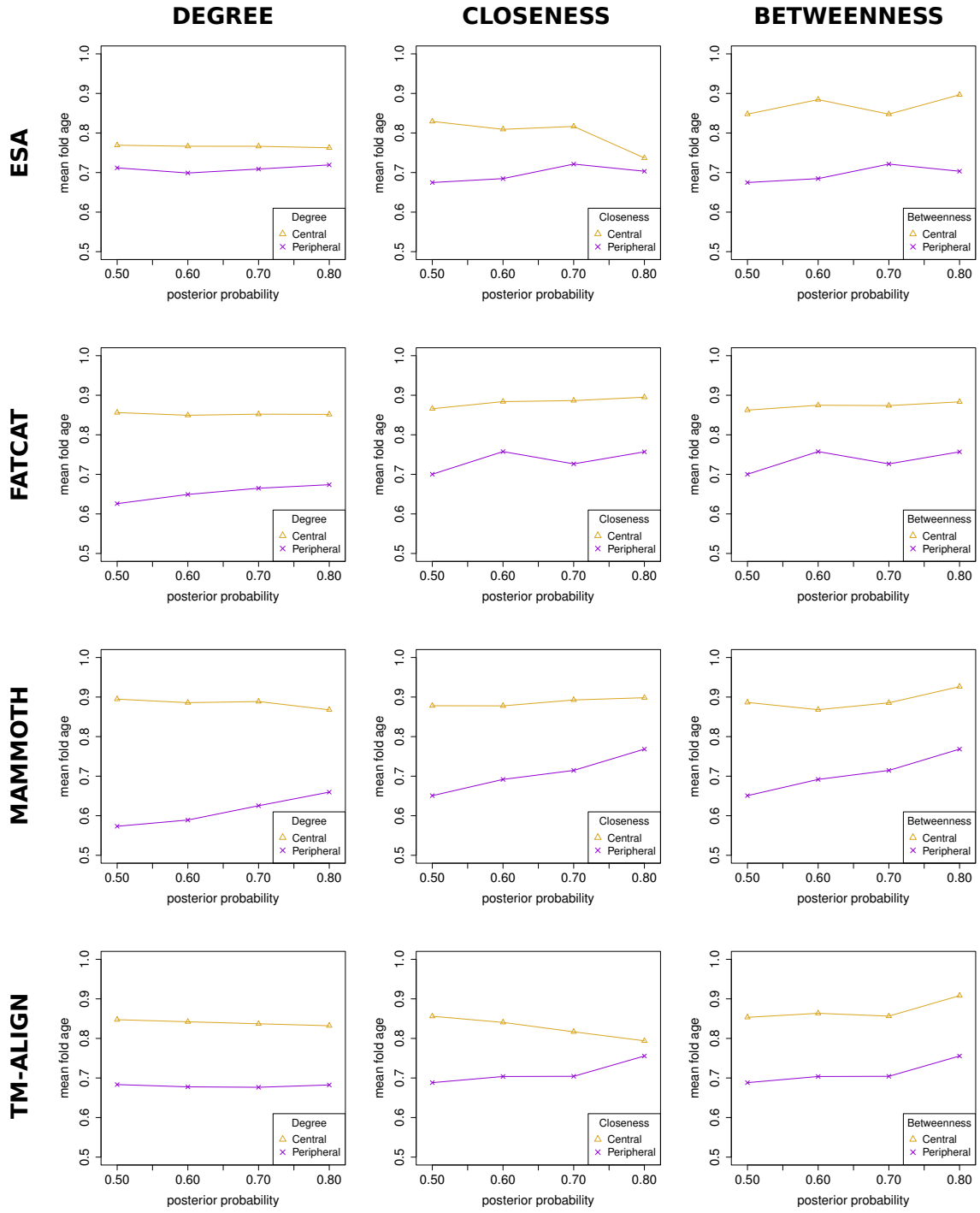


Figure 4.15: Plots showing the mean fold age of the central and peripheral folds within each network. These populations were identified for each network as in the Methods. In each case the central nodes are found to be older than the peripheral nodes (significant at the 0.01 level with the Mann-Whitney U test). For simplicity, only points corresponding to networks with one giant component containing all SCOP classes, are included. This is due to the need to consider connected components separately for closeness and betweenness centralities. However, the same signal is seen for the different components of the remaining networks: separate networks at a threshold of 0.9 and consensus networks at all thresholds.

majority of nodes having a degree close to zero and a small minority of nodes with high degree. We considered nodes of low degree to be those with either zero or one structural neighbour in the network. Nodes of high degree were the top 30% of nodes when ordered by their degree within the network. Apart from in the ESA network it was with degree centrality that the age differences between central and peripheral nodes was most evident. The reason for this diminished signal in ESA appeared to be the reduced connectivity between the α/β folds compared to the other classes in this network. However the differences in ages was still significant in this network.

Closeness centrality Probably the least intuitive measure of node centrality is the closeness, which is the inverse of the average shortest path lengths for each node. However it is possible to consider each path length as a series of small, detectable structural changes. In this context the closeness centrality of a fold captures its ability to cover the structural diversity of fold space (or, at least, the connected component it appears in). As laid out in the methods, we calculated closeness centralities for nodes in each connected component separately and compared only central and peripheral folds of the same component. Figure 4.15 shows the mean fold ages of central and peripheral nodes in the largest connected component of each network. The distributions of closeness centralities in these networks were more symmetric than the other centrality measures, with roughly equal numbers of nodes assigned low and high values. We thus took the top and bottom 30% of nodes when ordered by their closeness values, to be those with high and low centralities respectively.

Betweenness centrality High betweenness centralities, in the context of social networks, control flow within the network. They appear more often on shortest paths between other nodes. The concept of flow is non-transferable to these structure similarity networks however. A more appropriate understanding is that, by lying more often on shortest paths, nodes with high betweenness will occupy significant positions between otherwise isolated groups of folds. In other words, these nodes form crucial bridges between highly populated areas within fold space. Sim-

ilarly to the closeness values, we calculated betweenness centralities for nodes in each connected component separately. Like degrees, distributions of betweenness centralities across these networks were highly skewed. Large numbers of nodes had a betweenness of zero but a few folds were calculated to have centralities in the 10,000s. Different networks tended to share similar distributions for their betweenness centralities. We therefore uniformly took nodes with a betweenness of 0 to be those with low centrality and the top 30% of folds ordered by betweenness to have high centrality.

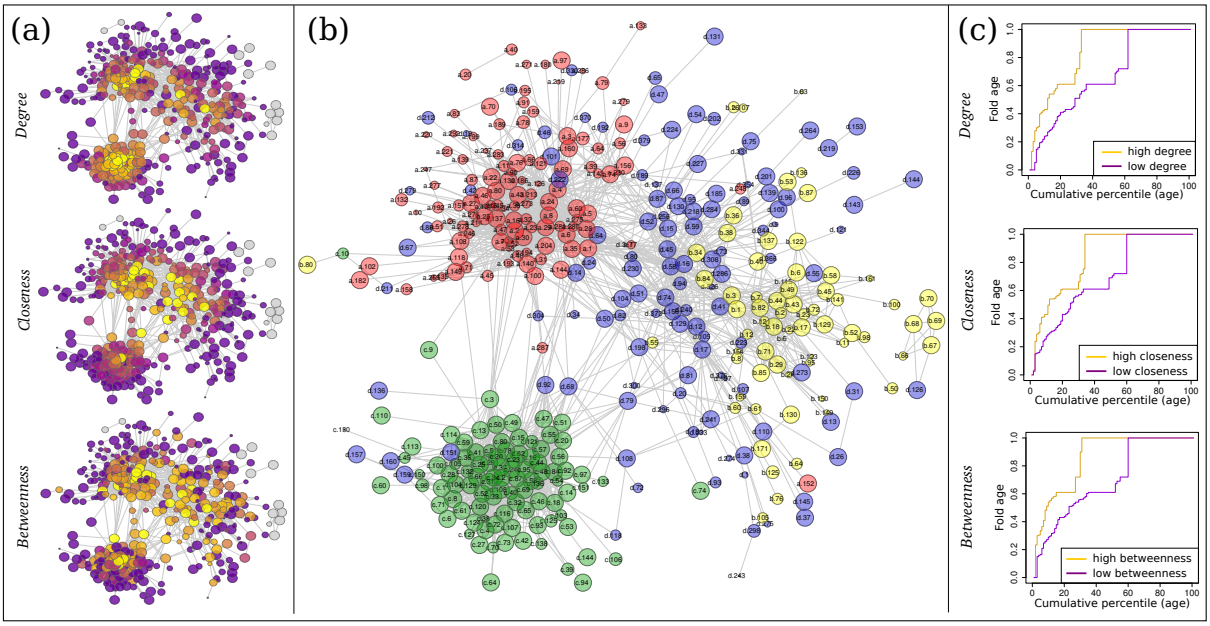


Figure 4.16: Node centralities in the TM-align network at a probability threshold of 0.6. (a) For ease of comparison heatmaps of centralities in the main connected component of the network are displayed with the 40% most peripheral nodes according to the measure in indigo and the top 5% of nodes in yellow. The remaining nodes are scaled by centrality from purple to orange. (b) For reference, nodes are coloured by their SCOP class and annotated by their fold name. Fold nodes are given a size proportional to their age estimate. (c) Cumulative percentile plots for the fold ages of central and peripheral nodes. The sets of folds with high and low centralities are identified as described in the Methods.

Using these measures central nodes represent folds containing popular structural features, those connecting together disparate groups of folds, or those which can access several highly dissimilar structures by only a few small structural changes. In all these cases, central nodes

tend to have a higher age distribution than their less central counterparts.

4.6.5 Pivotal nodes

The above network centrality analysis exposed certain *pivotal* folds, which were calculated as central in all networks, including the consensus networks. Figure 4.17 shows the centralities of nodes in the consensus network at a threshold of 0.5. This network consists of two prominent

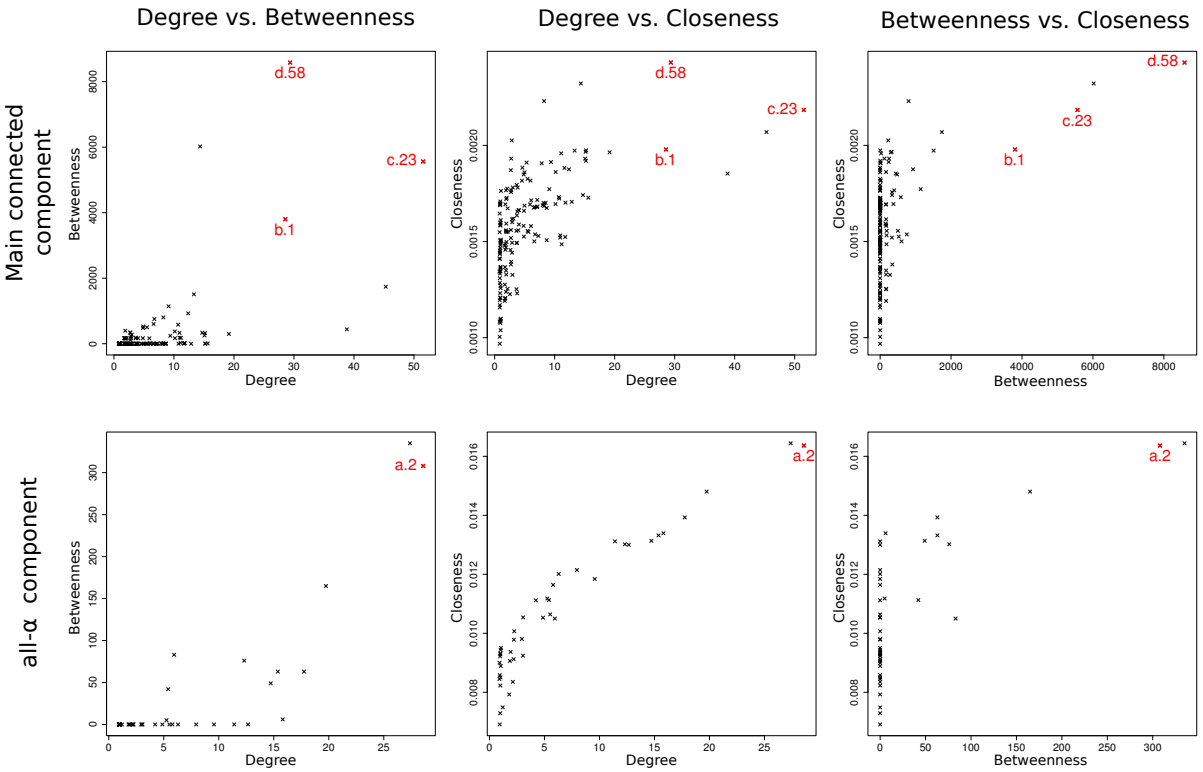


Figure 4.17: Pivotal nodes in the consensus network. Scatter plots comparing centrality measures in the two largest connected components of the network. Pivotal nodes, which were manually identified as nodes central by all three measures and across all networks, are highlighted and labelled in red.

connected components: one consisting of a majority of all-α folds, the other representing the three other classes. These connected components are considered separately as the values of the betweenness and closeness centralities are non-comparable between different components. Pivotal nodes are highlighted in red. Nodes were identified as pivotal if they appeared as highly

central in all networks. Therefore, some of the nodes in the top right hand corner of the plots in Figure 4.17 are not considered pivotal as they were not identified in the other networks. Moreover, we felt it was valuable to consider examples in depth from across all four SCOP classes. It is therefore important to bear in mind that the following is not an exhaustive list, and is the result of a manual analysis of nodes and their centralities. In the following sections we examine four pivotal folds: the Ferrodoxin-like fold (d.58), the Flavodoxin-like fold (c.23), the Immunoglobulin-like β-sandwich fold (b.1) and the long α-hairpin (a.2).

4.6.5.1 a.2: the long α-hairpin

The long α-hairpin fold contains 20 superfamilies and consists of two long anti-parallel α-helices connected by a short loop segment with a left-handed twist. Figure 4.18 shows a cartoon representation of a domain classified as this fold and the subnetwork of the consensus network (at a threshold of 0.5) centred on a.2. In this fold network a.2 is connected to the majority of

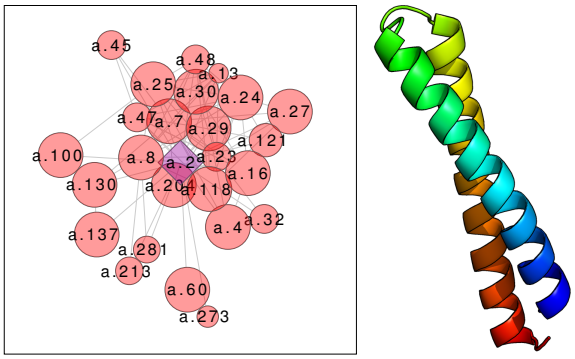


Figure 4.18: Pivotal fold a.2: the long α-hairpin. a.2 is shown in the consensus network as a purple diamond and is central to the all-α component of the network, connected to 23 other α-helical folds but none from the other classes.

folds in the all-α densely clustered group. It is not connected to any fold from another class. Topologically this fold is relatively simple, however the majority of the alignments resulting in its neighbourhood match both the helical hairpin and also the relative orientations of the two helices brought about by the left-handed twist in a.2.

4.6.5.2 b.1: the Immunoglobulin-like β-sandwich

The Immunoglobulin-like β-sandwich fold contains 28 different superfamilies, all similar topologically to the anti-parallel β-sandwich which comprises the core of the antibody, or Immunoglobulin (Ig), domain. This fold consists of two β-sheets with at least one Greek key motif connecting strands across the two sheets. Figure 4.19 shows a cartoon representation of a domain classified as b.1, a topological diagram indicating the strand connectivity between the two β-sheets, and the subnetwork of the consensus network at a probability threshold of 0.5 and centred at b.1. In the fold network b.1 is connected to a large number of other all-β folds. In the majority

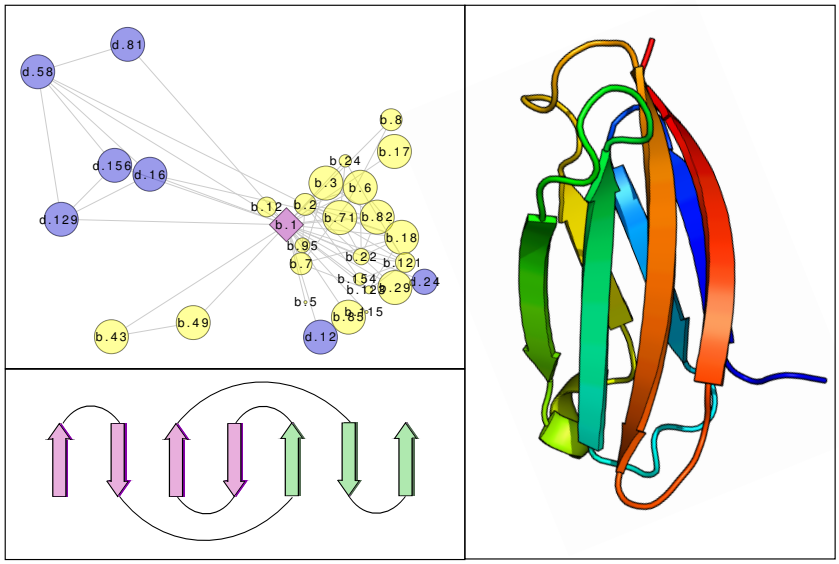


Figure 4.19: Pivotal fold b.1: the Immunoglobulin-like β-sandwich. b.1 is shown in the consensus network as a purple diamond and connects sandwich folds, including 23 other allβ folds and 7 $\alpha + \beta$ folds. The topological diagram shows strands within the two sheets (purple and green) of the Ig-like fold, and their connectivity.

of these cases neighbouring folds are β-sandwiches containing a Greek key motif. However, b.1 also aligns with two Greek key barrel folds: b.43 (Reductase/isomerase/elongation factor common domain) and b.49 (Domain of alpha and beta subunits of F1 ATP synthase-like). The Ig-like fold also neighbours several jelly roll topologies: b.18 (Galactose-binding domain-like), b.22 (TNF-like), b.85 (beta-clip) and b.123 (Hypothetical protein TM1070). b.1 is also con-

nected to seven $\alpha + \beta$ folds across different regions in the network: d.12 (Ribosomal proteins S24e, L23 and L15e), d.16 (FAD-linked reductases, C-terminal domain), d.24 (Pili subunits), d.58, d.81 (FwdE/GAPDH domain-like), d.129 (TBP-like) and d.156 (S-adenosylmethionine decarboxylase). These are all folds consisting of a β-sheet packed against a helical layer. The connectivities in the sheet are similar to that of the Ig-like fold but the alignment of the helical layer to the second sheet of b.1 varies substantially.

4.6.5.3 c.23: the Flavodoxin-like fold

The Flavodoxin-like fold contains 15 different SCOP superfamilies. It is made up of three layers: a 5-stranded parallel β-sheet packed against a helical layer on both sides. The strand order in the sheet is 21345, making it a doubly-wound β-sheet. In terms of secondary structure, the chain alternates between β-strand units and α-helical units. Thus the helices all run anti-parallel to the strand direction, and parallel to each other. The two helices closest to the N terminus and C terminus respectively form one of the helical layers. The other is formed of the three interior helices. Figure 4.20 shows a cartoon representation of a domain classified as c.23 under SCOP, a simple topology string representation of its architecture, as well as the subnetwork of the consensus network at a probability threshold of 0.5 and centred on c.23. In the fold network c.23 is connected to the majority of the other α/β folds which make up the densely packed group of this SCOP class. It has high centrality by all three measures in the MAMMOTH, FATCAT and TM-align networks. It is comparatively less central, as are all α/β folds, in the ESA network. However, it is still regarded as one of the most central α/β folds. The majority of these connected folds also consist of three layers, with a parallel doubly wound β-sheet as the second layer. The β-α-β motifs which connect consecutive strands remain largely conserved across these structural siblings, with more variation in the longer linking regions between the non-hydrogen bonded strands. c.23 is also connected to two $\alpha + \beta$ folds: d.79 (Bacillus chorismate mutase-like) and d.92 (Zincin-like). Both of these folds are formed of a mixed β-sheet layer and a helical layer, and contain β-α-β motifs which align well to domains

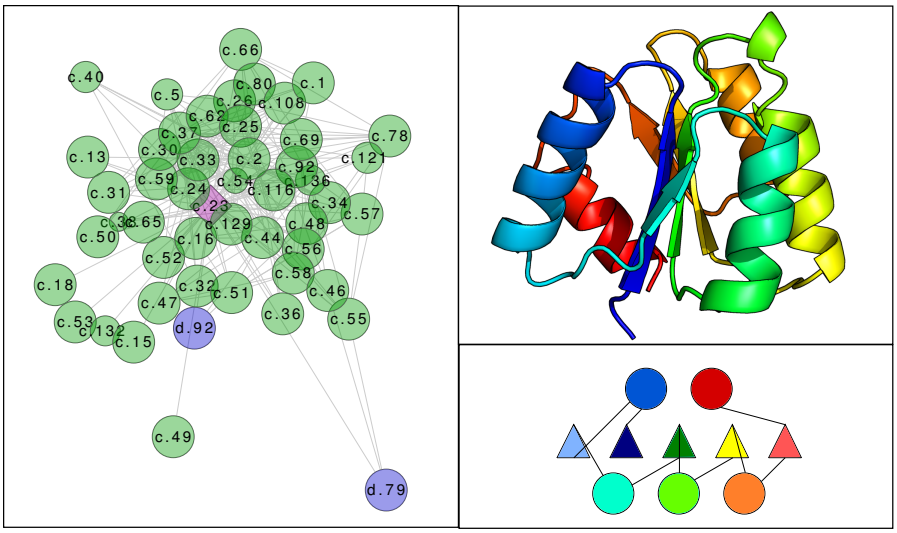


Figure 4.20: Pivotal fold c.23: the Flavodoxin-like fold. In the consensus network at 0.5 threshold, c.23 is connected to 44 other α/β folds and two $\alpha+\beta$ folds. The fold typically resembles a three layer doubly wound β-sheet constructed of β-α-β units, one going from right to left along the sheet (incorporating the blue helix) and two in the opposite direction (green and orange helices).

in c.23. They are also both connected to other α/β folds.

4.6.5.4 d.58: the Ferrodoxin-like fold

The Ferrodoxin-like fold is one of the most populated SCOP folds, containing 40 superfamilies and over 1000 PBD entries classified as ferredoxin-like under SCOP. Structurally, it consists of a sandwich: an anti-parallel β-sheet packed against two α-helices. At its core, it contains two repeats of an β-α-β switch motif, where two strands in the sheet are connected via an α-helix running anti-parallel to the strand direction. These two motifs are interlocked together in the fold, involving alternate strands in the β-sheet, and one running from left to right and the other from right to left. Figure 4.21 shows a cartoon representation of a domain from this fold and also the subnetwork of the consensus network including d.58 and its neighbours (folds directly connected to d.58). In the fold network d.58 is connected to 25 other $\alpha+\beta$ folds, several of which also contain two layers and are made up of β-α-β switches. It is also significantly aligned to two α/β folds: c.58 (Amino acid dehydrogenase-like, N-terminal domain) and c.131

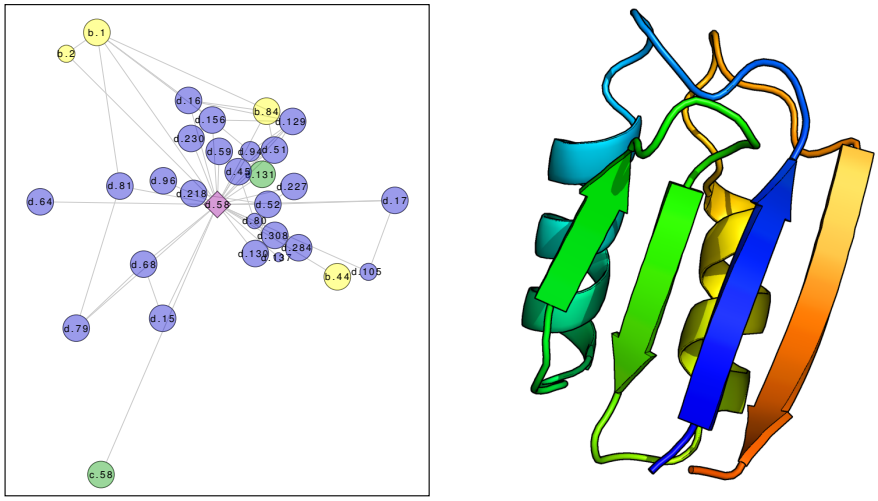


Figure 4.21: Pivotal fold d.58: the Ferrodoxin-like fold. In the consensus network, d.58 is connected to 25 $\alpha + \beta$, 4 all-β sandwich folds and 2 α/β folds. Topologically it consists of two interlocking β-α-β switches.

(Peptidyl-tRNA hydrolase II), both of which contain three layers, but contain β-α-β units at their core. Moreover, d.58 is further connected to four all-β sandwich domains. In these cases, the two α-helices in the helical layer of b.58 are aligned to the second β-sheet in the sandwich domains.

In all networks apart from the MAMMOTH network d.58 is assigned the highest betweenness centrality (in the MAMMOTH network it is the second highest after a.2), mostly due to its strong connections with b.1 (Immunoglobulin-like beta-sandwich) and c.58: both of which are central in the groups of all-β and α/β folds respectively.

4.6.6 Structural siblings

To conclude this chapter we examine two pairs of folds which were identified as structural siblings by all four alignment methods. These case studies provide specific examples of the power of this methodology in examining structural links across fold space.

4.6.6.1 Cystatin-like (d.17) and PH domain-like barrel (b.55) folds

Notable for its absence in the pivotal folds identified in Section 4.6.5 is the β-meander: a common anti-parallel β-sheet motif made up of consecutive strands. Central to the alignment of d.17 and b.55 is their composition of a β-meander topology. The bridge between d.17 and b.55 is recognised by all four alignment methods with posterior probabilities ranging from 0.705 to 0.883. The Cystatin-like fold (d.17) consists of a single α-helix packed against a 4-stranded coiled β-sheet meander. The PH domain-like barrel (b.55) is a partially open barrel of 6 strands, capped by an α-helix. Figure 4.22 shows cartoon representations of domains from both of these folds and their alignment using TM-align. This alignment illustrates a link between two

Figure 4.22: Structural siblings: d.17 and b.55. A representative domain of d.17 is shown in purple and b.55 is shown in orange. Their structural alignment using TM-align is also shown.

disparate structural features: the sheet and the barrel. d.17 has an estimated fold age of 1.0 and b.55 has an age estimate of 0.61. b.55 occurs across 386 different species, but only 10 of these are bacterial species leading to its lower age estimate. Of these bacterial genomes though it appears in genomes where d.17 has not been assigned.

4.6.6.2 Barrel-sandwich hybrid (b.84) and TBP-like (d.129) folds

Bridges between b.84 and d.129 are found in the structure networks with probabilities ranging from 0.623 in ESA alignments, to 0.859 using TM-align. The barrel-sandwich hybrid fold (b.84)

is formed from a sandwich of half-barrel shaped β-sheets. Its first two superfamilies are the single hybrid motif superfamily consisting of the biotinyl/lipoyl-carrier proteins and domains, and the rudiment single hybrid motif superfamily containing several families thought to be probable rudiment forms of the biotinyl-carrier domain. Members of this second superfamily are functionally similar to the single hybrid motif superfamily but differ structurally in that they contain helices instead of their second β-sheet. So while this domain consists of a helical layer as well as a strand layer it is still classified as an all-β fold. The TBP-like fold (d.129) is formed of two layers: an anti-parallel 5-stranded β-sheet packed against two helices. The secondary structure elements, in order of sequence, are β-α-β(4)-α. Figure 4.23 shows cartoon representations of domains from both of these folds and their alignment using TM-align. The core topologies of

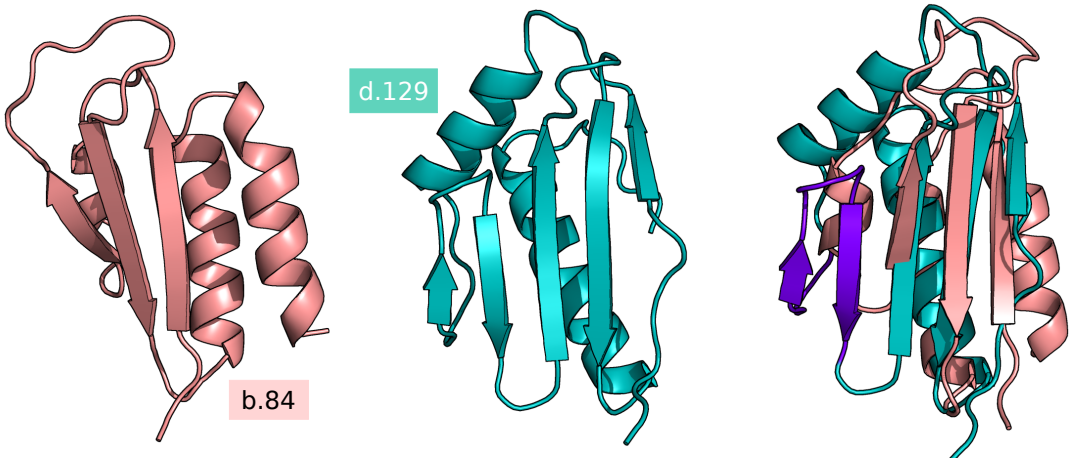


Figure 4.23: Structural siblings: b.84 and d.129. Representative domains of these folds are shown in peach (d3etja1) and blue (d1p5dx4) respectively. Their structural alignment using TM-align is also shown, with the unaligned β-meander of d.129 highlighted in purple.

these domains are very similar, sharing an β-α-β(2)-α organisation. The b.84 domain (d3etja1) has an extra C-terminal α-helix which is unaligned to the d.129 domain (d1p5dx4). There are also two extra interior strands in the sheet of the d.129 domain, highlighted in purple in Figure 4.23. These strands alter the orientation of the first α-helix in d.129. However, as the strands are connected via a meander their addition doesn't further disrupt the overall topology of the structure. This extended sheet in the d.129 domain also forms a cavity which is the active site

of the protein [Mosyak *et al.*, 2000].

The age of both of these folds is 1.0 and they both occur in every single genome of the species tree.

4.7 Conclusions

We have proposed and constructed a dynamic network representation of fold space to capture variations in its organisation resulting from different methodologies and similarity thresholds. While a vast array of different techniques have been applied to visualise the structural organisation of the global protein universe, very little has been done to ensure such landscapes are robust to differences in the alignment methodology which generates them. One of the reasons behind this is that rarely can the similarity scores produced by different methods be comparable. Here a posterior probability has been attached to similarity scores resulting in landscapes which can be compared and contrasted between different methods. We have shown that, in terms of network representations using four dissimilar methods, there are several disagreements as to where bridges between different folds in the global space lie. We also found that these disagreements cannot be overcome by simply increasing the threshold at which a structural bridge is determined for each method. It is likely therefore, that these disagreements remain artefacts of the different alignment methods used to generate the networks and that the use of consensus networks is an important part of the analysis of a global structure space.

On the other hand, a possible limitation of using the posterior probability is that it uses the SCOP database as a gold standard, and calibrates appropriate score thresholds against this scheme. Explicitly this measure estimates the probability of two structures being related when the similarity score from their alignment is above a certain threshold. But the probability of two structures being related is taken from SCOP's classification which is not guided purely by structural similarity, and is also limited by biases in annotation and available structural data.

Structural bridges could exist for a variety of reasons. It is possible they are the result of a

misannotation of fold boundaries, or indeed that fold space is wrongly assumed to be discrete. They may also be the result of convergent evolution to a particularly favourable confirmation. They could also represent the structural relic of an evolutionary transition from one fold to another. Whatever their cause, such inter-fold similarities are deserving of further study, to illuminate the overall structure and dynamic of naturally occurring fold space. Moreover, the significant number of these bridges, especially in consensus networks representing the agreement of all four methods, suggests that structural classification, while an important and useful construct, might be a misrepresention of the true nature of the protein universe.

Another feature of the structure spaces is the population of ancestral folds at highly central positions within its landscape. Using each different alignment method separately, as well as in consensus, and at different levels of significance, we examined the age distributions of central and peripheral folds. We calculated the centrality of folds as nodes in each network using three different centrality measures, each with a different interpretation of the priority of a node within the landscape. In all these cases, key locations within the landscapes tend to be occupied by older folds than those at the periphery of the space. A previous study identified a functionally diverse core within fold space [Osadchy and Kolodny, 2011]. This core was predominantly characterised by α/β folds, which have also been identified as predominantly ancient [Winstanley *et al.*, 2005; Choi and Kim, 2006]. The central folds we identify here, on the other hand, represent all four SCOP classes, and form key structural bridges both within and between the class communities. To illustrate this diversity we identified four highly central pivotal folds. These folds represent dominant structural features, such as the Greek key motif, the doubly-wound β-sheet, the β-α-β switch and the α hairpin which act as key attractors within our landscapes.

Structural alignment in general remains an unsolved problem, and much has been written about the inaccuracies of current methodologies. For example a recent study demonstrated a high level of evolutionary inconsistency when comparing several alignment methods, including MAMMOTH, FATCAT and TM-align [Sadowski and Taylor, 2012]. However, despite their

limitations, these alignments can give us clues as to a global structure space, in ways in which common classification systems cannot. The representations we present here cannot be claimed to be accurate depictions of this global space. However, there is a well defined core to this space where different alignment methods agree on the architecture of fold space. This landscape maintains a previously observed distinction between the four different secondary structure classes, with an additional discrimination between all-β sandwiches and barrels. It also reveals novel features such as a propensity for bridges across this space to connect folds of similar evolutionary ages.

CHAPTER 5

Tunnels through sequence space

In this chapter hidden Markov models representing different portions of fold space are aligned and compared. Each model is constructed using domains in the same SCOP superfamily and their homologues, and where possible is seeded by a multiple structural alignment of related domains. By examining significant alignments between these models a network of relationships based on sequence patterns can be postulated, within and across fold space.

5.1 Motivation

Ever since the seminal paper of Chothia and Lesk [1986] demonstrating the evolutionary conservation of protein structures, folds have been thought to supersede sequences in terms of evolutionary information. Much of the work in previous chapters of this thesis has been instructed by this argument. In Chapter 2 it was shown that structural superfamilies and folds can be viewed as evolutionary units and are capable of largely reconstructing a tree of sequenced life. Using sequence based orthologues instead of these structural units fails to detect an evolutionary signal and produces a nonsensical tree topology [Lin and Gerstein, 2000].

However, it is also known that this conjecture is not absolute. There are an increasing number of examples which show closely related sequences differing dramatically in structure. For example, He *et al.* [2012] designed a progressive series of sequences, each differing from the last by only a single amino acid, yet which alternated between a $4\alpha + \beta$ and a 3α conformation. Furthermore, for some proteins, even attributing a single tertiary structure is inappropriate [Bryan and Orban, 2010]. A study by Burmann *et al.* [2012] identified an example of such fold switching in the C terminal domain of the protein RfaH, which adopts an all-α conformation when in complex with its N-terminal domain, yet when unbound refolds into a β-barrel.

Moreover, approaches for identifying evolutionary signals through sequence patterns have drastically improved in recent years. While the growth of novel structural data has been slowing ever since 1995 [Levitt, 2007], the expansion of sequence data remains exponential [Finn *et al.*, 2014]. This vast resource provides a valuable avenue in collating evolutionary information for proteins. Alignments of related sequences can capture the constraints affecting the evolution of a protein family. In some cases such constraints can even be used in *de novo* structure prediction [Marks *et al.*, 2011].

As we have discussed in previous chapters, formulating these evolutionary constraints from sequence data is a particular strength of hidden Markov models (HMMs) [Eddy, 1998]. These models capture the statistical profile of an alignment and thus contain the signal generated

by the observed constraints. The pairwise alignment of HMMs has been found to be more successful at identifying homologous sequences than the alignment of a single sequence to the model (for example, see Söding [2005]). HMM alignment has been used to demonstrate the power of this new generation of sequence methods in inter-fold relationships. Alva *et al.* [2010] constructed a map of the protein universe based on pairwise HMM comparisons in which a majority of domains sat in a single connected component. This was examined in Chapter 4 and is shown in Figure 4.1b. More recently, Farías-Rico *et al.* [2014] used HMM alignment to suggest homology between two well known superfolds: the TIM barrel and the Flavodoxin-like folds.

5.2 Overview

In this chapter 1,728 hidden Markov models are constructed and compared. The models represent structures from the four main SCOP classes (all-α, all-β, α/β and $\alpha+\beta$). Where possible, multiple structure alignments guide the profiles which constrain the model, and an iterative procedure is employed to ensure that every structure in the dataset contributes to a model. The pairwise alignment of these models produce a landscape of sequence links, or tunnels, within and between the profiles of different folds.

In Chapter 3 of this thesis similar relationships were constructed using structural alignment between domains. The links between models are compared to these structural bridges. Links between different folds identified during this method are dominated by two clusters: one of Rossmannoid folds, and one of β-propellers.

This chapter looks first at extending the theory behind hidden Markov models to allow for their pairwise alignment. This alignment can be understood as a path of pairwise states in a joint model of the two HMMs and can be assessed and scored by comparing the joint model to a null model. The methods section then describes how HMMs were constructed and seeded using multiple structural alignments of domains within the same superfamily. Structure alignments

are assessed by introducing an average core (Avcore) which modifies the strict core generated by the alignment algorithm. An iterative procedure is described by which domains representing a superfamily are organised into seeds for the models. Models are generated using the SAM-T2K method which was outlined in Section 1.6.2.3, and their alignments were assessed using an E-value. The resultant landscapes of sequence links derived from significant alignments are visualised. Inter-fold sequence links are then compared to the structural bridges described in Chapter 4. Finally, specific inter-fold sequence links are examined in detail alongside their structural alignments. These include a Rossmannoid cluster, a β-propeller cluster and two superfamilies containing Ferrodoxin proteins.

5.3 Hidden Markov models

In section 1.6.2.2 of this thesis, HMMs were introduced as a powerful tool to determine homology. The age estimates of superfamilies and folds introduced in Chapter 2 rely heavily on the use of HMMs to predict superfamily annotations on protein sequences from completely sequenced genomes. These predictions are derived from determining an alignment between a hidden Markov model and the protein sequence as described in section 1.6.2.3. In this chapter, potential evolutionary links between different folds are identified through the alignment between two HMMs. This alignment can be calculated in an analogous way to that between a model and a sequence. In this section the general theory behind HMMs will be recalled and methods for their pairwise alignment will be introduced.

Hidden Markov models can be interpreted as finite state machines capturing the statistical profile of a set of protein sequences. Figure 5.1 gives an example of such a machine. At each position i of the model, the machine can be in either a match state, an insert state or a delete state. In a match or insert state the machine will emit an amino acid. The particular amino acid emitted depends on a position and state specific set of emission probabilities, $f(\alpha \mid X_i)$. From a certain state, the machine can transition to another state according to another set

of probabilities, transition probabilities $p(X \to Y)$. The allowed transitions for the machine depend on its topology. For example, if the machine described in Figure 5.1a is in state M_2, it can transition to states I_2, M_3 or D_3. A diagram of the transitions allowed for this HMM is shown in Figure 5.1b.

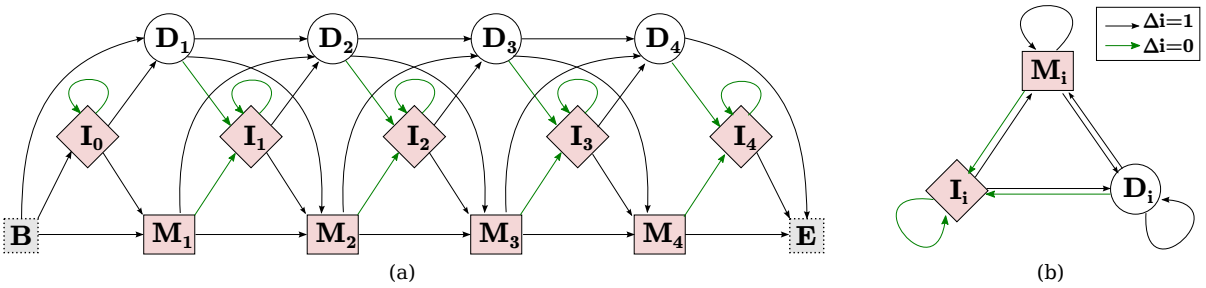


Figure 5.1: An example of the topology of a hidden Markov model for protein sequences. (a) Four positions in a HMM, showing the transitions between the different states. States involving the emission of an amino acid are coloured in red. (b) State transitions of the HMM displayed in a different format. The three states: match, insert and delete are displayed only once. Transitions between these states are shown by arrows coloured by the shift in position of that transformation. For example, transitions from I_i end at I_i, D_{i+1} and M_{i+1}.

5.3.1 Aligning HMMs

Aligning a sequence to an HMM can be understood as identifying the most likely path through the HMM which could have emitted that sequence. In a similar way, aligning an HMM to another HMM can be interpreted as finding the two paths through the HMMs which, jointly, maximise the probability of emitting every possible sequence. An example of a pairwise path through two models is shown in Figure 5.2. The two paths shown can both emit sequences of six residues, and any such sequence $x_1, x_2, \ldots, x_6$ can be scored according to a joint probability. In fact, the emission and transition probabilities of the joint model can be derived explicitly from those of each individual model.

Another way of visualising the pairwise paths shown in Figure 5.2 is as a single path of pair states. In other words, a joint model of pairwise states can be constructed. If the state of one model on the path is an emission state (i.e. it is a match or insert state), the other must be as

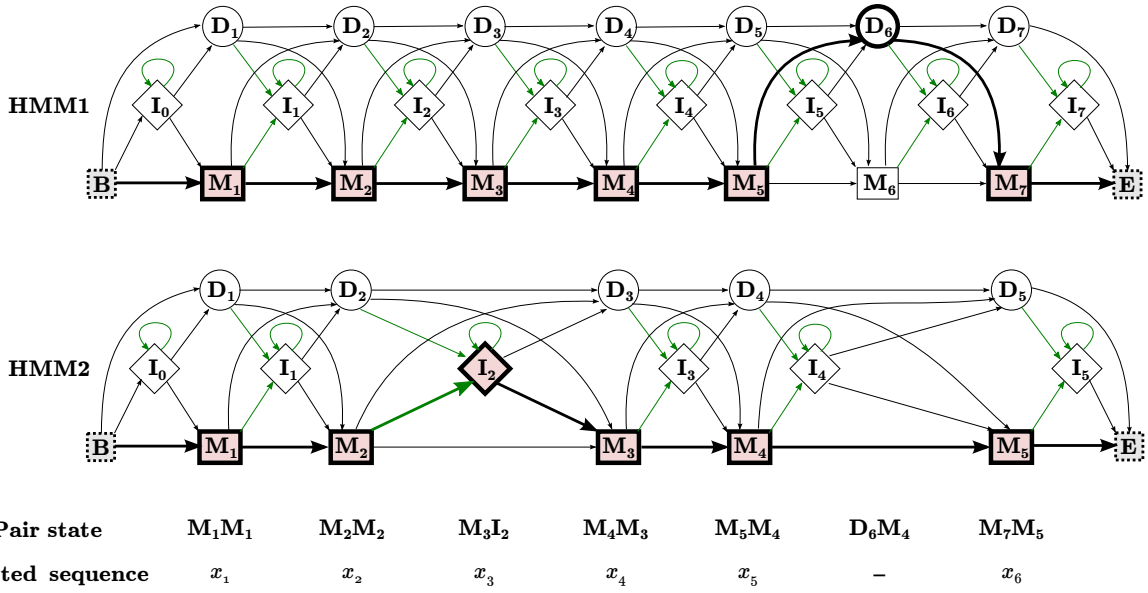


Figure 5.2: Aligning two HMMs. Any path through a single HMM can emit a sequence with a certain probability. Aligning two HMMs is equivalent to finding the paths (of the same length) which maximise the probability of emitting any sequence. Here, two HMMs are shown with paths in bold of length six. Any given sequence $x_1, x2, \ldots, x_6$ can be emitted by both HMM1 and HMM2. The two paths can also be represented as a single path through pair states. Emission states (match and insert states) can only be aligned with each other and are shaded in red. Delete states can be aligned with another delete state or a gap in the other model.

well. Moreover if one model's path transitions to a delete state, the other model must pause. For clarity, we refer to such a pause as a gap state G. The exception to this rule is where both models transition to a delete state. The pairwise states of the joint model and their possible transitions are summarised in Figure 5.3.

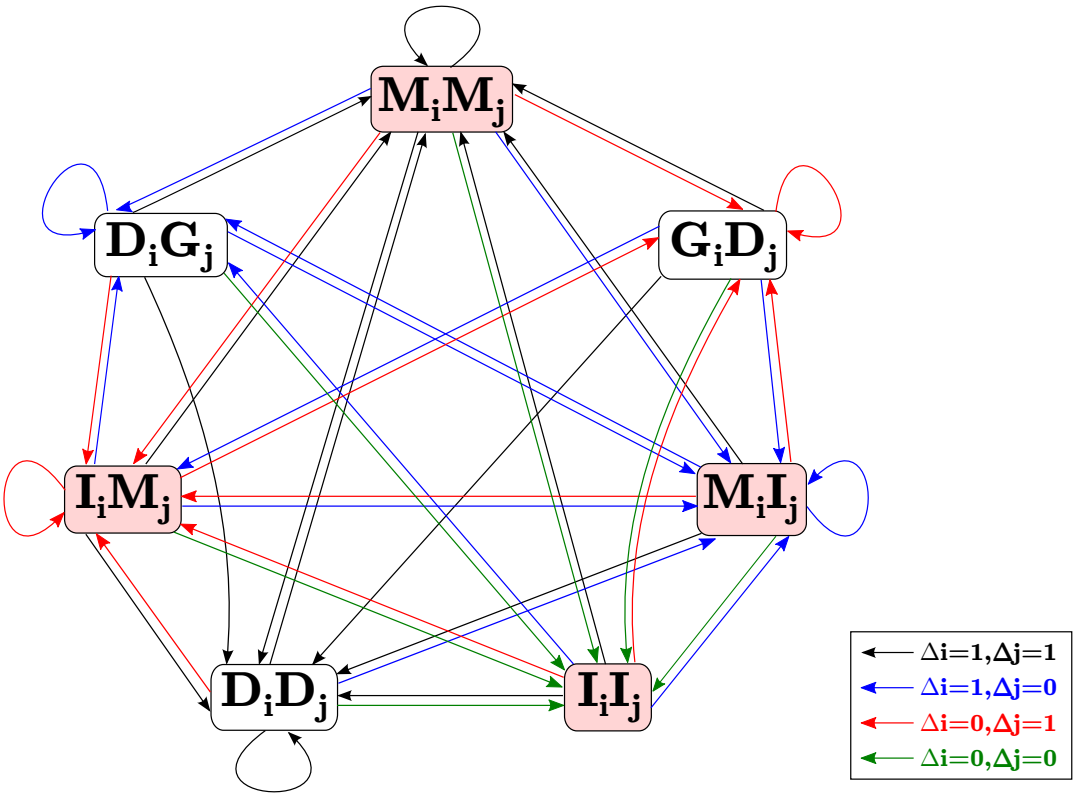


Figure 5.3: Paired states in aligned HMMs. The figure shows all seven states which form a joint model derived from the legitimate pairwise combinations of the three states (as well as a gap state) of the individual HMMs. Transitions are also shown and coloured according to their shift in position, derived from the shifts within the single HMMs. For example the transitions $M_i \rightarrow I_i$ in HMM1 and $M_j \rightarrow M_{j+1}$ in HMM2 will result in the pairwise transition $M_iM_j \rightarrow I_iM_{j+1}$. Pairwise emission states are made up of the combination of an emission state from each model and coloured in red. This figure has been motivated by work in Madera [2008].

In this chapter we will use the profile compiler (PRC) software to align and assess the pairwise comparison of our HMMs [Madera, 2008].

5.3.1.1 Pairwise probabilities

The emission probabilities from states M_iM_j, M_iI_j, I_iM_j and I_iI_j, and the transition probabilities between these and the other pairwise states are calculated from their corresponding probabilities in the individual models, and by assuming their independence. Explicitly, emission probabilities of an amino acid α from a state $X_iY_j \in \{M_iM_j, M_iI_j, I_iM_j, I_iI_j\}$ in the pair model become:

$$f^{\text{pair}}(\alpha \mid X_iY_j) = f^{\text{HMM1}}(\alpha \mid X_i) \cdot f^{\text{HMM2}}(\alpha \mid Y_j) \tag{5.1}$$

and transition probabilities between any pair state W_iX_j to another state Y_kZ_l are:

$$p^{\text{pair}}(W_iX_j \to Y_kZ_l) = p^{\text{HMM1}}(W_i \to Y_k) \cdot p^{\text{HMM2}}(X_j \to Z_l). \tag{5.2}$$

In the event of one model transitioning to a gap state (i.e. when the other model is in a delete state) this gap is really a pause and no transition is required. Therefore in these cases the transition probabilities will just derive from the single transition to a delete state. For example

$$p^{\text{pair}}(M_iI_j \to G_iD_{j+1}) = p^{\text{HMM2}}(I_j \to D_{j+1}). \tag{5.3}$$

In this way the probability of co-emitting any sequence along the pairwise path can be calculated from the constituent models.

5.3.1.2 Scoring the alignment

For simplicity, in previous sections, alignments of HMMs have been referred to in terms of maximising the probability of emitting any possible sequence along a path. While this captures the overall intention of the alignment process, in practice the probability of emitting a sequence needs to be compared to a null model in order to normalise for biasing factors such as the length of the model, and generally over-represented residues.

The null model used for a model of length M is shown in Figure 5.4. The null model for

a given HMM will emit sequences with an expected length M, where M is the length of the HMM. Furthermore the emission probabilities at each position depend on the overall propensity for each amino acid across the whole model.

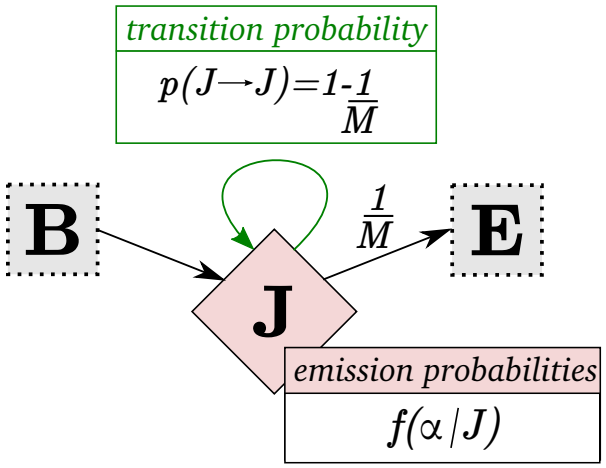


Figure 5.4: The null model for scoring HMMs. The model has a single emission state J and a transition probability of $p(J \rightarrow J) = 1 - 1/M$ where M is the length of the HMM being compared to the null model. Emission probabilities of an amino acid α from J are calculated as the average of the match emission probabilities for α across all positions in the HMM. This figure has been motivated by work in Madera [2008].

Pairwise log-odds scores are introduced to normalise the probabilities given above in equations 5.1 and 5.2 by the pair null model. Explicitly, transition scores for any pair state W_iX_j to another state Y_kZ_l become

$$S^{\text{pair}}(W_iX_j \rightarrow Y_kZ_l) = \log \frac{p^{\text{HMM1}}(W_i \rightarrow Y_k) \cdot p^{\text{HMM2}}(X_j \rightarrow Z_l)}{p^{\text{null1}}(W_i \rightarrow Y_k) \cdot p^{\text{null2}}(X_j \rightarrow Z_l)} \tag{5.4}$$

where the null transition probability from state X to state Y is defined

$$p^{\text{null}}(X \rightarrow Y) = \begin{cases} p^{\text{null}}(J \rightarrow J) & \text{if } Y \in \{M, I\} \\ 1 & \text{if } Y \in \{D\} \end{cases} \tag{5.5}$$

Emission scores at a pair state X_iY_j of the joint model are summed across all possible amino

acids and are normalised by the average probability of emission in the null models.

$$S^{\text{pair}}(X_i Y_j) = \log \sum_{\alpha} \frac{f^{\text{HMM1}}(\alpha \mid X_i) \cdot f^{\text{HMM2}}(\alpha \mid Y_j)}{0.5(f^{\text{null1}}(\alpha \mid J) + f^{\text{null2}}(\alpha \mid J))}. \tag{5.6}$$

To assess an alignment of two HMMs these log-odds scores can be summed along the pairwise path. The optimal alignment between two models is the pairwise path which maximises this score.

As stated above, the probability of emitting a sequence along a pairwise path in the joint model is calculated under the assumption of independence between the two HMMs. In other words, the score for a sequence emitted by HMM1 must not depend on the score for the same sequence to be emitted by HMM2. When HMM1 and HMM2 cover evolutionarily related sections of sequence space this assumption is clearly untrue as a high scoring sequence on one model will also score well on any related model. However, the exact nature of the dependencies between a given two models is unknown and would be hard to implement into a method aligning different HMMs.

5.4 Methods

In this chapter, 1,728 hidden Markov models are constructed from structural alignments of superfamily siblings. These models are then aligned to identify potentially distant homologous relationships.

5.4.1 Domain dataset

The domain dataset from which the structural alignment seeds were constructed are the same 4,098 domains as were used in Chapter 4. To recall, this set consists of structures from the ASTRAL database, filtered to $\leq$ 40% sequence identity [Brenner *et al.*, 2000]. Structures consisting of purely C_α coordinates, as well as those containing chain breaks were omitted. These domains represent a total of 1,025 different SCOP superfamilies and 631 folds from the

four main secondary structure classes (all-α, all-β, α/β and $\alpha+\beta$). Domains within this dataset follow ASTRAL's naming convention, a seven character long string of the form dppppcx, where pppp represents the PDB code of the protein the domain is found in, c represents its chain identifier as given in the PDB file, and x represents the domain within that chain as assigned by SCOP.

5.4.2 Multiple structural alignments as seeds for the HMMs

Multiple structural alignments were constructed from superfamily siblings amongst the domain dataset using MAMMOTH-mult [Lupyan *et al.*, 2005]. For each of the 1,025 superfamilies represented by at least one structure in the domain dataset the pipeline shown in Figure 5.5 was used to construct one or more structural seeds. Where possible, these seeds represented multiple structural alignments. The criteria for accepting a structural alignment made use of the strict core generated by MAMMOTH-mult's algorithm. This core represents the columns in the final alignment with no gaps and where, in the structural superposition, each residue is placed within 4Å of every other residue [Lupyan *et al.*, 2005]. In MAMMOTH-mult this core is considered as a percentage of the length of the shortest structure. However, to avoid the inclusion of a single short domain resulting in an unfairly high core percentage, here we calculated the Avcore: defined as the percentage of the average domain length which was part of the strict core. Multiple structure alignments were accepted as seeds for the subsequent HMMs if this Avcore was at least 50%.

If a structural alignment resulted in an Avcore of less than 50% an iterative pruning procedure was used to split the domains into smaller groups. This procedure made use of the bifurcating dendogram generated by MAMMOTH-mult. This tree of domains is the result of applying an average linkage clustering algorithm to the set of pairwise MAMMOTH Z-scores [Lupyan *et al.*, 2005]. If the alignment did not result in a large enough Avcore this tree was pruned by splitting the two most recently joined clusters. The largest of these clusters was then realigned. In the case where the clusters were of equal sizes a random choice was made as to

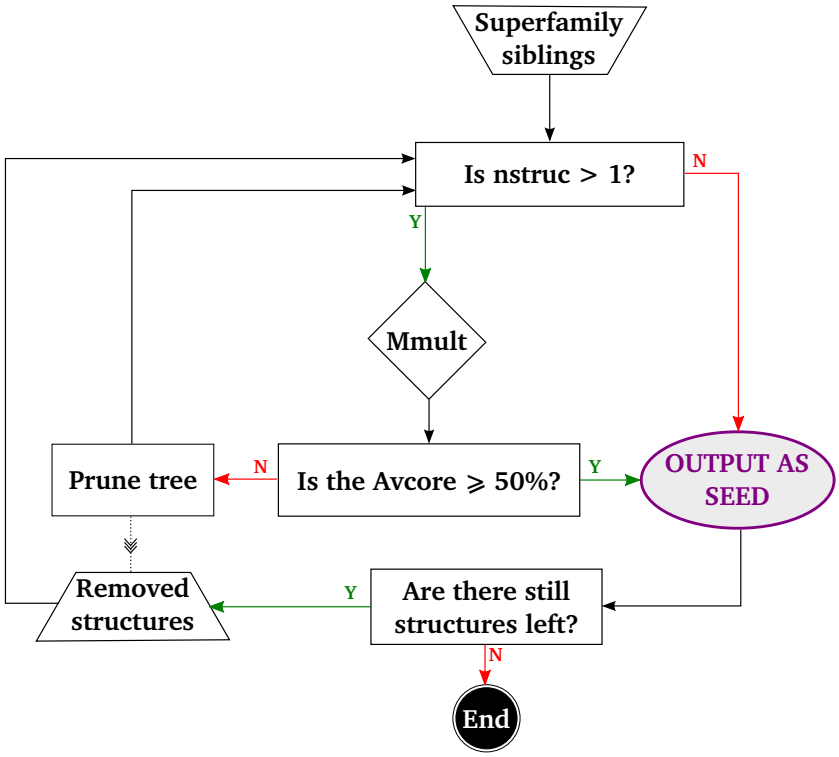


Figure 5.5: Pipeline for establishing the structural seeds for each HMM. The input to this pipeline are the structures in the domain dataset representing a single superfamily according to SCOP. These domains are known as superfamily siblings. Each set of superfamilies are divided into structural seeds, which represent either a multiple structure alignment with an Avcore of at least 50% (see text for details) or a single sequence if such an alignment is unavailable.

which cluster to realign. The domains from the cleaved cluster were set aside temporarily. If the Avcore of the realignment was at least 50% this alignment is accepted as a seed for HMM construction. If the Avcore is still not high enough the new tree is pruned in the same way until the remaining domains result in an acceptable alignment. After a seed has been generated for a superfamily any domains which have been removed during the pruning process are realigned and the process is repeated. An example of this process is shown in Figure 5.6.

Using the above method 1,728 seeds were generated from the 4,098 domains in the dataset.

5.4.3 Constructing the HMMs

Each of these 1,728 structural seeds was used to generate a hidden Markov model using the SAM-T2K pipeline outlined in Section 1.6.2.3. This method gradually added homologous sequences to the original alignment, iteratively training the model to capture the statistical profile of the growing alignment. Where the seed was a multiple structural alignment, the model remains constrained by that initial alignment. In the case where the seed was a single sequence no such constraints can be applied. SAM-T2K version 3.5 was used with default options [Karplus *et al.*, 1998]. Weights for the observed frequencies in each alignment column relative to background distributions of amino acids were 0.8 in the first iteration, 0.7 in the second, 0.6 in the third and 0.5 in the fourth. E-values for the selection of homologues to add to the alignment at each iteration were 0.0001, 0.002, 0.001, and 0.005 respectively. The NCBI's non-redundant protein database was used for the homologue search, although the training set was further filtered by a sequence identity threshold as part of the SAM-T2K procedure.

5.4.4 Comparing the HMMs

HMMs were aligned using PRC [Madera, 2008]. Each profile was tested against the library of all 1,728 models and E-values were calculated for each run based on fitting a two-parameter sigmoidal distribution to the reverse score as suggested in Karplus *et al.* [2005]. E-values were non-symmetric as they are model specific. In other words an alignment between two HMMs

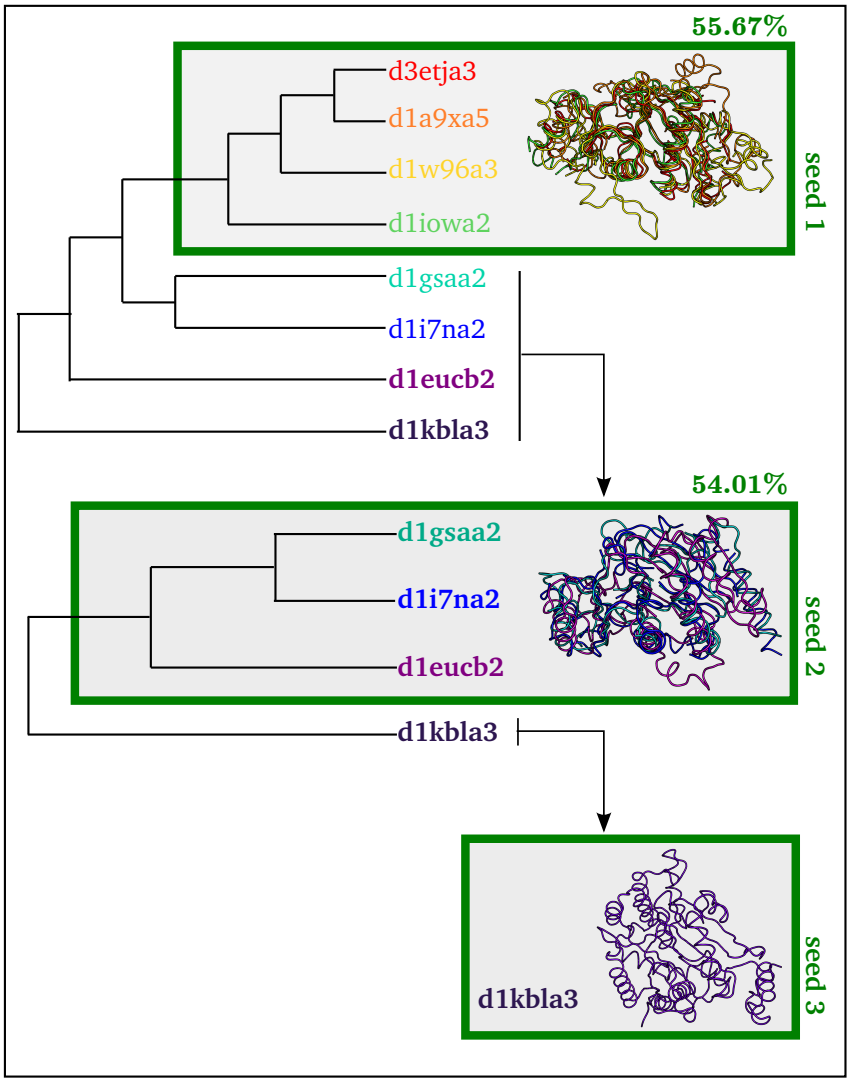


Figure 5.6: Pruning the MAMMOTH-mult tree to generate structural seeds for the HMMs. The figure shows an example of the pipeline described above for the superfamily d.142.1. This superfamily contains 8 representative structures in the domain dataset. The initial alignment contains a core of 58 residues and an average domain length of 246 resulting in an Avcore of 23.6%. To improve the alignment the tree is pruned by removing branches until the remaining domains produce an Avcore of at least 50%. At each stage, the two branches at the root are cleaved and the smallest of these branches is then removed. In the first example, the first stage removes the branch containing d1kbla3. Before an Avcore of 50% is seen for d.142.1 two further cleavages are required: first the single domain d1eucb2, then the branch containing d1gsaa2 and d1i7na2. After these domains have been removed the remaining four structures are assigned an alignment with an Avcore of 55.7%. This alignment becomes the first structural seed for d.142.1. The four domains which were cleaved from the tree are then realigned into the second tree. This alignment generates an Avcore of 32.3% so the tree is cleaved. In this case, removing d1kbla3 results in an alignment with Avcore of 54% which becomes the second seed for d.142.1. There is only one domain left so no further alignments can take place. In this case the single sequence is used as the third seed for the superfamily.

will receive different E-values depending on which model was compared to the library. We thus considered the E-value attached to an alignment as the highest of these two possible values. Relationships between models were examined when their alignment resulted in an E-value $< 3 \times 10^{-4}$, as suggested by Madera [2008] as indicative of homology.

5.4.5 Structural comparison of related folds

Structural bridges were calculated as in Chapter 4. All four alignment methods were considered.

5.4.6 Fold ages

Fold ages were estimated as in Chapter 2. For simplicity the set of ages used the NCBI tree with branch lengths estimated using occurrences of superfamilies and maximum parsimony to estimate gain and loss events.

5.5 Results

We constructed 1,728 models representing 1,025 different superfamilies and 631 folds. Of these models, 897 were constructed from a multiple structural alignment and 831 from a single sequence. The structure alignments which formed the seeds of the 897 models contained a maximum of 23, and a mean average of 3.64 different structures.

5.5.1 Relationships between different models

Pairwise sequence links were established between models whose alignment had an E-value $< 3 \times 10^{-4}$. 1,082 such links were found. Of these 908 (83.9%) were between different models within the same superfamily, 68 (6.3%) were between models in different superfamilies but the same fold, and 106 (9.8%) were between unrelated models. These proportions remained robust to the choice of E-value as a cutoff. Figure 5.7 shows how the number of links and their type alter between E-values of 0.05 and 1e-06.

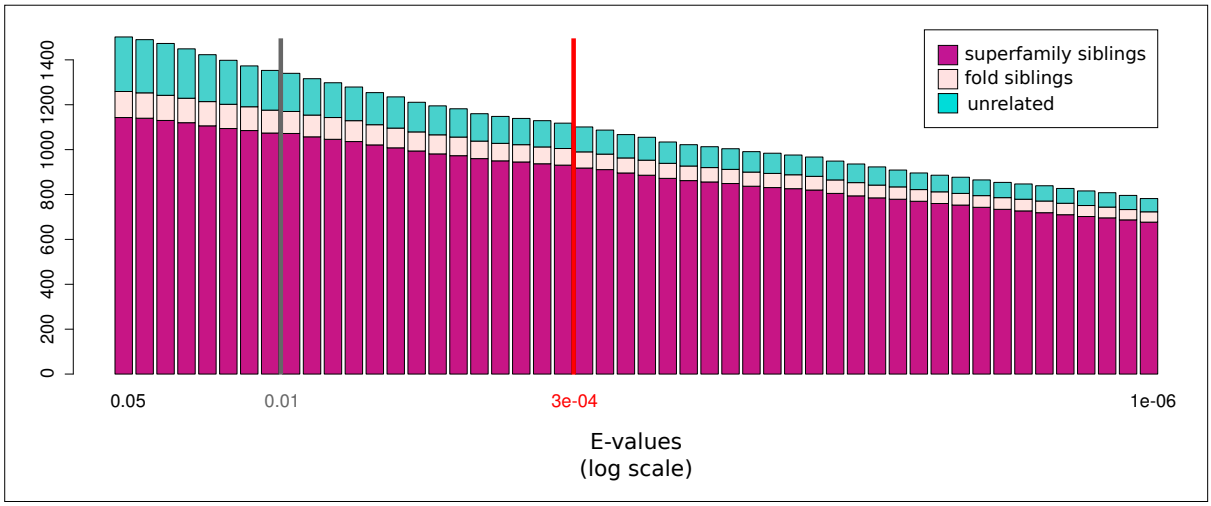


Figure 5.7: Sequence links between models and their E-values. A stacked barchart of the number of links found with E-values below a certain threshold (log scale). Each bar is split into links between models representing the same SCOP superfamily, those between models of different superfamilies but the same fold, and those with no such relationship.

Roughly equal numbers of the models were seeded by multiple structural alignments (897) and single sequences (831). However, of the 1,082 links between the models, 558 (52%) linked models constructed by structure alignments and only 207 (19%) linked two models both seeded by single sequences. The remaining 317 (29%) of the links were between one model seeded by an alignment and one seeded by a sequence. There was, however, no discernible difference in the distribution of E-values attached to links on the basis of their seed type.

5.5.2 Relationships between different folds

Relationships established between different models were collapsed into those linking different folds. Each pairwise combination of folds was attached an E-value of the most significant alignment between their representative models.

5.5.2.1 Comparisons to structural bridges

The 106 links between models of different folds were translated into 36 inter-fold links between 51 folds. A list of all these sequence links can be found in Supplementary Table D1. Of these 36 links, only 18 (50%) are linked by a structural bridge in any of the networks presented in Chapter 4. The remaining 18 pairs do not appear together in any of the structure networks even at the most lenient threshold. Of the 18 pairs with evidence for a structural relationship five lie between different β-propeller folds and seven connect different Rossmannoid folds. These cases will be discussed in more detail in future sections. Three of the remaining pairs are only detected as structurally similar by one of the four alignment methods. The remaining three sequence and structurally linked folds are shown in Table 5.1.

fold1	fold2	ESA	FATCAT	MAMMOTH	TM-ALIGN
a.1	d.58	0.5	-	0.7	0.8
b.1	b.95	0.7	0.7	0.9	0.9
d.58	d.90	-	0.8	0.8	0.5

Table 5.1: Structurally similar linked folds. The five folds, linked by HMM alignment and not β-propellers or Rossmannoids, which share a likely structural relationship. The value of the highest posterior probability threshold for which the two folds are connected by each of the four alignment methods is shown. Where a value is omitted (-) the pair is not joined by that method at a threshold of 0.5. For these folds' names see the Supplementary Table A2

5.5.2.2 Fold ages

Fold ages were used to examine the links between folds uncovered by this analysis. The 51 folds which were found to link to another fold were predominantly ancient. 36 of these folds had a fold age estimate of 1.0. Only one such fold was counted as a new-born. The Ganglioside M2 activator fold (b.95) has a fold age of 0.38 and is only found in certain Eukaryotes. In the network it is linked to the ancient Immunoglobulin-like fold (b.1).

Of the 36 inter-fold links representing significant alignments of the HMMs, only three did not contain an ancient fold at either endpoint. To compare this frequency to a random set of

36 fold pairs we performed 5,000 simulations where 36 fold pairs were taken at random from the 631 folds represented by at least one HMM in the dataset. For each simulation the number of pairs out of 36 containing no ancient folds were recorded. Of the 5,000 simulations of 36 random fold pairs, only 21 (0.42%) have at most 3 pairs with no ancient participant.

5.5.3 Tunnels through sequence space

The links uncovered by the HMM alignments are shown in Figure 5.8. For simplicity only components containing at least one inter fold link are shown. Each node in the figure represents a different HMM and are arranged according to their fold. Links are drawn between nodes where their alignment results in an E-value of < 0.01. Significant E-values $< 3 \times 10^{-4}$ are shown in black. Non-significant E-values are shown in light grey.

Several discernible clusters are evident, distinguished by multiple links between two different folds. In particular, a Rossmannoid cluster can be identified as one of the most densely connected groups (Figure 5.8i) incorporating the classical Rossmann-like folds c.2, c.4 and c.5 as well as other, topologically similar folds c.66 and c.79. The β-propeller folds (b.67-70) are also strongly linked (Figure 5.8k), as they are in the structure networks. The multiheme cytochrome fold (a.138) is also linked to cytochrome c (a.3) demonstrating their shared evolutionary history (Figure 5.8a).

As well as relationships between folds Figure 5.8 also exhibits the links between models representing the same fold. Some folds demonstrate suggestions of a monophyletic origin for their different superfamilies by displaying several links between these models. For example the TIM barrel fold (c.1). On the other hand, other folds show few links between models representing different superfamilies within the same fold, for example the Ferredoxin-like fold (d.58).

Specific clusters and inter-fold relationships will be further examined in the following sections.

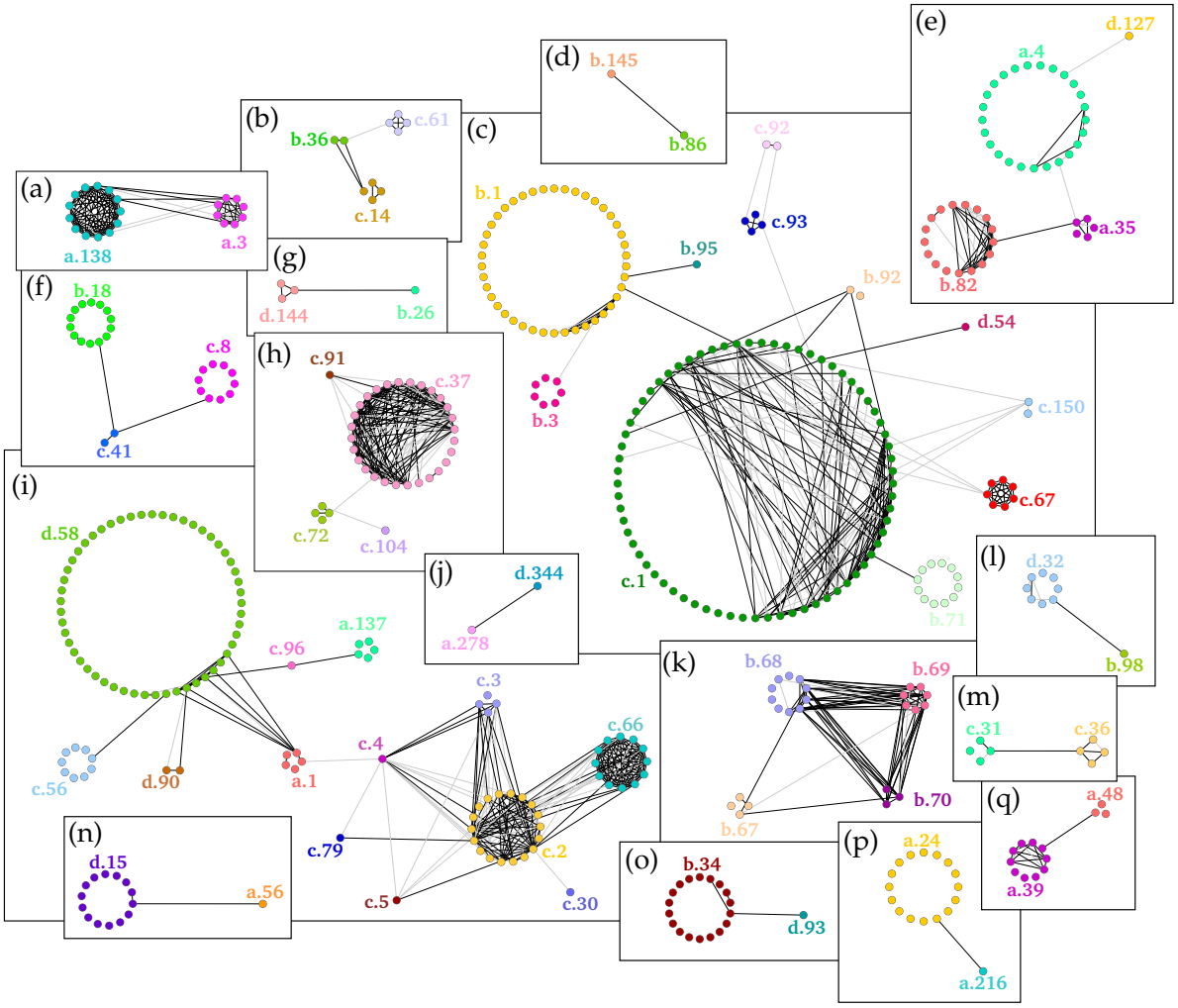


Figure 5.8: Tunnels through sequence space. Network representation of the results of HMM alignment. Each node represents a model seeded by one or more domains from the structural domain dataset. Models representing the same fold are arranged in a circle. Light grey edges represent alignment of the models resulting in an E-value < 0.01. Edges in black are significant and have E-value $< 3 \times 10^{-4}$.

5.5.4 Rossmannoid relationships

Immediately evident in Figure 5.8i is a Rossmannoid cluster centred on the NAD(P)-binding Rossmann fold (c.2). Multiple models representing c.2 are linked to models from other Rossmannoid folds: FAD/NAD(P) binding domains (c.3), Nucleotide-binding domains (c.4) and S-adenosyl-L-methionine-dependent methyltransferases (c.66). c.2 is also linked to the MurCD N-terminal domain (c.5) and the Tryptophan synthase beta subunit-like PLP-dependent enzymes (c.79). There are also multiple links joining c.3 and c.4. The classical Rossmann fold has an architecture consisting of α-β-α layers. The central layer is a parallel, doubly-wound β sheet of six strands.

There are seven different folds annotated as Rossmannoids in SCOP version 1.75 for which HMMs were constructed. They are c.2, c.4, c.5, c.27 (Nucleoside phosphorylase/ phosphoribosyltransferase catalytic domain), c.28 (Cryptochrome/photolyase N-terminal domain), c.30 (PreATP-grasp domain) and c.31 (DHS-like NAD/FAD-binding domain). Models representing c.27 and c.28 were not found to align significantly to any other folds. c.30 aligns to c.2 at a non-significant E-value < 0.01. c.31 appears in the network (Figure 5.8m) but is linked to the THDP-binding fold (c.36). c.36 has a similar three layer architecture to the Rossmannoids but the strand ordering within the β-sheet is 213465 as opposed to the doubly-wound Rossmann sheet (order 321456).

S-adenosyl-L-methionine-dependent methyltransferases (c.66) are not annotated as Rossmannoids under SCOP due to the insertion of an anti-parallel strand in the interior of the central β-sheet layer. c.66 has strand order 3214576 where strand 7 is anti-parallel to the rest of the sheet. Despite this topological difference S-adenosyl-L-methionine-dependent methyltransferases are classified in the same fold as the Rossmann folds in CATH [Orengo *et al.*, 1997]. A structural alignment of domain d1ej0a_ from c.66 and d1yb5a2 from c.2 is shown in Figure 5.9. The inserted anti-parallel β strand of d1ej0a_ is shown highlighted in blue. It is aligned, in sequence though not in structure, to a helix in d1yb5a2 which sits outside of the helical layers.

A notable topological outlier in the Rossmannoid cluster in Figure 5.8i is the FAD/NAD(P)

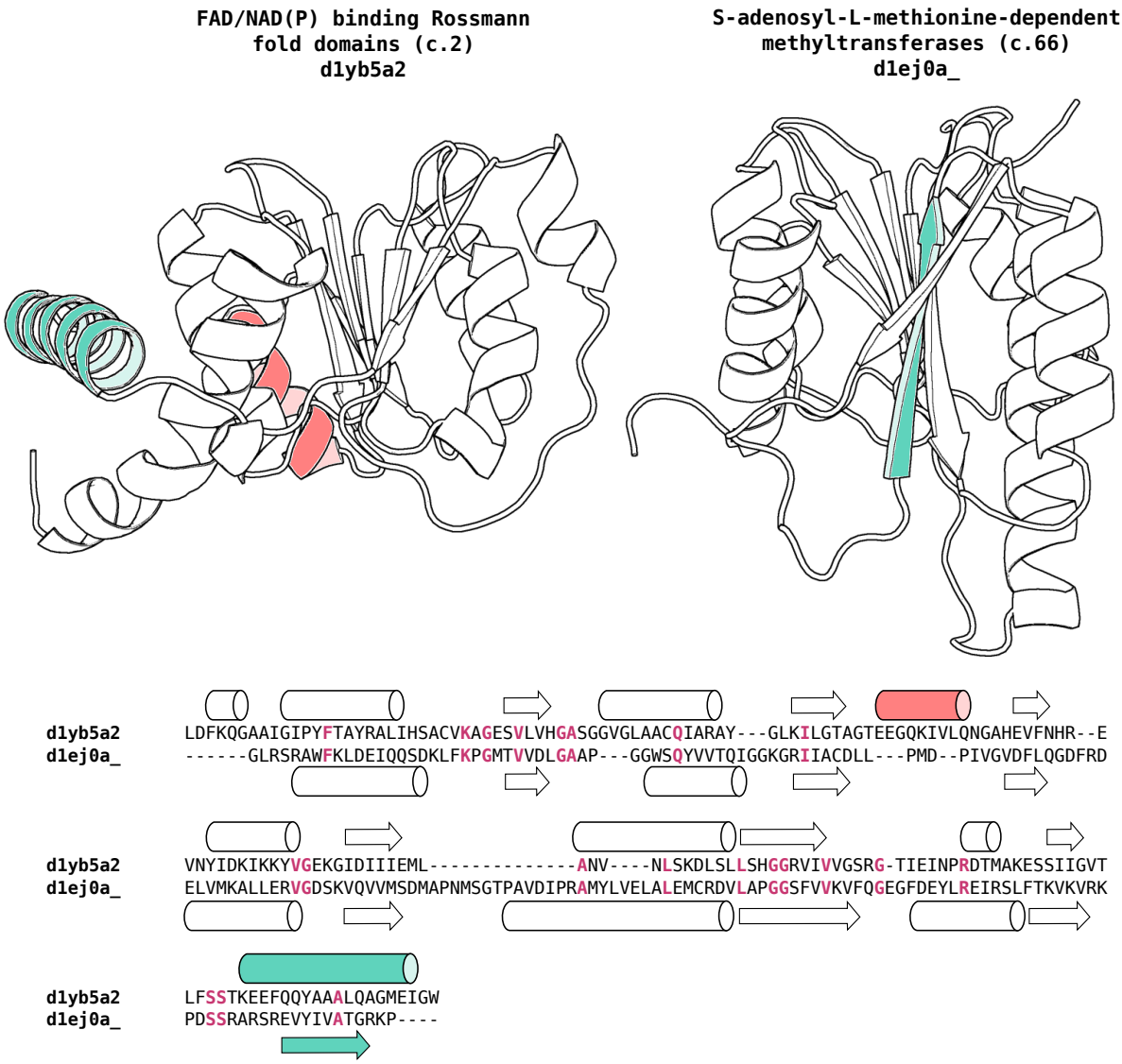


Figure 5.9: Structural alignment between domains from c.2 and c.66. Structural alignment was performed using MAMMOTH-mult. Structures were superposed from this alignment then separated for clarity. The sequence alignment generated by MAMMOTH-mult is also included and secondary structure is annotated using JOY [Mizuguchi *et al.*, 1998]. α-helices are shown as cylinders, and β-strands as arrows. The anti-parallel strand (highlighted in blue) of d1ej0a_ which distinguishes members of c.66 is inserted between two parallel strands. It is aligned, with sequence similarity, to a helix in d1yb5a2 which lies outside its core 3 layer fold. An inserted helix in d1yb5a2 is also shown in red. Identical residues in the sequence alignment are coloured in purple.

binding domains (c.3). c.3 consists of a three layered architecture with a doubly-wound β-sheet at its core. However, the upper layer is a β-sheet meander rather than a helical layer. It is therefore a β-β-α layered fold, differing from the Rossmannoid folds it is linked to (c.2 and c.4). A structural alignment of domain d1rp0a1 from c.3 and d1djqa3 from c.4 is shown in Figure 5.10. The β-sheet meander is shown in yellow and is aligned with a single helix in d1djqa3. d1rp0a1 also has an additional anti-parallel strand at the end of the central sheet coloured in purple, although this is atypical of the fold.

Domains from c.3 are strongly linked both to c.4 and to c.2.

5.5.5 β-propellers

Propellers are all-β folds consisting of a number of blade-like β-meander sheets. Propellers are classified into folds on the basis of the number of blades they contain. In our dataset we included models for 4,5,6,7 and 8 bladed propellers (b.66-70 respectively). Figure 5.11 shows the structure of a 6-bladed propeller. Each blade consists of four anti-parallel β-strands arranged from the centre of the propeller to the outside (strands a-d in Figure 5.11). Each blade shares sequence similarity to the other blades in the protein and corresponding strands are aligned. For example the sequence coding for strands a-d can be aligned to the sequence coding for strands a'-d'. This is also true for propellers with different numbers of blades. As each blade is a linear section of sequence and exhibits particular sequence patterns, propellers with different numbers of blades can be aligned with just a single insertion or deletion of one or more blades. Evolutionary relationships between these folds have previously been suggested [Chaudhuri *et al.*, 2008].

5.5.6 Inter-class ferredoxins

A notable sequence link in Figure 5.8i is that between the Ferredoxin-like fold (d.58) and the Globin-like fold (a.1). In particular, the links appear only between the superfamilies d.58.1 (4Fe-4S ferredoxins) and a.1.2 (alpha-helical ferrodoxin). These folds represent ferredoxins of

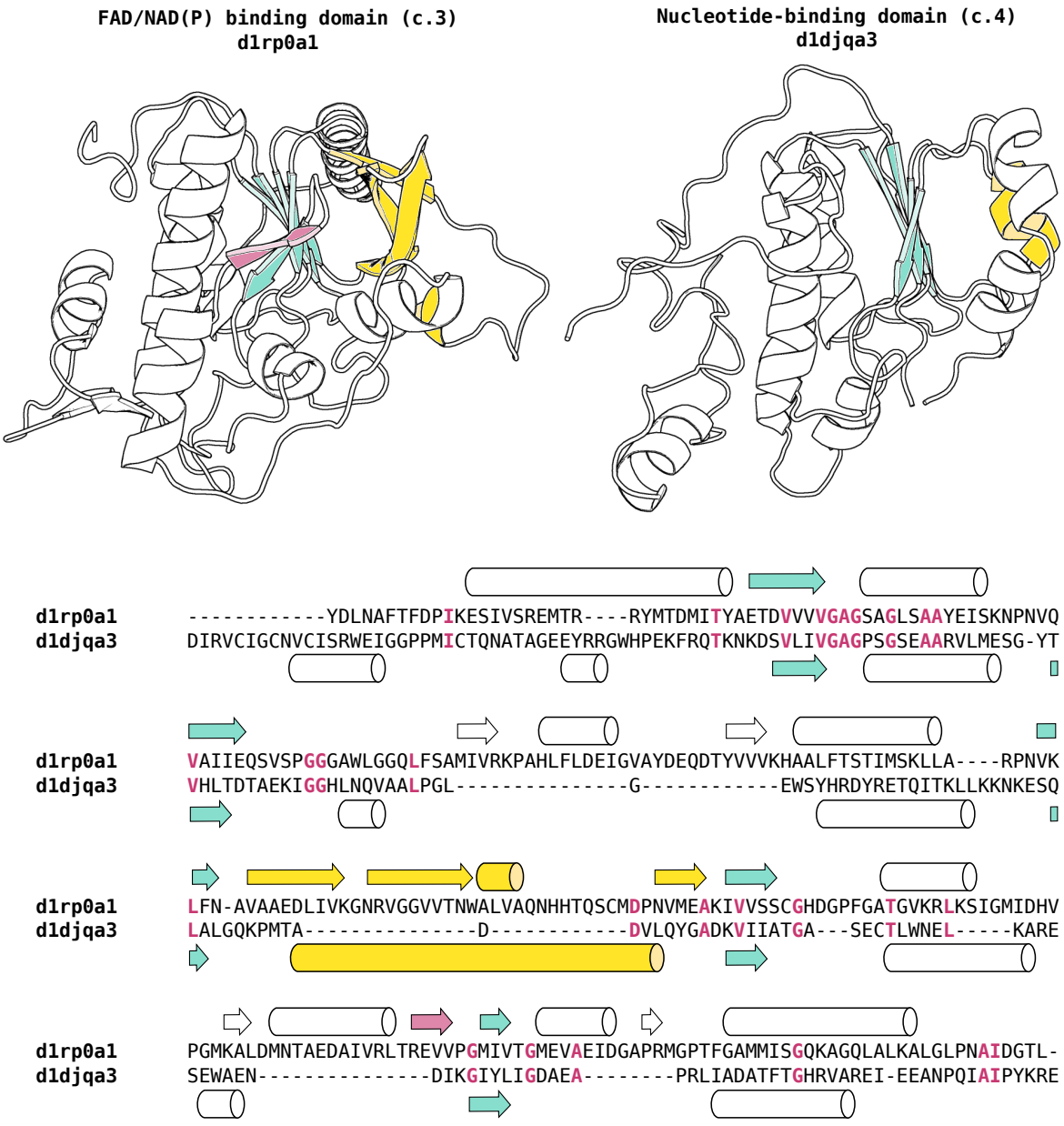


Figure 5.10: Structural alignment between domains from c.3 and c.4. Structural alignment was performed using MAMMOTH-mult. Structures were superposed from this alignment then separated for clarity. The sequence alignment generated by MAMMOTH-mult is also included and secondary structure is annotated using JOY [Mizuguchi *et al.*, 1998]. α-helices are shown as cylinders, and β-strands as arrows. The β-sheet replacing one of the α layers in c.3 is aligned with a single helix in c.4 (shown in yellow). The central β-sheet is highlighted in blue, with an inserted anti-parallel strand in c.3 shown in red. Identical residues in this alignment are coloured in purple.

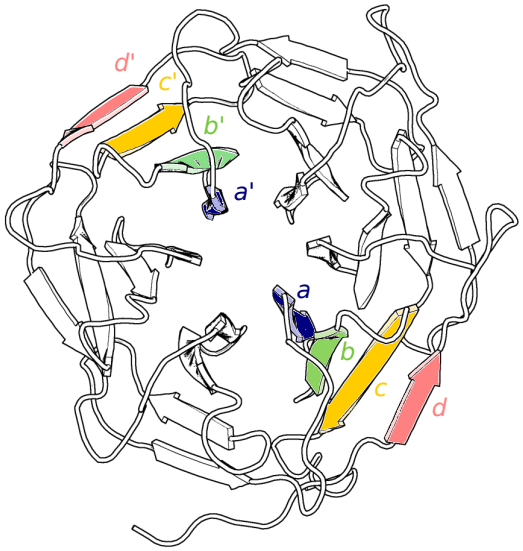


Figure 5.11: An example structure of a six bladed β-propeller. Two blades of this propeller are coloured according to sequence conserved segments within the protein (equivalent strands are labelled as (a,a'), (b,b'), (c,c') and (d,d')).

completely different structures, one a helical bundle and the other an $\alpha + \beta$ sandwich. A structure alignment of two domains representing these folds is shown in Figure 5.12.

The alignment has at its core the metal binding sites of two Iron-Sulphur clusters. The clusters are held in both folds below two helices and cystine residues on each chain bind to the sulphur atoms in the clusters. The sequence alignment of the domains in Figure 5.12 are also shown. The cystine residues which bind to the clusters are conserved (shown in orange) in the alignment. The alignment also demonstrates a correspondence between the β-α-β units making up d3c8ya3 and the helical units in d2bs2b1. In the structure of d3c8ya3 small helical turns augment the β side of the fold.

5.6 Conclusions

In this chapter we constructed and compared a set of hidden Markov models representing a large portion of fold space. Models were seeded by multiple structure alignments where such align-

ments were of good quality, and single sequences where structural similarity was missing. Links between models were established where their pairwise alignment returned significant E-values. The landscape of links generated by this method produces a much sparser set of relationships than the structural networks discussed in the previous chapter. Moreover, there does not appear to be much evidence for the inter-fold links to be supported by structural similarity.

It was interesting to see that more inter-fold sequence links than would be expected by chance connected at least one ancient fold. One possible reason for this is that the links found during the HMM alignment may also represent possible candidates for false positives in the prediction models used to assign superfamilies to genome sequences, on which the age estimates are based. It is therefore possible that some of these folds have been incorrectly classified as ancient as a result of sequences representing other folds scoring significantly against SUPERFAMILY models.

Two strongly supported clusters are evident in the set of inter-fold links: a Rossmannoid cluster incorporating classical Rossmann folds as well as the topologically distinct FAD/NAD(P) binding domains (c.3) which consist of a second β sheet in place of one of the Rossmannoid fold layers. This fold is strongly linked to the cluster however, and seems to be more closely linked to the classical Rossmann fold than other topological Rossmann folds like the DHS-like fold (c.31) which has a link to a structurally dissimilar domain from the THDP-binding fold (c.36). Other topological Rossmannoids do not appear connected to any other folds. The other evident cluster is the β-propellers, which consist of β-meander blades.

Another interesting example of linked folds is found in the relationship between the α-helical ferredoxin and the $2\alpha + 4\beta$ ferredoxin fold. Similarity between these two superfamilies centres on the metal binding site, in particular, the cystine residues which bind to the iron-sulphur cluster through their sulphur atoms.

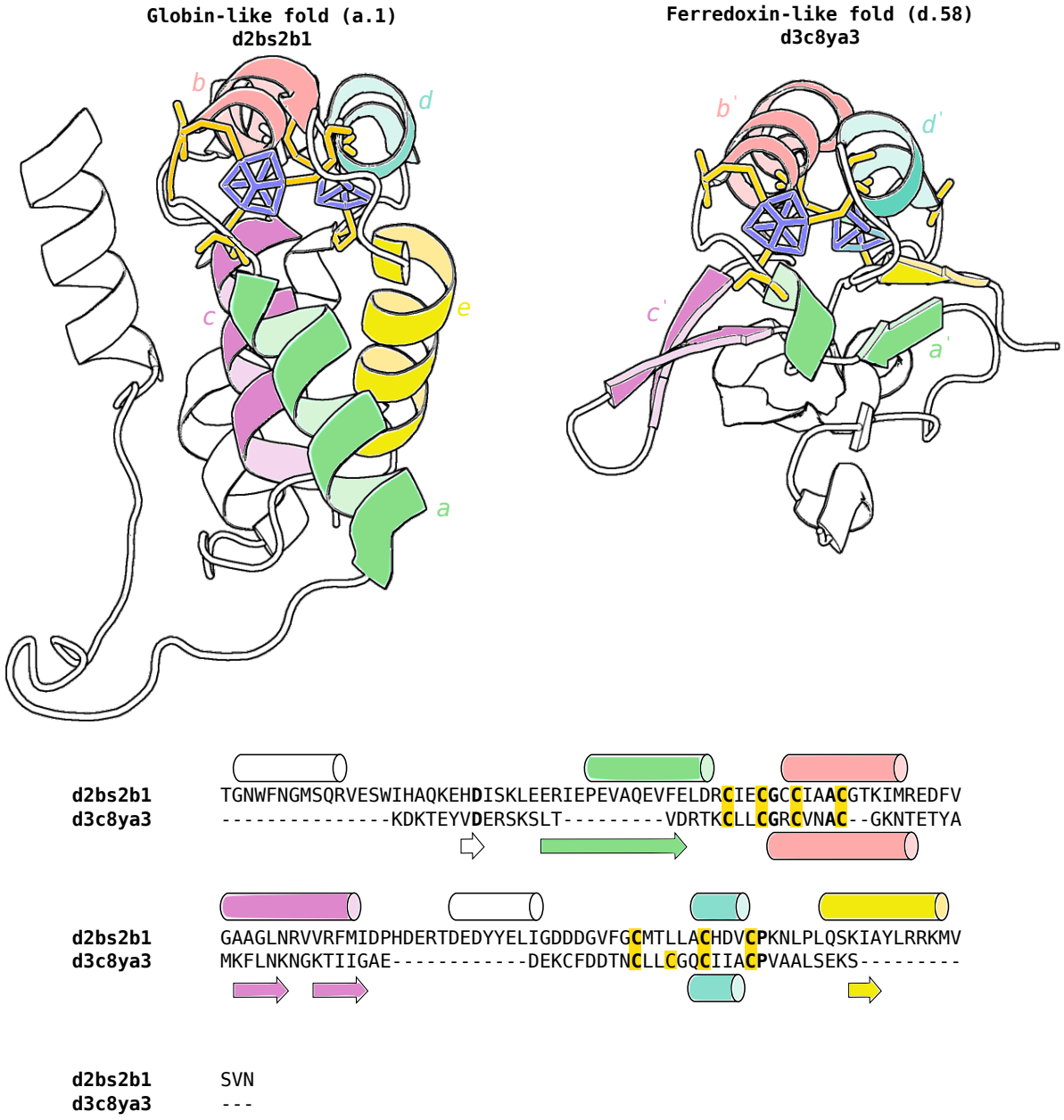


Figure 5.12: Structural alignment between domains from a.1 and d.58. Structural alignment was performed using MAMMOTH-mult. Structures were superposed from this alignment then separated for clarity. The sequence alignment generated by MAMMOTH-mult is also included and secondary structure is annotated using JOY [Mizuguchi *et al.*, 1998]. Iron-sulphur (4Fe-4S) clusters are shown in blue. Side chains for the cystine residues which bind to the sulphur in the cluster are shown in orange. Aligned secondary structures in the lower body of the folds are shown in green, pink and yellow.

CHAPTER 6

Conclusions, context and future directions

This thesis has examined different themes relating to the evolutionary and structural landscapes of proteins. Because of the nature of its subject matter, many of the questions that arise from the work presented here are open ended and relate to the overall nature of the protein universe. In this final chapter I will discuss some of these questions, attempt to place the work in previous chapters into a global context and outline further directions for this work.

6.1 Superfamilies and folds: evolutionary and structural units

In this thesis I have examined different ways of comparing the evolutionary and structural properties of proteins. Within specific chapters I have looked at either superfamily or fold units. In essence I have tried to maintain a distinction: using superfamilies to represent the evolutionary units, and folds the structural units, of the protein universe. Thus, Chapters 2, 3 and 5 predominantly focus on superfamilies: estimating their ages, preferences and constructing their sequence profiles. In Chapter 4 however, the emphasis was on constructing a map of structure space so I presented the work using folds as units instead.

However, I did also perform the analysis in these chapters using the different units. This work supported the results presented in these chapters. For example, the properties of ancient and new-born folds revealed a similar preference for less elaborate structures in new-born folds, and in networks of superfamilies, ancient nodes tended to be more central.

While SCOP folds are classified as structural, rather than evolutionary, units, it is possible that in some, or even most, cases, these groupings represent evolutionary relationships which have been missed, for example, by insufficient functional annotation. Thus the age estimates of folds may provide useful information.

6.2 Age estimates

6.2.1 Evolutionary ancestors or relics of biased annotation?

Chapter 2 introduced the concept of ages determined by the superfamily and fold content of completely sequenced genomes. This chapter also discussed different interpretations for the age estimates. In the subsequent chapters ages are referred to as estimates for the origin of a superfamily's ancestor. However, the nature of this origin is uncertain. It is possible that origin of some superfamilies was, in essence, *de novo*, possibly derived from expression of previously non-coding sections of the genome. Other possibilities include a transition from an already existing superfamily or fold to a separate structural and functional unit, possibly through gene

duplication and subsequent mutation.

This in itself raises the question of whether superfamily units are truly distinct. It is still possible to reconcile the concept of discrete superfamily units even if there are evolutionary transitions linking superfamilies. This is because the functional constraints applicable to each unit will differ but remain relatively consistent within the superfamily. However, the repercussions of an ancestral evolutionary connection between different superfamilies may result in errors in classification or assignment of proteins to one or other of two linked superfamilies. If there are families of proteins effectively midway between different superfamilies as a result of the ancestral relationship, as has been recently observed [Farías-Rico *et al.*, 2014], classification and assignment of members of these families may be incorrect, particularly if specific structural and functional annotation is not available for these proteins.

Lack of structural and functional annotation for different protein families is thus not an insignificant problem. In particular, annotation is not only incomplete, but unequal. Biases have been identified in the resolution of annotation of both structure and function [Peng *et al.*, 2004; Schnoes *et al.*, 2013]. This means that such errors are likely to affect particular superfamilies to different extents.

On the other hand, while there is potential for errors in classification and assignment to threaten the validity of the occurrence profiles upon which the ages are based, Chapter 2 also demonstrated that these profiles could be used to reproduce viable topologies for the tree of life. This fact supports the likelihood of superfamilies to be relatively well classified evolutionary units and the occurrence profiles to be accurate summaries of these units on the species considered.

Figure 6.1 shows six different scenarios which could explain a given occurrence profile derived from superfamily assignments. These scenarios are:

(a) Convergent evolution

(b) Widespread domain loss

(c) Lateral gene transfer

(d) False positive assignment or classification

(e) False negative assignment or classification

(f) Errors in the species tree topology

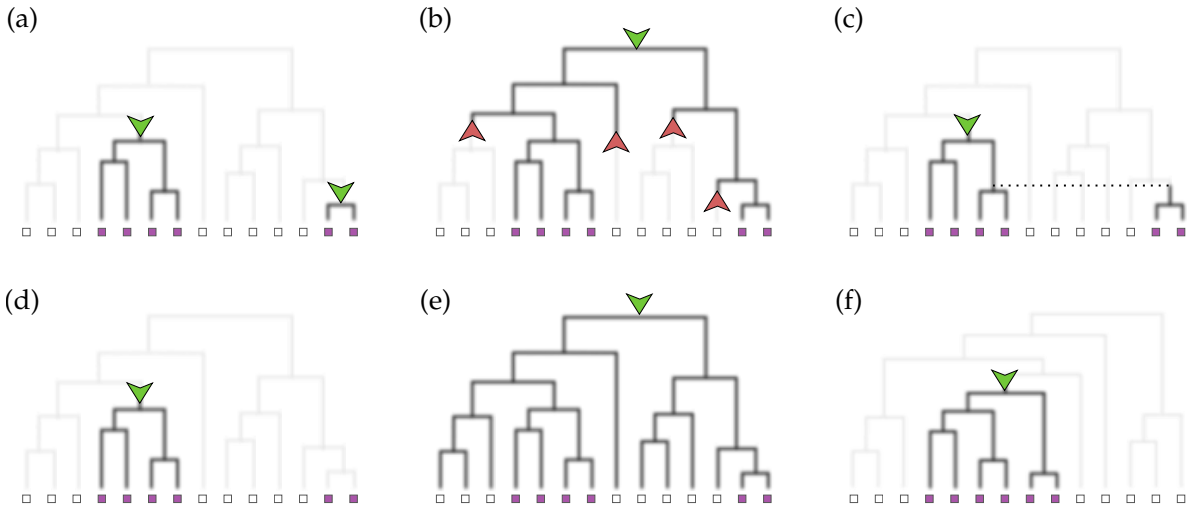


Figure 6.1: Six scenarios to explain an occurrence profile. Evolutionary gain events are shown in green, and loss events are shown in red. See text for details.

Under a maximum parsimony model in this particular example, age estimates would be be derived from scenarios shown in Figure 6.1 (a), (c) and (d). That is, evolution by convergence, lateral gene transfer or false positives. These types of scenarios tend to minimise the age estimate. Under a Dollo parsimony model, more likely scenarios would be domain loss and false negative assignments or classification errors (Figure 6.1 (b) and (e)). These scenarios tend to push the age estimate towards 1.0. Finally, the underlying tree topology can produce a different age to either of these scenarios. Interestingly, a maximum parsimony algorithm tends to produce the lowest age estimate in this scenario, although it is easy to imagine a more recent ancestor and widespread lateral gene transfer resulting in the observed pattern.

In the subsequent chapters of this thesis, superfamilies or folds tended to be partitioned into ancients and new-borns. One particular reason for this is that the scenarios leading to different age estimates for these superfamilies seem less likely than for the middle-aged superfamilies with more complex occurrence profiles. For example, a superfamily with an ancient estimate will appear on a wide range of different genomes from right across the tree of life. Possible scenarios leading to an incorrect ancient estimate would be almost universal lateral gene transfer, even across several Eukaryotes, high numbers of false positives in the assignment of that superfamily to the genome sequences or incorrect classification of unrelated proteins as belonging to that superfamily. New-born superfamilies will only appear on either a single genome, or a few closely related species. Incorrectly assigned new-borns will be the result of widespread domain loss from several different species, missed assignments on large numbers of species, or an over classification of a family of proteins as a separate evolutionary unit. They could also be the result of an incomplete tree of life. These scenarios do seem more likely than those which could have lead to an incorrect ancient age estimate. It is easy to imagine an ancestral superfamily developing a highly specialised function and being lost in a large number of species save a few.

On the other hand, the types of superfamilies that are included in the new-born population seem to support the hypothesis that they evolved later along the evolutionary timeline. In particular, as I looked at in Chapter 3, superfamilies with disulphide bonds are enriched in this set. Disulphide bonds are often found in proteins which require additional stabilising features, such as those that exist in an extracellular environment. Because the origin of multicellular life was relatively recent on an evolutionary scale, it seems unlikely that such superfamilies are in fact ancient.

6.2.2 Future directions for age estimation

In Chapter 2 I attempted to look at the robustness of the age estimates in various ways. By altering the relative likelihoods of gain and loss events and the topology of the phylogenetic tree, different combinations of the events illustrated in Figure 6.1 can be considered for each

superfamily. However, due to limited time I was not able to fully investigate robustness of the method to further variation.

For example, calculating ages based on CATH classifications, as opposed to SCOP, could indicate how far these values depend on a given classification scheme. There are also other methods for estimating evolutionary histories, rather than using parsimony to predict gain and loss events. For example, using maximum likelihood to infer the node of origin for a superfamily's ancestor, which can incorporate uncertainty in this estimate as well as uncertainty in the phylogenetic tree [Pagel *et al.*, 2004].

One important factor in further work must be the consideration of differential rates in the evolution of superfamilies. I have already spoken about altering the rate at which evolutionary events occur across different portions of the tree of life. In this thesis, a simple implementation of this idea was used to consider the Eukaryotic subtree as subject to different dynamics than the Prokaryotic branches of the tree, in the formulation of a fusion parsimony model. It is likely that further resolution of differential evolutionary dynamics across the tree may improve the model (for examples of differential evolution in metazoa and primates see [Ekman *et al.*, 2007; Hahn *et al.*, 2007]).

It is also possible that different superfamily units themselves may evolve differently [Wilkins *et al.*, 2012]. In fact, this is highly plausible as so much of their evolutionary drift is mitigated by functional, and therefore structural, constraints [Choi *et al.*, 2007]. As such, it may be beneficial to consider adopting gain weights specific to different superfamilies. In order for this approach to be relevant, superfamily centric evolutionary dynamics would need to be interpreted in terms of the relative likelihoods of loss and gain events.

Another possible future application for the work on superfamily ages would be to express each estimate in terms of years. In this way, superfamily evolution could be examined in the context of other evolutionary events and key functional innovations. This could be attempted by annotating internal nodes of the tree with known divergence dates. For example, TimeTree is a tool which collates a consensus of divergence times for species from the literature [Hedges

et al., 2006].

6.3 Preferences of ancient and new-born superfamilies

In Chapter 3 significant preferences of the ancient and new-born populations of superfamilies were examined. What were termed middle aged superfamilies were largely ignored in this section as I postulated that these might be the superfamilies with a less reliable estimate. Structural properties of ancient and new-born superfamilies were assessed in this chapter to give a general overview of the preferences of these populations. As such the structural summaries were calculated rather crudely. I took summary properties for a superfamily based on the mean across its representative domains. Although, not explicitly shown in the results here I did also look at other summary statistics, such as the median, maximum and minimum of the values found in representative domains. Using each of these different values the same signal was shown as was reported in the main body of this thesis. However, it would also be interesting to look at how conserved the properties were across representative domains, which I did not look at.

There are also other properties which examine tertiary packing structure in a more detailed way. For example, defining a binary classification for buried and surface residues ignores some of the more complex relationships each residue has with its environment. As a result it is possible to estimate a set of shells within the structure (called Voroni shells), derived from the distance between atoms and the boundary of the molecule, which quantify *how* buried that residue is [Bouvier *et al.*, 2009]. Another property I would have liked to look at was the overall *roundness* of the domain, using a method which examines how well the shape can be approximated by some central spheres [Cazals *et al.*, 2014].

For many of the structural properties examined in Chapter 3, a dependence on the length of the protein presented a problem. For example, the proportion of buried residues and non-local contacts will both correlate with length in globular proteins. As such, an important part of this work was maintaining that these properties were strictly preferences amongst ancient

superfamilies, rather than being residually correlated with age due to the strong length bias.

One particular question which is prompted by the structural preferences presented in Chapter 3 is whether there might be some correlation between preferred properties and the types of bias which might induce incorrect age estimation. As an example, could shorter superfamilies be harder to recognise using the SUPERFAMILY models, and thus have higher rates of false negatives on assignments to the genomes leading to artificially lower ages? Questions of this type are very complex and hard to validate. However, I did examine in Chapter 2 whether SUPERFAMILY E-values showed any dependence on the average length of a superfamily and saw no such signal. Moreover as I mentioned in the previous section, other properties like enriched disulphide bonds in the new-born set of superfamilies do suggest that their age estimates are, at the very least, not ancient.

Additionally, however, is the consideration that both the structurally solved universe and the phylogenetic tree are incomplete. In Chapter 2 it was shown that on average, across the 1,014 completely sequenced genomes, on average 65% of a species' protein sequences were assigned a known superfamily. The remaining unannotated sections of the proteomes could either belong to an existing superfamily, and therefore be a false negative assignment, or belong to a novel superfamily, with its own set of structural properties which could affect the preferences of the two populations.

As well as structural preferences I also examined the functional annotations of new-born and ancient superfamilies. In particular, it was evident that these populations were more effectively partitioned using their structural, rather than their functional properties. It is important to note that this does not necessarily mean that evolutionary structural constraints are stronger than functional ones. In fact, a large part of the argument behind structural conservation in superfamilies is the presence of functional constraints. However, it is most likely symptomatic of the difficulties present in the field of functional annotation, which relies on manual curation [Schnoes *et al.*, 2013].

6.4 Fold space networks

In Chapter 4 of this thesis I attempted to investigate the structural landscape of proteins which was exposed in Chapter 3 in a global manner by constructing networks derived from structural alignments between domains of different folds.

I believe there are some important strengths to the method I have presented in this chapter when compared to other attempts to illustrate fold space. In particular, network representations, consisting of a set of edges, or structural bridges, between folds, do not require the underlying relationships between structures to be transitive. On the other hand, multidimensional scaling methods and principal component analysis, which attempt to project structure space onto a lower dimensional space, assume that similarity is transitive [Ben-Tal and Kolodny, 2014]. Moreover these projected spaces are derived from every single pairwise comparison, regardless of its significance. While constructing a network requires the introduction of a threshold above which an edge is drawn, this choice also allows for missense alignments, which correspond to the majority of the pairwise comparisons, to be interpreted as a lack of a relationship, rather than the exact score, which is often meaningless to contribute information on the space.

In order to allow for flexibility in the threshold I chose several different possible values, rebuilding the networks at each stage. Moreover, the introduction of the posterior probability in defining these thresholds allowed for a way of establishing comparative spaces across different alignment methods, and in particular in the generation of consensus networks between these methods. Using these consensus spaces allows for the identification of bridges between folds which are supported by multiple different methods, and those which may simply artefacts of a particular algorithm. In particular it was shown that up to half of the bridges identified by each method were unsupported by the others, suggesting there are still inconsistencies in the field of structural alignment. In particular, the unsupported portion of each method's bridges could not be diminished by increasing the threshold for the alignment score, which suggests that inconsistent alignments may be artefacts of the algorithms themselves. This further highlights

the need to improve structure alignment algorithms.

Elastic shape analysis (ESA) shows great promise in this area, especially as it is a mathematically rigorous method which uses a metric, rather than a normalised RMSD or statistical value, to measure the distortion between two chains [Liu *et al.*, 2011]. However, at present it has not been designed to compare a large number of domains of varying sizes. In particular, it establishes an initial alignment between the chains by simply picking equidistant points along each domain. Subsequent refinements and calculation of the metric thus lack a biologically relevant starting point which can lead to improper alignments between domains of different lengths.

Despite the weaknesses of protein structure alignment, another question which is prompted by the results of this chapter is whether the nature of fold space is truly discrete as classification schemes suggest. This discreteness assumption could be rephrased as domains within a fold falling into an equivalence class within structure space. Certainly, by the proportion of bridges evident, even in consensus networks at high threshold values, the number of structural similarities between different folds is substantial. Moreover, for most alignment methods used in this chapter, the networks consisted of a majority connected component at probability thresholds less than 0.9, which could suggest that at least large portions of structure space may not be discrete. We have also shown that similar network landscapes cannot be derived by taking the average similarities between different folds. An explanation for this observed behaviour could be found in the different ways in which structural alignment compares proteins to the fold classification procedure. In particular, an almost explicit assumption of the manual classification method is that conservation of the core structure is more meaningful than periphery elements. However, alignment methods tend not to make such a distinction. In order to examine this hypothesis it would be interesting to look at the differences between inter-fold and intra-fold alignments in terms of the structural core. Similarly, I would also like to examine other features of the alignments between different folds. For example, looking at the conservation of secondary structure or motifs along the bridges of the networks.

6.5 Sequence models and HMMs

In Chapter 5 a different approach was adopted in detecting relationships between folds. In particular, I wanted to investigate if it was possible to detect definitive evolutionary signals supporting any of the relationships exposed by structural bridges in the previous chapter. Seeding models with multiple structure alignments appeared to produce broader profiles as was indicated by the increased number of links identified by these models, which makes them good potential candidates in searches for more distant homologues.

It was interesting to see in this chapter how few of the identified sequence links also corresponded to structural bridges. The structurally similar portions of the sequence landscape were heavily dominated by the Rossmannoid cluster, and the β-propellers, and half of the sequence links showed no detectable structural similarity at all.

The sequence links which were identified offer a great potential for discussing not only inter-fold relationships, but intra-fold links as well. In particular, for studying the evolution of different superfamilies within the same fold. Comparing profiles of superfamilies which fall under the same fold can indicate how likely the phenomenon of convergent evolution to the same structural architecture takes place [Gough, 2005].

There is a great potential for further work developing from this chapter. I would like to investigate how sensitive the HMM profiles were to the structural alignment method used (MAMMOTH-mult) and the Avcore 50% cutoff for accepting a multiple structure alignment. Additionally, I would like to examine more recent advances in improving hidden Markov models. For example, using patterns of correlated mutations evident in the sequence alignments which construct HMMs, to guide alignments between them [Deng and Cheng, 2014].

6.6 CATH vs SCOP: classification schemes

The exclusive use of SCOP as a classification scheme throughout this thesis is the final point I wish to mention here. Recalculating ages, preferences and fold space networks using CATH's

superfamilies and topologies would be an important step in validating the claim that the work presented here is not a residual effect of classification bias. CATH and SCOP are derived from completely dissimilar procedures. Even their domain assignments are independent. Nevertheless, estimating ages using CATH classifications has been done before, revealing similar pictures of protein evolution as those estimated using SCOP [Abeln, 2007; Bukhari and Caetano-Anollés, 2011].

The principal reason why the SCOP scheme was used in this thesis is that it attempts to encapsulate more evolutionary information, whereas the focus of CATH is more structural. As a result, SCOP maintains a distinction, which this thesis also supports, between α/β and $\alpha+\beta$ superfamilies.

6.7 Closing remarks

In this thesis, I have examined several structural and evolutionary relationships between protein superfamilies and folds. The age estimates of superfamilies derive their origin from their occurrence profile across sequenced life. Despite the potential for error in this process, I believe the age estimates calculated here remain incredibly valuable tools. In this thesis I have used them to analyse widespread structural, sequence and functional preferences, as well as more global structural relationships.

While, for the most part, this thesis has been concerned with global properties and spaces, one of its greatest potentials is for age estimates, structure maps and sequence links to be used to examine either specific motifs, folds, or even proteins of interest. Knowledge of where these molecules appear on the tree of life, their neighbourhood of similar architectures, and any possible distant evolutionary relationships can open new avenues of understanding.

References

Abeln, S. (2007). *Protein fold evolution on completed genomes: distinguishing between young and old folds.* Ph.D. thesis, University of Oxford. 42, 43, 75, 192

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410. 43

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**(17), 3389–3402. 29

Alva, V., Remmert, M., Biegert, A., Lupas, A. N., and Söding, J. (2010). A galaxy of folds. *Protein Sci.*, **19**(1), 124–130. 106, 157

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.*, **62**(11), 1182–1197. 13

Andreeva, A. and Murzin, A. G. (2006). Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.*, **16**(3), 399–408. 15, 16

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**(Database issue), D310–D314. 21

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29. 83

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**(Database issue), D154–D159. 83

Ben-Tal, N. and Kolodny, R. (2014). Representation of the Protein Universe using Classifications, Maps, and Networks. *Isr. J. Chem.*, **31905**, 1286–1292. 189

Benner, S., Cohen, M., and Gonnet, G. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.*, **229**, 1065–1082. 12

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235–242. 8

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.*, page P10008. 118

Bouvier, B., Grünberg, R., Nilges, M., and Cazals, F. (2009). Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins*, **76**(3), 677–692. 187

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.*, **7**(3), 369–376. 74

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**(1), 254–256. 23, 78, 120, 164

Bryan, P. N. and Orban, J. (2010). Proteins that switch folds. *Curr. Opin. Struct. Biol.*, **20**(4), 482–8. 18, 156

Buchan, D. W. A., Shepherd, A. J., Lee, D., Pearl, F. M. G., Rison, S. C. G., Thornton, J. M., and Orengo, C. A. (2002). Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.*, **12**(3), 503–514. 43

Bukhari, S. A. and Caetano-Anollés, G. (2011). Evolution of protein architectures inferred from phylogenomic analysis of CATH. *IEEE Int. Conf. Bioinform. Biomed. Workshops*, pages 1029–1031. 192

Burmann, B. M., Knauer, S. H., Sevostyanova, A., Schweimer, K., Mooney, R. A., Landick, R., Artsimovitch, I., and Rösch, P. (2012). An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, **150**(2), 291–303. 18, 156

Caetano-Anollés, G. and Caetano-Anollés, D. (2003). An evolutionarily structured universe of protein architecture. *Genome Res.*, **13**(7), 1563–1571. 41, 48

Camin, J. H. and Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution*, **19**(3), 311–326. 69

Capra, J. A., Williams, A. G., and Pollard, K. S. (2012). ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput. Biol.*, **8**(6), e1002567. 41, 42, 75, 87

Cazals, F., Dreyfus, T., Sachdeva, S., and Shah, N. (2014). Greedy Geometric Algorithms for Collection of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining. *Comput. Graphics Forum*, **33**(6), 1–17. 187

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**(Database issue), D189–92. 23

Chaudhuri, I., Söding, J., and Lupas, A. N. (2008). Evolution of the beta-propeller fold. *Proteins*, **71**(2), 795–803. 176

Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.*, **26**(7), 1523–1531. 12

Cheng, H., Kim, B.-H., and Grishin, N. V. (2008). Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J. Mol. Biol.*, **377**(4), 1265–1278. 18

Cheng, S. and Brooks, C. L. (2013). Viral capsid proteins are segregated in structural fold space. *PLoS Comput. Biol.*, **9**(2), e1002905. 98

Choi, I.-G. and Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proc. Natl. Acad. Sci. U.S.A.*, **103**(38), 14056–14061. 41, 46, 49, 74, 75, 85, 87, 153

Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Mol. Biol. Evol.*, **24**(8), 1769–82. 186

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**(4), 823–826. 13, 41, 156

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423. 120

Cooper, G. M., Brudno, M., Stone, E. A., Dubchak, I., Batzoglou, S., and Sidow, A. (2004). Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res.*, **14**(4), 539–548. 12

Cordes, M., Burton, R., Walsh, N., McKnight, C., and Sauer, R. (2000). An evolutionary bridge to a new protein fold. *Nat. Struct. Biol.*, **7**, 1129–1132. 16, 17

Day, R., Beck, D., Armen, R., and Daggett, V. (2003). A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.*, pages 2150–2160. 20

Deng, X. and Cheng, J. (2014). Enhancing HMM-based protein profile-profile alignment with structural features and evolutionary coupling information. *BMC Bioinform.*, **15**(1), 252. 191

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numer. Math*, **1**(1), 269–271. 118

Doolittle, W. F. and Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(7), 2043–2049. 13, 43

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, pages 755–763. 28, 29, 43, 156

Edwards, H., Abeln, S., and Deane, C. M. (2013). Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput. Biol.*, **9**(11), e1003325. 37, 38, 50, 51, 57, 73, 74, 86, 87, 89, 92, 97, 99

Ekman, D., Björklund, A. K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *J. Mol. Biol.*, **372**(5), 1337–1348. 186

Farías-Rico, J. A., Schmidt, S., and Höcker, B. (2014). Evolutionary relationship of two ancient protein superfolds. *Nat. Chem. Biol.*, **10**, 710–715. 157, 183

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res.*, **40**(Database issue), D136–D143. 44, 52

Fedoroff, N. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758–767. 12

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376. 46

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**(2), 164–166. 52, 69

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res.*, **42**(Database issue), D222–D230. 156

Friedberg, I. and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. *Structure*, **13**(8), 1213–1224. 76, 104, 107

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf : Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics*, **19**(1), 163–164. 14

Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, **3**(9), 679–687. 12, 13

Goldstein, R. A. (2008). The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.*, **18**(2), 170–127. 74

Goodsell, D. S. and Olson, A. J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 105–153. 8

Gough, J. (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics*, **21**(8), 1464–71. 41, 42, 191

Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**(1), 268–272. 29

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (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. 25, 31, 43, 49, 50, 58, 83, 100

Greenwald, J. and Riek, R. (2012). On the Possible Amyloid Origin of Protein Folds. *J. Mol. Biol.*, **421**, 417–426. 76

Grishin, N. V. (2001). Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**(2-3), 167–185. 15, 16, 17

Guzzetta, A. (2000). Ionsource.com mass spectrometry resource, amino acid chart. `http://www.ionsource.com/virtit/VirtualIT/aainfo.htm`. 4

Hadley, C. and Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**(9), 1099–1112. 71

Hahn, M. W., Demuth, J. P., and Han, S.-G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics*, **177**(3), 1941–1949. 12, 186

Hasegawa, H. and Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**(3), 341–348. 108

Hauptman, H. A. (1991). The phase problem of x-ray crystallography. *Rep. Prog. Phys.*, pages 1427–1454. 9

He, Y., Chen, Y., Alexander, P. A., Bryan, P. N., and Orban, J. (2012). Mutational tipping points for switching protein folds and functions. *Structure*, **20**(2), 283–291. 18, 156

Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**(23), 2971–2. 186

Hill, J. R., Kelm, S., Shi, J., and Deane, C. M. (2011). Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, **27**(13), i15–i23. 27

Hollup, S. M., Sadowski, M. I., Jonassen, I., and Taylor, W. R. (2011). Exploring the limits of fold discrimination by structural alignment: a large scale benchmark using decoys of known fold. *Comput. Biol. Chem.*, **35**(3), 174–188. 108

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138. 106

Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, **273**(5275), 595–603. 19, 107

Hou, J., Sims, G. E., Zhang, C., and Kim, S.-H. (2003). A global representation of the protein fold space. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(5), 2386–2390. 85, 106

Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(10), 3651–3656. 76

Hughey, R., Karplus, K., and Krogh, A. (2003). SAM: Sequence Alignment and Modeling Software System. `http://compbio.soe.ucsc.edu/papers/sam_doc/node4.html`. Accessed: 15/08/2014. 30

Hutchinson, E. G. and Thornton, J. M. (1993). The Greek key motif: extraction, classification and analysis. *Protein Eng.*, **6**(3), 233–245. 84, 98

Hutchinson, E. G. and Thornton, J. M. (1996). PROMOTIF–a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220. 80, 84

Irving, J. A., Whisstock, J. C., and Lesk, A. M. (2001). Protein structural alignments and functional genomics. *Proteins*, **42**(3), 378–382. 108

Jeltsch, A. (1999). Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.*, **49**(1), 161–164. 12

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**(10), 846–856. 29, 43, 167

Karplus, K., Karchin, R., Shackelford, G., and Hughey, R. (2005). Calibrating E-values for hidden Markov models using reverse-sequence null models. *Bioinformatics*, **21**(22), 4107–4115. 167

Kedem, K., Chew, L. P., and Elber, R. (1999). Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, **37**(4), 554–564. 109, 110

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, **30**(1), 81–93. 64

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., and Wyckoff, H. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662–666. 8

Kessel, A. and Ben-Tal, N. (2011). *Introduction to proteins: structure, function and motion*. CRC Press. 3, 5, 7, 8

Kim, H. S., Mittenthal, J. E., and Caetano-Anollés, G. (2006). MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinform.*, **7**(351). 68

Kim, K. M. and Caetano-Anollés, G. (2011). The proteomic complexity and rise of the primordial ancestor of diversified life The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evol. Biol.*, **11**(140). 48, 67, 68

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**, 624–626. 13

Kluge, A. (1969). Quantitative phyletics and the evolution of anurans. *Syst. Biol.*, **18**(1), 1–32. 45

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.*, **323**(2), 297–307. 75

Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, **420**(6912), 218–223. 32, 76

Kyrpides, N. (1999). Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**(9), 773–774. 49

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**(1), 105–132. 4, 82, 93

Levitt, M. (2007). Growth of novel protein structural data. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(9), 3183–3188. 156

Lin, J. and Gerstein, M. (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. 38, 42, 156

Liu, W., Srivastava, A., and Zhang, J. (2011). A mathematical framework for protein structure comparison. *PLoS Comput. Biol.*, **7**(2), e1001075. 108, 109, 115, 116, 190

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**(15), 3255–3263. 165

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden markov models. *Bioinformatics*, **24**(22), 2630–2631. 161, 163, 167, 169

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**(1), 50–60. 79, 129

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS One*, **6**(12), e28766. 15, 156

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., and Koonin, E. V. (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**(2). 39, 46, 47, 55, 56, 65

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**(7), 617–623. 82, 83, 175, 177, 180

Mosyak, L., Zhang, Y., Glasfeld, E., Haney, S., Stahl, M., Seehra, J., and Somers, W. S. (2000). The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *EMBO J.*, **19**(13), 3179–3191. 152

Murray, R. G. and Schleifer, K. H. (1994). Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. *Int. J. Syst. Bacteriol.*, **44**(1), 174–6. 51

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.*, pages 380–387. 15, 16

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**(4), 536–540. 19, 20, 33, 104

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2014). Global view of the protein universe. *Proc. Natl. Acad. Sci. U.S.A.*, **111**(32), 11691–11696. 106

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, **246**, 96–98. 13

Omland, K. E. (1999). The assumptions and challenges of ancestral state reconstructions. *Syst. Biol.*, **48**(3), 604–611. 65

Orengo, C. and Taylor, W. (1993). A local alignment method for protein structure motifs. *J. Mol. Biol.*, **233**, 488–497. 24

Orengo, C., Jones, D., and Thornton, J. (1994). Protein superfamilies and domain superfolds. *Nature*. 76, 86

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. (1997). CATH– a hierarchic classification of protein domain structures. *Structure*, **5**(8), 1093–1108. 20, 23, 104, 174

Ortiz, A. R., Strauss, C. E. M., and Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Sci.*, **11**, 2606–2621. 109

Osadchy, M. and Kolodny, R. (2011). Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl. Acad. Sci. U.S.A.*, **208**(30), 12301–12306. 107, 153

Overington, J., Donnelly, D., and Johnson, M. (1992). Environment specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci.*, pages 216–226. 91

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.*, **53**(5), 673–684. 186

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nat. Rev. Genet.*, **7**(5), 337–348. 12

Pascual-García, A., Abia, D., Ortiz, A. R., and Bastolla, U. (2009). Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.*, **5**(3), e1000331. 20

Passner, J. M., Schultz, S. C., and Steitz, T. A. (2000). Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 A resolution. *J. Mol. Biol.*, **304**, 847–859. 6

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling–a review. *Gene*, **238**(1), 103–114. 12

Peng, K., Obradovic, Z., and Vucetic, S. (2004). Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, **446**, 435–446. 183

Plaxco, K. W., Simons, K. T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**(4), 985–994. 82

Ponting, C. P. and Russell, R. R. (2002). The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 45–71. 38, 42

Ptitsyn, O. B. and Finkelstein, A. V. (1980). Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys.*, **13**(3), 339–386. 33

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and

Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res.*, **40**(Database issue), D290–D301. 25

Rawiso, M. (1999). From intensity to structure in physical chemistry of polymers. *J. Phys. IV*, **9**(P1), 147–195. 81

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**. 15, 32, 86

Rito, T., Deane, C. M., and Reinert, G. (2012). The importance of age and high degree, in protein-protein interaction networks. *J. Comput. Biol.*, **19**(6), 785–795. 96

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**(4), 401–407. 84

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147. 54, 63

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., and Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Briefings Bioinf.*, **6**(2), 118–134. 39, 48, 66

Ross, C. A. and Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nat. Med.*, **10**(Suppl), S10–S17. 11

Sadowski, M. I. and Taylor, W. R. (2010). On the evolutionary origins of "Fold Space Continuity": a study of topological convergence and divergence in mixed alpha-beta domains. *J. Struct. Biol.*, **172**(3), 244–252. 33, 91, 104

Sadowski, M. I. and Taylor, W. R. (2012). Evolutionary inaccuracy of pairwise structural alignments. *Bioinformatics*, **28**(9), 1209–1215. 108, 153

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**(4), 406–425. 44, 45

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol.*, **9**(5), e1003063. 183, 188

Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1. 10

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**(11), 2498–2504. 133

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. 4th edition. 79

Shindyalov, I. N. and Bourne, P. E. (2000). An alternative view of protein fold space. *Proteins*, **38**(3), 247–260. 104

Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**(9), 776–785. 110

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. 46

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7), 951–960. 157

Stirk, H., Woolfson, D., Hutchinson, E., and Thornton, J. (1992). Depicting topology and handedness in jellyroll structures. *FEBS Lett.*, **308**(1), 1–3. 98

Sweet, R. M. and Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.*, **171**(4), 479–488. 82

Taylor, W. R. (2002). A 'periodic table' for protein structures. *Nature*, **416**(6881), 657–660. 107

Toll-Riera, M., Bostick, D., Albà, M. M., and Plotkin, J. B. (2012). Structure and age jointly influence rates of protein evolution. *PLoS Comput. Biol.*, **8**(5), e1002542. 41, 42, 43, 100

Trifonov, E. N. (2004). The triplet code from first principles. *J. Biomol. Struct. Dyn.*, **22**(1), 1–11. 76, 94

Valas, R. E., Yang, S., and Bourne, P. E. (2009). Nothing about protein structure classification makes sense except in the light of evolution. *Curr. Opin. Struct. Biol.*, **19**(3), 329–334. 41

Wang, M., Boca, S. M., Kalelkar, R., Mittenthal, J. E., and Caetano-Anollés, G. (2006). A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. *Complexity*, **12**(1), 27–40. 48

Wang, M., Kurland, C., and Caetano-Anollés, G. (2011). Reductive evolution of proteomes and protein structures. *Proc. Natl. Acad. Sci. U.S.A.*, **108**(29), 119954–11958. 75

Wilkins, A. D., Bachman, B. J., Erdin, S., and Lichtarge, O. (2012). The use of evolutionary patterns in protein annotation. *Curr. Opin. Struct. Biol.*, **22**(3), 316–325. 186

Winstanley, H. F., Abeln, S., and Deane, C. M. (2005). How old is your fold? *Bioinformatics*, **21**(Suppl 1), i449–458. 41, 43, 46, 68, 76, 85, 153

Wong, J. W. H., Ho, S. Y. W., and Hogg, P. J. (2011). Disulfide bond acquisition through eukaryotic protein evolution. *Mol. Biol. Evol.*, **28**(1), 327–334. 92

Worth, C. L., Gong, S., and Blundell, T. L. (2009). Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.*, **10**(10), 709–720. 11, 13

Wrabl, J. O. and Grishin, N. V. (2008). Statistics of random protein superpositions: p-values for pairwise structure alignment. *J. Comput. Biol.*, **15**(3), 317–355. 108

Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**(7), 889–895. 124

Yaffe, M. (2005). X-ray crystallography and structural biology. *Crit. Care Med.*, **22**, S435–S440. 9

Yang, A.-S. and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**(3), 665–678. 104

Yang, S. and Bourne, P. E. (2009). The evolutionary history of protein domains viewed by species phylogeny. *PloS One*, **4**(12), e8378. 46

Ye, Y. and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(Suppl 2), ii246–ii255. 109, 111, 112, 113, 131

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.*, **18**(6), 292–298. 12

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4), 702–710. 108

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**(7), 2302–2309. 109, 113, 114

Appendix A: Material supplementary to Chapter 1

Table A1: List of superfamilies under SCOP (1.75). Superfamlies are given for reference by their SCOP concise classification strings (sccs) as an identifier which is used throughout this thesis and their names according to the scheme

SCCS	Superfamily
a.1.1	Globin-like
a.1.2	alpha-helical ferredoxin
a.2.1	GreA transcript cleavage protein, N-terminal domain
a.2.2	Ribosomal protein L29 (L29p)
a.2.3	Chaperone J-domain
a.2.5	Prefoldin
a.2.6	HR1 repeat
a.2.7	tRNA-binding arm
a.2.8	Eukaryotic DNA topoisomerase I, dispensable insert domain
a.2.9	C-terminal UvrC-binding domain of UvrB
a.2.10	Epsilon subunit of F1F0-ATP synthase C-terminal domain
a.2.11	Fe,Mn superoxide dismutase (SOD), N-terminal domain
a.2.12	Sporulation inhibitor Sda
a.2.13	Transcriptional repressor TraM
a.2.14	DnaK suppressor protein DksA, alpha-hairpin domain
a.2.15	ISY1 domain-like
a.2.16	Calcyclin-binding protein-like
a.2.17	Endosomal sorting complex assembly domain
a.2.18	SPy1572-like
a.2.19	Rabenosyn-5 Rab-binding domain-like
a.2.20	MxiH-like
a.2.21	YnzC-like
a.3.1	Cytochrome c
a.4.1	Homeodomain-like
a.4.2	Methylated DNA-protein cysteine methyltransferase, C-terminal domain
a.4.3	ARID-like
a.4.5	"Winged helix" DNA-binding domain
a.4.6	C-terminal effector domain of the bipartite response regulators
a.4.7	Ribosomal protein L11, C-terminal domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.4.8	Ribosomal protein S18
a.4.9	Polynucleotide phosphorylase/guanosine pentaphosphate synthase (PNPase/GPSI), domain 3
a.4.10	N-terminal Zn binding domain of HIV integrase
a.4.11	RNA polymerase subunit RPB10
a.4.12	TrpR-like
a.4.13	Sigma3 and sigma4 domains of RNA polymerase sigma factors
a.4.14	KorB DNA-binding domain-like
a.4.15	Rps17e-like
a.5.1	DNA helicase RuvA subunit, C-terminal domain
a.5.2	UBA-like
a.5.3	CRAL/TRIO N-terminal domain
a.5.4	Elongation factor TFIIS domain 2
a.5.6	Double-stranded DNA-binding domain
a.5.7	post-HMGL domain-like
a.5.8	Hypothetical protein AF0491, middle domain
a.5.9	HBS1-like domain
a.5.10	FGAM synthase PurL, linker domain
a.6.1	Putative DNA-binding domain
a.7.1	Spectrin repeat
a.7.2	Enzyme IIa from lactose specific PTS, IIa-lac
a.7.3	Succinate dehydrogenase/fumarate reductase flavoprotein C-terminal domain
a.7.4	Smac/diablo
a.7.5	Tubulin chaperone cofactor A
a.7.6	Ribosomal protein S20
a.7.7	BAG domain
a.7.8	GAT-like domain
a.7.10	Glycogen synthesis protein GlgS
a.7.11	Alpha-hemoglobin stabilizing protein AHSP
a.7.12	PhoU-like
a.7.13	XseB-like
a.7.14	MIT domain
a.7.15	PPK N-terminal domain-like
a.7.16	MIT domain-like
a.7.17	Efb C-domain-like
a.8.1	Bacterial immunoglobulin/albumin-binding domains
a.8.2	Plasmid maintenance system epsilon/zeta, antidote epsilon subunit
a.8.3	Families 57/38 glycoside transferase middle domain
a.8.4	Heat shock protein 70kD (HSP70), C-terminal subdomain
a.8.5	Phosphoprotein XD domain
a.8.6	Staphylocoagulase
a.8.7	Typo IV secretion system protein TraC
a.8.8	Avirulence protein AvrPto
a.8.9	Coronavirus NSP7-like
a.8.10	Vng1086c-like
a.8.11	AF1782-like
a.9.1	Peripheral subunit-binding domain of 2-oxo acid dehydrogenase complex
a.10.1	Protozoan pheromone proteins
a.10.2	Hypothetical membrane protein Ta0354, soluble domain
a.11.1	Acyl-CoA binding protein
a.11.2	Second domain of FERM
a.12.1	Kix domain of CBP (creb binding protein)
a.13.1	RAP domain-like
a.14.1	VHP, Villin headpiece domain
a.15.1	TAF(II)230 TBP-binding fragment
a.16.1	S15/NS1 RNA-binding domain
a.17.1	Cysteine alpha-hairpin motif

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.18.1	T4 endonuclease V
a.19.1	Fertilization protein
a.20.1	PGBD-like
a.21.1	HMG-box
a.22.1	Histone-fold
a.23.1	HSC20 (HSCB), C-terminal oligomerisation domain
a.23.2	Diol dehydratase, gamma subunit
a.23.3	Methane monooxygenase hydrolase, gamma subunit
a.23.4	Mitochondrial import receptor subunit Tom20
a.23.5	Hemolysin expression modulating protein HHA
a.23.6	EF2458-like
a.23.7	YvfG-like
a.24.1	Apolipoprotein
a.24.2	Aspartate receptor, ligand-binding domain
a.24.3	Cytochromes
a.24.4	Hemerythrin-like
a.24.5	TMV-like viral coat proteins
a.24.7	FKBP12-rapamycin-binding domain of FKBP-rapamycin-associated protein (FRAP)
a.24.8	Proteasome activator
a.24.9	alpha-catenin/vinculin-like
a.24.10	Histidine-containing phosphotransfer domain, HPT domain
a.24.11	Bacterial GAP domain
a.24.12	Outer surface protein C (OspC)
a.24.13	Domain of the SRP/SRP receptor G-proteins
a.24.14	FAT domain of focal adhesion kinase
a.24.15	FAD-dependent thiol oxidase
a.24.16	Nucleotidyltransferase substrate binding subunit/domain
a.24.17	Group V grass pollen allergen
a.24.18	Oxygen-evolving enhancer protein 3,
a.24.19	Flagellar export chaperone FliS
a.24.20	Colicin D immunity protein
a.24.21	RecG, N-terminal domain
a.24.22	Nickel-containing superoxide dismutase, NiSOD
a.24.23	Mannose-6-phosphate receptor binding protein 1 (Tip47), C-terminal domain
a.24.24	Domain from hypothetical 2610208m17rik protein
a.24.25	TrmE connector domain
a.24.26	YppE-like
a.24.27	MW0975(SA0943)-like
a.24.28	VPS28 C-terminal domain-like
a.24.29	TM1646-like
a.25.1	Ferritin-like
a.25.2	Cobalamin adenosyltransferase-like
a.25.3	EsxAB dimer-like
a.25.4	PE/PPE dimer-like
a.25.5	HP0062-like
a.25.6	SO2669-like
a.26.1	4-helical cytokines
a.27.1	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
a.28.1	ACP-like
a.28.2	Colicin E immunity proteins
a.28.3	Retrovirus capsid dimerization domain-like
a.29.2	Bromodomain
a.29.3	Acyl-CoA dehydrogenase C-terminal domain-like
a.29.5	alpha-ketoacid dehydrogenase kinase, N-terminal domain
a.29.6	Plant invertase/pectin methylesterase inhibitor
a.29.7	Mob1/phocein

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.29.8	Bacteriocin immunity protein-like
a.29.9	LemA-like
a.29.10	MAST3 pre-PK domain-like
a.29.11	PA2201 N-terminal domain-like
a.29.12	Nqo1C-terminal domain-like
a.29.13	Bacillus cereus metalloprotein-like
a.29.14	Ta0600-like
a.29.15	DsbB-like
a.29.16	IVS-encoded protein-like
a.29.17	YqcC-like
a.30.1	ROP protein
a.30.2	Homodimeric domain of signal transducing histidine kinase
a.30.3	Nonstructural protein ns2, Nep, M1-binding domain
a.30.4	Dimerisation domain of CENP-B
a.30.5	Hypothetical protein D-63
a.30.6	HP1531-like
a.30.7	BAS1536-like
a.30.8	FHV B2 protein-like
a.31.1	Dimerization-anchoring domain of cAMP-dependent PK regulatory subunit
a.32.1	Transcription factor IIA (TFIIA), alpha-helical domain
a.33.1	Ectatomin subunits
a.34.1	SinR repressor dimerisation domain-like
a.34.2	Dimerization cofactor of HNF-1 alpha
a.34.3	Docking domain A of the erythromycin polyketide synthase (DEBS)
a.34.4	Phenylalanine zipper
a.35.1	lambda repressor-like DNA-binding domains
a.36.1	Signal peptide-binding domain
a.37.1	A DNA-binding domain in eukaryotic transcription factors
a.38.1	HLH, helix-loop-helix DNA-binding domain
a.38.2	Docking domain B of the erythromycin polyketide synthase (DEBS)
a.39.1	EF-hand
a.39.2	Insect pheromone/odorant-binding proteins
a.39.3	Cloroperoxidase
a.39.4	Hypothetical protein MTH865
a.40.1	Calponin-homology domain, CH-domain
a.40.2	X-Prolyl dipeptidyl aminopeptidase PepX, N-terminal domain
a.40.3	Hook domain
a.41.1	Domain of poly(ADP-ribose) polymerase
a.42.1	SWIB/MDM2 domain
a.43.1	Ribbon-helix-helix
a.45.1	GST C-terminal domain-like
a.46.1	Methionine synthase domain
a.46.2	Nucleoside phosphorylase/phosphoribosyltransferase N-terminal domain
a.46.3	TM0693-like
a.47.1	STAT
a.47.2	t-snare proteins
a.47.3	Cag-Z
a.47.4	CAPPD, an extracellular domain of amyloid beta A4 protein
a.47.5	FlgN-like
a.47.6	MukF C-terminal domain-like
a.48.1	N-terminal domain of cbl (N-cbl)
a.48.2	Transferrin receptor-like dimerisation domain
a.48.3	Conserved domain common to transcription factors TFIIS, elongin A, CRSP70
a.48.4	HIV integrase-binding domain
a.48.5	PG0775 C-terminal domain-like
a.49.1	C-terminal domain of B transposition protein

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.50.1	Anaphylotoxins (complement system)
a.51.1	Cytochrome c oxidase subunit h
a.52.1	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin
a.53.1	p53 tetramerization domain
a.54.1	Domain of early E2A DNA-binding protein, ADDBP
a.55.1	IHF-like DNA-binding proteins
a.56.1	CO dehydrogenase ISP C-domain like
a.57.1	Protein HNS-dependent expression A; HdeA
a.58.1	Chemotaxis receptor methyltransferase CheR, N-terminal domain
a.59.1	PAH2 domain
a.60.1	SAM/Pointed domain
a.60.2	RuvA domain 2-like
a.60.3	C-terminal domain of RNA polymerase alpha subunit
a.60.4	Rad51 N-terminal domain-like
a.60.5	Barrier-to-autointegration factor, BAF
a.60.6	DNA polymerase beta, N-terminal domain-like
a.60.7	5 to 3 exonuclease, C-terminal subdomain
a.60.8	HRDC-like
a.60.9	lambda integrase-like, N-terminal domain
a.60.10	Enzyme I of the PEP:sugar phosphotransferase system HPr-binding (sub)domain
a.60.11	Hypothetical protein YjbJ
a.60.12	PsbU/PolX domain-like
a.60.13	Putative methyltransferase TM0872, insert domain
a.60.14	eIF2alpha middle domain-like
a.60.15	YozE-like
a.60.16	GspK insert domain-like
a.61.1	Retroviral matrix proteins
a.62.1	Hepatitis B viral capsid (hbcag)
a.63.1	Apolipophorin-III
a.64.1	Saposin
a.64.2	Bacteriocin AS-48
a.65.1	Annexin
a.66.1	Transducin (alpha subunit), insertion domain
a.68.1	Wiscott-Aldrich syndrome protein, WASP, C-terminal domain
a.69.1	C-terminal domain of alpha and beta subunits of F1 ATP synthase
a.69.2	Ypt/Rab-GAP domain of gyp1p
a.69.3	1-deoxy-D-xylulose-5-phosphate reductoisomerase, C-terminal domain
a.69.4	BH3980-like
a.70.1	N-terminal domain of the delta subunit of the F1F0-ATP synthase
a.70.2	AF1862-like
a.71.1	ERP29 C domain-like
a.71.2	Helical domain of Sec23/24
a.72.1	Functional domain of the splicing factor Prp18
a.73.1	Retrovirus capsid protein, N-terminal core domain
a.74.1	Cyclin-like
a.75.1	Ribosomal protein S7
a.76.1	Iron-dependent repressor protein, dimerization domain
a.77.1	DEATH domain
a.78.1	GntR ligand-binding domain-like
a.79.1	NusB-like
a.80.1	post-AAA+ oligomerization domain-like
a.81.1	N-terminal domain of DnaB helicase
a.83.1	Guanido kinase N-terminal domain
a.84.1	Scaffolding protein gpD of bacteriophage procapsid
a.85.1	Hemocyanin, N-terminal domain
a.86.1	Di-copper centre-containing domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.87.1	DBL homology domain (DH-domain)
a.88.1	LigA subunit of an aromatic-ring-opening dioxygenase LigAB
a.89.1	Methyl-coenzyme M reductase alpha and beta chain C-terminal domain
a.90.1	Transcription factor STAT-4 N-domain
a.91.1	Regulator of G-protein signaling, RGS
a.92.1	Carbamoyl phosphate synthetase, large subunit connection domain
a.93.1	Heme-dependent peroxidases
a.94.1	Ribosomal protein L19 (L19e)
a.95.1	Influenza virus matrix protein M1
a.96.1	DNA-glycosylase
a.97.1	An anticodon-binding domain of class I aminoacyl-tRNA synthetases
a.98.1	R1 subunit of ribonucleotide reductase, N-terminal domain
a.99.1	Cryptochrome/photolyase FAD-binding domain
a.100.1	6-phosphogluconate dehydrogenase C-terminal domain-like
a.101.1	Uteroglobin-like
a.102.1	Six-hairpin glycosidases
a.102.2	Seven-hairpin glycosidases
a.102.3	Chondroitin AC/alginate lyase
a.102.4	Terpenoid cyclases/Protein prenyltransferases
a.102.5	Family 10 polysaccharide lyase
a.102.6	LanC-like
a.103.1	Citrate synthase
a.104.1	Cytochrome P450
a.108.1	Ribosomal protein L7/12, oligomerisation (N-terminal) domain
a.109.1	Class II MHC-associated invariant chain ectoplasmic trimerization domain
a.110.1	Aldehyde ferredoxin oxidoreductase, C-terminal domains
a.111.1	Acid phosphatase/Vanadium-dependent haloperoxidase
a.113.1	DNA repair protein MutS, domain III
a.114.1	Interferon-induced guanylate-binding protein 1 (GBP1), C-terminal domain
a.115.1	A virus capsid protein alpha-helical domain
a.116.1	GTPase activation domain, GAP
a.117.1	Ras GEF
a.118.1	ARM repeat
a.118.3	Sec7 domain
a.118.4	Lipovitellin-phosvitin complex, superhelical domain
a.118.5	Bacterial muramidases
a.118.6	Protein prenylyltransferase
a.118.7	14-3-3 protein
a.118.8	TPR-like
a.118.9	ENTH/VHS domain
a.118.11	Cytochrome c oxidase subunit E
a.118.12	Ran-GTPase activating protein 1 (RanGAP1), C-terminal domain
a.118.13	Arp2/3 complex 16 kDa subunit ARPC5
a.118.14	FliG
a.118.15	Aconitase B, N-terminal domain
a.118.16	Translin
a.118.17	Cullin repeat-like
a.118.18	HCP-like
a.118.19	C-terminal domain of Ku80
a.118.20	Hypothetical protein ST1625
a.118.21	Chemosensory protein Csp2
a.118.22	IP3 receptor type 1 binding core, domain 2
a.118.23	USP8 N-terminal domain-like
a.118.24	Pseudo ankyrin repeat-like
a.118.25	TROVE domain-like
a.118.26	MgtE N-terminal domain-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.119.1	Lipoxigenase
a.120.1	gene 59 helicase assembly protein
a.121.1	Tetracyclin repressor-like, C-terminal domain
a.123.1	Nuclear receptor ligand-binding domain
a.124.1	Phospholipase C/P1 nuclease
a.126.1	Serum albumin-like
a.127.1	L-aspartase-like
a.128.1	Terpenoid synthases
a.129.1	GroEL equatorial domain-like
a.130.1	Chorismate mutase II
a.131.1	Peridinin-chlorophyll protein
a.132.1	Heme oxygenase-like
a.133.1	Phospholipase A2, PLA2
a.134.1	Fungal elicitin
a.135.1	Tetraspanin
a.136.1	FinO-like
a.137.1	Ribosomal protein L39e
a.137.2	Methanol dehydrogenase subunit
a.137.3	Transducin (heterotrimeric G protein), gamma chain
a.137.4	Fe-only hydrogenase smaller subunit
a.137.5	Moesin tail domain
a.137.7	Proteinase A inhibitor IA3
a.137.8	Epsilon subunit of mitochondrial F1F0-ATP synthase
a.137.9	Quinohemoprotein amine dehydrogenase C chain
a.137.10	Stathmin
a.137.11	Anti-sigma factor FlgM
a.137.12	Glu-tRNAGln amidotransferase C subunit
a.137.13	RelB-like
a.137.14	Lag-3 N-terminal region
a.137.15	Lipase chaperone-like
a.138.1	Multiheme cytochromes
a.139.1	Type I dockerin domain
a.140.1	LEM domain
a.140.2	SAP domain
a.140.3	Rho N-terminal domain-like
a.140.4	Recombination endonuclease VII, C-terminal and dimerization domains
a.140.5	DNA-binding domain of EIN3-like
a.140.6	PRP4-like
a.141.1	Frizzled cysteine-rich domain
a.142.1	PTS-regulatory domain, PRD
a.143.1	RPB6/omega subunit-like
a.144.1	PABC (PABP) domain
a.144.2	Ribosomal protein L20
a.145.1	Flagellar transcriptional activator FlhD
a.146.1	Telomeric repeat binding factor (TRF) dimerisation domain
a.147.1	Bcr-Abl oncoprotein oligomerization domain
a.148.1	Arp2/3 complex 21 kDa subunit ARPC3
a.149.1	RNase III domain-like
a.150.1	Anti-sigma factor AsiA
a.151.1	Glutamyl tRNA-reductase dimerization domain
a.152.1	AhpD-like
a.153.1	Nuclear receptor coactivator interlocking domain
a.154.1	Variable surface antigen VlsE
a.155.1	H-NS histone-like proteins
a.156.1	S13-like H2TH domain
a.157.1	Skp1 dimerisation domain-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.158.1	F-box domain
a.159.1	Protein serine/threonine phosphatase 2C, C-terminal domain
a.159.2	FF domain
a.159.3	B-form DNA mimic Ocr
a.159.4	DEK C-terminal domain
a.159.5	IscX-like
a.160.1	PAP/OAS1 substrate-binding domain
a.161.1	beta-catenin-interacting protein ICAT
a.162.1	Pre-protein crosslinking domain of SecA
a.163.1	Crustacean CHH/MIH/GIH neurohormone
a.164.1	C-terminal domain of DFF45/ICAD (DFF-C domain)
a.165.1	Myosin phosphatase inhibitor 17kDa protein, CPI-17
a.166.1	RuBisCo LSMT C-terminal, substrate-binding domain
a.168.1	SopE-like GEF domain
a.169.1	BEACH domain
a.170.1	BRCA2 helical domain
a.171.1	BRCA2 tower domain
a.172.1	Helical scaffold and wing domains of SecA
a.173.1	Poly A polymerase C-terminal region-like
a.174.1	Double Clp-N motif
a.175.1	Orange carotenoid protein, N-terminal domain
a.176.1	N-terminal domain of bifunctional PutA protein
a.177.1	Sigma2 domain of RNA polymerase sigma factors
a.178.1	Soluble domain of poliovirus core protein 3a
a.179.1	Replisome organizer (g39p helicase loader/inhibitor protein)
a.180.1	N-terminal, cytoplasmic domain of anti-sigmaE factor RseA
a.181.1	Antibiotic binding domain of TipA-like multidrug resistance regulators
a.182.1	GatB/YqeY motif
a.183.1	Nop domain
a.184.1	TorD-like
a.185.1	Gametocyte protein Pfg27
a.186.1	KaiA/RbsU domain
a.187.1	HAND domain of the nucleosome remodeling ATPase ISWI
a.188.1	PWI domain
a.189.1	XPC-binding domain
a.190.1	Flavivirus capsid protein C
a.191.1	Methenyltetrahydrofolate cyclohydrolase-like
a.192.1	N-terminal domain of adenylylcyclase associated protein, CAP
a.193.1	GRIP domain
a.194.1	L27 domain
a.195.1	YutG-like
a.196.1	Invasion protein A (SipA) , C-terminal actin binding domain
a.198.1	YcfC-like
a.199.1	YgfB-like
a.200.1	Hypothetical protein MTH393
a.202.1	Superantigen MAM
a.203.1	Putative anticodon-binding domain of alanyl-tRNA synthetase (AlaRS)
a.204.1	all-alpha NTP pyrophosphatases
a.205.1	Hsp90 co-chaperone CDC37
a.206.1	P40 nucleoprotein
a.207.1	Formin homology 2 domain (FH2 domain)
a.208.1	DhaL-like
a.209.1	ADP-ribosylglycohydrolase
a.210.1	Eukaryotic initiation factor 4f subunit eIF4g, eIF4e-binding domain
a.211.1	HD-domain/PDEase-like
a.212.1	KRAB domain (Kruppel-associated box)

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.213.1	DinB/YfiT-like putative metalloenzymes
a.214.1	NblA-like
a.215.1	A middle domain of Talin 1
a.216.1	I/LWEQ domain
a.217.1	Surp module (SWAP domain)
a.218.1	YgfY-like
a.219.1	Hypothetical protein YhaI
a.220.1	Hypothetical protein At3g22680
a.221.1	Lissencephaly-1 protein (Lis-1, PAF-AH alpha) N-terminal domain
a.222.1	VPS9 domain
a.223.1	Triger factor/SurA peptide-binding domain-like
a.224.1	Glycolipid transfer protein, GLTP
a.225.1	Hypothetical protein MG354
a.226.1	Her-1
a.227.1	ERO1-like
a.228.1	GDNF receptor-like
a.229.1	Hypothetical protein YqbG
a.230.1	YugE-like
a.231.1	EspA/CesA-like
a.232.1	RNA-binding protein She2p
a.233.1	YfbU-like
a.234.1	Hypothetical protein MPN330
a.235.1	ATP-dependent DNA ligase DNA-binding domain
a.236.1	DNA primase DnaG, C-terminal domain
a.237.1	DNA polymerase III theta subunit-like
a.238.1	BAR/IMD domain-like
a.239.1	ChaB-like
a.240.1	BSD domain-like
a.241.1	TraM-like
a.242.1	Dcp2 domain-like
a.243.1	Type III secretion system domain
a.244.1	EF2947-like
a.245.1	EB1 dimerisation domain-like
a.246.1	Hyaluronidase post-catalytic domain-like
a.246.2	TTHA0068-like
a.246.3	FLJ32549 domain-like
a.247.1	YoaC-like
a.248.1	SP0561-like
a.249.1	YfmB-like
a.250.1	IpaD-like
a.251.1	Phage replication organizer domain
a.252.1	Mediator hinge subcomplex-like
a.253.1	AF0941-like
a.254.1	PA2201 C-terminal domain-like
a.255.1	Rv1873-like
a.256.1	RUN domain-like
a.257.1	SipA N-terminal domain-like
a.258.1	PG0816-like
a.259.1	YidB-like
a.260.1	Rhabdovirus nucleoprotein-like
a.261.1	GUN4-like
a.262.1	PriB N-terminal domain-like
a.263.1	DNA terminal protein
a.264.1	Duffy binding domain-like
a.265.1	Fic-like
a.266.1	Indolic compounds 2,3-dioxygenase-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
a.267.1	Topoisomerase V catalytic domain-like
a.268.1	PTPA-like
a.269.1	FtsH protease domain-like
a.270.1	Hermes dimerisation domain
a.271.1	SOCS box-like
a.272.1	YqgQ-like
a.273.1	Orange domain-like
a.274.1	HAMP domain-like
a.275.1	DnaD domain-like
a.276.1	BH2638-like
a.277.1	TAFH domain-like
a.278.1	GINS helical bundle-like
a.279.1	Jann4075-like
a.280.1	RbcX-like
a.281.1	YheA/YmcA-like
a.282.1	RPA2825-like
a.283.1	ENT-like
a.284.1	YejL-like
a.285.1	MtlR-like
a.286.1	Sama2622-like
a.287.1	TerB-like
a.288.1	UraD-Like
a.289.1	Sec63 N-terminal domain-like
a.290.1	PSPTO4464-like
a.291.1	MG296-like
a.292.1	HP0242-like
a.293.1	SMc04008-like
a.294.1	Tex N-terminal region-like
a.295.1	AGR C 984p-like
a.296.1	PMT central region-like
b.1.1	Immunoglobulin
b.1.2	Fibronectin type III
b.1.3	PKD domain
b.1.4	beta-Galactosidase/glucuronidase domain
b.1.5	Transglutaminase, two C-terminal domains
b.1.6	Cadherin-like
b.1.7	Actinoxanthin-like
b.1.8	Cu,Zn superoxide dismutase-like
b.1.9	CBD9-like
b.1.10	Clathrin adaptor appendage domain
b.1.11	PapD-like
b.1.12	Purple acid phosphatase, N-terminal domain
b.1.13	Superoxide reductase-like
b.1.14	Invasin/intimin cell-adhesion fragments
b.1.15	Integrin domains
b.1.16	Lamin A/C globular tail domain
b.1.17	Thiol:disulfide interchange protein DsbD, N-terminal domain (DsbD-alpha)
b.1.18	E set domains
b.1.19	Antigen MPT63/MPB63 (immunoprotective extracellular protein)
b.1.20	Tp47 lipoprotein, middle and C-terminal domains
b.1.21	Fungal immunomodulatory protein, FIP
b.1.22	ASF1-like
b.1.23	ApaG-like
b.1.24	Accessory protein X4 (ORF8, ORF7a)
b.1.25	LEA14-like
b.1.26	ICP-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.1.27	CalX-like
b.1.28	NEAT domain-like
b.2.1	Diphtheria toxin, C-terminal domain
b.2.2	Carbohydrate-binding domain
b.2.3	Bacterial adhesins
b.2.4	Alpha-macroglobulin receptor domain
b.2.5	p53-like transcription factors
b.2.6	Cytochrome f, large domain
b.2.7	Second domain of Mu2 adaptin subunit (ap50) of ap2 adaptor
b.2.8	beta-sandwich domain of Sec23/24
b.2.9	Peptidylarginine deiminase Pad4, middle domain
b.2.10	DR1885-like metal-binding protein
b.3.1	Starch-binding domain-like
b.3.2	Carboxypeptidase regulatory domain-like
b.3.3	VHL
b.3.4	Transthyretin (synonym: prealbumin)
b.3.5	Cna protein B-type domain
b.3.6	Aromatic compound dioxygenase
b.3.7	Hypothetical protein PA1324
b.4.1	HSP40/DnaJ peptide-binding domain
b.5.1	alpha-Amylase inhibitor tendamistat
b.6.1	Cupredoxins
b.6.2	Major surface antigen p30, SAG1
b.7.1	C2 domain (Calcium/lipid-binding domain, CaLB)
b.7.2	Periplasmic chaperone C-domain
b.7.3	PHL pollen allergen
b.7.4	Rab geranylgeranyltransferase alpha-subunit, insert domain
b.7.5	Smr-associated domain-like
b.8.1	TRAF domain-like
b.9.1	Neurophysin II
b.11.1	gamma-Crystallin-like
b.12.1	Lipase/lipooxygenase domain (PLAT/LH2 domain)
b.14.1	Calpain large subunit, middle domain (domain III)
b.15.1	HSP20-like chaperones
b.16.1	Ecotin, trypsin inhibitor
b.17.1	PEBP-like
b.18.1	Galactose-binding domain-like
b.19.1	Viral protein domain
b.20.1	ENV polyprotein, receptor-binding domain
b.21.1	Virus attachment protein globular domain
b.22.1	TNF-like
b.23.1	Spermadhesin, CUB domain
b.23.2	Collagen-binding domain
b.23.3	Acetamidase/Formamidase-like
b.24.1	Hyaluronate lyase-like, C-terminal domain
b.25.1	Osmotin, thaumatin-like protein
b.26.1	SMAD/FHA domain
b.27.1	Soluble secreted chemokine inhibitor, VCCI
b.28.1	Baculovirus p35 protein
b.29.1	Concanavalin A-like lectins/glucanases
b.30.2	Amine oxidase catalytic domain
b.30.5	Galactose mutarotase-like
b.30.6	V-region of surface antigen I/II (SA I/II, PAC)
b.31.1	EV matrix protein
b.32.1	gp9
b.33.1	ISP domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.34.1	C-terminal domain of transcriptional repressors
b.34.2	SH3-domain
b.34.3	Myosin S1 fragment, N-terminal domain
b.34.4	Electron transport accessory proteins
b.34.5	Translation proteins SH3-like domain
b.34.6	Cell growth inhibitor/plasmid maintenance toxic component
b.34.7	DNA-binding domain of retroviral integrase
b.34.8	Fumarylacetoacetate hydrolase, FAH, N-terminal domain
b.34.9	Tudor/PWWP/MBT
b.34.10	Cap-Gly domain
b.34.11	Prokaryotic SH3-related domain
b.34.12	BAH domain
b.34.13	Chromo domain-like
b.34.14	PAZ domain
b.34.15	Hypothetical protein YfhH
b.34.16	Kinase-associated protein B-like
b.34.17	YccV-like
b.34.18	CarD-like
b.34.19	Mib/herc2 domain-like
b.34.20	YorP-like
b.34.21	Plus3-like
b.35.1	GroES-like
b.35.2	SacY-like RNA-binding domain
b.36.1	PDZ domain-like
b.37.1	N-terminal domains of the minor coat protein g3p
b.38.1	Sm-like ribonucleoproteins
b.38.2	YhbC-like, C-terminal domain
b.38.3	GatD N-terminal domain-like
b.38.4	Dom34/Pelota N-terminal domain-like
b.38.5	TrmB C-terminal domain-like
b.39.1	Ribosomal protein L14
b.40.1	Staphylococcal nuclease
b.40.2	Bacterial enterotoxins
b.40.3	TIMP-like
b.40.4	Nucleic acid-binding proteins
b.40.5	Inorganic pyrophosphatase
b.40.6	MOP-like
b.40.7	CheW-like
b.40.8	gp5 N-terminal domain-like
b.40.9	Heme chaperone CcmE
b.40.10	Hypothetical protein YgiW
b.40.11	TM0957-like
b.40.12	NfeD domain-like
b.40.13	BC4932-like
b.40.14	HupF/HypC-like
b.40.15	EutN/CcmL-like
b.40.16	HIN-2000 domain-like
b.41.1	PRC-barrel domain
b.42.1	Cytokine
b.42.2	Ricin B-like lectins
b.42.3	Agglutinin
b.42.4	STI-like
b.42.5	Actin-crosslinking proteins
b.42.6	MIR domain
b.42.7	DNA-binding protein LAG-1 (CSL)
b.42.8	AbfB domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.43.2	FucI/AraA C-terminal domain-like
b.43.3	Translation proteins
b.43.4	Riboflavin synthase domain-like
b.43.5	Riboflavin kinase-like
b.44.1	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
b.44.2	Aminomethyltransferase beta-barrel domain
b.45.1	FMN-binding split barrel
b.45.2	PilZ domain-like
b.45.3	YkvR-like
b.46.1	FMT C-terminal domain-like
b.47.1	Trypsin-like serine proteases
b.48.1	mu transposase, C-terminal domain
b.49.1	N-terminal domain of alpha and beta subunits of F1 ATP synthase
b.49.2	Alanine racemase C-terminal domain-like
b.49.3	Aminopeptidase/glucanase lid domain
b.50.1	Acid proteases
b.51.1	ValRS/IleRS/LeuRS editing domain
b.52.1	Barwin-like endoglucanases
b.52.2	ADC-like
b.53.1	Ribosomal protein L25-like
b.54.1	Core binding factor beta, CBF
b.55.1	PH domain-like
b.55.2	PA2021-like
b.56.1	Transcription factor IIA (TFIIA), beta-barrel domain
b.57.1	Herpes virus serine proteinase, assemblin
b.58.1	PK beta-barrel domain-like
b.59.1	XRCC4, N-terminal domain
b.60.1	Lipocalins
b.61.1	Avidin/streptavidin
b.61.2	beta-Barrel protease inhibitors
b.61.3	D-aminopeptidase, middle and C-terminal domains
b.61.4	Quinohemoprotein amine dehydrogenase A chain, domain 3
b.61.5	Dipeptidyl peptidase I (cathepsin C), exclusion domain
b.61.6	YceI-like
b.61.7	Extracellular hemoglobin linker subunit, receptor domain
b.61.8	YdhA-like
b.62.1	Cyclophilin-like
b.63.1	Oncogene products
b.64.1	Mannose 6-phosphate receptor domain
b.65.1	Rap30/74 interaction domains
b.66.1	Hemopexin-like domain
b.67.1	Tachylectin-2
b.67.2	Arabinanase/levansucrase/invertase
b.67.3	Apyrase
b.68.1	Sialidases
b.68.2	Soluble quinoprotein glucose dehydrogenase
b.68.3	Thermostable phytase (3-phytase)
b.68.4	TolB, C-terminal domain
b.68.5	YWTD domain
b.68.6	Calcium-dependent phosphotriesterase
b.68.7	Tricorn protease N-terminal domain
b.68.8	Fucose-specific lectin
b.68.9	NHL repeat
b.68.10	GyrA/ParC C-terminal domain-like
b.68.11	Kelch motif
b.69.1	Galactose oxidase, central domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.69.2	YVTN repeat-like/Quinoprotein amine dehydrogenase
b.69.3	Nitrous oxide reductase, N-terminal domain
b.69.4	WD40 repeat-like
b.69.5	RCC1/BLIP-II
b.69.6	Clathrin heavy-chain terminal domain
b.69.7	Peptidase/esterase gauge domain
b.69.8	Integrin alpha N-terminal domain
b.69.9	Tricorn protease domain 2
b.69.10	3-carboxy-cis,cis-mucoante lactonizing enzyme
b.69.11	Putative isomerase YbhE
b.69.12	Sema domain
b.69.13	Oligoxyloglucan reducing end-specific cellobiohydrolase
b.69.14	Nucleoporin domain
b.70.1	Quinoprotein alcohol dehydrogenase-like
b.70.2	C-terminal (heme d1) domain of cytochrome cd1-nitrite reductase
b.70.3	DPP6 N-terminal domain-like
b.71.1	Glycosyl hydrolase domain
b.72.1	WW domain
b.72.2	Carbohydrate binding domain
b.72.3	Pym (Within the bgcn gene intron protein, WIBG), N-terminal domain
b.73.1	Head domain of nucleotide exchange factor GrpE
b.74.1	Carbonic anhydrase
b.75.1	Bacteriochlorophyll A protein
b.76.1	Outer surface protein
b.76.2	Histone H3 K4-specific methyltransferase SET7/9 N-terminal domain
b.77.1	Vitelline membrane outer protein-I (VMO-I)
b.77.2	delta-Endotoxin (insectocide), middle domain
b.77.3	Mannose-binding lectins
b.78.1	alpha-D-mannose-specific plant lectins
b.80.1	Pectin lyase-like
b.80.2	Insect cysteine-rich antifreeze protein
b.80.3	Cell-division inhibitor MinC, C-terminal domain
b.80.4	Alpha subunit of glutamate synthase, C-terminal domain
b.80.5	C-terminal domain of adenylylcyclase associated protein
b.80.6	Stabilizer of iron transporter SufD
b.80.7	beta-Roll
b.80.8	Pentapeptide repeat-like
b.81.1	Trimeric LpxA-like enzymes
b.81.2	An insect antifreeze protein
b.81.3	Adhesin YadA, collagen-binding domain
b.81.4	Guanosine diphospho-D-mannose pyrophosphorylase/mannose-6-phosphate isomerase linker domain
b.82.1	RmlC-like cupins
b.82.2	Clavaminate synthase-like
b.82.3	cAMP-binding domain-like
b.82.4	Regulatory protein AraC
b.82.5	TRAP-like
b.82.6	Thiamin pyrophosphokinase, substrate-binding domain
b.82.7	Calcium ATPase, transduction domain A
b.83.1	Fibre shaft of virus attachment proteins
b.84.1	Single hybrid motif
b.84.2	Rudiment single hybrid motif
b.84.3	Duplicated hybrid motif
b.84.4	Ribosomal L27 protein-like
b.85.1	AFP III-like domain
b.85.2	Head decoration protein D (gpD, major capsid protein D)
b.85.3	Urease, beta-subunit

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.85.4	dUTPase-like
b.85.5	Tlp20, baculovirus telokin-like protein
b.85.6	MoeA C-terminal domain-like
b.85.7	SET domain
b.86.1	Hedgehog/intein (Hint) domain
b.87.1	LexA/Signal peptidase
b.88.1	Mss4-like
b.89.1	Cyanovirin-N
b.90.1	Head-binding domain of phage P22 tailspike protein
b.91.1	E2 regulatory, transactivation domain
b.92.1	Composite domain of metallo-dependent hydrolases
b.93.1	Epsilon subunit of F1F0-ATP synthase N-terminal domain
b.94.1	Olfactory marker protein
b.95.1	Ganglioside M2 (gm2) activator
b.96.1	Nicotinic receptor ligand binding domain-like
b.97.1	Cytolysin/lectin
b.98.1	Leukotriene A4 hydrolase N-terminal domain
b.100.1	Sortase
b.101.1	Ribonuclease domain of colicin E3
b.102.1	Methuselah ectodomain
b.103.1	MoeA N-terminal region -like
b.104.1	P-domain of calnexin/calreticulin
b.105.1	Penicillin-binding protein associated domain
b.106.1	Phage tail proteins
b.107.1	Urease metallochaperone UreE, N-terminal domain
b.108.1	Phage fibre proteins
b.109.1	Cell wall binding repeat
b.110.1	Cloacin translocation domain
b.111.1	Small protein B (SmpB)
b.112.1	C-terminal domain of mollusc hemocyanin
b.113.1	N-terminal domain of MutM-like DNA repair proteins
b.114.1	N-utilization substance G protein NusG, insert domain
b.115.1	Calcium-mediated lectin
b.116.1	Viral chemokine binding protein m3
b.117.1	Obg GTP-binding protein N-terminal domain
b.118.1	FAS1 domain
b.119.1	C-terminal autoproteolytic domain of nucleoporin nup98
b.120.1	Tp47 lipoprotein, N-terminal domain
b.121.1	PHM/PNGase F
b.121.2	Group II dsDNA viruses VP
b.121.3	Nucleoplasmin-like core domain
b.121.4	Positive stranded ssRNA viruses
b.121.5	ssDNA viruses
b.121.6	Group I dsDNA viruses
b.121.7	Satellite viruses
b.122.1	PUA domain-like
b.123.1	Hypothetical protein TM1070
b.124.1	HesB-like domain
b.125.1	Prokaryotic lipoproteins and lipoprotein localization factors
b.126.1	Adsorption protein p2
b.127.1	Baseplate structural protein gp8
b.128.1	Hypothetical protein YojF
b.129.1	AbrB/MazE/MraZ-like
b.129.2	AF2212/PG0164-like
b.130.1	Heat shock protein 70kD (HSP70), peptide-binding domain
b.131.1	SPOC domain-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
b.132.1	Supernatant protein factor (SPF), C-terminal domain
b.133.1	Dextranase, N-terminal domain
b.134.1	Smp-1-like
b.135.1	Superantigen (mitogen) Ypm
b.136.1	SspB-like
b.137.1	Rof/RNase P subunit-like
b.138.1	Hydrophobin II, HfbII
b.139.1	Surface presentation of antigens (SPOA)
b.140.1	Replicase NSP9
b.141.1	Bacterial fluorinating enzyme, C-terminal domain
b.142.1	DNA-binding pseudobarrel domain
b.143.1	NAC domain
b.144.1	Trimeric adhesin
b.145.1	AXH domain
b.146.1	Ctag/Cox11
b.147.1	BTV NS2-like ssRNA-binding domain
b.148.1	Coronavirus RNA-binding domain
b.149.1	Beta-galactosidase LacA, domain 3
b.150.1	Putative glucosidase YicI, C-terminal domain
b.151.1	CsrA-like
b.152.1	Flagellar hook protein flgE
b.153.1	PheT/TilS domain
b.154.1	Agglutinin HPA-like
b.155.1	L21p-like
b.156.1	Atu1913-like
b.157.1	Hcp1-like
b.158.1	BH3618-like
b.159.1	Allene oxide cyclase-like
b.159.2	SO1590-like
b.160.1	L,D-transpeptidase catalytic domain-like
b.161.1	PTSIIA/GutA-like
b.162.1	At5g01610-like
b.163.1	Bacteriophage trimeric proteins domain
b.164.1	SARS ORF9b-like
b.165.1	MOSC N-terminal domain-like
b.166.1	MAL13P1.257-like
b.167.1	FimD N-terminal domain-like
b.168.1	HisI-like
b.169.1	MFPT repeat-like
b.170.1	WSSV envelope protein-like
b.171.1	Trm112p-like
b.172.1	YopX-like
b.173.1	NifT/FixU-like
b.174.1	YopT-like
b.175.1	FomD-like
b.176.1	AttH-like
b.177.1	YmcC-like
b.178.1	PA1994-like
c.1.1	Triosephosphate isomerase (TIM)
c.1.2	Ribulose-phoshate binding barrel
c.1.3	Thiamin phosphate synthase
c.1.4	FMN-linked oxidoreductases
c.1.5	Inosine monophosphate dehydrogenase (IMPDH)
c.1.6	PLP-binding barrel
c.1.7	NAD(P)-linked oxidoreductase
c.1.8	(Trans)glycosidases

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
c.1.9	Metallo-dependent hydrolases
c.1.10	Aldolase
c.1.11	Enolase C-terminal domain-like
c.1.12	Phosphoenolpyruvate/pyruvate domain
c.1.13	Malate synthase G
c.1.14	RuBisCo, C-terminal domain
c.1.15	Xylose isomerase-like
c.1.16	Bacterial luciferase-like
c.1.17	Nicotinate/Quinolinate PRTase C-terminal domain-like
c.1.18	PLC-like phosphodiesterases
c.1.19	Cobalamin (vitamin B12)-dependent enzymes
c.1.20	tRNA-guanine transglycosylase
c.1.21	Dihydropteroate synthetase-like
c.1.22	UROD/MetE-like
c.1.23	FAD-linked oxidoreductase
c.1.24	Pyridoxine 5-phosphate synthase
c.1.25	Monomethylamine methyltransferase MtmB
c.1.26	Homocysteine S-methyltransferase
c.1.27	(2r)-phospho-3-sulfolactate synthase ComA
c.1.28	Radical SAM enzymes
c.1.29	GlpP-like
c.1.30	CutC-like
c.1.31	ThiG-like
c.1.32	TM1631-like
c.1.33	EAL domain-like
c.2.1	NAD(P)-binding Rossmann-fold domains
c.3.1	FAD/NAD(P)-binding domain
c.4.1	Nucleotide-binding domain
c.5.1	MurCD N-terminal domain
c.6.1	Glycosyl hydrolases family 6, cellulases
c.6.2	Glycoside hydrolase/deacetylase
c.6.3	PHP domain-like
c.7.1	PFL-like glycyl radical enzymes
c.8.1	Phosphohistidine domain
c.8.2	LeuD/IlvD-like
c.8.3	Carbamoyl phosphate synthetase, small subunit N-terminal domain
c.8.4	PA domain
c.8.5	GroEL apical domain-like
c.8.6	Swiveling domain of dehydratase reactivase alpha subunit
c.8.7	RraA-like
c.8.8	Putative cyclase
c.8.9	FumA C-terminal domain-like
c.8.10	LD-carboxypeptidase A C-terminal domain-like
c.9.1	Barstar-related
c.9.2	Ribosomal protein L32e
c.10.1	RNI-like
c.10.2	L domain-like
c.10.3	Outer arm dynein light chain 1
c.12.1	Ribosomal proteins L15p and L18e
c.13.1	CRAL/TRIO domain
c.13.2	SpoIIaa-like
c.14.1	ClpP/crotonase
c.15.1	BRCT domain
c.16.1	Lumazine synthase
c.17.1	Caspase-like
c.18.1	Uracil-DNA glycosylase-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
c.19.1	FabD/lysophospholipase-like
c.20.1	Initiation factor IF2/eIF5b, domain 3
c.21.1	Ribosomal protein L13
c.22.1	Ribosomal protein L4
c.23.1	CheY-like
c.23.2	Toll/Interleukin receptor TIR domain
c.23.3	Hypothetical protein MTH538
c.23.4	Succinyl-CoA synthetase domains
c.23.5	Flavoproteins
c.23.6	Cobalamin (vitamin B12)-binding domain
c.23.8	N5-CAIR mutase (phosphoribosylaminoimidazole carboxylase, PurE)
c.23.10	SGNH hydrolase
c.23.11	Beta-D-glucan exohydrolase, C-terminal domain
c.23.12	Formate/glycerate dehydrogenase catalytic domain-like
c.23.13	Type II 3-dehydroquinate dehydratase
c.23.14	N-(deoxy)ribosyltransferase-like
c.23.15	Ribosomal protein S2
c.23.16	Class I glutamine amidotransferase-like
c.23.17	Precorrin-8X methylmutase CbiC/CobH
c.24.1	Methylglyoxal synthase-like
c.25.1	Ferredoxin reductase-like, C-terminal NADP-linked domain
c.26.1	Nucleotidylyl transferase
c.26.2	Adenine nucleotide alpha hydrolases-like
c.26.3	UDP-glucose/GDP-mannose dehydrogenase C-terminal domain
c.27.1	Nucleoside phosphorylase/phosphoribosyltransferase catalytic domain
c.28.1	Cryptochrome/photolyase, N-terminal domain
c.30.1	PreATP-grasp domain
c.31.1	DHS-like NAD/FAD-binding domain
c.32.1	Tubulin nucleotide-binding domain-like
c.33.1	Isochorismatase-like hydrolases
c.34.1	Homo-oligomeric flavin-containing Cys decarboxylases, HFCD
c.36.1	Thiamin diphosphate-binding fold (THDP-binding)
c.37.1	P-loop containing nucleoside triphosphate hydrolases
c.38.1	PTS IIb component
c.39.1	Nicotinate mononucleotide:5,6-dimethylbenzimidazole phosphoribosyltransferase (CobT)
c.40.1	Methylesterase CheB, C-terminal domain
c.41.1	Subtilisin-like
c.42.1	Arginase/deacetylase
c.43.1	CoA-dependent acyltransferases
c.44.1	Phosphotyrosine protein phosphatases I
c.44.2	PTS system IIB component-like
c.45.1	(Phosphotyrosine protein) phosphatases II
c.46.1	Rhodanese/Cell cycle control phosphatase
c.47.1	Thioredoxin-like
c.47.2	RNA 3-terminal phosphate cyclase, RPTC, insert domain
c.48.1	TK C-terminal domain-like
c.49.1	PK C-terminal domain-like
c.49.2	ATP synthase (F1-ATPase), gamma subunit
c.50.1	Macro domain-like
c.51.1	Class II aaRS ABD-related
c.51.2	TolB, N-terminal domain
c.51.3	B12-dependent dehydatase associated subunit
c.51.4	ITPase-like
c.51.5	CinA-like
c.51.6	XCC0632-like
c.52.1	Restriction endonuclease-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
c.52.2	tRNA-intron endonuclease catalytic domain-like
c.52.3	Eukaryotic RPB5 N-terminal domain
c.52.4	TBP-interacting protein-like
c.53.1	Resolvase-like
c.53.2	beta-carbonic anhydrase, cab
c.54.1	PTS system fructose IIA component-like
c.55.1	Actin-like ATPase domain
c.55.2	Creatinase/prolidase N-terminal domain
c.55.3	Ribonuclease H-like
c.55.4	Translational machinery components
c.55.5	Nitrogenase accessory factor-like
c.55.6	DNA repair protein MutS, domain II
c.55.7	Methylated DNA-protein cysteine methyltransferase domain
c.56.1	HybD-like
c.56.2	Purine and uridine phosphorylases
c.56.3	Peptidyl-tRNA hydrolase-like
c.56.4	Pyrrolidone carboxyl peptidase (pyroglutamate aminopeptidase)
c.56.5	Zn-dependent exopeptidases
c.56.6	LigB-like
c.56.7	AF0625-like
c.56.8	Cgl1923-like
c.57.1	Molybdenum cofactor biosynthesis proteins
c.58.1	Aminoacid dehydrogenase-like, N-terminal domain
c.59.1	MurD-like peptide ligases, peptide-binding domain
c.60.1	Phosphoglycerate mutase-like
c.61.1	PRTase-like
c.62.1	vWA-like
c.64.1	Pyruvate-ferredoxin oxidoreductase, PFOR, domain III
c.65.1	Formyltransferase
c.66.1	S-adenosyl-L-methionine-dependent methyltransferases
c.67.1	PLP-dependent transferases
c.67.2	PhnH-like
c.67.3	Dhaf3308-like
c.68.1	Nucleotide-diphospho-sugar transferases
c.69.1	alpha/beta-Hydrolases
c.70.1	Nucleoside hydrolase
c.71.1	Dihydrofolate reductase-like
c.72.1	Ribokinase-like
c.72.2	MurD-like peptide ligases, catalytic domain
c.72.3	CoaB-like
c.73.1	Carbamate kinase-like
c.74.1	AraD/HMP-PK domain-like
c.76.1	Alkaline phosphatase-like
c.77.1	Isocitrate/Isopropylmalate dehydrogenase-like
c.78.1	Aspartate/ornithine carbamoyltransferase
c.78.2	Aspartate/glutamate racemase
c.79.1	Tryptophan synthase beta subunit-like PLP-dependent enzymes
c.80.1	SIS domain
c.81.1	Formate dehydrogenase/DMSO reductase, domains 1-3
c.82.1	ALDH-like
c.83.1	Aconitase iron-sulfur domain
c.84.1	Phosphoglucomutase, first 3 domains
c.85.1	FucI/AraA N-terminal and middle domains
c.86.1	Phosphoglycerate kinase
c.87.1	UDP-Glycosyltransferase/glycogen phosphorylase
c.88.1	Glutaminase/Asparaginase

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
c.89.1	Phosphofructokinase
c.90.1	Tetrapyrrole methylase
c.91.1	PEP carboxykinase-like
c.92.1	Chelatase
c.92.2	"Helical backbone" metal receptor
c.92.3	PrpR receptor domain-like
c.93.1	Periplasmic binding protein-like I
c.94.1	Periplasmic binding protein-like II
c.95.1	Thiolase-like
c.96.1	Fe-only hydrogenase
c.97.1	Cytidine deaminase-like
c.97.3	JAB1/MPN domain
c.98.1	MurE/MurF N-terminal domain
c.98.2	HprK N-terminal domain-like
c.99.1	Dipeptide transport protein
c.100.1	Thiamin pyrophosphokinase, catalytic domain
c.101.1	Undecaprenyl diphosphate synthase
c.102.1	Cell-division inhibitor MinC, N-terminal domain
c.103.1	MTH938-like
c.104.1	YjeF N-terminal domain-like
c.105.1	2,3-Bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain
c.106.1	SurE-like
c.107.1	DHH phosphoesterases
c.108.1	HAD-like
c.109.1	PEP carboxykinase N-terminal domain
c.110.1	DTD-like
c.111.1	Activating enzymes of the ubiquitin-like proteins
c.112.1	Glycerol-3-phosphate (1)-acyltransferase
c.113.1	HemD-like
c.114.1	DsrEFH-like
c.115.1	Hypothetical protein MTH777 (MT0777)
c.116.1	alpha/beta knot
c.117.1	Amidase signature (AS) enzymes
c.118.1	GckA/TtuD-like
c.119.1	DAK1/DegV-like
c.120.1	PIN domain-like
c.121.1	Ribose/Galactose isomerase RpiB/AlsB
c.122.1	L-sulfolactate dehydrogenase-like
c.123.1	CoA-transferase family III (CaiB/BaiF)
c.124.1	NagB/RpiA/CoA transferase-like
c.125.1	Creatininase
c.126.1	DNA polymerase III psi subunit
c.127.1	F420-dependent methylenetetrahydromethanopterin dehydrogenase (MTD)
c.128.1	DNA polymerase III chi subunit
c.129.1	MCP/YpsA-like
c.130.1	Alpha-2,3/8-sialyltransferase CstII
c.131.1	Peptidyl-tRNA hydrolase II
c.132.1	Bacterial fluorinating enzyme, N-terminal domain
c.133.1	RbsD-like
c.134.1	LmbE-like
c.135.1	NIF3 (NGG1p interacting factor 3)-like
c.136.1	Toprim domain
c.138.1	Indigoidine synthase A-like
c.140.1	TTHA0583/YokD-like
c.141.1	Glycerate kinase I
c.142.1	Nqo1 FMN-binding domain-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
c.143.1	CofD-like
c.144.1	RibA-like
c.145.1	NadA-like
c.146.1	YgbK-like
c.147.1	CAC2185-like
c.148.1	ComB-like
c.149.1	AtpF-like
c.150.1	EreA/ChaN-like
c.151.1	CobE/GbiG C-terminal domain-like
c.152.1	CbiG N-terminal domain-like
c.153.1	YerB-like
c.154.1	CdCA1 repeat-like
d.1.1	Microbial ribonucleases
d.2.1	Lysozyme-like
d.3.1	Cysteine proteinases
d.4.1	His-Me finger endonucleases
d.5.1	RNase A-like
d.6.1	Prion-like
d.7.1	LysM domain
d.8.1	Urease, gamma-subunit
d.9.1	Interleukin 8-like chemokines
d.9.2	PhtA domain-like
d.10.1	DNA-binding domain
d.11.1	Penicillin-binding protein 2x (pbp-2x), c-terminal domain
d.12.1	Ribosomal proteins S24e, L23 and L15e
d.13.1	HIT-like
d.13.2	Rotavirus NSP2 fragment, C-terminal domain
d.14.1	Ribosomal protein S5 domain 2-like
d.15.1	Ubiquitin-like
d.15.2	CAD & PB1 domains
d.15.3	MoaD/ThiS
d.15.4	2Fe-2S ferredoxin-like
d.15.5	Staphylokinase/streptokinase
d.15.6	Superantigen toxins, C-terminal domain
d.15.7	Immunoglobulin-binding domains
d.15.8	Translation initiation factor IF3, N-terminal domain
d.15.9	Glutamine synthetase, N-terminal domain
d.15.10	TGS-like
d.15.11	Doublecortin (DC)
d.15.12	TmoB-like
d.15.13	Nqo1 middle domain-like
d.15.14	NSP3A-like
d.16.1	FAD-linked reductases, C-terminal domain
d.17.1	Cystatin/monellin
d.17.2	Amine oxidase N-terminal region
d.17.3	DsbC/DsbG N-terminal domain-like
d.17.4	NTF2-like
d.17.5	Uracil-DNA glycosylase inhibitor protein
d.17.6	Pre-PUA domain
d.17.7	Putative dsDNA mimic
d.18.1	ssDNA-binding transcriptional regulator domain
d.19.1	MHC antigen-recognition domain
d.20.1	UBC-like
d.21.1	Diaminopimelate epimerase-like
d.22.1	GFP-like
d.23.1	Tubby C-terminal domain-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.24.1	Pili subunits
d.25.1	Mitochondrial glycoprotein MAM33-like
d.26.1	FKBP-like
d.26.2	Colicin E3 immunity protein
d.26.3	Chitinase insertion domain
d.27.1	Ribosomal protein S16
d.28.1	Ribosomal protein S19
d.29.1	Ribosomal protein L31e
d.30.1	Allophycocyanin linker chain (domain)
d.31.1	Cdc48 domain 2-like
d.32.1	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
d.33.1	SecB-like
d.34.1	DNA-binding domain of Mlu1-box binding protein MBP1
d.35.1	Heme-binding protein A (HasA)
d.36.1	Chalcone isomerase
d.37.1	CBS-domain pair
d.38.1	Thioesterase/thiol ester dehydrase-isomerase
d.39.1	DLC
d.40.1	CI-2 family of serine protease inhibitors
d.41.1	CO dehydrogenase molybdoprotein N-domain-like
d.41.2	Nicotinate/Quinolinate PRTase N-terminal domain-like
d.41.3	Pyrimidine nucleoside phosphorylase C-terminal domain
d.41.4	Ribosomal protein L16p/L10e
d.41.5	Molybdopterin synthase subunit MoaE
d.42.1	POZ domain
d.43.1	Elongation factor Ts (EF-Ts), dimerisation domain
d.43.2	Band 7/SPFH domain
d.44.1	Fe,Mn superoxide dismutase (SOD), C-terminal domain
d.45.1	ClpS-like
d.47.1	Ribosomal L11/L12e N-terminal domain
d.48.1	RecA protein, C-terminal domain
d.49.1	Signal recognition particle alu RNA binding heterodimer, SRP9/14
d.50.1	dsRNA-binding domain-like
d.50.2	Porphobilinogen deaminase (hydroxymethylbilane synthase), C-terminal domain
d.50.3	YcfA/nrd intein domain
d.50.4	Peptidyl-tRNA hydrolase domain-like
d.50.5	Rv2632c-like
d.51.1	Eukaryotic type KH-domain (KH-domain type I)
d.52.1	Alpha-lytic protease prodomain
d.52.2	GMP synthetase C-terminal dimerisation domain
d.52.3	Prokaryotic type KH domain (KH-domain type II)
d.52.4	YhbC-like, N-terminal domain
d.52.5	Probable GTPase Der, C-terminal domain
d.52.6	BolA-like
d.52.7	Ribosome-binding factor A, RbfA
d.52.8	Fe-S cluster assembly (FSCA) domain-like
d.52.9	Cation efflux protein cytoplasmic domain-like
d.52.10	EspE N-terminal domain-like
d.53.1	Ribosomal protein S3 C-terminal domain
d.54.1	Enolase N-terminal domain-like
d.55.1	Ribosomal protein L22
d.56.1	GroEL-intermediate domain like
d.57.1	DNA damage-inducible protein DinI
d.58.1	4Fe-4S ferredoxins
d.58.2	Aspartate carbamoyltransferase, Regulatory-chain, N-terminal domain
d.58.3	Protease propeptides/inhibitors

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.58.4	Dimeric alpha+beta barrel
d.58.5	GlnB-like
d.58.6	Nucleoside diphosphate kinase, NDK
d.58.7	RNA-binding domain, RBD
d.58.8	Viral DNA-binding domain
d.58.9	RuBisCO, large subunit, small (N-terminal) domain
d.58.10	Acylphosphatase/BLUF domain-like
d.58.11	EF-G C-terminal domain-like
d.58.12	eEF-1beta-like
d.58.13	Anticodon-binding domain of PheRS
d.58.14	Ribosomal protein S6
d.58.15	Ribosomal protein S10
d.58.16	PAP/Archaeal CCA-adding enzyme, C-terminal domain
d.58.17	HMA, heavy metal-associated domain
d.58.18	ACT-like
d.58.19	Bacterial exopeptidase dimerisation domain
d.58.20	NAD-binding domain of HMG-CoA reductase
d.58.21	Molybdenum cofactor biosynthesis protein C, MoaC
d.58.22	TRADD, N-terminal domain
d.58.23	Probable ACP-binding domain of malonyl-CoA ACP transacylase
d.58.24	CheY-binding domain of CheA
d.58.25	Killer toxin KP6 alpha-subunit
d.58.26	GHMP Kinase, C-terminal domain
d.58.27	Translational regulator protein regA
d.58.28	Peptide methionine sulfoxide reductase
d.58.29	Nucleotide cyclase
d.58.30	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase, HPPK
d.58.31	Methyl-coenzyme M reductase subunits
d.58.32	FAD-linked oxidases, C-terminal domain
d.58.33	Formylmethanofuran:tetrahydromethanopterin formyltransferase
d.58.34	Formiminotransferase domain of formiminotransferase-cyclodeaminase.
d.58.36	Nitrite/Sulfite reductase N-terminal domain-like
d.58.38	Urease metallochaperone UreE, C-terminal domain
d.58.39	Glutamyl tRNA-reductase catalytic, N-terminal domain
d.58.40	D-ribose-5-phosphate isomerase (RpiA), lid domain
d.58.41	SEA domain
d.58.42	N-utilization substance G protein NusG, N-terminal domain
d.58.43	Mechanosensitive channel protein MscS (YggB), C-terminal domain
d.58.44	Multidrug efflux transporter AcrB pore domain; PN1, PN2, PC1 and PC2 subdomains
d.58.46	eEF1-gamma domain
d.58.47	Hypothetical protein VC0424
d.58.48	MTH1187/YkoF-like
d.58.49	YajQ-like
d.58.50	Hypothetical protein TT1725
d.58.51	eIF-2-alpha, C-terminal domain
d.58.52	Sporulation related repeat
d.58.53	CRISPR-associated protein
d.58.54	YbeD/HP0495-like
d.58.55	DOPA-like
d.58.56	CcmK-like
d.58.57	Transposase IS200-like
d.58.58	TTP0101/SSO1404-like
d.58.59	Rnp2-like
d.58.60	Bacterial polysaccharide co-polymerase-like
d.58.61	MTH889-like
d.58.62	Ribosomal protein L10-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.59.1	Ribosomal protein L30p/L7e
d.60.1	Probable bacterial effector-binding domain
d.61.1	LigT-like
d.62.1	Pepsin inhibitor-3
d.63.1	CYTH-like phosphatases
d.64.1	eIF1-like
d.64.2	TM1457-like
d.65.1	Hedgehog/DD-peptidase
d.66.1	Alpha-L RNA-binding motif
d.67.1	ThrRS/AlaRS common domain
d.67.2	Arginyl-tRNA synthetase (ArgRS), N-terminal additional domain
d.67.3	Ribosome recycling factor, RRF
d.67.4	General secretion pathway protein M, EpsM
d.68.1	Translation initiation factor IF3, C-terminal domain
d.68.2	EPT/RTPC-like
d.68.3	SirA-like
d.68.4	YhbY-like
d.68.5	C-terminal domain of ProRS
d.68.6	AlbA-like
d.68.7	R3H domain
d.68.8	SMR domain-like
d.70.1	Yeast killer toxins
d.71.1	Cell division protein MinE topological specificity domain
d.72.1	Cyanase C-terminal domain
d.73.1	RuBisCO, small subunit
d.74.1	PCD-like
d.74.2	C-terminal domain of arginine repressor
d.74.3	RBP11-like subunits of RNA polymerase
d.74.4	GAD domain-like
d.74.5	PH0987 N-terminal domain-like
d.75.1	tRNA-intron endonuclease N-terminal domain-like
d.75.2	DNA repair protein MutS, domain I
d.76.1	GYF domain
d.76.2	BRK domain-like
d.77.1	RL5-like
d.78.1	RPB5-like RNA polymerase subunit
d.79.1	YjgF-like
d.79.2	Tubulin C-terminal domain-like
d.79.3	L30e-like
d.79.4	PurM N-terminal domain-like
d.79.5	IpsF-like
d.79.6	Holliday junction resolvase RusA
d.79.7	OmpA-like
d.79.8	YueI-like
d.79.9	BB2672-like
d.80.1	Tautomerase/MIF
d.81.1	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain
d.81.2	Serine metabolism enzymes domain
d.81.3	FwdE-like
d.81.4	V-type ATPase subunit E-like
d.82.1	Copper amine oxidase, domain N
d.82.2	Frataxin/Nqo15-like
d.82.3	Hypothetical protein c14orf129, hspc210
d.82.4	GAS2 domain-like
d.82.5	GK1464-like
d.83.1	Bactericidal permeability-increasing protein, BPI

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.83.2	Activator of Hsp90 ATPase, Aha1
d.84.1	Subtilisin inhibitor
d.85.1	RNA bacteriophage capsid protein
d.86.1	eIF4e-like
d.87.1	FAD/NAD-linked reductases, dimerisation (C-terminal) domain
d.87.2	CO dehydrogenase flavoprotein C-terminal domain-like
d.88.1	SRF-like
d.89.1	Origin of replication-binding domain, RBD-like
d.90.1	FMN-dependent nitroreductase-like
d.91.1	N-terminal domain of eukaryotic peptide chain release factor subunit 1, ERF1
d.92.1	Metalloproteases ("zincins"), catalytic domain
d.92.2	beta-N-acetylhexosaminidase-like domain
d.93.1	SH2 domain
d.94.1	HPr-like
d.94.2	Putative transcriptional regulator TM1602, C-terminal domain
d.95.1	Glucose permease domain IIB
d.95.2	Homing endonucleases
d.96.1	Tetrahydrobiopterin biosynthesis enzymes-like
d.96.2	ApbE-like
d.97.1	Cell cycle regulatory proteins
d.98.1	beta-lactamase-inhibitor protein, BLIP
d.98.2	BT0923-like
d.99.1	Ribosomal protein L9 C-domain
d.100.1	L9 N-domain-like
d.100.2	MbtH-like
d.101.1	Ribonuclease PH domain 2-like
d.102.1	Regulatory factor Nef
d.103.1	CytB endotoxin-like
d.104.1	Class II aaRS and biotin synthetases
d.105.1	Subdomain of clathrin and coatomer appendage domain
d.106.1	SCP-like
d.107.1	Mog1p/PsbP-like
d.108.1	Acyl-CoA N-acyltransferases (Nat)
d.109.1	Actin depolymerizing proteins
d.109.2	C-terminal, gelsolin-like domain of Sec23/24
d.109.3	FLJ32549 C-terminal domain-like
d.110.1	Profilin (actin-binding protein)
d.110.2	GAF domain-like
d.110.3	PYP-like sensor domain (PAS domain)
d.110.4	SNARE-like
d.110.5	Pheromone-binding domain of LuxR-like quorum-sensing transcription factors
d.110.6	Sensory domain-like
d.110.7	Roadblock/LC7 domain
d.110.8	YeeU-like
d.110.9	GlcG-like
d.110.10	YNR034W-A-like
d.111.1	PR-1-like
d.112.1	Phoshotransferase/anion transport protein
d.113.1	Nudix
d.114.1	5-nucleotidase (syn. UDP-sugar hydrolase), C-terminal domain
d.115.1	YrdC/RibB
d.116.1	YbaK/ProRS associated domain
d.117.1	Thymidylate synthase/dCMP hydroxymethylase
d.118.1	N-acetylmuramoyl-L-alanine amidase-like
d.120.1	Cytochrome b5-like heme/steroid binding domain
d.121.1	DNA topoisomerase I domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.122.1	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase
d.123.1	Sporulation response regulatory protein Spo0B
d.124.1	Ribonuclease Rh-like
d.125.1	Ornithine decarboxylase C-terminal domain
d.126.1	Pentein
d.127.1	Creatinase/aminopeptidase
d.128.1	Glutamine synthetase/guanido kinase
d.129.1	TATA-box binding protein-like
d.129.2	Phosphoglucomutase, C-terminal domain
d.129.3	Bet v1-like
d.129.4	Cell-division protein ZipA, C-terminal domain
d.129.5	MoaD-related protein, C-terminal domain
d.129.6	KA1-like
d.129.7	TT1751-like
d.129.8	Rbstp2229 protein
d.129.9	RalF, C-terminal domain
d.129.10	YwmB-like
d.129.11	YugN-like
d.130.1	S-adenosylmethionine synthetase
d.131.1	DNA clamp
d.133.1	Molybdenum cofactor-binding domain
d.134.1	Nitrite and sulphite reductase 4Fe-4S domain-like
d.135.1	The spindle assembly checkpoint protein mad2
d.136.1	Phospholipase D/nuclease
d.137.1	Monooxygenase (hydroxylase) regulatory protein
d.139.1	PurM C-terminal domain-like
d.140.1	Ribosomal protein S8
d.141.1	Ribosomal protein L6
d.142.1	Glutathione synthetase ATP-binding domain-like
d.142.2	DNA ligase/mRNA capping enzyme, catalytic domain
d.143.1	SAICAR synthase-like
d.144.1	Protein kinase-like (PK-like)
d.145.1	FAD-binding/transporter-associated domain-like
d.146.1	Uridine diphospho-N-Acetylenolpyruvylglucosamine reductase, MurB, C-terminal domain
d.147.1	Methenyltetrahydromethanopterin cyclohydrolase
d.148.1	Hect, E3 ligase catalytic domain
d.149.1	Nitrile hydratase alpha chain
d.150.1	4-phosphopantetheinyl transferase
d.151.1	DNase I-like
d.152.1	Aldehyde ferredoxin oxidoreductase, N-terminal domain
d.153.1	N-terminal nucleophile aminohydrolases (Ntn hydrolases)
d.153.2	Archaeal IMP cyclohydrolase PurO
d.154.1	DmpA/ArgJ-like
d.155.1	Pyruvoyl-dependent histidine and arginine decarboxylases
d.156.1	S-adenosylmethionine decarboxylase
d.157.1	Metallo-hydrolase/oxidoreductase
d.159.1	Metallo-dependent phosphatases
d.160.1	Carbon-nitrogen hydrolase
d.161.1	ADC synthase
d.162.1	LDH C-terminal domain-like
d.163.1	DNA breaking-rejoining enzymes
d.164.1	SMAD MH1 domain
d.165.1	Ribosome inactivating proteins (RIP)
d.166.1	ADP-ribosylation
d.167.1	Peptide deformylase
d.168.1	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.169.1	C-type lectin-like
d.170.1	SRCR-like
d.170.2	A heparin-binding domain
d.171.1	Fibrinogen C-terminal domain-like
d.172.1	gp120 core
d.173.1	Methionine synthase activation domain-like
d.174.1	Nitric oxide (NO) synthase oxygenase domain
d.175.1	Penicillin binding protein dimerisation domain
d.176.1	Oxidoreductase molybdopterin-binding domain
d.177.1	FAH
d.178.1	Aromatic aminoacid monoxygenases, catalytic and oligomerization domains
d.179.1	Substrate-binding domain of HMG-CoA reductase
d.180.1	Conserved core of transcriptional regulatory protein vp16
d.181.1	Insert subdomain of RNA polymerase alpha subunit
d.182.1	Baseplate structural protein gp11
d.183.1	Major capsid protein gp5
d.184.1	Non-globular alpha+beta subunits of globular proteins
d.185.1	LuxS/MPP-like metallohydrolase
d.186.1	Head-to-tail joining protein W, gpW
d.186.2	XkdW-like
d.187.1	Photosystem I subunit PsaD
d.188.1	Prokaryotic ribosomal protein L17
d.189.1	PX domain
d.190.1	Chorismate lyase-like
d.192.1	YlxR-like
d.193.1	Hsp33 domain
d.194.1	CNF1/YfiH-like putative cysteine hydrolases
d.195.1	YopH tyrosine phosphatase N-terminal domain
d.196.1	Outer capsid protein sigma 3
d.197.1	Protein-L-isoaspartyl O-methyltransferase, C-terminal domain
d.198.1	Type III secretory system chaperone-like
d.198.2	Arp2/3 complex subunits
d.198.3	YjbR-like
d.198.4	YdhG-like
d.198.5	YgaC/TfoX-N like
d.199.1	DNA-binding C-terminal domain of the transcription factor MotA
d.200.1	Integrin beta tail domain
d.201.1	SRP19
d.202.1	Transcription factor NusA, N-terminal domain
d.203.1	DsrC, the gamma subunit of dissimilatory sulfite reductase
d.204.1	Ribosome binding protein Y (YfiA homologue)
d.205.1	GTP cyclohydrolase I feedback regulatory protein, GFRP
d.206.1	YggU-like
d.207.1	Thymidylate synthase-complementing protein Thy1
d.208.1	MTH1598-like
d.209.1	LCCL domain
d.210.1	Argininosuccinate synthetase, C-terminal domain
d.211.1	Ankyrin repeat
d.211.2	Plakin repeat
d.212.1	TolA/TonB C-terminal domain
d.213.1	VSV matrix protein
d.214.1	Hypothetical protein MTH1880
d.215.1	Smc hinge domain
d.216.1	Rotavirus NSP2 fragment, N-terminal domain
d.217.1	SAND domain-like
d.218.1	Nucleotidyltransferase

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.219.1	PP2C-like
d.220.1	Metal cation-transporting ATPase, ATP-binding domain N
d.221.1	Nuclease A inhibitor (NuiA)
d.222.1	YbaB-like
d.223.1	Polo-box domain
d.224.1	SufE/NifU
d.225.1	Multidrug efflux transporter AcrB TolC docking domain; DN and DC subdomains
d.226.1	GIY-YIG endonuclease
d.227.1	OsmC-like
d.228.1	Replication modulator SeqA, C-terminal DNA-binding domain
d.229.1	MesJ substrate recognition domain-like
d.230.1	N-terminal, heterodimerisation domain of RBP7 (RpoE)
d.230.2	Dodecin-like
d.230.3	Amyloid beta a4 protein copper binding domain (domain 2)
d.230.4	D-lysine 5,6-aminomutase beta subunit KamE, N-terminal domain
d.230.5	YbjQ-like
d.230.6	YdgH-like
d.231.1	Receptor-binding domain of short tail fibre protein gp12
d.232.1	Mago nashi protein
d.233.1	Inhibitor of vertebrate lysozyme, Ivy
d.234.1	Proguanylin
d.235.1	FYSH domain
d.236.1	DNA-binding protein Tfx
d.237.1	Hypothetical protein YjiA, C-terminal domain
d.238.1	Hypothetical protein TM0875
d.239.1	GCM domain
d.240.1	Lesion bypass DNA polymerase (Y-family), little finger domain
d.241.1	Translation initiation factor 2 beta, aIF2beta, N-terminal domain
d.241.2	Trigger factor ribosome-binding domain
d.242.1	Obg GTP-binding protein C-terminal domain
d.243.1	Colicin D/E5 nuclease domain
d.244.1	Cell division protein ZapA-like
d.245.1	NSFL1 (p97 ATPase) cofactor p47, SEP domain
d.246.1	mRNA decapping enzyme DcpS N-terminal domain
d.247.1	Chromosomal protein MC1
d.248.1	Coproporphyrinogen III oxidase
d.249.1	Hypothetical protein Ta1206
d.250.1	Folate-binding domain
d.251.1	Hypothetical protein YwqG
d.252.1	CheC-like
d.253.1	Hypothetical protein YoaG
d.254.1	Nucleocapsid protein dimerization domain
d.255.1	Tombusvirus P19 core protein, VP19
d.256.1	Ta1353-like
d.257.1	Hypothetical protein TM0160
d.258.1	Chorismate synthase, AroC
d.259.1	Hypothetical protein HI1480
d.260.1	Suppressor of Fused, N-terminal domain
d.261.1	Hypothetical protein PH1602
d.262.1	NinB
d.263.1	Hypothetical protein Yml108w
d.264.1	Prim-pol domain
d.265.1	Pseudouridine synthase
d.266.1	Hypothetical protein MTH677
d.267.1	Hypothetical protein SAV1430
d.268.1	ParB/Sulfiredoxin

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.269.1	Gamma-glutamyl cyclotransferase-like
d.270.1	2-isopropylmalate synthase LeuA, allosteric (dimerisation) domain
d.271.1	HSP90 C-terminal domain
d.272.1	Dystroglycan, domain 2
d.273.1	YjbQ-like
d.274.1	Hypothetical protein PF0899
d.275.1	Hut operon positive regulatory protein HutP
d.276.1	Hypothetical protein yfbM
d.277.1	Bacillus phage protein
d.278.1	Ligand-binding domain in the NO signalling and Golgi transport
d.279.1	YggX-like
d.280.1	Sulfolobus fructose-1,6-bisphosphatase-like
d.281.1	Hemolytic lectin CEL-III, C-terminal domain
d.282.1	SSo0622-like
d.283.1	Putative modulator of DNA gyrase, PmbA/TldD
d.284.1	PurS-like
d.285.1	DNA-binding domain of intron-encoded endonucleases
d.286.1	TrkA C-terminal domain-like
d.287.1	DNA methylase specificity domain
d.288.1	GTF2I-like repeat
d.289.1	WWE domain
d.290.1	AF0104/ALDC/Ptd012-like
d.291.1	YehU-like
d.292.1	DNA mismatch repair protein MutL
d.293.1	Phosphoprotein M1, C-terminal domain
d.294.1	EndoU-like
d.295.1	TFB5-like
d.296.1	YktB/PF0168-like
d.297.1	WGR domain-like
d.298.1	RelE-like
d.299.1	Ns1 effector domain-like
d.300.1	Kinetochore globular domain
d.301.1	L35p-like
d.302.1	Coronavirus NSP8-like
d.303.1	BB1717-like
d.304.1	TTHA1013/TTHA0281-like
d.305.1	NAP-like
d.306.1	YefM-like
d.307.1	Nqo5-like
d.308.1	THUMP domain-like
d.309.1	AMMECR1-like
d.310.1	VC0467-like
d.311.1	ImmE5-like
d.312.1	TM1622-like
d.313.1	Prenyltransferase-like
d.314.1	PUG domain-like
d.315.1	TRCF domain-like
d.316.1	MK0786-like
d.317.1	YkuJ-like
d.318.1	SARS receptor-binding domain-like
d.319.1	TTHA1528-like
d.320.1	YojJ-like
d.321.1	STIV B116-like
d.322.1	PHP14-like
d.323.1	Phage tail protein-like
d.324.1	DUSP-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.325.1	L28p-like
d.326.1	XisI-like
d.327.1	DeoB insert domain-like
d.328.1	CorA soluble domain-like
d.329.1	PF0523-like
d.330.1	ERH-like
d.331.1	NE0471 N-terminal domain-like
d.332.1	RGC domain-like
d.333.1	UbiD C-terminal domain-like
d.334.1	IlvD/EDD N-terminal domain-like
d.335.1	L,D-transpeptidase pre-catalytic domain-like
d.336.1	YbiA-like
d.337.1	AF2331-like
d.338.1	Oxysterol-binding protein-like
d.339.1	ORC1-binding domain
d.340.1	CofE-like
d.341.1	Peptidoglycan deacetylase N-terminal noncatalytic region
d.342.1	PH1570-like
d.343.1	MM3350-like
d.344.1	PriA/YqbF domain
d.345.1	NRDP1 C-terminal domain-like
d.346.1	SARS Nsp1-like
d.347.1	Acetoacetate decarboxylase-like
d.348.1	YegP-like
d.349.1	CPE0013-like
d.350.1	YcgL-like
d.351.1	NMB0488-like
d.352.1	FlaG-like
d.353.1	AMPKBI-like
d.354.1	Shew3726-like
d.355.1	RplX-like
d.356.1	SP0830-like
d.357.1	NosL/MerB-like
d.358.1	YdfO-like
d.359.1	BH3703-like
d.360.1	PG1857-like
d.361.1	PB2 C-terminal domain-like
d.362.1	BLRF2-like
d.363.1	NMB0513-like
d.364.1	PA1123-like
d.365.1	Ava3019-like
d.366.1	SpoVG-like
d.367.1	EscU C-terminal domain-like
d.368.1	YonK-like
d.369.1	SMI1/KNR4-like
d.370.1	BTG domain-like
d.371.1	YehR-like
d.372.1	YqaI-like
d.373.1	gpW/gp25-like
d.374.1	TTHC002-like
d.375.1	NE1680-like
d.376.1	Lp2179-like
d.377.1	Rv2827c C-terminal domain-like
d.378.1	Phosphoprotein oligomerization domain-like
d.379.1	Taf5 N-terminal domain-like
d.380.1	Jann2411-like

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
d.381.1	ATP12-like
d.382.1	PSTPO5379-like
d.383.1	PG1388-like
e.1.1	Serpins
e.2.1	Replication terminator protein (Tus)
e.3.1	beta-lactamase/transpeptidase-like
e.5.1	Heme-dependent catalase-like
e.6.1	Acyl-CoA dehydrogenase NM domain-like
e.7.1	Carbohydrate phosphatase
e.8.1	DNA/RNA polymerases
e.10.1	Prokaryotic type I DNA topoisomerase
e.11.1	Type II DNA topoisomerase
e.12.1	DNA topoisomerase IV, alpha subunit
e.13.1	DNA primase core
e.15.1	Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment
e.17.1	D-aminoacid aminotransferase-like PLP-dependent enzymes
e.18.1	HydB/Nqo4-like
e.19.1	HydA/Nqo6-like
e.22.1	Dehydroquinate synthase-like
e.23.1	Acetyl-CoA synthetase-like
e.24.1	Ribosomal protein L1
e.25.1	Sec1/munc18-like (SM) proteins
e.26.1	Prismane protein-like
e.27.1	Upper collar protein gp10 (connector protein)
e.28.1	Reovirus inner layer core protein p3
e.29.1	beta and beta-prime subunits of DNA dependent RNA-polymerase
e.32.1	Phase 1 flagellin
e.34.1	NSP3 homodimer
e.35.1	Membrane penetration protein mu1
e.37.1	Siroheme synthase middle domains-like
e.38.1	Release factor
e.39.1	YebC-like
e.40.1	Cullin homology domain
e.41.1	Adenylylcyclase toxin (the edema factor)
e.42.1	L-A virus major coat protein
e.43.1	Subunits of heterodimeric actin filament capping protein Capz
e.44.1	2-methylcitrate dehydratase PrpD
e.45.1	Antivirulence factor
e.46.1	Virulence-associated V antigen
e.47.1	39 kda initiator binding protein, IBP39, C-terminal domains
e.48.1	Major capsid protein VP5
e.49.1	Recombination protein RecR
e.50.1	AF1104-like
e.51.1	Urocanase
e.52.1	NAD kinase/diacylglycerol kinase-like
e.53.1	QueA-like
e.54.1	CbiD-like
e.55.1	Rap/Ran-GAP
e.56.1	YaeB-like
e.57.1	Vacuolar ATP synthase subunit C
e.58.1	Viral ssDNA binding protein
e.59.1	FdhE-like
e.60.1	Thermophilic metalloprotease-like
e.61.1	ImpE-like
e.62.1	Heme iron utilization protein-like
e.63.1	E2F-DP heterodimerization region

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
e.64.1	FlhC-like
e.65.1	VPA0735-like
e.66.1	Api92-like
e.67.1	PH0156-like
e.68.1	YacF-like
e.69.1	Poly(A) polymerase catalytic subunit-like
e.70.1	MalF N-terminal region-like
e.71.1	AF1531-like
e.72.1	SSO1389-like
e.73.1	CV3147-like
e.74.1	HI0933 insert domain-like
e.75.1	flu NP-like
e.76.1	Viral glycoprotein ectodomain-like
f.1.1	Colicin
f.1.2	Diphtheria toxin, middle domain
f.1.3	delta-Endotoxin (insectocide), N-terminal domain
f.1.4	Bcl-2 inhibitors of programmed cell death
f.1.5	Exotoxin A, middle domain
f.3.1	Light-harvesting complex subunits
f.4.1	OMPA-like
f.4.2	Outer membrane phospholipase A (OMPLA)
f.4.3	Porins
f.4.4	OMPT-like
f.4.5	Autotransporter
f.4.6	Tsx-like channel
f.5.1	Outer membrane efflux proteins (OEP)
f.6.1	Leukocidin-like
f.7.1	Lipovitellin-phosvitin complex; beta-sheet shell regions
f.8.1	Aerolisin/ETX pore-forming domain
f.9.1	Perfringolysin
f.10.1	Viral glycoprotein, central and dimerisation domains
f.11.1	Anthrax protective antigen
f.12.1	Head and neck region of the ectodomain of NDV fusion glycoprotein
f.13.1	Family A G protein-coupled receptor-like
f.14.1	Voltage-gated potassium channels
f.15.1	Small-conductance potassium channel
f.16.1	Gated mechanosensitive channel
f.17.1	F1F0 ATP synthase subunit C
f.17.2	Cytochrome c oxidase subunit II-like, transmembrane region
f.17.3	Magnesium transport protein CorA, transmembrane region
f.17.4	Htr2 transmembrane domain-like
f.17.5	PsbZ-like
f.18.1	F1F0 ATP synthase subunit A
f.19.1	Aquaporin-like
f.20.1	Clc chloride channel
f.21.1	Transmembrane di-heme cytochromes
f.21.2	Fumarate reductase respiratory complex transmembrane subunits
f.21.3	Respiratory nitrate reductase 1 gamma chain
f.22.1	ABC transporter involved in vitamin B12 uptake, BtuC
f.23.1	Mitochondrial cytochrome c oxidase subunit IV
f.23.2	Mitochondrial cytochrome c oxidase subunit VIa
f.23.3	Mitochondrial cytochrome c oxidase subunit VIc
f.23.4	Mitochondrial cytochrome c oxidase subunit VIIa
f.23.5	Mitochondrial cytochrome c oxidase subunit VIIb
f.23.6	Mitochondrial cytochrome c oxidase subunit VIIc (aka VIIIa)
f.23.7	Mitochondrial cytochrome c oxidase subunit VIIIb (aka IX)

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
f.23.8	Bacterial aa3 type cytochrome c oxidase subunit IV
f.23.9	Bacterial ba3 type cytochrome c oxidase subunit IIa
f.23.10	Photosystem II reaction centre subunit H, transmembrane region
f.23.11	Cytochrome c1 subunit of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase), transmembrane anchor
f.23.12	ISP transmembrane anchor
f.23.13	Ubiquinone-binding protein QP-C of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.23.14	Subunit X (non-heme 7 kDa protein) of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.23.15	Subunit XI (6.4 kDa protein) of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.23.16	Subunit III of photosystem I reaction centre, PsaF
f.23.17	Subunit VIII of photosystem I reaction centre, PsaI
f.23.18	Subunit IX of photosystem I reaction centre, PsaJ
f.23.19	Subunit XII of photosystem I reaction centre, PsaM
f.23.20	Subunit PsaX of photosystem I reaction centre
f.23.21	F1F0 ATP synthase subunit B, membrane domain
f.23.22	Iron-sulfur subunit of formate dehydrogenase N, transmembrane anchor
f.23.23	Cytochrome f subunit of the cytochrome b6f complex, transmembrane anchor
f.23.24	PetL subunit of the cytochrome b6f complex
f.23.25	PetM subunit of the cytochrome b6f complex
f.23.26	PetG subunit of the cytochrome b6f complex
f.23.27	PetN subunit of the cytochrome b6f complex
f.23.28	Preprotein translocase SecE subunit
f.23.29	Sec-beta subunit
f.23.30	Oligosaccharyltransferase subunit ost4p
f.23.31	Photosystem II reaction center protein L, PsbL
f.23.32	Photosystem II reaction center protein J, PsbJ
f.23.33	Photosystem II 10 kDa phosphoprotein PsbH
f.23.34	Photosystem II reaction center protein T, PsbT
f.23.35	Photosystem II reaction center protein M, PsbM
f.23.36	Photosystem II reaction center protein K, PsbK
f.23.37	Photosystem II reaction center protein I, PsbI
f.23.38	Cytochrome b559 subunits
f.24.1	Cytochrome c oxidase subunit I-like
f.25.1	Cytochrome c oxidase subunit III-like
f.26.1	Bacterial photosystem II reaction centre, L and M subunits
f.27.1	14 kDa protein of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.28.1	Non-heme 11 kDa protein of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.29.1	Photosystem I subunits PsaA/PsaB
f.30.1	Photosystem I reaction center subunit X, PsaK
f.31.1	Photosystem I reaction center subunit XI, PsaL
f.32.1	a domain/subunit of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.33.1	Calcium ATPase, transmembrane domain M
f.34.1	Mechanosensitive channel protein MscS (YggB), transmembrane region
f.35.1	Multidrug efflux transporter AcrB transmembrane domain
f.36.1	Neurotransmitter-gated ion-channel transmembrane pore
f.37.1	ABC transporter transmembrane region
f.38.1	MFS general substrate transporter
f.39.1	Multidrug resistance efflux transporter EmrE
f.40.1	V-type ATP synthase subunit C
f.41.1	Preprotein translocase SecY subunit
f.42.1	Mitochondrial carrier
f.43.1	Chlorophyll a-b binding protein
f.44.1	Ammonium transporter
f.45.1	Mitochondrial ATP synthase coupling factor 6
f.46.1	HlyD-like secretion proteins
f.47.1	VP4 membrane interaction domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
f.48.1	OmpH-like
f.49.1	Proton glutamate symport protein
f.50.1	Connexin43
f.51.1	Rhomboid-like
f.52.1	ATP synthase B chain-like
f.53.1	ATP synthase D chain-like
f.54.1	SNF-like
f.55.1	Photosystem II antenna protein-like
f.56.1	MAPEG domain-like
f.57.1	MgtE membrane domain-like
f.58.1	MetI-like
f.59.1	Cation efflux protein transmembrane domain-like
g.1.1	Insulin-like
g.2.1	Heat-stable enterotoxin B
g.2.2	Neurotoxin B-IV
g.2.3	Pollen allergen ole e 6
g.2.4	VhTI-like
g.3.1	Plant lectins/antimicrobial peptides
g.3.2	Plant inhibitors of proteinases and amylases
g.3.3	Cyclotides
g.3.4	Gurmarin-like
g.3.5	Agouti-related protein
g.3.6	omega toxin-like
g.3.7	Scorpion toxin-like
g.3.8	Cellulose-binding domain
g.3.9	Growth factor receptor domain
g.3.10	Colipase-like
g.3.11	EGF/Laminin
g.3.12	Bromelain inhibitor VI (cysteine protease inhibitor)
g.3.13	Bowman-Birk inhibitor, BBI
g.3.14	Elafin-like
g.3.15	Leech antihemostatic proteins
g.3.16	Granulin repeat
g.3.17	Satiety factor CART (cocaine and amphetamine regulated transcript)
g.3.18	DPY module
g.3.19	Bubble protein
g.4.1	PMP inhibitors
g.5.1	Midkine
g.6.1	Amb V allergen
g.7.1	Snake toxin-like
g.8.1	BPTI-like
g.9.1	Defensin-like
g.10.1	Hairpin loop containing domain-like
g.11.1	Neurotoxin III (ATX III)
g.12.1	LDL receptor-like module
g.13.1	Crambin-like
g.14.1	Kringle-like
g.16.1	Trefoil
g.16.2	Plexin repeat
g.16.3	Variant surface glycoprotein MITAT 1.2, VSG 221, C-terminal domain
g.17.1	Cystine-knot cytokines
g.18.1	Complement control module/SCR domain
g.19.1	Crisp domain-like
g.20.1	Blood coagulation inhibitor (disintegrin)
g.21.1	Methylamine dehydrogenase, L chain
g.22.1	Serine protease inhibitors

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
g.23.1	TB module/8-cys domain
g.24.1	TNF receptor-like
g.25.1	Heparin-binding domain from vascular endothelial growth factor
g.26.1	Antifungal protein (AGAFP)
g.27.1	FnI-like domain
g.28.1	Thyroglobulin type-1 domain
g.29.1	Type X cellulose binding domain, CBDX
g.30.1	Carboxypeptidase inhibitor
g.31.1	Invertebrate chitin-binding proteins
g.32.1	GLA-domain
g.33.1	Cholecystokinin A receptor, N-domain
g.34.1	HIV-1 VPU cytoplasmic domain
g.35.1	HIPIP (high potential iron protein)
g.36.1	Ferredoxin thioredoxin reductase (FTR), catalytic beta chain
g.37.1	beta-beta-alpha zinc fingers
g.38.1	Zn2/Cys6 DNA-binding domain
g.39.1	Glucocorticoid receptor-like (DNA-binding domain)
g.40.1	Retrovirus zinc finger-like domains
g.41.1	Methionyl-tRNA synthetase (MetRS), Zn-domain
g.41.2	Microbial and mitochondrial ADK, insert "zinc finger" domain
g.41.3	Zinc beta-ribbon
g.41.4	Casein kinase II beta subunit
g.41.5	Rubredoxin-like
g.41.6	Hypothetical protein MTH1184
g.41.7	Aspartate carbamoyltransferase, Regulatory-chain, C-terminal domain
g.41.8	Zn-binding ribosomal proteins
g.41.9	RNA polymerase subunits
g.41.10	Zn-finger domain of Sec23/24
g.41.11	Ran binding protein zinc finger-like
g.41.13	Hypothetical protein Ta0289 C-terminal domain
g.41.14	NADH pyrophosphatase intervening domain
g.41.15	NOB1 zinc finger-like
g.41.16	Nop10-like SnoRNP
g.41.17	CSL zinc finger
g.41.18	YfgJ-like
g.42.1	Ribosomal protein L36
g.43.1	B-box zinc-binding domain
g.44.1	RING/U-box
g.45.1	ArfGap/RecO-like zinc finger
g.46.1	Metallothionein
g.47.1	Zinc domain conserved in yeast copper-regulated transcription factors
g.48.1	Ada DNA repair protein, N-terminal domain (N-Ada 10)
g.49.1	Cysteine-rich domain
g.50.1	FYVE/PHD zinc finger
g.51.1	Zn-binding domains of ADDBP
g.52.1	Inhibitor of apoptosis (IAP) repeat
g.53.1	TAZ domain
g.54.1	DnaJ/Hsp40 cysteine-rich domain
g.55.1	Cellulose docking domain, dockering
g.58.1	Pheromone ER-23
g.59.1	Zinc-binding domain of translation initiation factor 2 beta
g.60.1	TSP-1 type 1 repeat
g.61.1	Apical membrane antigen 1
g.62.1	Cysteine-rich DNA binding domain, (DM domain)
g.63.1	Mollusk pheromone
g.64.1	Somatomedin B domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
g.65.1	Notch domain
g.66.1	CCCH zinc finger
g.67.1	Zinc finger domain of DNA polymerase-alpha
g.68.1	Kazal-type serine protease inhibitors
g.69.1	Plant proteinase inhibitors
g.70.1	Necrosis inducing protein 1, NIP1
g.71.1	Mini-collagen I, C-terminal domain
g.72.1	SBT domain
g.73.1	CCHHC domain
g.74.1	Sec-C motif
g.75.1	TSP type-3 repeat
g.76.1	Hormone receptor domain
g.77.1	Resistin
g.78.1	YAP1 redox domain
g.79.1	WRKY DNA-binding domain
g.80.1	AN1-like Zinc finger
g.81.1	HSP33 redox switch-like
g.82.1	Expressed protein At2g23090/F21P24.15
g.83.1	Tim10-like
g.84.1	Vanabin-like
g.85.1	HIT/MYND zinc finger-like
g.86.1	Coronavirus NSP10-like
g.87.1	Viral leader polypeptide zinc finger
g.88.1	TSP9-like
g.89.1	CHY zinc finger-like
g.90.1	E6 C-terminal domain-like
g.91.1	E7 C-terminal domain-like
g.92.1	T-antigen specific domain-like
g.93.1	Zinc hairpin stack
h.1.1	Triple coiled coil domain of C-type lectins
h.1.2	Trimerization domain of TRAF
h.1.3	Leucine zipper domain
h.1.4	Inovirus (filamentous phage) major coat protein
h.1.5	Tropomyosin
h.1.6	Chicken cartilage matrix protein
h.1.7	Assembly domain of cartilage oligomeric matrix protein
h.1.8	Fibrinogen coiled-coil and central regions
h.1.9	Coiled-coil domain of nucleotide exchange factor GrpE
h.1.10	Coiled-coil dimerization domain from cortexillin I
h.1.11	XRCC4, C-terminal oligomerization domain
h.1.12	Delta-sleep-inducing peptide immunoreactive peptide
h.1.13	Rotavirus nonstructural proteins
h.1.14	Multimerization domain of the phosphoprotein from sendai virus
h.1.15	SNARE fusion complex
h.1.16	Outer membrane lipoprotein
h.1.17	Fibritin
h.1.18	N-terminal coiled coil domain from apc
h.1.19	Tetrabrachion
h.1.20	Intermediate filament protein, coiled coil region
h.1.21	Eea1 homodimerisation domain
h.1.22	Mitotic arrest deficient-like 1, Mad1
h.1.23	Dimerization motif of sir4
h.1.24	Head morphogenesis protein gp7
h.1.25	Troponin coil-coiled subunits
h.1.26	Myosin rod fragments
h.1.27	G protein-binding domain

Continued on next page

Table A1 – continued from previous page

SCCS	Superfamily
h.1.28	Geminin coiled-coil domain
h.1.29	Vasodilator-stimulated phosphoprotein, VASP, tetramerisation domain
h.1.30	MPN010-like
h.1.31	Eferin C-derminal domain-like
h.1.32	Heterotrimerisation domain of extracellular hemoglobin linker subunits
h.1.33	Sec2 N-terminal region
h.1.34	RILP dimerisation region
h.2.1	Tetramerization domain of the Mnt repressor
h.3.1	Influenza hemagglutinin (stalk)
h.3.2	Virus ectodomain
h.3.3	Coronavirus S2 glycoprotein
h.4.1	Variant surface glycoprotein (N-terminal domain)
h.4.2	Clostridium neurotoxins, "coiled-coil" domain
h.4.3	Colicin Ia, N-terminal domain
h.4.4	Bacterial hemolysins
h.4.5	Methyl-accepting chemotaxis protein (MCP) signaling domain
h.4.6	Oligomerization domain of hepatitis delta antigen
h.4.8	F1 ATPase inhibitor, IF1, C-terminal domain
h.4.9	Colicin E3 receptor domain
h.4.10	C-terminal domain of PLC-beta
h.4.11	Chemotaxis phosphatase CheZ
h.4.12	Rad50 coiled-coil Zn hook
h.4.13	Tumor suppressor gene product Apc
h.4.15	Proline/betaine transporter ProP, C-terminal cytoplasmic domain
h.4.16	Fzo-like conserved region
h.4.17	occludin/ELL-like
h.4.18	Gam-like
h.4.19	PspA lactotransferrin-binding region
h.5.1	Apolipoprotein A-I
h.6.1	Apolipoprotein A-II
h.7.1	Synuclein

Table A2: List of folds under SCOP (1.75). Folds are given for reference by their SCOP concise classification strings (sccs) as an identifier which is used throughout this thesis and their names according to the scheme

SCCS	Fold
a.1	Globin-like
a.2	Long alpha-hairpin
a.3	Cytochrome c
a.4	DNA/RNA-binding 3-helical bundle
a.5	RuvA C-terminal domain-like
a.6	Putative DNA-binding domain
a.7	Spectrin repeat-like
a.8	immunoglobulin/albumin-binding domain-like
a.9	Peripheral subunit-binding domain of 2-oxo acid dehydrogenase complex
a.10	Protozoan pheromone-like
a.11	Acyl-CoA binding protein-like
a.12	Kix domain of CBP (creb binding protein)
a.13	RAP domain-like
a.14	VHP, Villin headpiece domain

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
a.15	TAF(II)230 TBP-binding fragment
a.16	S15/NS1 RNA-binding domain
a.17	Cysteine alpha-hairpin motif
a.18	T4 endonuclease V
a.19	Fertilization protein
a.20	PGBD-like
a.21	HMG-box
a.22	Histone-fold
a.23	Open three-helical up-and-down bundle
a.24	Four-helical up-and-down bundle
a.25	Ferritin-like
a.26	4-helical cytokines
a.27	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
a.28	Acyl carrier protein-like
a.29	Bromodomain-like
a.30	ROP-like
a.31	Dimerization-anchoring domain of cAMP-dependent PK regulatory subunit
a.32	Transcription factor IIA (TFIIA), alpha-helical domain
a.33	Ectatomin subunits
a.34	Dimerisation interlock
a.35	lambda repressor-like DNA-binding domains
a.36	Signal peptide-binding domain
a.37	A DNA-binding domain in eukaryotic transcription factors
a.38	HLH-like
a.39	EF Hand-like
a.40	CH domain-like
a.41	Domain of poly(ADP-ribose) polymerase
a.42	SWIB/MDM2 domain
a.43	Ribbon-helix-helix
a.45	GST C-terminal domain-like
a.46	Methionine synthase domain-like
a.47	STAT-like
a.48	N-cbl like
a.49	C-terminal domain of B transposition protein
a.50	Anaphylotoxins (complement system)
a.51	Cytochrome c oxidase subunit h
a.52	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin
a.53	p53 tetramerization domain
a.54	Domain of early E2A DNA-binding protein, ADDBP
a.55	IHF-like DNA-binding proteins
a.56	CO dehydrogenase ISP C-domain like
a.57	Protein HNS-dependent expression A; HdeA
a.58	Chemotaxis receptor methyltransferase CheR, N-terminal domain
a.59	PAH2 domain
a.60	SAM domain-like
a.61	Retroviral matrix proteins
a.62	Hepatitis B viral capsid (hbcag)
a.63	Apolipophorin-III
a.64	Saposin-like
a.65	Annexin
a.66	Transducin (alpha subunit), insertion domain
a.68	Wiscott-Aldrich syndrome protein, WASP, C-terminal domain
a.69	Left-handed superhelix
a.70	ATPD N-terminal domain-like
a.71	ERP29 C domain-like
a.72	Functional domain of the splicing factor Prp18

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
a.73	Retrovirus capsid protein, N-terminal core domain
a.74	Cyclin-like
a.75	Ribosomal protein S7
a.76	Iron-dependent repressor protein, dimerization domain
a.77	DEATH domain
a.78	GntR ligand-binding domain-like
a.79	NusB-like
a.80	post-AAA+ oligomerization domain-like
a.81	N-terminal domain of DnaB helicase
a.83	Guanido kinase N-terminal domain
a.84	Scaffolding protein gpD of bacteriophage procapsid
a.85	Hemocyanin, N-terminal domain
a.86	Di-copper centre-containing domain
a.87	DBL homology domain (DH-domain)
a.88	LigA subunit of an aromatic-ring-opening dioxygenase LigAB
a.89	Methyl-coenzyme M reductase alpha and beta chain C-terminal domain
a.90	Transcription factor STAT-4 N-domain
a.91	Regulator of G-protein signaling, RGS
a.92	Carbamoyl phosphate synthetase, large subunit connection domain
a.93	Heme-dependent peroxidases
a.94	Ribosomal protein L19 (L19e)
a.95	Influenza virus matrix protein M1
a.96	DNA-glycosylase
a.97	An anticodon-binding domain of class I aminoacyl-tRNA synthetases
a.98	R1 subunit of ribonucleotide reductase, N-terminal domain
a.99	Cryptochrome/photolyase FAD-binding domain
a.100	6-phosphogluconate dehydrogenase C-terminal domain-like
a.101	Uteroglobin-like
a.102	alpha/alpha toroid
a.103	Citrate synthase
a.104	Cytochrome P450
a.108	Ribosomal protein L7/12, oligomerisation (N-terminal) domain
a.109	Class II MHC-associated invariant chain ectoplasmic trimerization domain
a.110	Aldehyde ferredoxin oxidoreductase, C-terminal domains
a.111	Acid phosphatase/Vanadium-dependent haloperoxidase
a.113	DNA repair protein MutS, domain III
a.114	Interferon-induced guanylate-binding protein 1 (GBP1), C-terminal domain
a.115	A virus capsid protein alpha-helical domain
a.116	GTPase activation domain, GAP
a.117	Ras GEF
a.118	alpha-alpha superhelix
a.119	Lipoxigenase
a.120	gene 59 helicase assembly protein
a.121	Tetracyclin repressor-like, C-terminal domain
a.123	Nuclear receptor ligand-binding domain
a.124	Phospholipase C/P1 nuclease
a.126	Serum albumin-like
a.127	L-aspartase-like
a.128	Terpenoid synthases
a.129	GroEL equatorial domain-like
a.130	Chorismate mutase II
a.131	Peridinin-chlorophyll protein
a.132	Heme oxygenase-like
a.133	Phospholipase A2, PLA2
a.134	Fungal elicitin
a.135	Tetraspanin

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
a.136	FinO-like
a.137	Non-globular all-alpha subunits of globular proteins
a.138	Multiheme cytochromes
a.139	Type I dockerin domain
a.140	LEM/SAP HeH motif
a.141	Frizzled cysteine-rich domain
a.142	PTS-regulatory domain, PRD
a.143	RPB6/omega subunit-like
a.144	PABP domain-like
a.145	Flagellar transcriptional activator FlhD
a.146	Telomeric repeat binding factor (TRF) dimerisation domain
a.147	Bcr-Abl oncoprotein oligomerization domain
a.148	Arp2/3 complex 21 kDa subunit ARPC3
a.149	RNase III domain-like
a.150	Anti-sigma factor AsiA
a.151	Glutamyl tRNA-reductase dimerization domain
a.152	AhpD-like
a.153	Nuclear receptor coactivator interlocking domain
a.154	Variable surface antigen VlsE
a.155	H-NS histone-like proteins
a.156	S13-like H2TH domain
a.157	Skp1 dimerisation domain-like
a.158	F-box domain
a.159	Another 3-helical bundle
a.160	PAP/OAS1 substrate-binding domain
a.161	beta-catenin-interacting protein ICAT
a.162	Pre-protein crosslinking domain of SecA
a.163	Crustacean CHH/MIH/GIH neurohormone
a.164	C-terminal domain of DFF45/ICAD (DFF-C domain)
a.165	Myosin phosphatase inhibitor 17kDa protein, CPI-17
a.166	RuBisCo LSMT C-terminal, substrate-binding domain
a.168	SopE-like GEF domain
a.169	BEACH domain
a.170	BRCA2 helical domain
a.171	BRCA2 tower domain
a.172	Helical scaffold and wing domains of SecA
a.173	Poly A polymerase C-terminal region-like
a.174	Double Clp-N motif
a.175	Orange carotenoid protein, N-terminal domain
a.176	N-terminal domain of bifunctional PutA protein
a.177	Sigma2 domain of RNA polymerase sigma factors
a.178	Soluble domain of poliovirus core protein 3a
a.179	Replisome organizer (g39p helicase loader/inhibitor protein)
a.180	N-terminal, cytoplasmic domain of anti-sigmaE factor RseA
a.181	Antibiotic binding domain of TipA-like multidrug resistance regulators
a.182	GatB/YqeY motif
a.183	Nop domain
a.184	TorD-like
a.185	Gametocyte protein Pfg27
a.186	KaiA/RbsU domain
a.187	HAND domain of the nucleosome remodeling ATPase ISWI
a.188	PWI domain
a.189	XPC-binding domain
a.190	Flavivirus capsid protein C
a.191	Methenyltetrahydrofolate cyclohydrolase-like
a.192	N-terminal domain of adenylylcyclase associated protein, CAP

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
a.193	GRIP domain
a.194	L27 domain
a.195	YutG-like
a.196	Invasion protein A (SipA) , C-terminal actin binding domain
a.198	YcfC-like
a.199	YgfB-like
a.200	Hypothetical protein MTH393
a.202	Superantigen MAM
a.203	Putative anticodon-binding domain of alanyl-tRNA synthetase (AlaRS)
a.204	all-alpha NTP pyrophosphatases
a.205	Hsp90 co-chaperone CDC37
a.206	P40 nucleoprotein
a.207	Formin homology 2 domain (FH2 domain)
a.208	DhaL-like
a.209	ADP-ribosylglycohydrolase
a.210	Eukaryotic initiation factor 4f subunit eIF4g, eIF4e-binding domain
a.211	HD-domain/PDEase-like
a.212	KRAB domain (Kruppel-associated box)
a.213	DinB/YfiT-like putative metalloenzymes
a.214	NblA-like
a.215	A middle domain of Talin 1
a.216	I/LWEQ domain
a.217	Surp module (SWAP domain)
a.218	YgfY-like
a.219	Hypothetical protein YhaI
a.220	Hypothetical protein At3g22680
a.221	Lissencephaly-1 protein (Lis-1, PAF-AH alpha) N-terminal domain
a.222	VPS9 domain
a.223	Triger factor/SurA peptide-binding domain-like
a.224	Glycolipid transfer protein, GLTP
a.225	Hypothetical protein MG354
a.226	Her-1
a.227	ERO1-like
a.228	GDNF receptor-like
a.229	Hypothetical protein YqbG
a.230	YugE-like
a.231	EspA/CesA-like
a.232	RNA-binding protein She2p
a.233	YfbU-like
a.234	Hypothetical protein MPN330
a.235	ATP-dependent DNA ligase DNA-binding domain
a.236	DNA primase DnaG, C-terminal domain
a.237	DNA polymerase III theta subunit-like
a.238	BAR/IMD domain-like
a.239	ChaB-like
a.240	BSD domain-like
a.241	TraM-like
a.242	Dcp2 domain-like
a.243	Type III secretion system domain
a.244	EF2947-like
a.245	EB1 dimerisation domain-like
a.246	Hyaluronidase domain-like
a.247	YoaC-like
a.248	SP0561-like
a.249	YfmB-like
a.250	IpaD-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
a.251	Phage replication organizer domain
a.252	Mediator hinge subcomplex-like
a.253	AF0941-like
a.254	PA2201 C-terminal domain-like
a.255	Rv1873-like
a.256	RUN domain-like
a.257	SipA N-terminal domain-like
a.258	PG0816-like
a.259	YidB-like
a.260	Rhabdovirus nucleoprotein-like
a.261	GUN4-like
a.262	PriB N-terminal domain-like
a.263	DNA terminal protein
a.264	Duffy binding domain-like
a.265	Fic-like
a.266	Indolic compounds 2,3-dioxygenase-like
a.267	Topoisomerase V catalytic domain-like
a.268	PTPA-like
a.269	FtsH protease domain-like
a.270	Hermes dimerisation domain
a.271	SOCS box-like
a.272	YqgQ-like
a.273	Orange domain-like
a.274	HAMP domain-like
a.275	DnaD domain-like
a.276	BH2638-like
a.277	TAFH domain-like
a.278	GINS helical bundle-like
a.279	Jann4075-like
a.280	RbcX-like
a.281	YheA-like
a.282	RPA2825-like
a.283	ENT-like
a.284	YejL-like
a.285	MtlR-like
a.286	Sama2622-like
a.287	TerB-like
a.288	UraD-like
a.289	Sec63 N-terminal domain-like
a.290	PSPTO4464-like
a.291	MG296-like
a.292	HP0242-like
a.293	SMc04008-like
a.294	Tex N-terminal region-like
a.295	AGR C 984p-like
a.296	PMT central region-like
b.1	Immunoglobulin-like beta-sandwich
b.2	Common fold of diphtheria toxin/transcription factors/cytochrome f
b.3	Prealbumin-like
b.4	HSP40/DnaJ peptide-binding domain
b.5	alpha-Amylase inhibitor tendamistat
b.6	Cupredoxin-like
b.7	C2 domain-like
b.8	TRAF domain-like
b.9	Neurophysin II
b.11	gamma-Crystallin-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
b.12	Lipase/lipooxygenase domain (PLAT/LH2 domain)
b.14	Calpain large subunit, middle domain (domain III)
b.15	HSP20-like chaperones
b.16	Ecotin, trypsin inhibitor
b.17	PEBP-like
b.18	Galactose-binding domain-like
b.19	Viral protein domain
b.20	ENV polyprotein, receptor-binding domain
b.21	Virus attachment protein globular domain
b.22	TNF-like
b.23	CUB-like
b.24	Hyaluronate lyase-like, C-terminal domain
b.25	Osmotin, thaumatin-like protein
b.26	SMAD/FHA domain
b.27	Soluble secreted chemokine inhibitor, VCCI
b.28	Baculovirus p35 protein
b.29	Concanavalin A-like lectins/glucanases
b.30	Supersandwich
b.31	EV matrix protein
b.32	gp9
b.33	ISP domain
b.34	SH3-like barrel
b.35	GroES-like
b.36	PDZ domain-like
b.37	N-terminal domains of the minor coat protein g3p
b.38	Sm-like fold
b.39	Ribosomal protein L14
b.40	OB-fold
b.41	PRC-barrel domain
b.42	beta-Trefoil
b.43	Reductase/isomerase/elongation factor common domain
b.44	Elongation factor/aminomethyltransferase common domain
b.45	Split barrel-like
b.46	FMT C-terminal domain-like
b.47	Trypsin-like serine proteases
b.48	mu transposase, C-terminal domain
b.49	Domain of alpha and beta subunits of F1 ATP synthase-like
b.50	Acid proteases
b.51	ValRS/IleRS/LeuRS editing domain
b.52	Double psi beta-barrel
b.53	Ribosomal protein L25-like
b.54	Core binding factor beta, CBF
b.55	PH domain-like barrel
b.56	Transcription factor IIA (TFIIA), beta-barrel domain
b.57	Herpes virus serine proteinase, assemblin
b.58	PK beta-barrel domain-like
b.59	XRCC4, N-terminal domain
b.60	Lipocalins
b.61	Streptavidin-like
b.62	Cyclophilin-like
b.63	Oncogene products
b.64	Mannose 6-phosphate receptor domain
b.65	triple barrel
b.66	4-bladed beta-propeller
b.67	5-bladed beta-propeller
b.68	6-bladed beta-propeller

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
b.69	7-bladed beta-propeller
b.70	8-bladed beta-propeller
b.71	Glycosyl hydrolase domain
b.72	WW domain-like
b.73	Head domain of nucleotide exchange factor GrpE
b.74	Carbonic anhydrase
b.75	Bacteriochlorophyll A protein
b.76	open-sided beta-meander
b.77	beta-Prism I
b.78	beta-Prism II
b.80	Single-stranded right-handed beta-helix
b.81	Single-stranded left-handed beta-helix
b.82	Double-stranded beta-helix
b.83	Triple beta-spiral
b.84	Barrel-sandwich hybrid
b.85	beta-clip
b.86	Hedgehog/intein (Hint) domain
b.87	LexA/Signal peptidase
b.88	Mss4-like
b.89	Cyanovirin-N
b.90	Head-binding domain of phage P22 tailspike protein
b.91	E2 regulatory, transactivation domain
b.92	Composite domain of metallo-dependent hydrolases
b.93	Epsilon subunit of F1F0-ATP synthase N-terminal domain
b.94	Olfactory marker protein
b.95	Ganglioside M2 (gm2) activator
b.96	Nicotinic receptor ligand binding domain-like
b.97	Cytolysin/lectin
b.98	Leukotriene A4 hydrolase N-terminal domain
b.100	Sortase
b.101	Ribonuclease domain of colicin E3
b.102	Methuselah ectodomain
b.103	MoeA N-terminal region -like
b.104	P-domain of calnexin/calreticulin
b.105	Penicillin-binding protein associated domain
b.106	Phage tail proteins
b.107	Urease metallochaperone UreE, N-terminal domain
b.108	Triple-stranded beta-helix
b.109	beta-hairpin stack
b.110	Cloacin translocation domain
b.111	Small protein B (SmpB)
b.112	C-terminal domain of mollusc hemocyanin
b.113	N-terminal domain of MutM-like DNA repair proteins
b.114	N-utilization substance G protein NusG, insert domain
b.115	Calcium-mediated lectin
b.116	Viral chemokine binding protein m3
b.117	Obg-fold
b.118	FAS1 domain
b.119	C-terminal autoproteolytic domain of nucleoporin nup98
b.120	Tp47 lipoprotein, N-terminal domain
b.121	Nucleoplasmin-like/VP (viral coat and capsid proteins)
b.122	PUA domain-like
b.123	Hypothetical protein TM1070
b.124	HesB-like domain
b.125	LolA-like prokaryotic lipoproteins and lipoprotein localization factors
b.126	Adsorption protein p2

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
b.127	Baseplate structural protein gp8
b.128	Hypothetical protein YojF
b.129	Double-split beta-barrel
b.130	Heat shock protein 70kD (HSP70), peptide-binding domain
b.131	SPOC domain-like
b.132	Supernatant protein factor (SPF), C-terminal domain
b.133	Dextranase, N-terminal domain
b.134	Smp-1-like
b.135	Superantigen (mitogen) Ypm
b.136	SspB-like
b.137	Rof/RNase P subunit-like
b.138	Hydrophobin II, HfbII
b.139	Surface presentation of antigens (SPOA)
b.140	Replicase NSP9
b.141	Bacterial fluorinating enzyme, C-terminal domain
b.142	DNA-binding pseudobarrel domain
b.143	NAC domain
b.144	Trimeric adhesin
b.145	AXH domain
b.146	Ctag/Cox11
b.147	BTV NS2-like ssRNA-binding domain
b.148	Coronavirus RNA-binding domain
b.149	Beta-galactosidase LacA, domain 3
b.150	Putative glucosidase YicI, C-terminal domain
b.151	CsrA-like
b.152	Flagellar hook protein flgE
b.153	PheT/TilS domain
b.154	HPA-like
b.155	L21p-like
b.156	Atu1913-like
b.157	Hcp1-like
b.158	BH3618-like
b.159	AOC barrel-like
b.160	L,D-transpeptidase catalytic domain-like
b.161	PTSIIA/GutA-like
b.162	At5g01610-like
b.163	Pseudo beta-prism
b.164	SARS ORF9b-like
b.165	MOSC N-terminal domain-like
b.166	MAL13P1.257-like
b.167	FimD N-terminal domain-like
b.168	HisI-like
b.169	MFPT repeat-like
b.170	WSSV envelope protein-like
b.171	Trm112p-like
b.172	YopX-like
b.173	NifT/FixU barrel-like
b.174	YopT-like
b.175	FomD barrel-like
b.176	AttH-like
b.177	YmcC-like
b.178	Spiral beta-roll
c.1	TIM beta/alpha-barrel
c.2	NAD(P)-binding Rossmann-fold domains
c.3	FAD/NAD(P)-binding domain
c.4	Nucleotide-binding domain

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
c.5	MurCD N-terminal domain
c.6	7-stranded beta/alpha barrel
c.7	PFL-like glycyl radical enzymes
c.8	The swivelling beta/beta/alpha domain
c.9	Barstar-like
c.10	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)
c.12	Ribosomal proteins L15p and L18e
c.13	SpoIIaa-like
c.14	ClpP/crotonase
c.15	BRCT domain
c.16	Lumazine synthase
c.17	Caspase-like
c.18	Uracil-DNA glycosylase-like
c.19	FabD/lysophospholipase-like
c.20	Initiation factor IF2/eIF5b, domain 3
c.21	Ribosomal protein L13
c.22	Ribosomal protein L4
c.23	Flavodoxin-like
c.24	Methylglyoxal synthase-like
c.25	Ferredoxin reductase-like, C-terminal NADP-linked domain
c.26	Adenine nucleotide alpha hydrolase-like
c.27	Nucleoside phosphorylase/phosphoribosyltransferase catalytic domain
c.28	Cryptochrome/photolyase, N-terminal domain
c.30	PreATP-grasp domain
c.31	DHS-like NAD/FAD-binding domain
c.32	Tubulin nucleotide-binding domain-like
c.33	Isochorismatase-like hydrolases
c.34	Homo-oligomeric flavin-containing Cys decarboxylases, HFCD
c.36	Thiamin diphosphate-binding fold (THDP-binding)
c.37	P-loop containing nucleoside triphosphate hydrolases
c.38	PTS IIb component
c.39	Nicotinate mononucleotide:5,6-dimethylbenzimidazole phosphoribosyltransferase (CobT)
c.40	Methylesterase CheB, C-terminal domain
c.41	Subtilisin-like
c.42	Arginase/deacetylase
c.43	CoA-dependent acyltransferases
c.44	Phosphotyrosine protein phosphatases I-like
c.45	(Phosphotyrosine protein) phosphatases II
c.46	Rhodanese/Cell cycle control phosphatase
c.47	Thioredoxin fold
c.48	TK C-terminal domain-like
c.49	Pyruvate kinase C-terminal domain-like
c.50	Macro domain-like
c.51	Anticodon-binding domain-like
c.52	Restriction endonuclease-like
c.53	Resolvase-like
c.54	PTS system fructose IIA component-like
c.55	Ribonuclease H-like motif
c.56	Phosphorylase/hydrolase-like
c.57	Molybdenum cofactor biosynthesis proteins
c.58	Aminoacid dehydrogenase-like, N-terminal domain
c.59	MurD-like peptide ligases, peptide-binding domain
c.60	Phosphoglycerate mutase-like
c.61	PRTase-like
c.62	vWA-like
c.64	Pyruvate-ferredoxin oxidoreductase, PFOR, domain III

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
c.65	Formyltransferase
c.66	S-adenosyl-L-methionine-dependent methyltransferases
c.67	PLP-dependent transferase-like
c.68	Nucleotide-diphospho-sugar transferases
c.69	alpha/beta-Hydrolases
c.70	Nucleoside hydrolase
c.71	Dihydrofolate reductase-like
c.72	Ribokinase-like
c.73	Carbamate kinase-like
c.74	AraD/HMP-PK domain-like
c.76	Alkaline phosphatase-like
c.77	Isocitrate/Isopropylmalate dehydrogenase-like
c.78	ATC-like
c.79	Tryptophan synthase beta subunit-like PLP-dependent enzymes
c.80	SIS domain
c.81	Formate dehydrogenase/DMSO reductase, domains 1-3
c.82	ALDH-like
c.83	Aconitase iron-sulfur domain
c.84	Phosphoglucomutase, first 3 domains
c.85	FucI/AraA N-terminal and middle domains
c.86	Phosphoglycerate kinase
c.87	UDP-Glycosyltransferase/glycogen phosphorylase
c.88	Glutaminase/Asparaginase
c.89	Phosphofructokinase
c.90	Tetrapyrrole methylase
c.91	PEP carboxykinase-like
c.92	Chelatase-like
c.93	Periplasmic binding protein-like I
c.94	Periplasmic binding protein-like II
c.95	Thiolase-like
c.96	Fe-only hydrogenase
c.97	Cytidine deaminase-like
c.98	MurF and HprK N-domain-like
c.99	Dipeptide transport protein
c.100	Thiamin pyrophosphokinase, catalytic domain
c.101	Undecaprenyl diphosphate synthase
c.102	Cell-division inhibitor MinC, N-terminal domain
c.103	MTH938-like
c.104	YjeF N-terminal domain-like
c.105	2,3-Bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain
c.106	SurE-like
c.107	DHH phosphoesterases
c.108	HAD-like
c.109	PEP carboxykinase N-terminal domain
c.110	DTD-like
c.111	Activating enzymes of the ubiquitin-like proteins
c.112	Glycerol-3-phosphate (1)-acyltransferase
c.113	HemD-like
c.114	DsrEFH-like
c.115	Hypothetical protein MTH777 (MT0777)
c.116	alpha/beta knot
c.117	Amidase signature (AS) enzymes
c.118	GckA/TtuD-like
c.119	DAK1/DegV-like
c.120	PIN domain-like
c.121	Ribose/Galactose isomerase RpiB/AlsB

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
c.122	L-sulfolactate dehydrogenase-like
c.123	CoA-transferase family III (CaiB/BaiF)
c.124	NagB/RpiA/CoA transferase-like
c.125	Creatininase
c.126	DNA polymerase III psi subunit
c.127	F420-dependent methylenetetrahydromethanopterin dehydrogenase (MTD)
c.128	DNA polymerase III chi subunit
c.129	MCP/YpsA-like
c.130	Alpha-2,3/8-sialyltransferase CstII
c.131	Peptidyl-tRNA hydrolase II
c.132	Bacterial fluorinating enzyme, N-terminal domain
c.133	RbsD-like
c.134	LmbE-like
c.135	NIF3 (NGG1p interacting factor 3)-like
c.136	Toprim domain
c.138	Indigoidine synthase A-like
c.140	TTHA0583/YokD-like
c.141	Glycerate kinase I
c.142	Nqo1 FMN-binding domain-like
c.143	CofD-like
c.144	RibA-like
c.145	NadA-like
c.146	YgbK-like
c.147	CAC2185-like
c.148	ComB-like
c.149	AtpF-like
c.150	EreA/ChaN-like
c.151	CobE/GbiG C-terminal domain-like
c.152	CbiG N-terminal domain-like
c.153	YerB-like
c.154	CdCA1 repeat-like
d.1	Microbial ribonucleases
d.2	Lysozyme-like
d.3	Cysteine proteinases
d.4	His-Me finger endonucleases
d.5	RNase A-like
d.6	Prion-like
d.7	LysM domain
d.8	Urease, gamma-subunit
d.9	IL8-like
d.10	DNA-binding domain
d.11	Penicillin-binding protein 2x (pbp-2x), c-terminal domain
d.12	Ribosomal proteins S24e, L23 and L15e
d.13	HIT-like
d.14	Ribosomal protein S5 domain 2-like
d.15	beta-Grasp (ubiquitin-like)
d.16	FAD-linked reductases, C-terminal domain
d.17	Cystatin-like
d.18	ssDNA-binding transcriptional regulator domain
d.19	MHC antigen-recognition domain
d.20	UBC-like
d.21	Diaminopimelate epimerase-like
d.22	GFP-like
d.23	Tubby C-terminal domain-like
d.24	Pili subunits
d.25	Mitochondrial glycoprotein MAM33-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.26	FKBP-like
d.27	Ribosomal protein S16
d.28	Ribosomal protein S19
d.29	Ribosomal protein L31e
d.30	Allophycocyanin linker chain (domain)
d.31	Cdc48 domain 2-like
d.32	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
d.33	SecB-like
d.34	DNA-binding domain of Mlu1-box binding protein MBP1
d.35	Heme-binding protein A (HasA)
d.36	Chalcone isomerase
d.37	CBS-domain pair
d.38	Thioesterase/thiol ester dehydrase-isomerase
d.39	DLC
d.40	CI-2 family of serine protease inhibitors
d.41	alpha/beta-Hammerhead
d.42	POZ domain
d.43	EF-Ts domain-like
d.44	Fe,Mn superoxide dismutase (SOD), C-terminal domain
d.45	ClpS-like
d.47	Ribosomal L11/L12e N-terminal domain
d.48	Anti-LPS factor/recA domain
d.49	Signal recognition particle alu RNA binding heterodimer, SRP9/14
d.50	dsRBD-like
d.51	Eukaryotic type KH-domain (KH-domain type I)
d.52	Alpha-lytic protease prodomain-like
d.53	Ribosomal protein S3 C-terminal domain
d.54	Enolase N-terminal domain-like
d.55	Ribosomal protein L22
d.56	GroEL-intermediate domain like
d.57	DNA damage-inducible protein DinI
d.58	Ferredoxin-like
d.59	Ribosomal protein L30p/L7e
d.60	Probable bacterial effector-binding domain
d.61	LigT-like
d.62	Pepsin inhibitor-3
d.63	CYTH-like phosphatases
d.64	eIF1-like
d.65	Hedgehog/DD-peptidase
d.66	Alpha-L RNA-binding motif
d.67	RRF/tRNA synthetase additional domain-like
d.68	IF3-like
d.70	Yeast killer toxins
d.71	Cell division protein MinE topological specificity domain
d.72	Cyanase C-terminal domain
d.73	RuBisCO, small subunit
d.74	DCoH-like
d.75	MutS N-terminal domain-like
d.76	GYF/BRK domain-like
d.77	RL5-like
d.78	RPB5-like RNA polymerase subunit
d.79	Bacillus chorismate mutase-like
d.80	Tautomerase/MIF
d.81	FwdE/GAPDH domain-like
d.82	N domain of copper amine oxidase-like
d.83	Aha1/BPI domain-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.84	Subtilisin inhibitor
d.85	RNA bacteriophage capsid protein
d.86	eIF4e-like
d.87	CO dehydrogenase flavoprotein C-domain-like
d.88	SRF-like
d.89	Origin of replication-binding domain, RBD-like
d.90	FMN-dependent nitroreductase-like
d.91	N-terminal domain of eukaryotic peptide chain release factor subunit 1, ERF1
d.92	Zincin-like
d.93	SH2-like
d.94	HPr-like
d.95	Homing endonuclease-like
d.96	T-fold
d.97	Cell cycle regulatory proteins
d.98	BLIP-like
d.99	Ribosomal protein L9 C-domain
d.100	MbtH/L9 domain-like
d.101	Ribonuclease PH domain 2-like
d.102	Regulatory factor Nef
d.103	CytB endotoxin-like
d.104	Class II aaRS and biotin synthetases
d.105	Subdomain of clathrin and coatomer appendage domain
d.106	SCP-like
d.107	Mog1p/PsbP-like
d.108	Acyl-CoA N-acyltransferases (Nat)
d.109	Gelsolin-like
d.110	Profilin-like
d.111	PR-1-like
d.112	Phoshotransferase/anion transport protein
d.113	Nudix
d.114	5-nucleotidase (syn. UDP-sugar hydrolase), C-terminal domain
d.115	YrdC/RibB
d.116	YbaK/ProRS associated domain
d.117	Thymidylate synthase/dCMP hydroxymethylase
d.118	N-acetylmuramoyl-L-alanine amidase-like
d.120	Cytochrome b5-like heme/steroid binding domain
d.121	DNA topoisomerase I domain
d.122	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase
d.123	Sporulation response regulatory protein Spo0B
d.124	Ribonuclease Rh-like
d.125	Ornithine decarboxylase C-terminal domain
d.126	Pentein, beta/alpha-propeller
d.127	Creatinase/aminopeptidase
d.128	Glutamine synthetase/guanido kinase
d.129	TBP-like
d.130	S-adenosylmethionine synthetase
d.131	DNA clamp
d.133	Molybdenum cofactor-binding domain
d.134	Nitrite and sulphite reductase 4Fe-4S domain-like
d.135	The spindle assembly checkpoint protein mad2
d.136	Phospholipase D/nuclease
d.137	Monooxygenase (hydroxylase) regulatory protein
d.139	PurM C-terminal domain-like
d.140	Ribosomal protein S8
d.141	Ribosomal protein L6
d.142	ATP-grasp

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.143	SAICAR synthase-like
d.144	Protein kinase-like (PK-like)
d.145	FAD-binding/transporter-associated domain-like
d.146	Uridine diphospho-N-Acetylenolpyruvylglucosamine reductase, MurB, C-terminal domain
d.147	Methenyltetrahydromethanopterin cyclohydrolase
d.148	Hect, E3 ligase catalytic domain
d.149	Nitrile hydratase alpha chain
d.150	4-phosphopantetheinyl transferase
d.151	DNase I-like
d.152	Aldehyde ferredoxin oxidoreductase, N-terminal domain
d.153	Ntn hydrolase-like
d.154	DmpA/ArgJ-like
d.155	Pyruvoyl-dependent histidine and arginine decarboxylases
d.156	S-adenosylmethionine decarboxylase
d.157	Metallo-hydrolase/oxidoreductase
d.159	Metallo-dependent phosphatases
d.160	Carbon-nitrogen hydrolase
d.161	ADC synthase
d.162	LDH C-terminal domain-like
d.163	DNA breaking-rejoining enzymes
d.164	SMAD MH1 domain
d.165	Ribosome inactivating proteins (RIP)
d.166	ADP-ribosylation
d.167	Peptide deformylase
d.168	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain
d.169	C-type lectin-like
d.170	SRCR-like
d.171	Fibrinogen C-terminal domain-like
d.172	gp120 core
d.173	Methionine synthase activation domain-like
d.174	Nitric oxide (NO) synthase oxygenase domain
d.175	Penicillin binding protein dimerisation domain
d.176	Oxidoreductase molybdopterin-binding domain
d.177	FAH
d.178	Aromatic aminoacid monoxygenases, catalytic and oligomerization domains
d.179	Substrate-binding domain of HMG-CoA reductase
d.180	Conserved core of transcriptional regulatory protein vp16
d.181	Insert subdomain of RNA polymerase alpha subunit
d.182	Baseplate structural protein gp11
d.183	Major capsid protein gp5
d.184	Non-globular alpha+beta subunits of globular proteins
d.185	LuxS/MPP-like metallohydrolase
d.186	gpW/XkdW-like
d.187	Photosystem I subunit PsaD
d.188	Prokaryotic ribosomal protein L17
d.189	PX domain
d.190	Chorismate lyase-like
d.192	YlxR-like
d.193	Hsp33 domain
d.194	CNF1/YfiH-like putative cysteine hydrolases
d.195	YopH tyrosine phosphatase N-terminal domain
d.196	Outer capsid protein sigma 3
d.197	Protein-L-isoaspartyl O-methyltransferase, C-terminal domain
d.198	Secretion chaperone-like
d.199	MotA C-terminal domain-like
d.200	Integrin beta tail domain

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.201	SRP19
d.202	Transcription factor NusA, N-terminal domain
d.203	DsrC, the gamma subunit of dissimilatory sulfite reductase
d.204	Ribosome binding protein Y (YfiA homologue)
d.205	GTP cyclohydrolase I feedback regulatory protein, GFRP
d.206	YggU-like
d.207	Thymidylate synthase-complementing protein Thy1
d.208	MTH1598-like
d.209	LCCL domain
d.210	Argininosuccinate synthetase, C-terminal domain
d.211	beta-hairpin-alpha-hairpin repeat
d.212	TolA/TonB C-terminal domain
d.213	VSV matrix protein
d.214	Hypothetical protein MTH1880
d.215	Smc hinge domain
d.216	Rotavirus NSP2 fragment, N-terminal domain
d.217	SAND domain-like
d.218	Nucleotidyltransferase
d.219	PP2C-like
d.220	Metal cation-transporting ATPase, ATP-binding domain N
d.221	Nuclease A inhibitor (NuiA)
d.222	YbaB-like
d.223	Polo-box domain
d.224	SufE/NifU
d.225	Multidrug efflux transporter AcrB TolC docking domain; DN and DC subdomains
d.226	GIY-YIG endonuclease
d.227	OsmC-like
d.228	Replication modulator SeqA, C-terminal DNA-binding domain
d.229	MesJ substrate recognition domain-like
d.230	Dodecin subunit-like
d.231	Receptor-binding domain of short tail fibre protein gp12
d.232	Mago nashi protein
d.233	Inhibitor of vertebrate lysozyme, Ivy
d.234	Proguanylin
d.235	FYSH domain
d.236	DNA-binding protein Tfx
d.237	Hypothetical protein YjiA, C-terminal domain
d.238	Hypothetical protein TM0875
d.239	GCM domain
d.240	Lesion bypass DNA polymerase (Y-family), little finger domain
d.241	Ribosome binding domain-like
d.242	Obg GTP-binding protein C-terminal domain
d.243	Colicin D/E5 nuclease domain
d.244	Cell division protein ZapA-like
d.245	NSFL1 (p97 ATPase) cofactor p47, SEP domain
d.246	mRNA decapping enzyme DcpS N-terminal domain
d.247	Chromosomal protein MC1
d.248	Coproporphyrinogen III oxidase
d.249	Hypothetical protein Ta1206
d.250	Folate-binding domain
d.251	Hypothetical protein YwqG
d.252	CheC-like
d.253	Hypothetical protein YoaG
d.254	Nucleocapsid protein dimerization domain
d.255	Tombusvirus P19 core protein, VP19
d.256	Ta1353-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.257	Hypothetical protein TM0160
d.258	Chorismate synthase, AroC
d.259	Hypothetical protein HI1480
d.260	Suppressor of Fused, N-terminal domain
d.261	Hypothetical protein PH1602
d.262	NinB
d.263	Hypothetical protein Yml108w
d.264	Prim-pol domain
d.265	Pseudouridine synthase
d.266	Hypothetical protein MTH677
d.267	Hypothetical protein SAV1430
d.268	ParB/Sulfiredoxin
d.269	Gamma-glutamyl cyclotransferase-like
d.270	2-isopropylmalate synthase LeuA, allosteric (dimerisation) domain
d.271	HSP90 C-terminal domain
d.272	Dystroglycan, domain 2
d.273	YjbQ-like
d.274	Hypothetical protein PF0899
d.275	Hut operon positive regulatory protein HutP
d.276	Hypothetical protein yfbM
d.277	Bacillus phage protein
d.278	Ligand-binding domain in the NO signalling and Golgi transport
d.279	YggX-like
d.280	Sulfolobus fructose-1,6-bisphosphatase-like
d.281	Hemolytic lectin CEL-III, C-terminal domain
d.282	SSo0622-like
d.283	Putative modulator of DNA gyrase, PmbA/TldD
d.284	PurS-like
d.285	DNA-binding domain of intron-encoded endonucleases
d.286	TrkA C-terminal domain-like
d.287	DNA methylase specificity domain
d.288	GTF2I-like repeat
d.289	WWE domain
d.290	AF0104/ALDC/Ptd012-like
d.291	YehU-like
d.292	DNA mismatch repair protein MutL
d.293	Phosphoprotein M1, C-terminal domain
d.294	EndoU-like
d.295	TFB5-like
d.296	YktB/PF0168-like
d.297	WGR domain-like
d.298	RelE-like
d.299	Ns1 effector domain-like
d.300	Kinetochore globular domain-like
d.301	L35p-like
d.302	Coronavirus NSP8-like
d.303	BB1717-like
d.304	TTHA1013/TTHA0281-like
d.305	NAP-like
d.306	YefM-like
d.307	Nqo5-like
d.308	THUMP domain
d.309	AMMECR1-like
d.310	VC0467-like
d.311	ImmE5-like
d.312	TM1622-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.313	Antiparallel beta/alpha barrel (PT-barrel)
d.314	PUG domain-like
d.315	TRCF domain-like
d.316	MK0786-like
d.317	YkuJ-like
d.318	SARS receptor-binding domain-like
d.319	TTHA1528-like
d.320	YojJ-like
d.321	STIV B116-like
d.322	PHP14-like
d.323	Phage tail protein-like
d.324	DUSP-like
d.325	L28p-like
d.326	XisI-like
d.327	DeoB insert domain-like
d.328	CorA soluble domain-like
d.329	PF0523-like
d.330	ERH-like
d.331	NE0471 N-terminal domain-like
d.332	RGC domain-like
d.333	UbiD C-terminal domain-like
d.334	IlvD/EDD N-terminal domain-like
d.335	L,D-transpeptidase pre-catalytic domain-like
d.336	YbiA-like
d.337	AF2331-like
d.338	Oxysterol-binding protein-like
d.339	ORC1-binding domain
d.340	CofE-like
d.341	Peptidoglycan deacetylase N-terminal noncatalytic region
d.342	PH1570-like
d.343	MM3350-like
d.344	GINS/PriA/YqbF domain
d.345	NRDP1 C-terminal domain-like
d.346	SARS Nsp1-like
d.347	Acetoacetate decarboxylase-like
d.348	YegP-like
d.349	CPE0013-like
d.350	YcgL-like
d.351	NMB0488-like
d.352	FlaG-like
d.353	AMPKBI-like
d.354	Shew3726-like
d.355	RplX-like
d.356	SP0830-like
d.357	NosL/MerB-like
d.358	YdfO-like
d.359	BH3703-like
d.360	PG1857-like
d.361	PB2 C-terminal domain-like
d.362	BLRF2-like
d.363	NMB0513-like
d.364	PA1123-like
d.365	Ava3019-like
d.366	SpoVG-like
d.367	EscU C-terminal domain-like
d.368	YonK-like

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
d.369	SMI1/KNR4-like
d.370	BTG domain-like
d.371	YehR-like
d.372	YqaI-like
d.373	gpW/gp25-like
d.374	TTHC002-like
d.375	NE1680-like
d.376	Lp2179-like
d.377	Rv2827c C-terminal domain-like
d.378	Phosphoprotein oligomerization domain-like
d.379	Taf5 N-terminal domain-like
d.380	Jann2411-like
d.381	ATP12-like
d.382	PSTPO5379-like
d.383	PG1388-like
e.1	Serpins
e.2	Replication terminator protein (Tus)
e.3	beta-lactamase/transpeptidase-like
e.5	Heme-dependent catalase-like
e.6	Acyl-CoA dehydrogenase NM domain-like
e.7	Carbohydrate phosphatase
e.8	DNA/RNA polymerases
e.10	Prokaryotic type I DNA topoisomerase
e.11	Type II DNA topoisomerase
e.12	DNA topoisomerase IV, alpha subunit
e.13	DNA primase core
e.15	Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment
e.17	D-aminoacid aminotransferase-like PLP-dependent enzymes
e.18	HydB/Nqo4-like
e.19	HydA/Nqo6-like
e.22	Dehydroquinate synthase-like
e.23	Acetyl-CoA synthetase-like
e.24	Ribosomal protein L1
e.25	Sec1/munc18-like (SM) proteins
e.26	Prismane protein-like
e.27	Upper collar protein gp10 (connector protein)
e.28	Reovirus inner layer core protein p3
e.29	beta and beta-prime subunits of DNA dependent RNA-polymerase
e.32	Phase 1 flagellin
e.34	NSP3 homodimer
e.35	Membrane penetration protein mu1
e.37	Siroheme synthase middle domains-like
e.38	Release factor
e.39	YebC-like
e.40	Cullin homology domain
e.41	Adenylylcyclase toxin (the edema factor)
e.42	L-A virus major coat protein
e.43	Subunits of heterodimeric actin filament capping protein Capz
e.44	2-methylcitrate dehydratase PrpD
e.45	Antivirulence factor
e.46	Virulence-associated V antigen
e.47	39 kda initiator binding protein, IBP39, C-terminal domains
e.48	Major capsid protein VP5
e.49	Recombination protein RecR
e.50	AF1104-like
e.51	Urocanase

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
e.52	NAD kinase/diacylglycerol kinase-like
e.53	QueA-like
e.54	CbiD-like
e.55	Rap/Ran-GAP
e.56	YaeB-like
e.57	Vacuolar ATP synthase subunit C
e.58	Viral ssDNA binding protein
e.59	FdhE-like
e.60	Thermophilic metalloprotease-like
e.61	ImpE-like
e.62	Heme iron utilization protein-like
e.63	E2F-DP heterodimerization region
e.64	FlhC-like
e.65	VPA0735-like
e.66	Api92-like
e.67	PH0156-like
e.68	YacF-like
e.69	Poly(A) polymerase catalytic subunit-like
e.70	MalF N-terminal region-like
e.71	AF1531-like
e.72	SSO1389-like
e.73	CV3147-like
e.74	HI0933 insert domain-like
e.75	Flu NP-like
e.76	Viral glycoprotein ectodomain-like
f.1	Toxins membrane translocation domains
f.3	Light-harvesting complex subunits
f.4	Transmembrane beta-barrels
f.5	Outer membrane efflux proteins (OEP)
f.6	Leukocidin-like
f.7	Lipovitellin-phosvitin complex; beta-sheet shell regions
f.8	Aerolisin/ETX pore-forming domain
f.9	Perfringolysin
f.10	Viral glycoprotein, central and dimerisation domains
f.11	Anthrax protective antigen
f.12	Head and neck region of the ectodomain of NDV fusion glycoprotein
f.13	Family A G protein-coupled receptor-like
f.14	Voltage-gated potassium channels
f.15	Small-conductance potassium channel
f.16	Gated mechanosensitive channel
f.17	Transmembrane helix hairpin
f.18	F1F0 ATP synthase subunit A
f.19	Aquaporin-like
f.20	Clc chloride channel
f.21	Heme-binding four-helical bundle
f.22	ABC transporter involved in vitamin B12 uptake, BtuC
f.23	Single transmembrane helix
f.24	Cytochrome c oxidase subunit I-like
f.25	Cytochrome c oxidase subunit III-like
f.26	Bacterial photosystem II reaction centre, L and M subunits
f.27	14 kDa protein of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.28	Non-heme 11 kDa protein of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)
f.29	Photosystem I subunits PsaA/PsaB
f.30	Photosystem I reaction center subunit X, PsaK
f.31	Photosystem I reaction center subunit XI, PsaL
f.32	a domain/subunit of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
f.33	Calcium ATPase, transmembrane domain M
f.34	Mechanosensitive channel protein MscS (YggB), transmembrane region
f.35	Multidrug efflux transporter AcrB transmembrane domain
f.36	Neurotransmitter-gated ion-channel transmembrane pore
f.37	ABC transporter transmembrane region
f.38	MFS general substrate transporter
f.39	Multidrug resistance efflux transporter EmrE
f.40	V-type ATP synthase subunit C
f.41	Preprotein translocase SecY subunit
f.42	Mitochondrial carrier
f.43	Chlorophyll a-b binding protein
f.44	Ammonium transporter
f.45	Mitochondrial ATP synthase coupling factor 6
f.46	HlyD-like secretion proteins
f.47	VP4 membrane interaction domain
f.48	OmpH-like
f.49	Proton glutamate symport protein
f.50	Connexin43
f.51	Rhomboid-like
f.52	ATP synthase B chain-like
f.53	ATP synthase D chain-like
f.54	SNF-like
f.55	Photosystem II antenna protein-like
f.56	MAPEG domain-like
f.57	MgtE membrane domain-like
f.58	MetI-like
f.59	Cation efflux protein transmembrane domain-like
g.1	Insulin-like
g.2	Toxic hairpin
g.3	Knottins (small inhibitors, toxins, lectins)
g.4	PMP inhibitors
g.5	Midkine
g.6	Amb V allergen
g.7	Snake toxin-like
g.8	BPTI-like
g.9	Defensin-like
g.10	Hairpin loop containing domain-like
g.11	Neurotoxin III (ATX III)
g.12	LDL receptor-like module
g.13	Crambin-like
g.14	Kringle-like
g.16	Trefoil/Plexin domain-like
g.17	Cystine-knot cytokines
g.18	Complement control module/SCR domain
g.19	Crisp domain-like
g.20	Blood coagulation inhibitor (disintegrin)
g.21	Methylamine dehydrogenase, L chain
g.22	Serine protease inhibitors
g.23	TB module/8-cys domain
g.24	TNF receptor-like
g.25	Heparin-binding domain from vascular endothelial growth factor
g.26	Antifungal protein (AGAFP)
g.27	FnI-like domain
g.28	Thyroglobulin type-1 domain
g.29	Type X cellulose binding domain, CBDX
g.30	Carboxypeptidase inhibitor

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
g.31	Invertebrate chitin-binding proteins
g.32	GLA-domain
g.33	Cholecystokinin A receptor, N-domain
g.34	HIV-1 VPU cytoplasmic domain
g.35	HIPIP (high potential iron protein)
g.36	Ferredoxin thioredoxin reductase (FTR), catalytic beta chain
g.37	beta-beta-alpha zinc fingers
g.38	Zn2/Cys6 DNA-binding domain
g.39	Glucocorticoid receptor-like (DNA-binding domain)
g.40	Retrovirus zinc finger-like domains
g.41	Rubredoxin-like
g.42	Ribosomal protein L36
g.43	B-box zinc-binding domain
g.44	RING/U-box
g.45	ArfGap/RecO-like zinc finger
g.46	Metallothionein
g.47	Zinc domain conserved in yeast copper-regulated transcription factors
g.48	Ada DNA repair protein, N-terminal domain (N-Ada 10)
g.49	Cysteine-rich domain
g.50	FYVE/PHD zinc finger
g.51	Zn-binding domains of ADDBP
g.52	Inhibitor of apoptosis (IAP) repeat
g.53	TAZ domain
g.54	DnaJ/Hsp40 cysteine-rich domain
g.55	Cellulose docking domain, dockering
g.58	Pheromone ER-23
g.59	Zinc-binding domain of translation initiation factor 2 beta
g.60	TSP-1 type 1 repeat
g.61	Apical membrane antigen 1
g.62	Cysteine-rich DNA binding domain, (DM domain)
g.63	Mollusk pheromone
g.64	Somatomedin B domain
g.65	Notch domain
g.66	CCCH zinc finger
g.67	Zinc finger domain of DNA polymerase-alpha
g.68	Kazal-type serine protease inhibitors
g.69	Plant proteinase inhibitors
g.70	Necrosis inducing protein 1, NIP1
g.71	Mini-collagen I, C-terminal domain
g.72	SBT domain
g.73	CCHHC domain
g.74	Sec-C motif
g.75	TSP type-3 repeat
g.76	Hormone receptor domain
g.77	Resistin
g.78	YAP1 redox domain
g.79	WRKY DNA-binding domain
g.80	AN1-like Zinc finger
g.81	HSP33 redox switch-like
g.82	Expressed protein At2g23090/F21P24.15
g.83	Tim10-like
g.84	Cysteine zipper
g.85	HIT/MYND zinc finger-like
g.86	Coronavirus NSP10-like
g.87	Viral leader polypeptide zinc finger
g.88	Intrinsically disordered proteins

Continued on next page

Table A2 – continued from previous page

SCCS	Fold
g.89	CHY zinc finger-like
g.90	E6 C-terminal domain-like
g.91	E7 C-terminal domain-like
g.92	T-antigen specific domain-like
g.93	Zinc hairpin stack
h.1	Parallel coiled-coil
h.2	Tetramerization domain of the Mnt repressor
h.3	Stalk segment of viral fusion proteins
h.4	Antiparallel coiled-coil
h.5	Apolipoprotein A-I
h.6	Apolipoprotein A-II
h.7	Synuclein

Appendix B: Material supplementary to Chapter 2

Table B1: List of manually removed genomes from the set downloaded from SUPERFAMILY (1.75). Species are listed by their full name as quoted by SUPERFAMILY. This list includes species manually removed from the dataset due to a poor resolution of the phylogenetic trees during tree building.

Manually removed taxa
Acholeplasma laidlawii PG-8A
Advenella kashmirensis WT001
Anaeromyxobacter dehalogenans 2CP-1
Anaplasma marginale str. Florida
Babesia bovis T2Bo
Blattabacterium sp. (Blattella germanica) str. Bge
Borrelia crocidurae str. Achema
Caldisericum exile AZM16c01
Cryptosporidium hominis
Cryptosporidium muris
Entamoeba dispar
Entamoeba histolytica
Entamoeba invadens
Francisella noatunensis subsp. orientalis str. Toba 04
Giardia lamblia ATCC 50803
Helicobacter bizzozeronii CIII-1
Helicobacter cetorum MIT 00-7128
Ketogulonicigenium vulgare WSH-001
Micavibrio aeruginosavorus ARL-13
Mycoplasma haemocanis str. Illinois
Mycoplasma pneumoniae 309
Mycoplasma suis KI3806
Mycoplasma wenyonii str. Massachusetts
Nanoarchaeum equitans Kin4-M
Nautilia profundicola AmH
Photorhabdus asymbiotica
Rickettsia montanensis str. OSU 85-930
Rickettsia parkeri str. Portsmouth
Rickettsia philipii str. 364D
Continued on next page

Table B1 – continued from previous page

Manually removed taxa
Rickettsia rhipicephali str. 3-7-female6-CWPP
Secondary endosymbiont of Ctenarytaina eucalypti
Serratia symbiotica str. Cinara cedri
Taylorella asinigenitalis MCE3
Taylorella equigenitalis MCE9
Theileria annulata
Theileria parva
Wigglesworthia glossinidia endo. of G. brevipalpis
Wolbachia endosymbiont of Culex quinquefasciatus Pel
Wolbachia sp. wRi

Table B2: Age estimates calculated using different methods. Maximum parsimony is used on ALLgenomes using the eight different phylogenetic trees. Fusion parsimony is calculated on ALLgenomes using the NCBI tree. Dollo parsimony ages use the NCBI tree of MULTIgenomes, where a superfamily has nonempty occurrence on MULTIgenomes. If the occurrence is empty a dash (-) is given. Finally, fold ages are calculated using a maximum parsimony algorithm on the NCBI tree of ALLgenomes. These ages are given for the first superfamily of a fold. Subsequent members are shown with a dash.

	Maximum parsimony								Fusion	Dollo	
sf	ncbi	ncbif	jacc	jaccf	bc	bcf	pars	parsf		MULTI	fold
a.1.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
a.1.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.1	0.72	0.72	0.78	0.79	0.74	0.76	0.84	0.65	0.72	-	1.00
a.2.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.6	0.54	0.53	0.56	0.55	0.44	0.49	0.53	0.55	0.61	0.91	-
a.2.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.8	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.54	0.71	-
a.2.9	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.51	-
a.2.10	0.56	0.56	0.72	0.79	0.64	0.69	0.59	0.65	0.56	0.71	-
a.2.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.2.12	0.29	0.29	0.60	0.56	0.53	0.54	0.41	0.36	0.29	-	-
a.2.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.2.14	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	0.72	0.00	-
a.2.15	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
a.2.16	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.61	0.70	-
a.2.17	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.2.18	0.18	0.18	0.39	0.39	0.31	0.33	0.22	0.18	0.18	-	-
a.2.19	0.54	0.53	0.54	0.54	0.45	0.49	0.48	0.55	0.54	0.91	-
a.2.20	0.12	0.12	0.23	0.20	0.19	0.34	0.10	0.08	0.12	-	-
a.2.21	0.50	0.49	0.69	0.68	0.64	0.58	0.42	0.53	0.50	-	-
a.3.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00
a.4.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.4.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.4.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.6	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.50	-
a.4.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.9	0.72	0.72	0.83	0.83	0.74	0.82	1.00	1.00	1.00	1.00	-
a.4.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	-
a.4.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.12	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.71	-
a.4.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.4.14	0.72	0.72	0.83	0.83	1.00	0.82	0.84	0.65	1.00	0.50	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.4.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.5.1	0.72	0.72	0.83	0.79	0.80	0.82	0.84	0.82	1.00	0.00	1.00
a.5.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.5.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.5.4	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.5.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.5.7	0.72	0.72	0.83	0.79	1.00	0.82	0.84	0.65	1.00	1.00	-
a.5.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.5.9	0.35	0.34	0.43	0.42	0.31	0.32	0.30	0.52	0.54	0.58	-
a.5.10	0.61	0.60	0.70	0.69	0.64	0.65	0.74	0.76	0.61	1.00	-
a.6.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.7.1	0.61	0.60	0.56	0.55	0.48	0.50	0.53	0.60	0.61	1.00	1.00
a.7.2	0.50	0.49	0.69	0.66	0.64	0.69	0.41	0.58	0.50	-	-
a.7.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.7.4	0.30	0.29	0.35	0.32	0.27	0.25	0.23	0.16	0.35	0.50	-
a.7.5	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.7.6	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	0.88	-
a.7.7	0.61	0.60	0.62	0.62	0.57	0.50	0.53	0.63	0.61	1.00	-
a.7.8	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.7.10	0.09	0.10	0.33	0.38	0.30	0.33	0.12	0.08	0.09	-	-
a.7.11	0.14	0.13	0.27	0.24	0.18	0.18	0.17	0.12	0.61	0.23	-
a.7.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	-
a.7.13	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.00	-
a.7.14	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.7.15	0.58	0.61	0.83	1.00	0.80	1.00	0.60	0.82	0.61	0.49	-
a.7.16	0.30	0.29	0.47	0.44	0.42	0.40	0.33	0.36	0.54	0.50	-
a.8.1	0.18	0.18	0.43	0.39	0.33	0.33	0.18	0.11	0.18	-	1.00
a.8.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.8.3	1.00	1.00	0.74	0.83	0.70	0.73	0.62	0.63	1.00	1.00	-
a.8.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.8.6	0.06	0.06	0.19	0.18	0.13	0.12	0.04	0.06	0.06	0.11	-
a.8.7	0.24	0.25	0.51	0.55	0.47	0.48	0.41	0.25	0.61	0.00	-
a.8.10	0.52	0.52	0.77	0.75	0.72	0.70	0.76	0.72	0.52	-	-
a.8.11	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	0.00	-
a.9.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
a.10.2	0.21	0.20	0.41	0.38	0.32	0.31	0.32	0.21	0.21	-	0.21
a.11.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.11.2	0.61	0.60	0.62	0.56	0.57	0.49	0.62	0.63	0.61	1.00	-
a.12.1	0.54	0.53	0.49	0.45	0.40	0.40	0.44	0.52	0.61	0.71	0.54
a.13.1	0.38	0.37	0.43	0.42	0.33	0.30	0.30	0.27	0.38	0.63	0.38
a.14.1	0.61	0.60	0.62	0.56	0.57	0.49	0.62	0.63	0.61	1.00	0.61
a.15.1	0.51	0.50	0.46	0.45	0.40	0.38	0.41	0.45	0.61	0.88	0.51
a.16.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.17.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.18.1	0.24	0.23	0.54	0.52	0.49	0.48	0.21	0.27	0.24	-	0.24
a.20.1	0.61	0.61	0.83	0.79	1.00	0.82	0.68	0.82	0.61	0.88	0.61
a.21.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.22.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.23.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.23.2	0.38	0.37	0.60	0.68	0.52	0.54	0.28	0.40	0.38	-	-
a.23.3	0.28	0.28	0.55	0.57	0.50	0.28	0.14	0.14	0.28	-	-
a.23.4	0.54	0.53	0.54	0.55	0.44	0.43	0.53	0.55	0.54	0.91	-
a.23.5	0.21	0.21	0.42	0.41	0.36	0.35	0.17	0.15	0.21	-	-
a.23.6	0.37	0.36	0.62	0.62	0.64	0.57	0.41	0.45	0.37	-	-
a.23.7	0.11	0.12	0.34	0.33	0.28	0.28	0.15	0.14	0.11	-	-
a.24.1	0.51	0.50	0.56	0.56	0.51	0.53	0.23	0.34	0.61	0.88	1.00
a.24.2	0.30	0.30	0.48	0.52	0.44	0.49	0.23	0.23	0.30	-	-
a.24.3	0.45	0.45	0.64	0.67	0.61	0.64	0.49	0.46	0.45	0.00	-
a.24.4	0.56	0.56	0.75	0.79	0.71	0.65	0.57	0.54	0.61	0.32	-
a.24.5	0.24	0.25	0.45	0.25	0.39	0.41	0.26	0.32	0.54	0.45	-
a.24.7	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.24.8	0.61	0.60	0.62	0.61	0.49	0.50	0.62	0.63	0.61	0.79	-
a.24.9	0.54	0.53	0.53	0.49	0.44	0.41	0.47	0.52	0.61	0.71	-
a.24.10	1.00	1.00	1.00	1.00	1.00	1.00	0.68	1.00	1.00	1.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.24.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-
a.24.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.24.14	0.35	0.34	0.40	0.36	0.31	0.32	0.27	0.27	0.54	0.58	-
a.24.15	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.24.16	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.65	0.72	0.71	-
a.24.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-
a.24.18	0.51	0.50	0.64	0.62	0.64	0.59	0.74	0.76	0.61	1.00	-
a.24.19	0.72	0.72	0.75	0.79	0.74	0.76	0.59	0.65	0.72	-	-
a.24.20	0.17	0.16	0.37	0.36	0.30	0.31	0.17	0.22	0.17	-	-
a.24.21	0.37	0.36	0.52	0.52	0.43	0.46	0.29	0.27	0.37	0.00	-
a.24.22	0.35	0.35	0.56	0.56	0.56	0.50	0.36	0.38	0.61	0.49	-
a.24.23	0.30	0.29	0.37	0.33	0.27	0.26	0.23	0.17	0.61	0.50	-
a.24.24	0.36	0.37	0.53	0.52	0.45	0.44	0.50	0.46	0.61	0.51	-
a.24.25	0.58	0.61	0.75	0.79	0.74	1.00	0.74	0.76	0.61	1.00	-
a.24.26	0.37	0.36	0.62	0.62	0.64	0.57	0.41	0.45	0.37	-	-
a.24.27	0.22	0.23	0.60	0.59	0.50	0.54	0.37	0.40	0.61	0.00	-
a.24.28	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.24.29	0.56	0.56	0.66	0.70	0.70	0.70	0.59	0.58	0.56	0.00	-
a.25.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.25.2	1.00	1.00	0.78	0.79	0.74	0.76	1.00	0.63	1.00	1.00	-
a.25.3	0.52	0.51	0.65	0.66	0.56	0.60	0.47	0.42	0.61	0.19	-
a.25.4	0.25	0.27	0.54	0.52	0.47	0.46	0.26	0.31	0.25	-	-
a.25.5	0.52	0.53	0.69	0.00	0.64	0.00	0.38	0.00	0.61	0.00	-
a.26.1	0.26	0.25	0.28	0.27	0.21	0.21	0.23	0.16	0.26	0.43	0.26
a.27.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.28.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.28.2	0.08	0.08	0.23	0.24	0.19	0.21	0.04	0.04	0.08	-	-
a.28.3	0.30	0.29	0.34	0.32	0.26	0.25	0.25	0.16	0.43	0.50	-
a.29.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
a.29.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.29.5	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.29.6	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.41	0.70	-
a.29.7	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.29.8	0.26	0.26	0.49	0.42	0.40	0.38	0.22	0.21	0.26	-	-
a.29.9	0.72	0.72	0.78	0.79	0.74	0.76	0.84	0.65	0.72	0.00	-
a.29.10	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.43	0.71	-
a.29.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-
a.29.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.29.13	0.26	0.26	0.62	0.58	0.54	0.56	0.28	0.43	0.26	-	-
a.29.14	0.69	0.67	0.71	0.69	0.69	0.62	0.69	0.62	0.69	-	-
a.29.15	0.54	0.53	0.65	0.70	0.64	0.64	0.54	0.49	0.61	0.00	-
a.29.16	0.58	0.61	0.75	0.79	0.71	0.76	0.60	0.54	0.61	0.00	-
a.29.17	0.32	0.32	0.56	0.58	0.50	0.54	0.41	0.32	0.32	-	-
a.30.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
a.30.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.30.4	0.14	0.13	0.22	0.21	0.18	0.14	0.13	0.12	0.14	0.23	-
a.30.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.30.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.30.7	0.37	0.36	0.65	0.62	0.64	0.57	0.50	0.53	0.37	0.00	-
a.31.1	0.61	0.60	0.63	0.61	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.32.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.34.1	0.22	0.23	0.51	0.50	0.46	0.45	0.31	0.27	0.22	-	0.30
a.34.2	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.38	0.32	-
a.34.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.34.4	0.30	0.29	0.36	0.36	0.27	0.29	0.20	0.21	0.38	0.50	-
a.35.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.36.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.37.1	0.54	0.53	0.47	0.45	0.40	0.40	0.30	0.33	0.54	0.71	0.54
a.38.1	0.54	0.53	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.54
a.38.2	0.25	0.26	0.51	0.51	0.44	0.47	0.20	0.24	0.25	-	-
a.39.1	0.61	0.60	0.74	0.79	1.00	1.00	0.74	0.76	0.61	1.00	0.61
a.39.2	0.26	0.26	0.36	0.34	0.28	0.26	0.19	0.19	0.26	0.44	-
a.39.3	0.32	0.32	0.49	0.46	0.43	0.49	0.40	0.46	0.61	0.66	-
a.39.4	0.39	0.39	0.77	0.75	0.72	0.70	0.51	0.72	0.39	-	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.40.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.40.2	0.26	0.26	0.49	0.48	0.40	0.42	0.31	0.27	0.26	-	-
a.40.3	0.54	0.53	0.59	0.55	0.49	0.49	0.62	0.63	0.61	1.00	-
a.41.1	0.61	0.60	0.63	0.54	0.49	0.50	0.62	0.63	0.61	1.00	0.61
a.42.1	0.61	0.60	0.70	0.74	0.64	0.71	0.74	0.76	0.61	1.00	0.61
a.43.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.72	0.00	1.00
a.45.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
a.46.1	0.72	0.72	0.83	0.83	1.00	0.82	0.68	0.82	1.00	1.00	1.00
a.46.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.46.3	0.37	0.36	0.48	0.49	0.40	0.43	0.29	0.27	0.37	-	-
a.47.1	0.43	0.42	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	0.61
a.47.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.47.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.47.4	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.27	0.43	0.71	-
a.47.5	0.56	0.56	0.72	0.79	0.71	0.76	0.59	0.65	0.56	-	-
a.47.6	0.28	0.30	0.53	0.52	0.46	0.48	0.32	0.23	0.61	0.00	-
a.48.1	0.54	0.53	0.47	0.45	0.40	0.40	0.30	0.33	0.54	0.71	0.61
a.48.2	0.43	0.42	0.70	0.70	0.65	0.68	0.74	0.76	0.61	1.00	-
a.48.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.48.4	0.43	0.42	0.40	0.36	0.27	0.32	0.25	0.23	0.54	0.71	-
a.48.5	0.29	0.31	0.68	0.69	0.62	0.62	0.37	0.41	0.29	-	-
a.49.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.50.1	0.26	0.25	0.28	0.27	0.21	0.21	0.23	0.16	0.35	0.43	0.26
a.51.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.52.1	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.61	0.70	0.41
a.53.1	0.30	0.29	0.40	0.36	0.31	0.29	0.30	0.20	0.54	0.50	0.30
a.55.1	0.72	0.72	0.83	0.83	1.00	0.82	0.76	1.00	1.00	0.72	0.72
a.56.1	0.61	0.60	0.78	0.79	0.71	0.76	0.60	0.63	0.61	1.00	0.61
a.57.1	0.00	0.00	0.29	0.26	0.24	0.23	0.09	0.08	0.00	0.00	0.00
a.58.1	0.72	0.72	1.00	1.00	1.00	1.00	0.76	1.00	0.72	0.00	0.72
a.59.1	0.61	0.60	0.62	0.56	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.60.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	1.00
a.60.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.3	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.30	-
a.60.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.5	0.43	0.42	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	-
a.60.6	0.61	0.60	0.77	0.79	0.72	0.74	0.62	0.72	0.61	1.00	-
a.60.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.9	0.54	0.53	0.74	0.79	0.70	0.74	0.54	0.54	0.61	0.72	-
a.60.10	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.00	-
a.60.11	0.46	0.47	0.74	0.79	0.70	0.82	0.50	0.54	0.61	0.77	-
a.60.12	0.61	0.60	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	-
a.60.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.60.15	0.50	0.49	0.55	0.53	0.55	0.47	0.34	0.30	0.50	-	-
a.60.16	0.52	0.53	0.75	0.79	0.71	0.70	0.57	0.54	0.52	0.00	-
a.61.1	0.04	0.04	0.15	0.19	0.11	0.14	0.13	0.09	0.61	0.06	0.04
a.63.1	0.16	0.15	0.36	0.38	0.31	0.34	0.17	0.17	0.61	0.26	0.16
a.64.1	0.61	0.60	0.62	0.56	0.55	0.45	0.51	0.54	0.61	0.88	0.61
a.64.2	0.00	0.00	0.19	0.16	0.15	0.14	0.08	0.04	0.00	-	-
a.65.1	0.54	0.53	0.62	0.62	0.49	0.49	0.62	0.63	0.61	1.00	0.54
a.66.1	0.61	0.60	0.62	0.56	0.49	0.49	0.62	0.63	0.61	1.00	0.61
a.68.1	0.54	0.53	0.56	0.55	0.44	0.41	0.53	0.55	0.61	0.91	0.54
a.69.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.69.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.69.3	0.72	0.72	1.00	0.83	1.00	0.82	1.00	1.00	1.00	1.00	-
a.69.4	0.32	0.33	0.59	0.58	0.55	0.52	0.31	0.33	0.32	-	-
a.70.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.70.2	0.48	0.47	0.71	0.70	0.64	0.64	0.50	0.49	0.48	-	-
a.71.1	0.61	0.60	0.62	0.61	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.71.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.72.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.73.1	0.15	0.14	0.27	0.23	0.18	0.18	0.17	0.11	0.61	0.25	0.15

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.74.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.75.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.76.1	1.00	1.00	0.78	0.79	0.74	0.76	1.00	0.72	1.00	-	1.00
a.77.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
a.78.1	0.72	0.72	0.78	0.79	0.74	0.76	0.60	0.65	1.00	0.00	0.72
a.79.1	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	0.88	0.72
a.80.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.81.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.00	0.72
a.83.1	0.44	0.43	0.62	0.52	0.56	0.45	0.50	0.50	0.61	0.79	0.44
a.85.1	0.24	0.23	0.36	0.34	0.27	0.26	0.17	0.19	0.24	0.40	0.24
a.86.1	0.54	0.53	0.62	0.62	0.52	0.52	0.62	0.63	0.61	1.00	0.54
a.87.1	0.61	0.60	0.56	0.69	0.48	0.64	0.53	0.76	0.61	1.00	0.61
a.88.1	0.26	0.26	0.54	0.56	0.50	0.51	0.37	0.25	0.26	0.00	0.26
a.89.1	0.39	0.39	0.70	0.69	0.69	0.62	0.58	0.62	0.39	-	0.39
a.90.1	0.43	0.42	0.47	0.49	0.40	0.41	0.47	0.49	0.61	0.71	0.43
a.91.1	0.61	0.60	0.62	0.62	0.49	0.50	0.62	0.63	0.61	1.00	0.61
a.92.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.93.1	0.61	0.61	0.75	1.00	0.71	1.00	0.74	1.00	0.61	1.00	0.61
a.94.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.96.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.97.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.98.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.99.1	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	1.00
a.100.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.101.1	0.17	0.15	0.27	0.26	0.43	0.17	0.19	0.18	0.54	0.34	0.17
a.102.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.102.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.102.3	0.54	0.53	0.75	0.79	0.71	0.74	0.53	0.47	0.61	0.77	-
a.102.4	0.61	0.60	0.75	1.00	1.00	1.00	0.74	1.00	0.61	1.00	-
a.102.5	0.35	0.38	0.64	0.79	0.59	0.72	0.36	0.31	0.61	0.54	-
a.102.6	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.103.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.104.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
a.108.1	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.109.1	0.18	0.17	0.28	0.27	0.21	0.21	0.20	0.16	0.20	0.30	0.18
a.110.1	0.69	0.67	0.77	0.79	0.72	0.70	0.69	0.62	0.69	0.00	0.69
a.111.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.113.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.114.1	0.61	0.60	0.62	0.61	0.57	0.49	0.62	0.63	0.61	1.00	0.61
a.116.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.117.1	0.61	0.60	0.56	0.55	0.48	0.50	0.53	0.60	0.61	0.91	0.61
a.118.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.118.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.118.4	0.43	0.42	0.53	0.63	0.45	0.59	0.44	0.33	0.61	0.71	-
a.118.5	0.54	0.53	0.65	0.66	0.61	0.64	0.60	0.61	0.61	0.00	-
a.118.6	0.61	0.60	1.00	1.00	0.69	1.00	0.74	0.76	0.61	1.00	-
a.118.7	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.118.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.118.9	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.118.11	0.54	0.53	0.54	0.55	0.44	0.43	0.53	0.55	0.54	0.91	-
a.118.12	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.23	0.43	0.71	-
a.118.13	0.61	0.60	0.62	0.61	0.49	0.50	0.62	0.63	0.61	1.00	-
a.118.14	0.72	0.72	0.78	0.79	0.74	0.76	0.68	0.65	1.00	0.00	-
a.118.15	0.54	0.53	0.64	0.65	0.58	0.60	0.49	0.61	0.61	0.49	-
a.118.16	1.00	1.00	0.71	0.69	0.69	0.64	1.00	0.76	1.00	1.00	-
a.118.17	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.118.18	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	-
a.118.19	0.61	0.60	0.62	0.62	0.49	0.49	0.53	0.63	0.61	1.00	-
a.118.20	0.32	0.31	0.51	0.48	0.43	0.41	0.33	0.21	0.32	-	-
a.118.21	0.24	0.23	0.36	0.34	0.27	0.26	0.17	0.19	0.24	0.40	-
a.118.22	0.61	0.60	0.62	0.49	0.44	0.41	0.50	0.49	0.61	0.71	-
a.118.23	0.54	0.53	0.56	0.55	0.45	0.49	0.53	0.60	0.61	0.91	-
a.118.24	0.61	0.60	0.62	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
a.118.25	0.54	0.53	0.61	0.67	0.56	0.63	0.47	0.52	0.61	0.71	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.118.26	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.72	-
a.119.1	0.51	0.50	0.63	0.59	0.57	0.56	0.56	0.63	0.61	1.00	0.51
a.121.1	0.72	0.72	1.00	1.00	1.00	1.00	0.84	0.82	1.00	0.00	0.72
a.123.1	0.43	0.42	0.43	0.42	0.37	0.32	0.30	0.32	0.61	0.71	0.43
a.124.1	0.61	0.61	0.73	0.79	0.69	0.74	0.62	0.63	0.61	0.77	0.61
a.126.1	0.26	0.25	0.28	0.27	0.21	0.21	0.23	0.16	0.35	0.43	0.26
a.127.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.128.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.129.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.130.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.132.1	0.69	0.67	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	0.69
a.133.1	0.43	0.42	0.62	0.66	0.57	0.62	0.56	0.63	0.61	0.71	0.43
a.134.1	0.28	0.28	0.49	0.46	0.42	0.40	0.35	0.41	0.61	0.51	0.28
a.135.1	0.43	0.42	0.49	0.49	0.42	0.41	0.36	0.52	0.61	0.91	0.43
a.136.1	0.32	0.32	0.55	0.56	0.50	0.59	0.43	0.35	0.32	-	0.32
a.137.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.137.2	0.28	0.28	0.55	0.57	0.50	0.49	0.30	0.17	0.28	-	-
a.137.3	0.54	0.53	0.54	0.54	0.44	0.49	0.53	0.60	0.61	0.91	-
a.137.4	0.23	0.21	0.57	0.60	0.50	0.51	0.41	0.33	0.61	0.00	-
a.137.5	0.54	0.53	0.47	0.45	0.40	0.40	0.30	0.33	0.54	0.71	-
a.137.7	0.09	0.10	0.31	0.30	0.23	0.25	0.07	0.08	0.09	-	-
a.137.8	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.137.9	0.00	0.00	0.53	0.00	0.48	0.00	0.00	0.00	0.00	-	-
a.137.10	0.43	0.42	0.40	0.42	0.31	0.32	0.30	0.25	0.54	0.71	-
a.137.11	0.56	0.56	0.72	0.79	0.71	0.76	0.59	0.65	0.56	-	-
a.137.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	-
a.137.13	0.00	0.00	0.41	0.35	0.00	0.29	0.11	0.08	0.00	-	-
a.137.14	0.15	0.14	0.37	0.32	0.25	0.25	0.15	0.17	0.15	0.25	-
a.137.15	0.25	0.24	0.60	0.64	0.55	0.62	0.26	0.30	0.25	0.00	-
a.138.1	0.72	0.72	0.78	0.79	0.74	0.76	0.84	0.65	1.00	1.00	0.72
a.139.1	0.52	0.53	0.71	0.75	0.72	0.70	0.59	0.57	0.54	0.71	0.52
a.140.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.61	0.71	1.00
a.140.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
a.140.3	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.64	-
a.140.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.140.5	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.61	0.70	-
a.140.6	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
a.141.1	0.44	0.43	0.59	0.52	0.48	0.44	0.36	0.41	0.61	1.00	0.44
a.142.1	0.50	0.49	0.69	0.68	0.64	0.69	0.50	0.58	0.61	0.00	0.50
a.143.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.144.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
a.144.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
a.145.1	0.43	0.43	0.52	0.52	0.44	0.49	0.33	0.27	0.43	-	0.43
a.146.1	0.26	0.24	0.34	0.32	0.26	0.25	0.20	0.16	0.38	0.42	0.26
a.147.1	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.35	0.32	0.20
a.148.1	0.61	0.60	0.56	0.56	0.49	0.49	0.53	0.63	0.61	1.00	0.61
a.149.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.151.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	1.00
a.152.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.153.1	0.30	0.29	0.38	0.36	0.28	0.28	0.25	0.23	0.38	0.50	0.30
a.154.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.155.1	0.45	0.45	0.59	0.59	0.54	0.59	0.49	0.41	0.45	0.00	0.45
a.156.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.157.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.158.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.159.1	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.25	0.43	0.71	0.61
a.159.2	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
a.159.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
a.159.4	0.54	0.53	0.62	0.56	0.64	0.64	0.74	0.76	0.61	1.00	-
a.159.5	0.43	0.43	0.59	0.59	0.53	0.59	0.46	0.31	0.43	-	-
a.160.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.161.1	0.54	0.53	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	0.54
a.162.1	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	0.72
a.163.1	0.30	0.29	0.43	0.34	0.32	0.26	0.19	0.19	0.30	0.49	0.30

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.164.1	0.20	0.19	0.28	0.32	0.21	0.21	0.20	0.16	0.35	0.32	0.20
a.165.1	0.43	0.42	0.40	0.42	0.31	0.32	0.30	0.27	0.43	0.71	0.43
a.166.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.169.1	0.61	0.60	0.63	0.61	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.170.1	0.61	0.60	0.62	0.56	0.55	0.45	0.62	0.63	0.61	1.00	0.61
a.171.1	0.43	0.42	0.43	0.42	0.37	0.34	0.36	0.39	0.61	0.71	0.43
a.172.1	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	0.72
a.173.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.174.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.175.1	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.41	-	0.41
a.176.1	0.45	0.45	0.65	0.66	0.61	0.59	0.54	0.46	0.61	0.00	0.45
a.177.1	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	0.72
a.179.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.180.1	0.45	0.45	0.61	0.63	0.56	0.59	0.45	0.41	0.45	-	0.45
a.181.1	0.57	0.57	0.65	0.68	0.64	0.69	0.47	0.53	0.57	-	0.57
a.182.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.183.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.184.1	0.57	0.57	0.78	0.79	0.64	0.66	0.59	0.57	0.57	0.00	0.57
a.186.1	0.41	0.39	0.63	0.62	0.64	0.59	0.37	0.61	0.41	-	0.41
a.187.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.188.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.189.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.191.1	0.47	0.47	0.72	0.77	0.64	0.70	0.53	0.53	0.61	0.50	0.47
a.192.1	0.61	0.60	0.62	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.193.1	0.43	0.42	0.40	0.40	0.33	0.38	0.40	0.43	0.61	0.71	0.43
a.194.1	0.43	0.42	0.43	0.42	0.35	0.36	0.30	0.33	0.54	0.71	0.43
a.195.1	0.58	0.61	0.77	0.79	0.72	0.76	0.59	0.72	0.58	-	0.58
a.198.1	0.45	0.45	0.61	0.62	0.54	0.59	0.45	0.39	0.61	0.00	0.45
a.199.1	0.45	0.45	0.61	0.63	0.56	0.59	0.49	0.41	0.61	0.00	0.45
a.200.1	0.33	0.33	0.64	0.54	0.49	0.45	0.41	0.38	0.33	-	0.33
a.202.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00
a.203.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.204.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.205.1	0.54	0.53	0.62	0.56	0.57	0.49	0.53	0.55	0.61	1.00	0.54
a.206.1	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.10	0.00	0.00
a.207.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.208.1	0.57	0.57	0.78	0.77	0.70	0.70	0.62	0.65	0.61	1.00	0.57
a.209.1	1.00	1.00	0.83	0.79	1.00	0.68	0.68	0.56	1.00	1.00	1.00
a.210.1	0.43	0.43	0.49	0.48	0.44	0.49	0.48	0.46	0.61	0.77	0.43
a.211.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.212.1	0.18	0.17	0.28	0.27	0.21	0.21	0.19	0.14	0.35	0.30	0.18
a.213.1	0.58	0.61	0.83	0.83	0.80	0.82	0.60	0.61	0.61	0.50	0.58
a.214.1	0.41	0.39	0.63	0.62	0.58	0.59	0.31	0.61	0.41	-	0.41
a.215.1	0.54	0.53	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	0.54
a.216.1	0.54	0.53	0.56	0.55	0.44	0.49	0.53	0.55	0.61	0.91	0.54
a.217.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.218.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.219.1	0.16	0.16	0.50	0.48	0.42	0.43	0.31	0.23	0.16	-	0.16
a.220.1	0.29	0.29	0.42	0.45	0.36	0.36	0.32	0.36	0.33	0.50	0.29
a.221.1	0.54	0.53	0.62	0.62	0.49	0.50	0.53	0.63	0.61	1.00	0.54
a.222.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.223.1	0.72	0.72	1.00	0.83	1.00	0.82	1.00	1.00	1.00	0.88	0.72
a.224.1	0.61	0.60	0.70	0.56	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.225.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
a.226.1	0.24	0.23	0.37	0.33	0.28	0.26	0.19	0.21	0.24	0.40	0.24
a.227.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.228.1	0.35	0.34	0.40	0.36	0.30	0.30	0.27	0.27	0.35	0.58	0.35
a.229.1	0.11	0.12	0.32	0.31	0.26	0.26	0.11	0.10	0.11	-	0.11
a.230.1	0.22	0.19	0.61	0.66	0.61	0.69	0.31	0.45	0.22	-	0.22
a.231.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.232.1	0.17	0.18	0.36	0.36	0.28	0.30	0.16	0.17	0.54	-	0.17
a.233.1	0.28	0.30	0.53	0.52	0.46	0.48	0.28	0.23	0.28	-	0.28
a.234.1	0.29	0.29	0.62	0.62	0.53	0.50	0.41	0.45	0.29	-	0.29
a.235.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
a.236.1	0.45	0.45	0.61	0.63	0.56	0.59	0.45	0.41	0.45	0.00	0.45
a.237.1	0.21	0.21	0.42	0.41	0.36	0.35	0.17	0.15	0.21	-	0.21
a.238.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.239.1	0.27	0.28	0.58	0.59	0.51	0.56	0.34	0.43	0.27	-	0.27
a.240.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.242.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.243.1	0.28	0.27	0.49	0.47	0.42	0.41	0.32	0.19	0.28	-	0.28
a.245.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.246.1	0.25	0.26	0.53	0.51	0.47	0.43	0.16	0.13	0.25	-	0.54
a.246.2	0.47	0.47	0.77	0.75	0.80	0.70	0.76	0.61	0.47	0.70	-
a.246.3	0.54	0.53	0.56	0.55	0.45	0.49	0.53	0.60	0.61	0.91	-
a.247.1	0.15	0.15	0.33	0.35	0.27	0.30	0.09	0.08	0.15	-	0.15
a.248.1	0.56	0.56	0.72	0.79	0.60	0.76	0.50	0.53	0.56	-	0.56
a.249.1	0.11	0.12	0.34	0.33	0.28	0.28	0.15	0.14	0.11	-	0.11
a.250.1	0.12	0.12	0.23	0.20	0.19	0.34	0.10	0.08	0.12	-	0.12
a.251.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
a.252.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
a.253.1	0.26	0.27	0.49	0.48	0.41	0.41	0.39	0.30	0.26	0.00	0.26
a.254.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00
a.255.1	0.32	0.34	0.58	0.56	0.58	0.57	0.45	0.34	0.61	0.00	0.32
a.256.1	0.61	0.60	0.54	0.52	0.45	0.44	0.50	0.49	0.61	0.79	0.61
a.258.1	0.29	0.31	0.57	0.61	0.51	0.54	0.27	0.36	0.29	-	0.29
a.259.1	0.30	0.30	0.59	0.70	0.49	0.51	0.32	0.39	0.30	0.00	0.30
a.260.1	0.00	0.00	0.36	0.34	0.27	0.25	0.00	0.00	0.30	0.00	0.00
a.261.1	0.51	0.50	0.64	0.62	0.64	0.59	0.74	0.76	0.61	1.00	0.51
a.262.1	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	1.00	0.42	0.69
a.263.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.264.1	0.18	0.18	0.39	0.35	0.26	0.25	0.18	0.15	0.61	0.00	0.18
a.265.1	0.72	0.72	0.83	1.00	1.00	1.00	0.84	0.82	1.00	0.71	0.72
a.266.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.267.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
a.268.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.269.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.270.1	0.33	0.33	0.43	0.45	0.42	0.36	0.36	0.41	0.61	0.63	0.33
a.271.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.54	0.71	0.43
a.272.1	0.50	0.49	0.69	0.62	0.64	0.57	0.41	0.45	0.50	-	0.50
a.273.1	0.43	0.42	0.40	0.42	0.31	0.32	0.30	0.25	0.43	0.71	0.43
a.274.1	0.72	0.72	0.83	1.00	0.80	0.82	0.84	0.82	1.00	0.24	0.72
a.275.1	0.56	0.56	0.72	0.68	0.64	0.69	0.59	0.58	0.56	0.00	0.56
a.276.1	0.50	0.49	0.69	0.62	0.64	0.57	0.41	0.45	0.50	-	0.50
a.277.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
a.278.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
a.279.1	0.36	0.35	0.60	0.61	0.51	0.56	0.32	0.24	0.36	-	0.36
a.280.1	0.51	0.50	0.64	0.62	0.64	0.59	0.51	0.61	0.61	0.88	0.51
a.281.1	0.56	0.56	0.69	0.68	0.64	0.69	0.59	0.58	0.56	0.00	0.56
a.282.1	0.23	0.22	0.52	0.50	0.46	0.47	0.27	0.21	0.23	0.00	0.23
a.283.1	0.51	0.50	0.57	0.56	0.55	0.45	0.51	0.54	0.61	0.88	0.51
a.284.1	0.32	0.32	0.53	0.56	0.46	0.51	0.35	0.30	0.32	-	0.32
a.285.1	0.28	0.30	0.53	0.52	0.46	0.48	0.28	0.23	0.28	-	0.28
a.286.1	0.25	0.24	0.42	0.45	0.36	0.40	0.24	0.20	0.25	-	0.25
a.287.1	0.58	0.61	0.83	0.83	0.80	0.82	0.60	0.61	0.61	0.00	0.58
a.288.1	0.51	0.50	0.62	0.64	0.57	0.60	0.62	0.63	0.61	1.00	0.51
a.289.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
a.290.1	0.51	0.50	0.63	0.66	0.59	0.62	0.51	0.54	0.61	0.88	0.51
a.292.1	0.35	0.34	0.57	0.58	0.51	0.52	0.35	0.38	0.35	-	0.35
a.293.1	0.32	0.32	0.58	0.60	0.53	0.58	0.38	0.39	0.61	0.00	0.32
a.294.1	0.61	0.60	0.83	0.79	1.00	1.00	0.74	1.00	0.61	1.00	0.61
a.295.1	0.30	0.28	0.58	0.60	0.53	0.58	0.40	0.39	0.30	-	0.30
a.296.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
b.1.1	0.43	0.42	0.65	0.65	0.61	0.60	0.41	0.50	0.61	0.79	1.00
b.1.2	1.00	1.00	0.78	0.79	0.74	0.76	0.84	0.65	1.00	1.00	-
b.1.3	0.58	0.61	0.77	0.79	0.72	0.74	0.76	0.72	0.61	0.79	-
b.1.4	0.61	0.61	0.78	0.79	0.74	0.76	0.62	0.63	0.61	1.00	-
b.1.5	0.43	0.42	0.43	0.42	0.43	0.32	0.30	0.25	0.43	0.71	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.1.6	0.58	0.61	0.75	0.79	0.71	0.76	0.60	0.60	0.61	0.91	-
b.1.7	0.23	0.23	0.49	0.48	0.43	0.40	0.18	0.17	0.23	-	-
b.1.8	0.61	0.60	0.69	0.74	0.70	0.71	0.62	0.63	0.61	1.00	-
b.1.9	0.61	0.61	0.78	0.79	0.71	0.76	0.62	0.63	0.61	0.91	-
b.1.10	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.1.11	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.1.12	0.61	0.60	0.70	0.79	0.64	0.71	0.74	0.76	0.61	1.00	-
b.1.13	0.56	0.56	0.70	0.79	0.69	0.73	0.58	0.62	0.61	-	-
b.1.14	0.56	0.56	0.75	0.79	0.72	0.76	0.76	0.72	0.61	0.71	-
b.1.15	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	-
b.1.16	0.58	0.61	0.75	0.79	0.71	0.74	0.60	0.57	0.61	1.00	-
b.1.17	0.54	0.53	0.61	0.63	0.56	0.59	0.49	0.41	0.61	0.00	-
b.1.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.1.19	0.14	0.16	0.38	0.36	0.32	0.31	0.14	0.14	0.14	-	-
b.1.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
b.1.21	0.18	0.19	0.42	0.37	0.31	0.33	0.25	0.25	0.35	0.31	-
b.1.22	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.1.23	0.61	0.60	0.74	0.79	0.70	0.76	0.74	0.76	0.61	1.00	-
b.1.25	0.56	0.56	0.72	0.79	0.67	0.62	0.48	0.49	0.61	0.77	-
b.1.26	0.39	0.39	0.70	0.69	0.69	0.62	0.51	0.62	0.61	-	-
b.1.27	0.54	0.53	0.65	0.79	0.60	0.65	0.50	0.61	0.61	0.79	-
b.1.28	0.17	0.17	0.60	0.51	0.53	0.46	0.26	0.25	0.17	-	-
b.2.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.61
b.2.2	0.57	0.57	0.73	0.79	0.72	0.71	0.59	0.53	0.61	0.70	-
b.2.3	0.50	0.49	0.69	0.62	0.64	0.57	0.45	0.46	0.61	0.00	-
b.2.4	0.43	0.42	0.59	0.64	0.55	0.62	0.30	0.28	0.61	0.71	-
b.2.5	0.54	0.53	0.56	0.55	0.44	0.49	0.53	0.55	0.61	0.91	-
b.2.6	0.41	0.39	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.64	-
b.2.7	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.2.8	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.2.9	0.30	0.29	0.37	0.35	0.31	0.23	0.23	0.16	0.54	0.50	-
b.2.10	0.54	0.53	0.74	0.70	0.70	0.67	0.60	0.54	0.54	0.00	-
b.3.1	0.61	0.61	0.75	1.00	0.71	1.00	0.74	0.76	0.61	1.00	1.00
b.3.2	1.00	1.00	1.00	0.79	1.00	0.74	0.76	0.72	1.00	0.71	-
b.3.3	0.44	0.43	0.62	0.67	0.64	0.63	0.74	0.76	0.61	1.00	-
b.3.4	0.51	0.50	0.62	0.64	0.57	0.60	0.62	0.63	0.61	1.00	-
b.3.5	0.58	0.61	0.74	0.79	0.70	0.68	0.50	0.49	0.61	0.71	-
b.3.6	0.44	0.43	0.66	0.79	0.59	0.72	0.41	0.49	0.61	0.79	-
b.3.7	0.47	0.47	0.75	0.79	0.71	0.74	0.50	0.56	0.61	0.24	-
b.4.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.5.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
b.6.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.6.2	0.18	0.19	0.32	0.32	0.23	0.24	0.22	0.18	0.18	-	-
b.7.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.7.2	0.30	0.29	0.59	0.50	0.46	0.48	0.34	0.28	0.61	0.00	-
b.7.3	0.51	0.50	0.60	0.66	0.53	0.60	0.51	0.54	0.61	0.88	-
b.7.4	0.26	0.24	0.34	0.32	0.26	0.25	0.20	0.16	0.35	0.42	-
b.7.5	0.46	0.47	0.64	0.65	0.57	0.65	0.45	0.36	0.46	-	-
b.8.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.9.1	0.35	0.34	0.38	0.36	0.27	0.28	0.25	0.23	0.35	0.58	0.35
b.11.1	0.47	0.47	0.59	0.64	0.55	0.62	0.47	0.45	0.61	0.43	0.47
b.12.1	0.54	0.53	0.57	0.56	0.64	0.45	0.51	0.54	0.61	0.88	0.54
b.14.1	0.61	0.60	0.63	0.62	0.49	0.50	0.62	0.63	0.61	1.00	0.61
b.15.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.16.1	0.35	0.35	0.51	0.51	0.38	0.41	0.28	0.24	0.61	0.00	0.35
b.17.1	1.00	1.00	1.00	0.79	0.72	0.74	1.00	0.72	1.00	0.91	1.00
b.18.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.20.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00
b.22.1	0.43	0.42	0.59	0.55	0.54	0.50	0.30	0.27	0.61	0.71	0.43
b.23.1	0.61	0.60	0.62	0.66	0.57	0.62	0.47	0.52	0.61	1.00	0.61
b.23.2	0.47	0.47	0.64	0.70	0.60	0.64	0.50	0.49	0.61	0.00	-
b.23.3	0.57	0.57	0.69	0.70	0.70	0.69	0.56	0.59	0.61	0.88	-
b.24.1	0.30	0.29	0.58	0.58	0.51	0.53	0.33	0.40	0.54	0.50	0.30
b.25.1	0.41	0.41	0.56	0.62	0.52	0.60	0.53	0.46	0.61	1.00	0.41

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.26.1	0.61	0.60	1.00	1.00	0.70	1.00	0.76	0.76	0.61	1.00	0.61
b.29.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.30.2	0.51	0.50	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
b.30.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.30.6	0.18	0.18	0.43	0.33	0.27	0.28	0.16	0.18	0.18	-	-
b.33.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.34.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.72	0.00	1.00
b.34.2	0.61	0.60	0.63	0.63	0.57	0.59	0.62	0.63	0.61	1.00	-
b.34.3	0.43	0.43	0.56	0.55	0.44	0.43	0.48	0.46	0.61	0.77	-
b.34.4	0.51	0.50	0.64	0.62	0.64	0.59	0.74	0.76	0.61	0.88	-
b.34.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.34.6	0.57	0.57	0.83	0.79	0.80	0.82	0.59	1.00	0.57	0.00	-
b.34.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	-
b.34.8	0.61	0.60	0.70	0.79	0.65	0.68	0.74	0.76	0.61	1.00	-
b.34.9	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.34.10	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.34.11	0.58	0.61	0.65	0.77	0.64	0.71	0.60	0.61	0.61	0.00	-
b.34.12	0.30	0.30	0.44	0.42	0.37	0.39	0.33	0.32	0.61	0.58	-
b.34.13	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.34.14	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
b.34.15	0.37	0.36	0.62	0.59	0.64	0.54	0.41	0.40	0.37	-	-
b.34.16	0.24	0.25	0.62	0.59	0.64	0.54	0.37	0.40	0.24	-	-
b.34.17	0.54	0.53	0.54	0.55	0.51	0.53	0.51	0.55	0.61	1.00	-
b.34.18	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.70	-
b.34.19	0.43	0.42	0.49	0.48	0.42	0.40	0.35	0.41	0.61	0.79	-
b.34.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
b.34.21	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.35.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.35.2	0.50	0.49	0.69	0.66	0.64	0.69	0.50	0.45	0.50	-	-
b.36.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.38.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.38.2	0.72	0.72	0.83	0.83	0.80	0.82	0.68	0.82	1.00	0.00	-
b.38.3	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	-	-
b.38.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.38.5	0.33	0.33	0.59	0.60	0.55	0.56	0.59	0.57	0.33	-	-
b.39.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.40.1	1.00	1.00	0.73	0.79	0.69	1.00	1.00	1.00	1.00	1.00	1.00
b.40.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
b.40.3	0.43	0.42	0.44	0.56	0.39	0.51	0.40	0.38	0.61	0.71	-
b.40.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.40.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.40.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	-
b.40.7	0.72	0.72	1.00	1.00	1.00	1.00	0.68	1.00	1.00	0.00	-
b.40.8	0.45	0.45	0.69	0.70	0.65	0.68	0.45	0.37	0.45	0.00	-
b.40.9	0.58	0.61	0.78	0.79	0.71	0.76	0.60	0.54	0.61	0.70	-
b.40.10	0.28	0.30	0.56	0.62	0.52	0.60	0.32	0.22	0.61	0.00	-
b.40.11	0.21	0.19	0.50	0.59	0.52	0.46	0.27	0.22	0.21	0.00	-
b.40.12	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.65	0.72	-	-
b.40.13	0.22	0.23	0.61	0.49	0.43	0.43	0.31	0.23	0.22	-	-
b.40.14	0.52	0.52	0.72	0.83	0.72	0.72	0.58	0.82	0.52	-	-
b.40.15	0.41	0.39	0.66	0.72	0.60	0.70	0.38	0.61	0.41	-	-
b.40.16	0.11	0.10	0.22	0.21	0.16	0.16	0.14	0.11	0.18	0.18	-
b.41.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
b.42.1	0.43	0.42	0.43	0.42	0.35	0.36	0.30	0.33	0.61	0.71	0.61
b.42.2	0.61	0.60	0.68	0.69	0.62	0.62	0.62	0.63	0.61	1.00	-
b.42.3	0.14	0.13	0.45	0.42	0.39	0.33	0.41	0.17	0.37	0.23	-
b.42.4	0.33	0.33	0.43	0.45	0.37	0.36	0.36	0.39	0.61	0.55	-
b.42.5	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.42.6	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.42.7	0.54	0.53	0.54	0.54	0.44	0.43	0.48	0.55	0.54	0.91	-
b.42.8	0.41	0.41	0.56	0.50	0.52	0.45	0.41	0.45	0.61	0.70	-
b.43.2	0.56	0.56	0.72	0.70	0.62	0.73	0.48	0.53	0.56	-	1.00
b.43.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.43.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.43.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.44.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.44.2	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	-
b.45.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.45.2	0.58	0.59	0.74	0.79	0.71	0.82	0.60	0.58	0.58	-	-
b.45.3	0.11	0.12	0.50	0.49	0.42	0.43	0.23	0.23	0.11	-	-
b.46.1	1.00	1.00	0.83	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.47.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.48.1	0.27	0.27	0.59	0.60	0.52	0.55	0.32	0.34	0.27	-	0.27
b.49.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.49.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.49.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.50.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	0.61
b.51.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.52.1	1.00	1.00	0.83	0.83	0.80	0.82	0.76	0.82	1.00	1.00	1.00
b.52.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.53.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.54.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
b.55.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.55.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
b.56.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.57.1	0.41	0.41	0.54	0.55	0.48	0.53	0.41	0.34	0.41	-	0.41
b.58.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.59.1	0.43	0.42	0.43	0.45	0.37	0.36	0.36	0.39	0.61	0.71	0.43
b.60.1	0.61	0.61	0.72	0.79	0.70	0.76	0.74	1.00	0.61	1.00	0.61
b.61.1	0.43	0.42	0.42	0.40	0.36	0.36	0.23	0.20	0.43	0.71	0.58
b.61.2	0.30	0.28	0.53	0.55	0.47	0.53	0.35	0.30	0.30	0.00	-
b.61.3	0.15	0.16	0.38	0.40	0.43	0.35	0.24	0.27	0.54	0.38	-
b.61.4	0.00	0.00	0.53	0.00	0.48	0.00	0.00	0.00	0.00	-	-
b.61.5	0.50	0.48	0.62	0.61	0.47	0.49	0.47	0.40	0.61	0.71	-
b.61.6	0.58	0.61	0.83	0.79	0.71	0.82	0.60	0.54	0.61	0.00	-
b.61.8	0.45	0.45	0.60	0.62	0.54	0.59	0.41	0.41	0.45	0.00	-
b.62.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.63.1	0.15	0.14	0.27	0.24	0.18	0.18	0.17	0.12	0.15	0.25	0.15
b.64.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.65.1	0.61	0.60	0.62	0.56	0.57	0.49	0.62	0.63	0.61	1.00	0.61
b.66.1	0.43	0.42	0.40	0.49	0.31	0.41	0.30	0.49	0.61	0.71	0.43
b.67.1	0.23	0.22	0.39	0.36	0.31	0.29	0.18	0.14	0.38	0.37	1.00
b.67.2	1.00	1.00	0.78	0.79	0.72	0.76	0.68	0.63	1.00	1.00	-
b.67.3	0.61	0.60	0.59	0.52	0.64	0.45	0.50	0.76	0.61	0.79	-
b.68.1	0.58	0.61	0.78	0.79	0.80	0.82	0.84	0.59	0.61	0.71	1.00
b.68.2	0.61	0.61	1.00	1.00	1.00	1.00	0.76	0.72	0.61	1.00	-
b.68.3	0.41	0.39	0.64	0.65	0.60	0.60	0.40	0.61	0.61	0.51	-
b.68.4	0.61	0.60	0.62	0.64	0.57	0.62	0.62	0.63	0.61	1.00	-
b.68.5	0.61	0.61	0.75	0.83	1.00	0.82	0.62	0.63	0.61	1.00	-
b.68.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.68.7	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.59	1.00	0.71	-
b.68.8	0.35	0.35	0.61	0.66	0.58	0.60	0.47	0.49	0.61	0.58	-
b.68.9	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	-
b.68.10	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.68.11	0.61	0.61	0.75	0.79	0.71	1.00	0.74	0.76	0.61	1.00	-
b.69.1	0.61	0.60	0.74	0.79	0.70	0.76	0.74	0.76	0.61	1.00	1.00
b.69.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	-
b.69.3	0.58	0.61	0.75	0.79	0.71	0.82	0.60	0.54	0.61	0.79	-
b.69.4	0.61	0.61	0.70	0.79	0.64	1.00	0.74	0.76	0.61	1.00	-
b.69.5	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.69.6	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.69.7	0.61	0.61	0.83	0.79	0.71	0.82	0.62	0.63	0.61	1.00	-
b.69.8	0.61	0.61	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	-
b.69.9	0.61	0.60	0.75	0.79	0.71	0.74	0.74	0.76	0.61	1.00	-
b.69.10	0.61	0.61	0.83	0.83	1.00	0.74	0.62	0.63	0.61	1.00	-
b.69.11	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.69.12	0.43	0.42	0.43	0.42	0.35	0.36	0.30	0.33	0.54	0.71	-
b.69.13	0.61	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.61	1.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.69.14	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
b.70.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.70.2	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.76	1.00	1.00	-
b.70.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.71.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.72.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.72.2	0.57	0.57	0.59	0.64	0.61	0.62	0.59	0.57	0.57	0.63	-
b.72.3	0.54	0.53	0.62	0.56	0.57	0.45	0.62	0.63	0.61	1.00	-
b.73.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.74.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.75.1	0.34	0.36	0.59	0.57	0.56	0.51	0.33	0.38	0.34	-	0.34
b.76.2	0.61	0.61	0.73	0.79	0.69	0.76	0.74	0.76	0.61	1.00	0.61
b.77.1	0.27	0.26	0.50	0.47	0.48	0.38	0.29	0.32	0.61	0.45	0.51
b.77.3	0.51	0.50	0.54	0.56	0.51	0.49	0.56	0.63	0.61	1.00	-
b.78.1	0.51	0.50	0.59	0.62	0.52	0.60	0.56	0.59	0.61	0.88	0.51
b.80.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.80.2	0.30	0.29	0.33	0.32	0.25	0.24	0.18	0.15	0.61	0.50	-
b.80.3	0.47	0.47	0.71	0.70	0.70	0.69	0.59	0.82	0.61	0.79	-
b.80.4	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.80.5	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
b.80.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.80.7	0.43	0.43	0.71	0.72	0.66	0.82	0.50	0.61	0.61	0.50	-
b.80.8	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	-
b.81.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.81.3	0.54	0.53	0.63	0.62	0.55	0.49	0.51	0.49	0.61	0.71	-
b.81.4	0.58	0.61	0.83	0.79	0.80	0.82	0.60	0.54	0.58	0.42	-
b.82.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.82.2	0.61	0.61	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	-
b.82.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.82.4	0.56	0.56	0.75	0.79	0.71	0.74	0.54	0.58	0.61	0.00	-
b.82.5	0.57	0.57	0.77	0.79	0.72	0.74	0.76	0.72	0.61	1.00	-
b.82.6	0.61	0.60	0.72	0.70	0.64	0.70	0.74	0.76	0.61	1.00	-
b.82.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.83.1	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.54	0.00	0.00
b.84.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.84.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.84.3	0.72	0.72	0.83	0.83	0.80	0.82	0.68	0.82	1.00	1.00	-
b.84.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.85.1	0.56	0.56	0.73	0.79	0.69	0.71	0.57	0.54	0.61	0.44	1.00
b.85.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
b.85.3	0.44	0.43	0.71	0.72	0.66	0.70	0.62	0.63	0.61	1.00	-
b.85.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.85.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
b.85.7	0.61	0.60	0.74	0.79	0.70	0.76	0.74	0.76	0.61	1.00	-
b.86.1	0.61	0.60	0.68	0.79	0.67	1.00	0.76	1.00	0.61	0.91	0.61
b.87.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.88.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.89.1	0.41	0.41	0.46	0.46	0.43	0.53	0.33	0.37	0.61	0.70	0.41
b.90.1	0.12	0.12	0.36	0.35	0.33	0.33	0.13	0.09	0.12	-	0.12
b.92.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.93.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.94.1	0.20	0.19	0.28	0.27	0.21	0.21	0.18	0.15	0.20	0.32	0.20
b.95.1	0.38	0.37	0.43	0.41	0.35	0.36	0.26	0.33	0.54	0.63	0.38
b.96.1	0.42	0.43	0.64	0.56	0.60	0.53	0.44	0.53	0.61	0.63	0.42
b.97.1	0.41	0.41	0.46	0.45	0.40	0.38	0.27	0.29	0.61	0.70	0.41
b.98.1	0.61	0.61	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	0.61
b.100.1	0.57	0.57	0.78	0.72	0.70	0.69	0.60	0.58	0.57	0.00	0.57
b.101.1	0.08	0.08	0.20	0.20	0.17	0.17	0.04	0.04	0.08	-	0.08
b.102.1	0.24	0.23	0.36	0.34	0.27	0.26	0.17	0.19	0.24	0.40	0.24
b.103.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.104.1	0.61	0.60	0.62	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.105.1	0.56	0.56	0.72	0.70	0.64	0.69	0.59	0.58	0.61	0.00	0.56
b.106.1	0.54	0.53	0.70	0.79	0.71	0.74	0.46	0.48	0.61	0.00	0.54
b.107.1	0.43	0.43	0.63	0.62	0.58	0.59	0.38	0.61	0.43	-	0.43

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.108.1	0.38	0.38	0.69	0.77	0.65	0.68	0.59	0.57	0.61	0.63	0.38
b.109.1	0.41	0.39	0.64	0.65	0.58	0.60	0.33	0.43	0.61	0.42	0.41
b.110.1	0.15	0.15	0.28	0.28	0.23	0.34	0.10	0.08	0.15	-	0.15
b.111.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	1.00	1.00	0.79	0.72
b.113.1	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	1.00
b.114.1	0.43	0.45	0.51	0.55	0.45	0.47	0.40	0.27	0.43	-	0.43
b.115.1	0.00	0.00	0.37	0.39	0.31	0.35	0.08	0.11	0.00	-	0.00
b.117.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.118.1	0.54	0.53	0.75	0.79	0.71	0.74	0.74	1.00	0.61	1.00	0.54
b.119.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.120.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
b.121.1	0.54	0.53	0.64	0.67	0.59	0.63	0.56	0.59	0.61	0.72	0.54
b.121.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	-
b.121.3	0.43	0.42	0.43	0.42	0.35	0.36	0.30	0.33	0.61	0.71	-
b.121.4	0.14	0.13	0.31	0.32	0.25	0.25	0.17	0.18	0.61	0.23	-
b.121.5	0.27	0.27	0.39	0.39	0.33	0.32	0.29	0.32	0.61	0.45	-
b.122.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.123.1	0.16	0.15	0.36	0.36	0.30	0.32	0.15	0.17	0.16	-	0.16
b.124.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
b.125.1	0.72	0.72	0.78	0.79	1.00	0.76	0.68	0.57	0.72	0.00	0.72
b.128.1	0.37	0.36	0.65	0.66	0.64	0.69	0.50	0.40	0.37	-	0.37
b.129.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.72	0.00	1.00
b.129.2	0.37	0.39	0.64	0.65	0.60	0.60	0.42	0.61	0.61	0.00	-
b.130.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.131.1	0.61	0.60	0.65	0.65	0.64	0.62	0.62	0.63	0.61	1.00	0.61
b.132.1	0.61	0.60	0.63	0.62	0.49	0.50	0.62	0.63	0.61	1.00	0.61
b.133.1	0.00	0.00	0.38	0.37	0.32	0.33	0.00	0.16	0.54	0.00	0.00
b.134.1	0.35	0.35	0.51	0.51	0.38	0.41	0.28	0.24	0.35	-	0.35
b.136.1	0.54	0.53	0.74	0.75	0.70	0.74	0.54	0.46	0.54	0.00	0.54
b.137.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.138.1	0.20	0.21	0.43	0.40	0.36	0.34	0.30	0.30	0.20	0.50	0.20
b.139.1	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	0.72	0.00	0.72
b.141.1	0.69	0.67	0.77	1.00	0.72	0.71	0.76	0.72	0.69	-	0.69
b.142.1	0.51	0.50	0.57	0.56	0.55	0.45	0.51	0.54	0.51	0.88	0.51
b.143.1	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.61	0.70	0.41
b.144.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	-	0.00
b.145.1	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.27	0.43	0.71	0.43
b.146.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
b.149.1	0.24	0.24	0.49	0.46	0.43	0.49	0.37	0.37	0.35	0.50	0.24
b.150.1	0.26	0.26	0.59	0.56	0.51	0.50	0.30	0.37	0.35	0.41	0.26
b.151.1	0.56	0.56	0.74	0.79	0.71	0.76	0.59	0.58	0.56	0.00	0.56
b.152.1	0.72	0.72	0.78	0.79	0.74	0.76	0.68	0.65	1.00	0.00	0.72
b.153.1	1.00	1.00	0.83	0.83	0.80	0.82	0.84	0.82	1.00	1.00	1.00
b.154.1	0.30	0.30	0.47	0.55	0.43	0.43	0.22	0.23	0.61	0.45	0.30
b.155.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.156.1	0.21	0.19	0.43	0.44	0.40	0.41	0.16	0.16	0.21	0.00	0.21
b.157.1	0.35	0.38	0.69	0.70	0.65	0.68	0.41	0.38	0.35	-	0.35
b.158.1	0.56	0.56	0.74	0.79	0.71	0.76	0.59	0.65	0.56	-	0.56
b.159.1	0.41	0.41	0.47	0.54	0.49	0.45	0.41	0.45	0.61	0.70	0.41
b.159.2	0.32	0.34	0.57	0.55	0.51	0.52	0.32	0.37	0.32	-	-
b.160.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.00	0.72
b.161.1	0.32	0.33	0.65	0.60	0.57	0.47	0.31	0.30	0.32	-	0.32
b.162.1	0.41	0.41	0.47	0.45	0.40	0.38	0.41	0.45	0.41	0.70	0.41
b.163.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
b.165.1	0.61	0.60	1.00	0.69	1.00	0.64	0.74	0.76	0.61	1.00	0.61
b.166.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
b.167.1	0.30	0.29	0.59	0.50	0.46	0.48	0.34	0.28	0.30	0.00	0.30
b.168.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
b.169.1	0.24	0.23	0.37	0.33	0.28	0.26	0.19	0.21	0.35	0.40	0.24
b.171.1	1.00	1.00	1.00	0.79	0.72	0.74	0.76	0.72	1.00	1.00	1.00
b.172.1	0.22	0.22	0.62	0.47	0.54	0.38	0.24	0.22	0.22	-	0.22
b.173.1	0.30	0.28	0.64	0.59	0.59	0.56	0.38	0.43	0.30	-	0.30
b.174.1	0.11	0.12	0.32	0.31	0.26	0.26	0.11	0.10	0.11	-	0.11
b.175.1	0.50	0.49	0.71	0.70	0.64	0.69	0.59	0.57	0.50	-	0.50

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
b.176.1	0.56	0.56	0.71	0.70	0.64	0.64	0.50	0.49	0.61	0.77	0.56
b.177.1	0.25	0.24	0.47	0.55	0.42	0.48	0.24	0.25	0.25	-	0.25
b.178.1	0.25	0.27	0.54	0.56	0.47	0.50	0.30	0.35	0.25	0.00	0.25
c.1.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.1.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.3	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.13	0.51	0.50	0.78	0.75	0.70	0.76	0.62	0.63	0.61	1.00	-
c.1.14	0.52	0.52	0.70	1.00	0.69	0.64	0.58	1.00	0.61	0.64	-
c.1.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.16	0.69	0.67	1.00	0.79	0.72	1.00	0.76	0.72	1.00	0.66	-
c.1.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.19	1.00	1.00	0.78	0.79	0.74	0.76	0.61	0.65	1.00	0.79	-
c.1.20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.24	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.61	1.00	0.00	-
c.1.25	0.25	0.27	0.54	0.52	0.49	0.48	0.39	0.49	0.25	-	-
c.1.26	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	-
c.1.27	0.46	0.47	0.65	0.79	0.69	0.71	0.44	0.45	0.61	0.70	-
c.1.28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.1.29	0.37	0.36	0.69	0.68	0.64	0.69	0.59	0.58	0.37	-	-
c.1.30	0.61	0.60	0.69	0.70	0.65	0.69	0.47	0.49	0.61	1.00	-
c.1.31	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.00	-
c.1.32	0.72	0.72	0.75	0.79	0.71	0.74	0.62	0.57	1.00	0.29	-
c.1.33	0.72	0.72	0.83	0.79	0.71	0.74	0.68	0.82	1.00	0.71	-
c.2.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.3.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.4.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.5.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.70	0.72
c.6.1	0.35	0.35	0.56	0.62	0.52	0.60	0.35	0.41	0.61	0.79	1.00
c.6.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.6.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.7.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.8.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.8.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.8.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.8.4	0.61	0.61	0.75	1.00	0.71	0.74	0.74	0.76	0.61	1.00	-
c.8.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.8.6	0.38	0.37	0.60	0.68	0.52	0.54	0.28	0.40	0.38	-	-
c.8.7	0.52	0.53	0.77	0.77	0.72	0.74	0.62	0.63	0.61	1.00	-
c.8.8	0.69	0.67	0.78	0.79	0.72	0.74	1.00	0.72	1.00	0.77	-
c.8.9	1.00	1.00	0.78	0.79	0.74	0.76	0.84	0.65	1.00	1.00	-
c.8.10	0.58	0.61	0.75	0.83	0.80	0.82	0.60	0.61	0.61	0.51	-
c.9.1	0.35	0.35	0.61	0.67	0.55	0.63	0.41	0.35	0.61	0.00	1.00
c.9.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.10.1	0.61	0.60	0.70	0.71	0.64	0.68	0.74	0.76	0.61	1.00	0.61
c.10.2	0.61	0.61	0.75	0.71	0.70	0.76	0.74	0.76	0.61	1.00	-
c.10.3	0.61	0.60	0.72	0.62	0.67	0.56	0.62	0.63	0.61	1.00	-
c.12.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.13.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
c.13.2	1.00	1.00	0.83	0.83	1.00	1.00	0.76	1.00	1.00	1.00	-
c.14.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.15.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
c.16.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.17.1	0.61	0.61	0.83	0.83	0.80	0.82	0.62	0.63	0.61	1.00	0.61
c.18.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.19.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.20.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.21.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.22.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.23.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.23.2	0.61	0.60	1.00	1.00	1.00	1.00	0.60	0.76	0.61	1.00	-
c.23.3	0.28	0.27	0.72	0.54	0.67	0.45	0.42	0.22	0.28	-	-
c.23.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.10	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	-
c.23.11	0.61	0.61	0.75	0.79	0.74	0.76	0.62	0.63	0.61	1.00	-
c.23.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.13	0.72	0.72	0.83	0.83	0.80	0.82	0.68	0.82	1.00	0.50	-
c.23.14	0.43	0.43	0.70	0.69	0.72	0.62	0.58	0.62	0.61	0.71	-
c.23.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.23.17	0.52	0.52	0.71	0.79	0.69	0.70	0.76	1.00	0.52	0.00	-
c.24.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.25.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.26.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.26.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.26.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.27.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.28.1	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	1.00
c.30.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.31.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.32.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.33.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.34.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.36.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.37.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.38.1	0.40	0.41	0.63	0.61	0.55	0.54	0.39	0.40	0.40	0.00	0.40
c.39.1	0.69	0.67	1.00	0.79	1.00	1.00	0.76	1.00	1.00	0.00	0.69
c.40.1	0.72	0.72	1.00	1.00	1.00	1.00	0.76	1.00	1.00	0.50	0.72
c.41.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
c.42.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.43.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
c.44.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.44.2	0.56	0.56	0.72	0.70	0.64	0.70	0.59	0.58	0.61	0.00	-
c.45.1	0.61	0.60	1.00	0.79	0.70	1.00	0.74	0.76	0.61	1.00	0.61
c.46.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.47.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.47.2	0.52	0.52	0.66	0.65	0.72	0.62	0.59	0.57	0.61	0.63	-
c.48.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.49.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.49.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.50.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.51.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.51.2	0.58	0.59	0.75	0.79	0.71	0.74	0.57	0.47	0.58	-	-
c.51.3	0.38	0.37	0.60	0.68	0.52	0.54	0.28	0.40	0.38	-	-
c.51.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.51.5	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.50	-
c.51.6	0.54	0.53	0.69	0.79	0.64	0.64	0.54	0.54	0.61	0.00	-
c.52.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.52.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.52.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
c.52.4	0.29	0.30	0.43	0.42	0.35	0.35	0.18	0.16	0.29	-	-
c.53.1	0.57	0.57	0.83	0.79	0.80	0.82	0.84	0.82	0.61	0.00	1.00
c.53.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
c.54.1	0.57	0.57	0.78	0.79	0.70	0.74	0.54	0.58	0.57	0.00	0.57
c.55.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.55.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.55.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.55.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.55.5	0.58	0.61	0.72	0.79	0.72	0.73	0.76	0.65	0.58	-	-
c.55.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.55.7	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.50	-
c.56.1	0.52	0.52	0.72	0.83	0.72	0.72	0.59	0.82	0.52	-	1.00
c.56.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.56.3	1.00	1.00	0.83	0.83	1.00	1.00	1.00	1.00	1.00	1.00	-
c.56.4	0.61	0.60	0.72	0.69	0.67	0.76	0.74	0.76	0.61	1.00	-
c.56.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.56.6	0.61	0.60	0.75	0.79	0.71	0.74	0.62	0.65	0.61	1.00	-
c.56.7	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	1.00	0.70	-
c.56.8	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	1.00	0.64	-
c.57.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.58.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.59.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.60.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.61.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.62.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.64.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77	1.00
c.65.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.66.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.67.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.67.2	0.24	0.23	0.54	0.55	0.48	0.57	0.28	0.34	0.24	-	-
c.67.3	0.48	0.47	0.66	0.79	0.61	0.62	0.47	0.50	0.48	-	-
c.68.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.69.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.70.1	0.61	0.60	0.73	0.72	0.70	0.70	0.74	0.76	0.61	1.00	0.61
c.71.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.72.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.72.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.72.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.73.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.74.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.76.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.77.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.78.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.78.2	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.70	-
c.79.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.80.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.81.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.82.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.83.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.84.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.85.1	0.56	0.56	0.75	0.79	0.71	0.76	0.60	0.53	0.61	0.00	0.56
c.86.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.87.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.88.1	1.00	1.00	0.77	0.79	0.72	1.00	0.76	1.00	1.00	1.00	1.00
c.89.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.90.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.91.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	0.76	1.00	1.00	1.00
c.92.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.92.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.92.3	0.30	0.30	0.62	0.62	0.54	0.52	0.28	0.34	0.30	-	-
c.93.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.94.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.95.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.96.1	0.61	0.60	0.70	0.79	0.64	0.70	0.74	0.76	0.61	1.00	0.61
c.97.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.97.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
c.98.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.88	0.72

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
c.98.2	0.72	0.72	0.83	0.83	0.80	0.82	0.76	1.00	1.00	0.79	-
c.99.1	0.56	0.56	0.71	0.60	0.56	0.56	0.36	0.38	0.56	-	0.56
c.100.1	1.00	1.00	0.73	0.79	0.70	0.71	0.74	1.00	1.00	1.00	1.00
c.101.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.102.1	0.37	0.36	0.52	0.52	0.43	0.46	0.29	0.27	0.37	-	0.37
c.103.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
c.104.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.105.1	0.61	0.60	0.75	1.00	1.00	1.00	0.60	1.00	0.61	1.00	0.61
c.106.1	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.72
c.107.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.108.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.109.1	1.00	1.00	1.00	1.00	0.74	1.00	0.76	0.76	1.00	1.00	1.00
c.110.1	1.00	1.00	1.00	0.79	0.74	0.76	1.00	0.76	1.00	1.00	1.00
c.111.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.112.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.113.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.114.1	1.00	1.00	0.78	0.79	1.00	0.76	1.00	1.00	1.00	0.00	1.00
c.115.1	0.35	0.33	0.70	0.69	0.49	0.62	0.41	0.50	0.35	-	0.35
c.116.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.117.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.118.1	0.69	0.67	0.77	0.75	0.70	0.74	0.76	0.72	1.00	0.71	0.69
c.119.1	0.61	0.60	0.78	0.77	0.70	0.70	0.62	0.65	0.61	1.00	0.61
c.120.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.121.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.70	0.72
c.122.1	0.56	0.56	0.74	0.79	0.71	0.74	0.49	0.50	0.61	0.79	0.56
c.123.1	0.61	0.60	0.78	0.72	0.70	0.74	0.60	0.60	0.61	1.00	0.61
c.124.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.125.1	0.69	0.67	0.77	0.75	0.70	0.74	0.76	0.61	1.00	0.32	0.69
c.126.1	0.28	0.30	0.53	0.52	0.46	0.48	0.28	0.26	0.28	-	0.28
c.127.1	0.52	0.52	0.70	0.69	0.69	0.62	0.58	0.62	0.52	-	0.52
c.128.1	0.54	0.53	0.65	0.66	0.61	0.64	0.54	0.46	0.61	0.00	0.54
c.129.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.130.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
c.131.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.132.1	0.69	0.67	0.77	1.00	0.72	0.71	0.76	0.72	0.69	-	0.69
c.133.1	0.56	0.56	0.72	0.72	0.66	0.70	0.59	0.50	0.61	0.79	0.56
c.134.1	0.61	0.61	0.83	0.79	0.71	0.82	0.62	0.82	0.61	1.00	0.61
c.135.1	0.72	0.72	0.78	0.79	0.74	1.00	0.76	0.72	1.00	0.91	0.72
c.136.1	0.69	0.67	0.72	0.70	0.69	0.69	0.69	0.62	0.69	0.72	0.69
c.138.1	0.61	0.60	0.71	0.70	0.64	0.69	0.74	0.76	0.61	1.00	0.61
c.140.1	0.56	0.56	0.65	0.66	0.64	0.69	0.50	0.51	0.56	0.24	0.56
c.141.1	0.57	0.57	0.78	0.72	0.74	0.74	0.59	0.53	0.61	0.45	0.57
c.142.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.143.1	0.69	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.69
c.144.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.145.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.146.1	0.41	0.41	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.79	0.41
c.147.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
c.148.1	0.58	0.61	0.73	0.79	0.69	0.71	0.50	0.61	0.58	-	0.58
c.149.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
c.150.1	0.51	0.50	0.70	0.79	0.64	0.65	0.74	0.76	0.61	1.00	0.51
c.151.1	0.52	0.52	0.71	0.79	0.64	0.70	0.76	1.00	0.52	0.00	0.52
c.152.1	0.52	0.52	0.71	0.79	0.64	0.70	0.76	1.00	0.52	0.00	0.52
c.153.1	0.47	0.47	0.78	0.72	0.70	0.66	0.60	0.50	0.47	-	0.47
c.154.1	0.19	0.19	0.44	0.38	0.41	0.32	0.26	0.28	0.61	0.32	0.19
d.1.1	0.31	0.31	0.59	0.57	0.53	0.61	0.38	0.37	0.35	0.66	0.31
d.2.1	1.00	1.00	0.83	0.83	0.80	0.82	0.76	0.82	1.00	1.00	1.00
d.3.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.4.1	0.61	0.60	0.75	0.79	0.70	0.82	0.62	0.63	0.61	1.00	0.61
d.5.1	0.20	0.19	0.28	0.27	0.21	0.20	0.20	0.16	0.20	0.32	0.20
d.6.1	0.20	0.19	0.28	0.27	0.21	0.20	0.20	0.16	0.20	0.32	0.20
d.7.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
d.8.1	0.44	0.43	0.71	0.72	0.66	0.70	0.62	0.63	0.61	1.00	0.44
d.9.1	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.61	0.32	0.20

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.9.2	0.18	0.18	0.39	0.47	0.31	0.31	0.22	0.16	0.18	-	-
d.10.1	0.51	0.50	0.62	0.63	0.64	0.64	0.74	0.76	0.61	1.00	0.51
d.11.1	0.58	0.61	0.75	0.79	0.70	0.76	0.59	0.65	0.58	-	0.58
d.12.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.13.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.14.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.15.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
d.15.2	0.61	0.60	0.62	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
d.15.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.15.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.15.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
d.15.8	1.00	1.00	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	-
d.15.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.15.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.15.11	0.61	0.60	0.62	0.61	0.47	0.50	0.58	0.60	0.61	0.71	-
d.15.12	0.15	0.14	0.61	0.58	0.31	0.53	0.26	0.31	0.15	-	-
d.15.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.16.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.17.1	0.61	0.60	0.69	0.69	0.62	0.62	0.56	0.63	0.61	1.00	1.00
d.17.2	0.51	0.50	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.17.3	0.45	0.45	0.61	0.79	0.57	0.59	0.49	0.41	0.61	0.00	-
d.17.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.17.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.17.7	0.28	0.30	0.53	0.52	0.46	0.48	0.28	0.23	0.28	-	-
d.18.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.19.1	0.29	0.29	0.42	0.45	0.36	0.36	0.32	0.36	0.61	0.50	0.29
d.20.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.21.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.22.1	0.38	0.37	0.40	0.36	0.31	0.30	0.27	0.27	0.61	0.63	0.38
d.23.1	0.61	0.60	0.63	0.62	0.57	0.60	0.62	0.63	0.61	1.00	0.61
d.24.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.00	0.72
d.25.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
d.26.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.26.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
d.26.3	0.54	0.53	0.64	0.79	0.59	0.62	0.53	0.55	0.61	0.77	-
d.27.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.28.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.29.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.30.1	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.41	-	0.41
d.31.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.32.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.33.1	0.54	0.53	0.66	0.79	0.61	0.64	0.54	0.47	0.54	0.00	0.54
d.34.1	0.43	0.43	0.54	0.54	0.45	0.49	0.48	0.46	0.61	0.77	0.43
d.36.1	0.61	0.60	0.70	0.74	0.64	0.71	0.74	0.76	0.61	1.00	0.61
d.37.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.38.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.39.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.40.1	0.41	0.41	0.59	0.54	0.49	0.45	0.56	0.59	0.61	0.70	0.41
d.41.1	0.61	0.60	0.78	0.79	0.71	0.76	0.60	0.63	0.61	1.00	1.00
d.41.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.41.3	0.56	0.56	0.78	0.79	0.71	0.76	0.60	0.65	0.61	0.71	-
d.41.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.41.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.42.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.43.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.43.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.44.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.45.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.47.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.48.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.79	0.72
d.49.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
d.50.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.50.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.50.3	0.56	0.56	0.70	0.79	0.72	0.62	0.59	0.57	0.56	-	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.50.4	0.61	0.60	0.65	0.67	0.57	0.74	0.62	0.63	0.61	1.00	-
d.50.5	0.25	0.26	0.56	0.51	0.42	0.47	0.39	0.27	0.25	-	-
d.51.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.52.1	0.11	0.12	0.50	0.47	0.42	0.44	0.12	0.11	0.11	-	1.00
d.52.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.52.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.52.4	0.72	0.72	0.83	0.83	0.80	0.82	0.68	0.82	1.00	0.70	-
d.52.5	0.21	0.20	0.62	0.66	0.57	0.62	0.28	0.27	0.21	-	-
d.52.6	0.61	0.60	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	-
d.52.7	1.00	1.00	1.00	0.83	1.00	0.82	1.00	1.00	1.00	1.00	-
d.52.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.52.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.52.10	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	0.72	0.00	-
d.53.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.54.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.55.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.56.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.57.1	0.21	0.21	0.48	0.47	0.41	0.43	0.21	0.18	0.21	-	0.21
d.58.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.58.2	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	0.00	-
d.58.3	0.61	0.60	0.69	0.64	0.57	0.62	0.62	0.63	0.61	1.00	-
d.58.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.7	0.61	0.61	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	-
d.58.9	0.52	0.52	0.70	1.00	0.69	0.64	0.58	1.00	0.61	0.64	-
d.58.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.20	1.00	1.00	0.77	0.75	0.72	0.71	0.76	0.72	1.00	1.00	-
d.58.21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.22	0.22	0.22	0.28	0.27	0.21	0.21	0.23	0.16	0.22	0.36	-
d.58.23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.24	0.56	0.56	0.70	0.70	0.72	0.70	0.59	0.58	0.56	-	-
d.58.26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.29	1.00	1.00	0.83	0.83	0.80	0.82	0.84	0.82	1.00	1.00	-
d.58.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.31	0.39	0.39	0.70	0.69	0.69	0.62	0.58	0.62	0.39	-	-
d.58.32	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.33	0.52	0.52	0.70	0.69	0.69	0.63	0.58	0.62	0.52	-	-
d.58.34	0.51	0.50	0.69	0.70	0.65	0.68	0.56	0.59	0.61	1.00	-
d.58.36	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.38	0.43	0.43	0.64	0.62	0.60	0.59	0.38	0.61	0.43	-	-
d.58.39	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.41	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.43	0.71	-
d.58.42	0.72	0.72	0.83	0.83	0.80	0.82	1.00	1.00	1.00	0.79	-
d.58.43	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	-
d.58.44	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.50	-
d.58.46	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.58.47	0.32	0.32	0.61	0.63	0.56	0.59	0.35	0.35	0.32	-	-
d.58.48	0.69	0.67	0.77	0.79	0.72	0.70	0.76	0.65	1.00	0.77	-
d.58.49	0.57	0.57	0.83	0.72	0.80	0.68	0.60	0.61	0.57	-	-
d.58.50	0.47	0.47	0.71	0.79	0.70	0.64	0.57	0.65	0.47	-	-
d.58.51	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.52	0.72	0.72	0.75	0.83	0.80	0.82	0.60	0.54	1.00	0.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.58.53	0.45	0.45	0.66	0.67	0.61	0.61	0.41	0.36	0.45	-	-
d.58.54	0.51	0.50	0.73	0.79	0.69	0.67	0.74	0.76	0.61	0.88	-
d.58.55	0.43	0.43	0.62	0.66	0.57	0.62	0.53	0.60	0.61	0.79	-
d.58.56	0.41	0.39	0.66	0.72	0.60	0.70	0.38	0.61	0.41	0.45	-
d.58.57	0.72	0.72	0.75	0.79	1.00	0.76	0.60	0.65	1.00	0.00	-
d.58.58	1.00	1.00	0.77	0.79	1.00	0.73	0.76	0.72	0.72	0.00	-
d.58.59	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.58.60	0.28	0.30	0.53	0.52	0.46	0.48	0.28	0.23	0.28	0.00	-
d.58.61	0.52	0.52	0.77	0.61	0.55	0.54	0.76	0.57	0.52	-	-
d.58.62	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.59.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.60.1	0.61	0.61	0.75	0.79	0.71	0.74	0.74	0.76	0.61	1.00	0.61
d.61.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.62.1	0.24	0.23	0.37	0.33	0.28	0.26	0.19	0.21	0.24	0.40	0.24
d.63.1	0.61	0.60	1.00	1.00	1.00	1.00	1.00	1.00	0.61	1.00	0.61
d.64.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.64.2	0.56	0.56	0.72	0.70	0.70	0.70	0.59	0.58	0.56	-	-
d.65.1	0.72	0.72	0.83	0.83	0.80	0.82	0.60	0.82	1.00	0.71	0.72
d.66.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.67.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.67.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.67.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.67.4	0.45	0.45	0.60	0.61	0.54	0.57	0.41	0.38	0.45	0.00	-
d.68.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.68.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.68.3	1.00	1.00	0.77	0.79	0.72	0.74	1.00	1.00	0.72	0.00	-
d.68.4	0.69	0.67	0.77	0.75	0.72	0.74	0.76	0.72	1.00	0.70	-
d.68.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.68.6	1.00	1.00	0.71	0.69	0.64	0.64	0.74	0.76	1.00	1.00	-
d.68.7	1.00	1.00	0.83	0.79	0.80	0.70	0.62	0.82	1.00	1.00	-
d.68.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.70.1	0.41	0.41	0.46	0.45	0.40	0.38	0.23	0.24	0.61	0.70	0.41
d.71.1	0.47	0.47	0.71	0.70	0.70	0.64	0.59	0.82	0.61	0.00	0.47
d.72.1	0.51	0.50	0.65	0.62	0.57	0.58	0.56	0.59	0.61	0.88	0.51
d.73.1	0.51	0.50	0.64	0.62	0.64	0.59	0.51	0.61	0.61	0.88	0.51
d.74.1	0.61	0.61	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	1.00
d.74.2	0.57	0.57	0.78	0.77	0.72	0.70	0.60	0.65	0.57	0.00	-
d.74.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.74.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	-
d.74.5	0.54	0.53	0.71	0.72	0.70	0.70	0.54	0.58	0.61	0.72	-
d.75.1	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	0.00	1.00
d.75.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.76.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.76.2	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.61	0.71	-
d.77.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.78.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.79.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.79.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.79.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.79.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.79.5	0.72	0.72	1.00	0.83	1.00	0.82	1.00	1.00	1.00	1.00	-
d.79.6	0.43	0.45	0.62	0.59	0.54	0.52	0.40	0.44	0.43	-	-
d.79.7	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.00	-
d.79.8	0.38	0.37	0.60	0.68	0.53	0.57	0.37	0.50	0.38	-	-
d.79.9	0.23	0.22	0.51	0.52	0.46	0.53	0.28	0.29	0.23	-	-
d.80.1	0.54	0.53	0.75	0.79	0.71	0.74	0.74	0.76	0.61	1.00	0.54
d.81.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.81.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.81.3	0.52	0.52	0.70	0.79	0.69	0.62	0.58	0.62	0.52	-	-
d.81.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.82.1	0.45	0.45	0.65	0.68	0.56	0.58	0.50	0.53	0.45	0.00	0.61
d.82.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.82.3	0.61	0.60	0.63	0.62	0.49	0.50	0.62	0.63	0.61	1.00	-
d.82.4	0.54	0.53	0.54	0.54	0.44	0.49	0.48	0.55	0.61	0.91	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.82.5	0.16	0.16	0.45	0.44	0.38	0.39	0.25	0.14	0.16	-	-
d.83.1	0.61	0.60	0.62	0.61	0.49	0.50	0.62	0.63	0.61	1.00	0.61
d.83.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.84.1	0.24	0.25	0.58	0.56	0.48	0.50	0.28	0.27	0.24	-	0.24
d.85.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d.86.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.87.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.87.2	0.61	0.60	0.78	0.79	0.70	0.76	0.60	0.63	0.61	1.00	-
d.88.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.89.1	0.48	0.47	0.58	0.66	0.61	0.59	0.34	0.35	0.61	0.00	0.48
d.90.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.91.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.92.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.92.2	0.61	0.61	0.75	0.79	0.71	0.76	0.62	0.63	0.61	0.91	-
d.93.1	0.61	0.60	0.62	0.56	0.57	0.45	0.62	0.63	0.61	1.00	0.61
d.94.1	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.00	0.72
d.94.2	0.56	0.56	0.69	0.68	0.64	0.69	0.50	0.53	0.56	-	-
d.95.1	0.56	0.56	0.72	0.70	0.64	0.70	0.47	0.58	0.56	0.00	0.57
d.95.2	0.57	0.57	0.72	0.77	0.70	0.70	0.68	0.82	0.61	0.88	-
d.96.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.96.2	0.72	0.72	0.78	0.79	0.74	0.76	0.84	0.65	1.00	0.00	-
d.97.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.98.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.56
d.98.2	0.56	0.56	0.72	0.70	0.67	0.65	0.45	0.54	0.56	0.00	-
d.99.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.88	0.72
d.100.1	1.00	1.00	0.83	0.83	0.80	0.82	0.84	0.82	1.00	1.00	1.00
d.100.2	0.35	0.35	0.59	0.64	0.55	0.62	0.33	0.37	0.35	-	-
d.101.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.103.1	0.11	0.11	0.43	0.27	0.24	0.24	0.10	0.09	0.24	0.00	0.11
d.104.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.105.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.106.1	0.61	0.60	1.00	0.75	0.70	0.74	0.62	0.63	0.61	1.00	0.61
d.107.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.108.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.109.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.109.2	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.109.3	0.54	0.53	0.56	0.55	0.45	0.49	0.53	0.60	0.61	0.91	-
d.110.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	1.00
d.110.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.110.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.110.4	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.110.5	0.41	0.41	0.54	0.55	0.48	0.54	0.45	0.35	0.61	0.00	-
d.110.6	0.72	0.72	0.75	0.79	0.71	0.74	0.59	0.82	1.00	0.50	-
d.110.7	0.61	0.60	0.71	0.79	0.64	0.64	0.69	0.63	0.61	1.00	-
d.110.8	0.15	0.15	0.00	0.00	0.00	0.00	0.07	0.00	0.15	-	-
d.110.9	0.56	0.56	0.74	0.70	0.70	0.64	0.60	0.46	0.61	0.77	-
d.110.10	0.17	0.18	0.36	0.36	0.28	0.30	0.16	0.17	0.17	-	-
d.111.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
d.112.1	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.71	0.72
d.113.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.114.1	1.00	1.00	0.78	0.79	0.74	0.76	0.62	0.82	1.00	1.00	1.00
d.115.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.116.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.117.1	0.61	0.60	0.78	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
d.118.1	0.57	0.57	0.83	0.83	0.80	0.82	0.60	0.61	0.61	0.58	0.57
d.120.1	0.61	0.60	0.70	0.69	0.64	0.65	0.74	0.76	0.61	1.00	0.61
d.121.1	0.38	0.39	0.65	0.65	0.59	0.62	0.39	0.33	0.38	-	0.38
d.122.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.123.1	0.45	0.45	0.69	0.66	0.64	0.59	0.50	0.51	0.45	-	0.45
d.124.1	0.61	0.60	0.62	0.62	0.57	0.58	0.62	0.63	0.61	1.00	0.61
d.125.1	0.56	0.56	0.72	0.68	0.64	0.69	0.59	0.61	0.61	0.88	0.56
d.126.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.127.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.128.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.129.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.129.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.129.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.129.4	0.45	0.45	0.61	0.63	0.56	0.59	0.49	0.41	0.45	-	-
d.129.5	0.18	0.19	0.45	0.41	0.38	0.36	0.18	0.18	0.18	-	-
d.129.6	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
d.129.7	0.45	0.45	0.75	0.79	0.71	0.67	0.59	0.57	0.61	0.32	-
d.129.8	0.24	0.25	0.58	0.56	0.51	0.54	0.31	0.20	0.24	-	-
d.129.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
d.129.10	0.37	0.35	0.65	0.62	0.64	0.58	0.50	0.53	0.37	-	-
d.129.11	0.37	0.36	0.62	0.59	0.64	0.54	0.41	0.40	0.37	-	-
d.130.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.131.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.133.1	0.61	0.60	0.78	0.79	0.71	0.76	0.60	0.63	0.61	1.00	0.61
d.134.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.135.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.136.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.137.1	0.28	0.28	0.61	0.58	0.50	0.53	0.26	0.31	0.28	0.00	0.28
d.139.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.140.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.141.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.142.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.142.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
d.143.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.144.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.145.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.146.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.70	0.72
d.147.1	0.52	0.52	0.77	0.75	0.72	0.70	0.76	0.72	0.52	-	0.52
d.148.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.149.1	0.36	0.37	0.62	0.64	0.56	0.62	0.37	0.46	0.61	0.00	0.36
d.150.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.151.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.152.1	0.69	0.67	0.77	0.79	0.72	0.70	0.69	0.62	0.69	0.00	0.69
d.153.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.153.2	0.35	0.33	0.64	0.61	0.56	0.56	0.59	0.57	0.35	-	-
d.154.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.155.1	0.69	0.67	0.77	0.79	0.72	0.70	0.76	0.72	0.69	0.00	0.69
d.156.1	1.00	1.00	0.75	0.79	0.70	0.76	0.74	0.76	1.00	1.00	1.00
d.157.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.159.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.160.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.161.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.162.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.163.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.164.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
d.165.1	0.27	0.27	0.41	0.39	0.34	0.32	0.29	0.32	0.61	0.45	0.27
d.166.1	0.61	0.60	0.70	0.72	0.70	0.76	0.62	0.63	0.61	1.00	0.61
d.167.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.168.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.169.1	0.61	0.61	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	0.61
d.170.1	0.43	0.42	0.53	0.52	0.48	0.40	0.34	0.32	0.61	0.71	0.43
d.170.2	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.27	0.43	0.71	-
d.171.1	0.43	0.42	0.74	0.66	0.70	0.76	0.42	0.33	0.61	0.71	0.43
d.173.1	1.00	1.00	0.83	0.83	0.80	0.82	0.68	0.82	1.00	1.00	1.00
d.174.1	0.43	0.42	0.62	0.59	0.64	0.54	0.41	0.40	0.61	0.71	0.43
d.175.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.71	0.72
d.176.1	0.61	0.60	1.00	1.00	1.00	1.00	1.00	1.00	0.61	1.00	0.61
d.177.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.178.1	0.61	0.60	0.73	0.79	0.69	0.71	0.74	0.76	0.61	1.00	0.61
d.179.1	1.00	1.00	0.77	0.75	0.72	0.71	0.76	0.72	1.00	1.00	1.00
d.181.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.183.1	0.56	0.56	0.72	0.70	0.70	0.65	0.47	0.53	0.56	0.00	0.56
d.184.1	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.26	0.32	0.20
d.185.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.186.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.11
d.186.2	0.11	0.12	0.32	0.31	0.26	0.26	0.11	0.10	0.11	-	-
d.187.1	0.51	0.50	0.64	0.62	0.64	0.59	0.51	0.61	0.61	0.88	0.51
d.188.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.189.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.190.1	0.57	0.57	0.78	0.79	0.70	0.82	0.76	0.82	0.61	0.72	0.57
d.192.1	0.56	0.56	0.78	0.79	0.70	0.82	0.60	0.82	0.61	0.49	0.56
d.193.1	0.61	0.60	0.74	0.75	0.70	1.00	0.74	1.00	0.61	1.00	0.61
d.194.1	0.72	0.72	1.00	1.00	1.00	1.00	0.76	1.00	0.72	0.32	0.72
d.197.1	0.16	0.15	0.19	0.19	0.15	0.16	0.12	0.07	0.16	-	0.16
d.198.1	0.41	0.39	0.72	0.64	0.60	0.62	0.36	0.61	0.61	-	1.00
d.198.2	0.61	0.60	0.62	0.61	0.57	0.50	0.62	0.63	0.61	1.00	-
d.198.3	0.46	0.47	0.69	0.79	0.65	0.76	0.47	0.48	0.46	-	-
d.198.4	0.47	0.47	0.75	0.79	0.71	0.74	0.54	0.49	0.61	0.00	-
d.198.5	0.56	0.56	0.72	0.70	0.62	0.74	0.46	0.41	0.56	-	-
d.200.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
d.201.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.202.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	0.72	0.00	0.72
d.203.1	0.45	0.45	0.70	0.79	0.61	0.62	0.44	0.48	0.45	0.00	0.45
d.204.1	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	0.72
d.205.1	0.26	0.24	0.40	0.36	0.31	0.32	0.27	0.21	0.43	0.42	0.26
d.206.1	1.00	1.00	0.83	0.83	0.80	0.82	0.84	0.62	1.00	0.88	1.00
d.207.1	1.00	1.00	0.83	0.83	0.80	0.82	1.00	0.82	0.72	-	1.00
d.208.1	1.00	1.00	0.77	0.79	1.00	0.70	1.00	0.72	1.00	0.72	1.00
d.209.1	0.43	0.43	0.63	0.54	0.49	0.49	0.48	0.63	0.61	0.70	0.43
d.210.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.211.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	0.61
d.211.2	0.35	0.34	0.43	0.42	0.32	0.30	0.27	0.27	0.35	0.58	-
d.212.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.00	0.72
d.214.1	0.35	0.33	0.57	0.54	0.49	0.45	0.41	0.38	0.35	-	0.35
d.215.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.217.1	0.43	0.42	0.57	0.49	0.55	0.43	0.44	0.47	0.61	0.72	0.43
d.218.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.219.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
d.220.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.221.1	0.41	0.39	0.54	0.54	0.48	0.49	0.26	0.32	0.41	0.45	0.41
d.222.1	1.00	1.00	0.83	0.83	1.00	0.82	1.00	1.00	1.00	1.00	1.00
d.223.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.224.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.225.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.50	0.72
d.226.1	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.72
d.227.1	1.00	1.00	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.45	1.00
d.228.1	0.32	0.32	0.53	0.56	0.46	0.51	0.35	0.30	0.32	-	0.32
d.229.1	0.57	0.57	0.78	0.79	0.70	0.82	0.84	0.82	0.61	0.45	0.57
d.230.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.230.2	0.34	0.36	0.70	0.79	0.67	0.62	0.62	0.57	0.34	-	-
d.230.3	0.38	0.37	0.40	0.36	0.31	0.30	0.30	0.27	0.38	0.63	-
d.230.4	0.37	0.36	0.56	0.62	0.52	0.60	0.25	0.44	0.37	-	-
d.230.5	0.61	0.60	0.75	0.79	0.80	0.74	1.00	0.76	0.61	0.79	-
d.230.6	0.21	0.21	0.42	0.41	0.36	0.35	0.17	0.15	0.21	-	-
d.231.1	0.43	0.42	0.64	0.67	0.59	0.74	0.36	0.44	0.61	0.71	0.43
d.232.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
d.233.1	0.18	0.18	0.41	0.44	0.37	0.41	0.22	0.16	0.18	-	0.18
d.234.1	0.15	0.14	0.27	0.24	0.18	0.18	0.17	0.13	0.20	0.25	0.15
d.235.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.236.1	0.39	0.39	0.63	0.63	0.55	0.56	0.51	0.50	0.39	-	0.39
d.237.1	0.61	0.60	1.00	1.00	1.00	1.00	0.74	0.76	0.61	1.00	0.61
d.238.1	0.16	0.15	0.19	0.19	0.15	0.16	0.12	0.07	0.16	-	0.16
d.239.1	0.38	0.37	0.40	0.36	0.31	0.30	0.25	0.23	0.54	0.63	0.38
d.240.1	0.61	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.61	1.00	0.61
d.241.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.241.2	0.72	0.72	0.83	0.83	1.00	0.82	1.00	1.00	1.00	0.88	-
d.242.1	0.57	0.57	0.78	0.77	0.70	0.70	0.74	0.76	0.61	0.88	0.57
d.243.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.244.1	0.58	0.61	0.75	0.79	0.74	0.76	0.68	0.65	0.58	0.00	0.58
d.245.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.246.1	0.43	0.42	0.56	0.55	0.51	0.49	0.53	0.46	0.61	0.91	0.43
d.247.1	0.39	0.39	0.62	0.60	0.72	0.56	0.59	0.72	0.39	-	0.39
d.248.1	0.61	0.60	0.75	0.79	0.71	1.00	0.74	0.76	0.61	1.00	0.61
d.249.1	0.39	0.39	0.65	0.69	0.55	0.62	0.50	0.56	0.61	0.00	0.39
d.250.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
d.251.1	0.27	0.25	0.61	0.63	0.58	0.59	0.31	0.34	0.61	0.45	0.27
d.252.1	0.72	0.72	0.78	0.79	0.74	0.76	0.68	0.72	0.72	-	0.72
d.253.1	0.15	0.15	0.33	0.35	0.27	0.30	0.11	0.10	0.15	-	0.15
d.256.1	0.52	0.52	0.71	0.63	0.64	0.56	0.54	0.44	0.61	0.00	0.52
d.257.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
d.258.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.260.1	0.43	0.42	0.60	0.58	0.53	0.54	0.30	0.25	0.61	0.71	0.43
d.261.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.262.1	0.21	0.21	0.42	0.35	0.27	0.30	0.12	0.13	0.21	-	0.21
d.263.1	0.23	0.25	0.40	0.36	0.31	0.30	0.18	0.22	0.23	-	0.23
d.264.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.265.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.266.1	0.35	0.33	0.57	0.45	0.49	0.45	0.29	0.25	0.35	-	0.35
d.267.1	0.61	0.61	0.73	0.79	0.69	0.71	0.74	0.76	0.61	1.00	0.61
d.268.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	0.88	1.00
d.269.1	0.61	0.60	0.71	1.00	1.00	1.00	1.00	0.76	0.61	1.00	0.61
d.270.1	0.72	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.72
d.271.1	0.61	0.61	1.00	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
d.272.1	0.30	0.29	0.40	0.40	0.31	0.29	0.25	0.23	0.43	0.50	0.30
d.273.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.274.1	0.29	0.30	0.43	0.42	0.35	0.35	0.18	0.16	0.29	-	0.29
d.275.1	0.30	0.30	0.60	0.60	0.52	0.53	0.33	0.50	0.30	-	0.30
d.276.1	0.26	0.26	0.56	0.58	0.50	0.54	0.24	0.24	0.26	-	0.26
d.278.1	0.61	0.60	0.70	0.79	0.69	0.69	0.74	0.76	0.61	1.00	0.61
d.279.1	0.45	0.45	0.74	0.75	0.70	0.74	0.49	0.41	0.45	-	0.45
d.280.1	0.69	0.67	0.71	0.69	0.61	0.64	0.69	0.53	0.69	-	0.69
d.281.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d.282.1	0.54	0.53	0.63	0.61	0.57	0.52	0.62	0.63	0.61	1.00	0.54
d.283.1	1.00	1.00	0.74	1.00	0.70	1.00	0.69	0.62	1.00	0.50	1.00
d.284.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.285.1	0.13	0.14	0.59	0.51	0.48	0.44	0.46	0.50	0.61	0.79	0.13
d.286.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
d.287.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	1.00	1.00	0.00	0.72
d.288.1	0.20	0.19	0.28	0.36	0.21	0.21	0.20	0.16	0.43	0.32	0.20
d.289.1	0.61	0.60	0.62	0.61	0.49	0.50	0.47	0.50	0.61	1.00	0.61
d.290.1	0.69	0.67	0.77	0.79	0.72	0.71	0.59	0.72	1.00	1.00	0.69
d.291.1	0.32	0.32	0.56	0.59	0.50	0.61	0.38	0.32	0.32	-	0.32
d.292.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.294.1	0.51	0.50	0.64	0.56	0.60	0.53	0.56	0.63	0.61	1.00	0.51
d.295.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.296.1	0.37	0.36	0.62	0.66	0.64	0.69	0.37	0.51	0.37	-	0.37
d.297.1	0.61	0.60	0.63	0.64	0.58	0.62	0.62	0.63	0.61	1.00	0.61
d.298.1	0.72	0.72	1.00	0.83	1.00	1.00	1.00	1.00	0.72	-	0.72
d.300.1	0.31	0.32	0.47	0.44	0.44	0.49	0.37	0.36	0.43	0.77	0.31
d.301.1	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.303.1	0.61	0.61	1.00	0.79	1.00	1.00	0.74	0.76	0.61	1.00	0.61
d.304.1	0.56	0.56	0.71	0.79	0.71	0.64	0.76	0.61	0.56	0.45	0.56
d.305.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
d.306.1	0.72	0.72	0.83	0.83	0.80	0.82	0.68	0.82	1.00	0.00	0.72
d.307.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00
d.308.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.309.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.310.1	0.61	0.61	0.75	0.79	0.74	0.76	0.74	0.76	0.61	1.00	0.61
d.311.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
d.312.1	0.19	0.16	0.45	0.45	0.37	0.39	0.24	0.22	0.19	-	0.19
d.313.1	0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.20	0.00	0.00
d.314.1	0.61	0.60	0.63	0.61	0.57	0.50	0.62	0.63	0.61	1.00	0.61

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
d.315.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.70	0.72
d.316.1	0.52	0.52	0.77	0.75	0.72	0.70	0.76	0.72	0.52	-	0.52
d.317.1	0.32	0.33	0.55	0.53	0.53	0.47	0.31	0.30	0.32	-	0.32
d.319.1	0.33	0.33	0.59	0.57	0.52	0.42	0.29	0.36	0.33	-	0.33
d.320.1	0.72	0.72	1.00	1.00	1.00	1.00	0.76	1.00	0.72	-	0.72
d.321.1	0.11	0.12	0.32	0.31	0.26	0.26	0.11	0.10	0.11	-	0.11
d.322.1	0.61	0.61	0.83	1.00	1.00	1.00	0.74	1.00	0.61	1.00	0.61
d.323.1	0.00	0.00	0.35	0.37	0.00	0.00	0.00	0.00	0.00	-	0.00
d.324.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
d.325.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.326.1	0.27	0.27	0.63	0.61	0.51	0.56	0.23	0.43	0.27	-	0.27
d.327.1	0.56	0.56	0.72	0.70	0.64	0.70	0.59	0.65	0.56	0.00	0.56
d.328.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.329.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.330.1	0.61	0.60	0.70	0.56	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.331.1	0.35	0.36	0.66	0.67	0.61	0.61	0.33	0.47	0.35	-	0.35
d.332.1	0.54	0.53	0.56	0.55	0.44	0.49	0.53	0.55	0.61	0.91	0.54
d.333.1	0.69	0.67	0.77	1.00	1.00	0.72	0.76	0.62	1.00	0.38	0.69
d.334.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d.335.1	0.28	0.29	0.62	0.62	0.54	0.52	0.33	0.46	0.28	-	0.28
d.336.1	0.61	0.60	0.64	0.74	0.59	0.71	0.62	0.63	0.61	1.00	0.61
d.337.1	0.26	0.27	0.49	0.48	0.41	0.41	0.39	0.30	0.26	-	0.26
d.338.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.339.1	0.17	0.18	0.36	0.36	0.28	0.30	0.13	0.13	0.17	-	0.17
d.340.1	0.69	0.67	0.78	0.75	0.72	0.70	0.76	0.72	0.69	0.00	0.69
d.341.1	0.18	0.18	0.39	0.37	0.31	0.31	0.22	0.16	0.18	-	0.18
d.342.1	0.29	0.30	0.43	0.42	0.35	0.35	0.18	0.16	0.29	-	0.29
d.343.1	0.46	0.47	0.66	0.67	0.61	0.65	0.45	0.53	0.61	0.38	0.46
d.344.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.345.1	0.43	0.42	0.43	0.42	0.33	0.32	0.25	0.23	0.43	0.71	0.43
d.347.1	0.51	0.50	0.74	0.67	0.70	0.61	0.59	0.57	0.61	0.88	0.51
d.348.1	0.35	0.36	0.64	0.65	0.59	0.65	0.59	0.57	0.61	0.00	0.35
d.349.1	0.38	0.37	0.66	0.68	0.60	0.70	0.53	0.53	0.38	-	0.38
d.350.1	0.45	0.45	0.61	0.62	0.54	0.59	0.41	0.39	0.45	-	0.45
d.351.1	0.18	0.18	0.36	0.35	0.29	0.30	0.11	0.14	0.18	-	0.18
d.352.1	0.54	0.53	0.66	0.79	0.64	0.64	0.59	0.65	0.54	-	0.54
d.353.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.354.1	0.32	0.32	0.53	0.53	0.46	0.48	0.35	0.30	0.32	-	0.32
d.355.1	1.00	1.00	1.00	1.00	1.00	0.64	1.00	1.00	1.00	1.00	1.00
d.356.1	0.35	0.35	0.69	0.79	0.65	0.72	0.45	0.42	0.61	0.00	0.35
d.357.1	0.58	0.61	0.73	0.79	0.69	0.71	0.59	0.57	0.58	0.00	0.58
d.358.1	0.27	0.27	0.44	0.65	0.38	0.60	0.24	0.27	0.27	-	0.27
d.359.1	0.25	0.25	0.58	0.59	0.50	0.52	0.28	0.27	0.25	-	0.25
d.360.1	0.38	0.37	0.60	0.68	0.53	0.54	0.38	0.40	0.38	-	0.38
d.362.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d.363.1	0.00	0.00	0.35	0.37	0.00	0.00	0.00	0.00	0.00	-	0.00
d.364.1	0.18	0.18	0.41	0.44	0.37	0.41	0.16	0.16	0.18	-	0.18
d.365.1	0.43	0.41	0.64	0.62	0.60	0.59	0.44	0.61	0.61	0.72	0.43
d.366.1	0.56	0.56	0.72	0.79	0.67	0.76	0.50	0.58	0.56	-	0.56
d.367.1	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	0.72	0.00	0.72
d.368.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
d.369.1	0.43	0.42	0.72	0.67	0.66	0.63	0.62	0.45	0.61	1.00	0.43
d.370.1	0.54	0.53	0.54	0.55	0.44	0.43	0.48	0.55	0.61	0.91	0.54
d.371.1	0.00	0.00	0.35	0.33	0.27	0.28	0.11	0.10	0.00	-	0.00
d.372.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
d.373.1	0.45	0.45	0.69	0.79	0.70	0.74	0.46	0.49	0.61	0.00	0.45
d.375.1	0.29	0.28	0.61	0.63	0.58	0.59	0.30	0.22	0.29	-	0.29
d.376.1	0.32	0.33	0.55	0.53	0.46	0.47	0.31	0.30	0.32	-	0.32
d.377.1	0.00	0.00	0.28	0.27	0.23	0.23	0.06	0.07	0.00	-	0.00
d.379.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.380.1	0.35	0.35	0.69	0.70	0.65	0.68	0.44	0.42	0.35	0.00	0.35
d.381.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
d.382.1	0.31	0.32	0.59	0.55	0.50	0.54	0.40	0.44	0.61	0.58	0.31
d.383.1	0.22	0.24	0.56	0.62	0.50	0.56	0.37	0.36	0.22	-	0.22

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
e.1.1	0.43	0.42	0.63	0.65	0.57	0.60	0.47	0.45	0.61	0.71	0.43
e.2.1	0.28	0.30	0.53	0.47	0.46	0.43	0.21	0.18	0.28	-	0.28
e.3.1	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	1.00
e.5.1	0.61	0.60	0.74	0.79	0.70	0.76	0.74	0.76	0.61	1.00	0.61
e.6.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.7.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.8.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.10.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.11.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.12.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.13.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.15.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
e.17.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.18.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00
e.19.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.22.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.23.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.24.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.25.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
e.26.1	0.52	0.52	0.72	0.79	0.69	0.70	0.58	0.65	0.61	0.70	0.52
e.28.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
e.29.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.32.1	0.72	0.72	0.78	0.79	0.74	0.76	0.68	0.65	1.00	0.42	0.72
e.37.1	0.69	0.67	0.77	0.79	0.74	0.74	0.76	0.72	0.69	0.77	0.69
e.38.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.39.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.40.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
e.41.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
e.42.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	-	0.00
e.43.1	0.61	0.60	0.63	0.61	0.49	0.50	0.62	0.63	0.61	1.00	0.61
e.44.1	0.50	0.48	0.72	0.79	0.66	0.71	0.62	0.42	0.61	0.66	0.50
e.46.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
e.47.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
e.49.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.13	0.72
e.50.1	0.61	0.60	0.70	0.79	0.69	0.64	0.74	0.76	0.61	1.00	0.61
e.51.1	0.58	0.61	0.78	0.79	0.71	0.74	0.60	0.60	0.61	0.79	0.58
e.52.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
e.53.1	0.72	0.72	0.83	1.00	0.80	0.82	0.84	0.82	1.00	0.00	0.72
e.54.1	0.52	0.52	0.71	0.79	0.69	0.64	0.69	0.62	0.52	-	0.52
e.55.1	0.61	0.60	0.56	0.69	0.48	0.64	0.53	0.76	0.61	0.91	0.61
e.56.1	0.61	0.60	0.77	1.00	1.00	0.73	1.00	1.00	0.61	1.00	0.61
e.57.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
e.59.1	0.48	0.47	0.66	0.79	0.61	0.64	0.47	0.53	0.61	0.00	0.48
e.60.1	0.72	0.72	0.78	0.79	0.71	0.76	0.60	0.65	1.00	0.00	0.72
e.61.1	0.28	0.28	0.54	0.63	0.50	0.59	0.24	0.26	0.28	-	0.28
e.62.1	0.41	0.39	0.64	0.62	0.60	0.59	0.38	0.53	0.61	0.00	0.41
e.63.1	0.61	0.60	0.62	0.56	0.64	0.45	0.74	0.76	0.61	1.00	0.61
e.64.1	0.43	0.43	0.52	0.52	0.44	0.49	0.33	0.27	0.43	-	0.43
e.65.1	0.35	0.38	0.63	0.71	0.59	0.68	0.45	0.34	0.61	0.00	0.35
e.66.1	0.12	0.12	0.00	0.35	0.00	0.00	0.00	0.00	0.12	0.00	0.12
e.67.1	0.21	0.21	0.43	0.42	0.35	0.35	0.14	0.12	0.21	-	0.21
e.68.1	0.45	0.45	0.61	0.63	0.56	0.59	0.49	0.41	0.45	0.00	0.45
e.70.1	0.52	0.51	0.78	0.72	0.70	0.70	0.60	0.58	0.52	-	0.52
e.71.1	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	0.00	0.69
e.72.1	0.48	0.47	0.66	0.66	0.53	0.60	0.54	0.39	0.48	-	0.48
e.73.1	0.43	0.42	0.62	0.58	0.51	0.53	0.37	0.40	0.61	0.71	0.43
e.74.1	0.58	0.61	0.78	0.79	0.71	0.74	0.74	1.00	0.61	1.00	0.58
e.76.1	0.15	0.14	0.37	0.32	0.25	0.25	0.12	0.10	0.24	0.25	0.15
f.1.1	0.12	0.12	0.23	0.20	0.19	0.17	0.05	0.04	0.61	-	0.43
f.1.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
f.1.3	0.30	0.29	0.39	0.37	0.30	0.30	0.19	0.18	0.61	0.26	-
f.1.4	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.27	0.54	0.71	-
f.3.1	0.26	0.26	0.54	0.56	0.49	0.53	0.27	0.31	0.26	-	0.26
f.4.1	0.58	0.61	0.83	0.83	1.00	0.82	0.74	0.76	0.61	1.00	0.72

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
f.4.2	0.45	0.45	0.61	0.62	0.56	0.59	0.49	0.41	0.61	0.00	-
f.4.3	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.42	-
f.4.4	0.38	0.39	0.66	0.67	0.61	0.61	0.32	0.35	0.38	-	-
f.4.5	0.57	0.57	0.75	0.79	0.71	0.76	0.60	0.54	0.61	0.00	-
f.4.6	0.32	0.32	0.62	0.66	0.57	0.62	0.35	0.28	0.32	0.00	-
f.5.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.45	0.72
f.6.1	0.25	0.27	0.54	0.52	0.47	0.46	0.26	0.31	0.25	-	0.25
f.7.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.61	0.71	0.43
f.8.1	0.38	0.37	0.46	0.49	0.42	0.43	0.33	0.34	0.61	0.63	0.38
f.9.1	0.21	0.21	0.61	0.54	0.56	0.50	0.22	0.22	0.21	-	0.21
f.11.1	0.61	0.60	0.75	0.79	0.71	0.76	0.60	0.60	0.61	1.00	0.61
f.13.1	0.61	0.60	0.69	0.62	0.64	0.56	0.62	0.63	0.61	1.00	0.61
f.14.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.15.1	0.35	0.34	0.43	0.40	0.30	0.30	0.27	0.27	0.35	0.58	0.35
f.16.1	0.58	0.61	0.78	0.83	0.80	0.82	0.68	0.61	0.61	0.72	0.58
f.17.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.17.2	0.58	0.61	1.00	0.83	1.00	0.82	0.60	0.61	0.61	0.70	-
f.17.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
f.17.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
f.17.5	0.41	0.39	0.63	0.62	0.58	0.59	0.41	0.61	0.61	0.64	-
f.18.1	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.88	0.72
f.19.1	0.61	0.60	0.78	0.79	0.74	0.74	0.76	0.76	0.61	1.00	0.61
f.20.1	1.00	1.00	1.00	1.00	1.00	1.00	0.74	1.00	1.00	1.00	1.00
f.21.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
f.21.2	1.00	1.00	1.00	0.79	1.00	1.00	1.00	0.76	1.00	1.00	-
f.21.3	0.58	0.61	0.78	0.79	0.74	0.76	0.60	0.72	0.58	0.00	-
f.22.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
f.23.1	0.54	0.53	0.59	0.55	0.48	0.49	0.53	0.60	0.61	0.91	1.00
f.23.2	0.61	0.60	0.62	0.56	0.64	0.64	0.74	0.76	0.61	1.00	-
f.23.3	0.30	0.29	0.43	0.42	0.33	0.32	0.27	0.27	0.61	0.50	-
f.23.4	0.30	0.29	0.39	0.42	0.31	0.32	0.25	0.23	0.54	0.50	-
f.23.5	0.18	0.17	0.28	0.27	0.21	0.21	0.19	0.14	0.30	0.30	-
f.23.6	0.54	0.53	0.49	0.54	0.44	0.38	0.44	0.52	0.54	0.91	-
f.23.7	0.30	0.29	0.27	0.27	0.20	0.25	0.18	0.15	0.35	0.50	-
f.23.8	0.41	0.41	0.54	0.55	0.48	0.54	0.45	0.34	0.41	-	-
f.23.9	0.33	0.33	0.52	0.49	0.45	0.42	0.29	0.26	0.33	-	-
f.23.10	0.26	0.26	0.54	0.56	0.49	0.53	0.27	0.31	0.26	-	-
f.23.11	0.61	0.60	0.70	0.56	0.64	0.64	0.74	0.76	0.61	1.00	-
f.23.12	0.61	0.60	0.70	0.56	0.64	0.64	0.74	0.76	0.61	1.00	-
f.23.13	0.54	0.53	0.70	0.69	0.48	0.64	0.53	0.76	0.61	0.91	-
f.23.14	0.61	0.60	0.70	0.56	0.64	0.64	0.74	0.76	0.61	1.00	-
f.23.15	0.30	0.29	0.47	0.45	0.40	0.40	0.24	0.52	0.61	0.49	-
f.23.16	0.51	0.50	0.64	0.62	0.64	0.59	0.51	0.61	0.61	0.88	-
f.23.17	0.41	0.39	0.63	0.62	0.58	0.59	0.31	0.61	0.61	0.45	-
f.23.18	0.41	0.39	0.63	0.62	0.58	0.59	0.31	0.61	0.61	0.00	-
f.23.19	0.41	0.39	0.63	0.62	0.58	0.59	0.31	0.61	0.61	0.00	-
f.23.20	0.27	0.27	0.50	0.59	0.49	0.56	0.15	0.43	0.27	-	-
f.23.21	0.72	0.72	0.83	0.83	0.80	0.82	0.84	0.82	1.00	0.00	-
f.23.22	0.20	0.19	0.46	0.50	0.40	0.53	0.23	0.22	0.20	-	-
f.23.23	0.41	0.39	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.64	-
f.23.24	0.27	0.27	0.31	0.36	0.26	0.32	0.09	0.16	0.33	0.00	-
f.23.25	0.51	0.50	0.63	0.62	0.58	0.59	0.51	0.61	0.51	0.88	-
f.23.26	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.00	-
f.23.27	0.43	0.41	0.64	0.62	0.64	0.59	0.44	0.61	0.61	0.72	-
f.23.28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
f.23.29	0.33	0.33	0.45	0.45	0.35	0.36	0.34	0.33	0.33	-	-
f.23.30	0.30	0.29	0.43	0.46	0.37	0.41	0.36	0.40	0.61	0.50	-
f.23.31	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.14	-
f.23.32	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.00	-
f.23.33	0.51	0.50	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.88	-
f.23.34	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.00	-
f.23.35	0.41	0.39	0.63	0.62	0.58	0.59	0.36	0.61	0.61	0.49	-
f.23.36	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.64	-
f.23.37	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.61	0.00	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
f.23.38	0.41	0.39	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.64	-
f.24.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.61	1.00	0.70	1.00
f.25.1	0.58	0.61	1.00	1.00	1.00	1.00	0.76	0.61	0.61	0.71	0.58
f.26.1	0.41	0.39	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.64	0.41
f.27.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
f.28.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
f.29.1	0.41	0.39	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.64	0.41
f.30.1	0.41	0.39	0.63	0.62	0.58	0.59	0.36	0.61	0.61	0.55	0.41
f.31.1	0.51	0.50	0.64	0.62	0.64	0.59	0.51	0.61	0.61	0.88	0.51
f.32.1	0.54	0.53	1.00	0.83	0.70	1.00	0.76	0.61	0.61	0.88	0.54
f.33.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.34.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
f.35.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.36.1	0.42	0.43	0.64	0.56	0.60	0.53	0.50	0.53	0.61	0.79	0.42
f.37.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.38.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.39.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.40.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.41.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.42.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
f.43.1	0.51	0.50	0.64	0.62	0.64	0.59	0.74	0.76	0.61	1.00	0.51
f.44.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.45.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.54	0.71	0.43
f.46.1	0.72	0.72	0.83	0.83	0.80	0.82	0.76	0.82	1.00	0.00	0.72
f.48.1	0.72	0.72	0.78	0.79	0.74	0.76	0.76	0.65	1.00	0.26	0.72
f.49.1	0.58	0.61	0.83	0.79	0.80	0.82	0.62	0.65	0.61	1.00	0.58
f.50.1	0.22	0.22	0.28	0.27	0.21	0.21	0.23	0.16	0.22	0.36	0.22
f.51.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
f.52.1	0.54	0.53	0.56	0.55	0.48	0.49	0.53	0.60	0.61	0.91	0.54
f.53.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
f.54.1	0.58	0.61	0.78	0.79	0.74	0.76	0.59	0.65	0.61	0.71	0.58
f.55.1	0.41	0.41	0.64	0.62	0.60	0.59	0.41	0.61	0.61	0.70	0.41
f.56.1	0.61	0.60	0.74	0.75	0.70	1.00	0.74	0.76	0.61	1.00	0.61
f.57.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00
f.58.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00
f.59.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
g.1.1	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.43	0.71	0.43
g.2.3	0.26	0.26	0.40	0.38	0.33	0.31	0.20	0.29	0.26	0.44	0.26
g.3.1	0.43	0.43	0.62	0.56	0.49	0.45	0.56	0.59	0.61	0.70	0.61
g.3.2	0.18	0.17	0.37	0.35	0.30	0.27	0.14	0.15	0.26	0.30	-
g.3.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-
g.3.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	-
g.3.5	0.16	0.16	0.27	0.27	0.20	0.20	0.20	0.16	0.20	0.27	-
g.3.6	0.26	0.26	0.35	0.34	0.27	0.25	0.15	0.12	0.61	0.44	-
g.3.7	0.33	0.33	0.43	0.45	0.37	0.36	0.36	0.39	0.61	0.55	-
g.3.8	0.31	0.32	0.54	0.48	0.47	0.49	0.36	0.41	0.61	0.79	-
g.3.9	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
g.3.10	0.26	0.25	0.28	0.27	0.21	0.21	0.23	0.16	0.26	0.43	-
g.3.11	0.61	0.60	0.63	0.64	0.57	0.62	0.62	0.63	0.61	1.00	-
g.3.13	0.33	0.33	0.43	0.39	0.37	0.32	0.29	0.32	0.33	0.55	-
g.3.14	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.54	0.71	-
g.3.15	0.54	0.53	0.49	0.41	0.40	0.36	0.40	0.43	0.61	0.71	-
g.3.16	0.54	0.53	0.53	0.47	0.51	0.40	0.51	0.54	0.61	0.88	-
g.3.17	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.20	0.32	-
g.3.18	0.24	0.23	0.36	0.34	0.27	0.26	0.15	0.19	0.24	0.40	-
g.3.19	0.08	0.08	0.35	0.35	0.29	0.31	0.15	0.14	0.20	0.21	-
g.4.1	0.43	0.42	0.39	0.36	0.31	0.32	0.23	0.20	0.43	0.71	0.43
g.5.1	0.30	0.29	0.34	0.36	0.27	0.25	0.23	0.16	0.35	0.50	0.30
g.7.1	0.38	0.37	0.43	0.42	0.33	0.30	0.30	0.27	0.61	0.63	0.38
g.8.1	0.43	0.42	0.50	0.47	0.48	0.38	0.30	0.33	0.61	0.71	0.43
g.9.1	0.14	0.13	0.27	0.24	0.18	0.17	0.17	0.12	0.38	0.23	0.14
g.10.1	0.54	0.53	0.56	0.55	0.47	0.49	0.50	0.60	0.61	0.79	0.54
g.12.1	0.61	0.60	0.53	0.49	0.44	0.41	0.47	0.52	0.61	0.71	0.61
g.13.1	0.27	0.27	0.39	0.39	0.33	0.32	0.29	0.32	0.33	0.45	0.27

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
g.14.1	0.54	0.53	0.62	0.52	0.48	0.44	0.47	0.50	0.61	0.71	0.54
g.16.1	0.43	0.42	0.53	0.52	0.51	0.44	0.50	0.46	0.61	0.71	0.43
g.16.2	0.43	0.42	0.43	0.49	0.35	0.41	0.37	0.46	0.61	0.71	-
g.16.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
g.17.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
g.18.1	0.54	0.53	0.53	0.52	0.45	0.44	0.50	0.49	0.61	0.71	0.54
g.19.1	0.24	0.23	0.36	0.42	0.27	0.26	0.23	0.19	0.35	0.40	0.24
g.20.1	0.54	0.53	0.59	0.64	0.55	0.62	0.53	0.55	0.61	0.91	0.54
g.21.1	0.15	0.16	0.50	0.42	0.45	0.38	0.12	0.21	0.15	-	0.15
g.22.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.61	0.71	0.43
g.23.1	0.38	0.37	0.40	0.36	0.31	0.30	0.30	0.25	0.38	0.63	0.38
g.24.1	0.43	0.42	0.42	0.41	0.35	0.36	0.30	0.33	0.61	0.71	0.43
g.25.1	0.22	0.22	0.28	0.27	0.21	0.21	0.23	0.16	0.22	0.36	0.22
g.26.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
g.27.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
g.28.1	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	0.43
g.29.1	0.00	0.00	0.49	0.48	0.43	0.44	0.16	0.18	0.00	-	0.00
g.31.1	0.38	0.37	0.56	0.49	0.48	0.41	0.40	0.44	0.61	0.63	0.38
g.32.1	0.30	0.29	0.36	0.34	0.27	0.26	0.23	0.18	0.38	0.50	0.30
g.33.1	0.00	0.00	0.19	0.18	0.13	0.12	0.06	0.04	0.11	0.00	0.00
g.35.1	0.35	0.35	0.54	0.53	0.49	0.50	0.31	0.27	0.35	-	0.35
g.36.1	0.51	0.50	0.70	0.79	0.69	0.64	0.74	0.76	0.61	0.88	0.51
g.37.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.38.1	0.43	0.43	0.62	0.55	0.48	0.49	0.53	0.55	0.61	0.77	0.43
g.39.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
g.40.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.41.1	1.00	1.00	1.00	1.00	0.72	1.00	1.00	1.00	1.00	1.00	1.00
g.41.2	0.61	0.60	0.62	0.60	0.64	0.51	0.74	0.76	0.61	1.00	-
g.41.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
g.41.4	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
g.41.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
g.41.6	0.35	0.33	0.64	0.57	0.55	0.53	0.31	0.39	0.35	-	-
g.41.7	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	0.69	0.00	-
g.41.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
g.41.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
g.41.10	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
g.41.11	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	-
g.41.13	0.12	0.13	0.27	0.27	0.20	0.21	0.18	0.13	0.12	0.00	-
g.41.15	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
g.41.16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
g.41.17	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
g.41.18	0.28	0.30	0.53	0.52	0.38	0.48	0.28	0.29	0.28	0.00	-
g.42.1	0.72	0.72	0.83	0.83	1.00	1.00	1.00	1.00	1.00	1.00	0.72
g.43.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.44.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.45.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
g.46.1	0.41	0.39	0.64	0.62	0.60	0.59	0.36	0.61	0.54	0.32	0.41
g.47.1	0.43	0.43	0.54	0.54	0.44	0.43	0.48	0.46	0.43	0.77	0.43
g.48.1	0.45	0.45	0.74	0.79	0.70	0.72	0.45	0.48	0.61	0.45	0.45
g.49.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.50.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.52.1	0.54	0.53	0.54	0.55	0.44	0.49	0.53	0.55	0.61	0.91	0.54
g.53.1	0.54	0.53	0.62	0.56	0.57	0.45	0.56	0.63	0.61	1.00	0.54
g.54.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
g.59.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
g.60.1	0.54	0.53	0.57	0.56	0.55	0.45	0.51	0.54	0.61	1.00	0.54
g.61.1	0.18	0.18	0.39	0.35	0.26	0.25	0.18	0.15	0.18	-	0.18
g.62.1	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.43	0.71	0.43
g.64.1	0.54	0.53	0.42	0.41	0.35	0.36	0.40	0.33	0.61	0.71	0.54
g.65.1	0.54	0.53	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	0.54
g.66.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.67.1	0.30	0.29	0.38	0.36	0.27	0.28	0.30	0.20	0.43	0.50	0.30
g.68.1	0.54	0.53	0.57	0.61	0.55	0.49	0.50	0.47	0.61	0.88	0.54
g.69.1	0.37	0.37	0.45	0.42	0.39	0.33	0.41	0.45	0.61	0.64	0.37

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
g.71.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g.72.1	0.51	0.50	0.57	0.56	0.55	0.45	0.51	0.54	0.51	0.88	0.51
g.73.1	0.43	0.42	0.43	0.48	0.33	0.32	0.40	0.43	0.54	0.71	0.43
g.74.1	0.54	0.53	0.74	0.79	0.71	0.74	0.62	0.59	0.61	0.88	0.54
g.75.1	0.46	0.47	0.68	0.79	0.62	0.65	0.59	0.57	0.61	0.79	0.46
g.76.1	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.54	0.71	0.43
g.77.1	0.30	0.29	0.27	0.35	0.18	0.29	0.17	0.12	0.38	0.50	0.30
g.78.1	0.35	0.35	0.49	0.48	0.43	0.43	0.45	0.40	0.43	0.77	0.35
g.79.1	0.51	0.50	0.57	0.56	0.55	0.45	0.51	0.54	0.61	0.88	0.51
g.80.1	0.61	0.60	0.70	1.00	0.64	1.00	1.00	1.00	0.61	1.00	0.61
g.81.1	0.61	0.60	0.74	0.75	0.70	1.00	0.74	1.00	0.61	1.00	0.61
g.82.1	0.61	0.60	0.63	0.62	0.57	0.50	0.62	0.63	0.61	1.00	0.61
g.83.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.84.1	0.14	0.13	0.33	0.33	0.24	0.24	0.19	0.17	0.14	0.22	0.14
g.85.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.88.1	0.33	0.33	0.43	0.45	0.37	0.36	0.36	0.39	0.33	0.55	0.33
g.89.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
g.90.1	0.00	0.00	0.19	0.20	0.15	0.16	0.08	0.11	0.00	-	0.00
g.93.1	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	0.61
h.1.1	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.20	0.32	1.00
h.1.2	0.30	0.29	0.36	0.33	0.27	0.24	0.23	0.17	0.38	0.50	-
h.1.3	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
h.1.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
h.1.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
h.1.6	0.22	0.22	0.28	0.27	0.21	0.21	0.23	0.16	0.22	0.36	-
h.1.7	0.20	0.19	0.28	0.27	0.21	0.21	0.20	0.16	0.26	0.32	-
h.1.8	0.26	0.25	0.39	0.36	0.31	0.29	0.25	0.18	0.61	0.43	-
h.1.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
h.1.10	0.30	0.29	0.39	0.37	0.30	0.30	0.19	0.18	0.61	0.00	-
h.1.11	0.51	0.50	0.62	0.62	0.57	0.45	0.51	0.54	0.61	1.00	-
h.1.12	0.43	0.42	0.43	0.40	0.33	0.32	0.30	0.27	0.61	0.71	-
h.1.15	0.61	0.60	0.70	0.69	0.64	0.64	0.74	0.76	0.61	1.00	-
h.1.16	0.28	0.30	0.53	0.52	0.46	0.48	0.35	0.26	0.61	0.00	-
h.1.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
h.1.18	0.30	0.29	0.34	0.32	0.26	0.25	0.20	0.16	0.61	0.50	-
h.1.20	0.54	0.53	0.53	0.49	0.44	0.41	0.47	0.49	0.61	0.71	-
h.1.21	0.30	0.29	0.39	0.36	0.31	0.29	0.23	0.21	0.43	0.50	-
h.1.22	0.61	0.60	0.59	0.56	0.49	0.45	0.62	0.63	0.61	1.00	-
h.1.23	0.09	0.10	0.31	0.30	0.23	0.25	0.12	0.13	0.17	-	-
h.1.25	0.35	0.34	0.43	0.42	0.35	0.36	0.27	0.27	0.61	0.58	-
h.1.26	0.61	0.60	0.70	0.74	0.64	0.71	0.74	0.76	0.61	1.00	-
h.1.27	0.43	0.42	0.40	0.40	0.31	0.32	0.30	0.27	0.43	0.71	-
h.1.28	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.43	0.71	-
h.1.29	0.38	0.37	0.47	0.45	0.40	0.40	0.25	0.25	0.61	0.63	-
h.1.30	0.12	0.11	0.25	0.21	0.17	0.18	0.13	0.12	0.15	0.20	-
h.1.31	0.43	0.42	0.42	0.41	0.35	0.36	0.30	0.33	0.54	0.71	-
h.1.33	0.54	0.53	0.54	0.54	0.44	0.49	0.48	0.46	0.61	0.91	-
h.1.34	0.35	0.34	0.43	0.42	0.31	0.30	0.27	0.27	0.38	0.58	-
h.2.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
h.3.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.19
h.3.2	0.16	0.16	0.27	0.27	0.20	0.20	0.14	0.13	0.61	0.27	-
h.3.3	0.19	0.20	0.40	0.43	0.34	0.34	0.21	0.18	0.19	-	-
h.4.1	0.22	0.22	0.47	0.48	0.33	0.37	0.19	0.17	0.22	-	1.00
h.4.4	0.48	0.47	0.66	0.61	0.61	0.56	0.50	0.60	0.61	0.71	-
h.4.5	0.72	0.72	1.00	1.00	1.00	1.00	0.76	1.00	1.00	0.91	-
h.4.8	0.54	0.53	0.49	0.48	0.40	0.38	0.44	0.52	0.54	0.71	-
h.4.10	0.43	0.42	0.43	0.42	0.33	0.32	0.30	0.27	0.61	0.71	-
h.4.11	0.54	0.53	0.68	0.79	0.64	0.67	0.57	0.54	0.61	0.00	-
h.4.12	0.69	0.67	0.77	0.75	0.72	0.70	0.76	0.72	1.00	0.58	-
h.4.13	0.30	0.29	0.34	0.32	0.26	0.25	0.20	0.16	0.61	0.50	-
h.4.15	0.21	0.21	0.42	0.41	0.36	0.35	0.17	0.15	0.61	0.00	-
h.4.16	0.43	0.42	0.40	0.36	0.31	0.32	0.30	0.27	0.43	0.71	-
h.4.17	0.43	0.42	0.43	0.42	0.36	0.34	0.32	0.36	0.61	0.71	-
h.4.18	0.22	0.21	0.62	0.63	0.56	0.57	0.21	0.24	0.22	-	-

Continued on next page

Table B2 – continued from previous page

	Maximum parsimony								Fusion	Dollo	
sf	**ncbi**	**ncbif**	**jacc**	**jaccf**	**bc**	**bcf**	**pars**	**parsf**		**MULTI**	**fold**
h.4.19	0.18	0.17	0.41	0.40	0.34	0.32	0.14	0.21	0.61	0.30	-
h.5.1	0.61	0.60	0.73	0.79	0.70	0.71	0.74	0.76	0.61	1.00	0.61
h.6.1	0.12	0.11	0.22	0.21	0.17	0.18	0.14	0.11	0.15	0.20	0.12
h.7.1	0.22	0.22	0.28	0.27	0.21	0.21	0.23	0.16	0.35	0.36	0.22

Appendix C: Material supplementary to Chapter 3

Table C1: Enriched functional terms for different age groups. GO terms that are found to be significantly enriched in new-born, middle-aged, or ancient superfamilies. Terms in italics are supported by analysis on annotations derived purely from single domain Uniprots only. These terms can be understood as domain-centric functional annotations but as they are more rare they lead to a less specific enrichment analysis.

GO term	Age	p-value
anatomical structure morphogenesis	middle	7.27e-03
positive regulation of cellular process	middle	3.46e-03
protein binding	middle	6.31e-03
regulation of cellular process	middle	9.50e-03
regulation of developmental process	middle	1.03e-03
tissue development	middle	6.45e-03
acetyl-CoA metabolic process	ancient	8.35e-05
adenyl nucleotide binding	ancient	1.69e-03
adenyl ribonucleotide binding	ancient	1.69e-03
amine biosynthetic process	ancient	3.26e-09
amine catabolic process	ancient	2.25e-03
amine metabolic process	ancient	5.41e-21
aspartate family amino acid metabolic process	ancient	1.93e-03
ATP metabolic process	ancient	7.44e-03
biosynthetic process	ancient	2.36e-34
carbohydrate biosynthetic process	ancient	1.45e-08
carbohydrate catabolic process	ancient	1.70e-04
carbohydrate derivative catabolic process	ancient	4.89e-04
carbohydrate derivative metabolic process	ancient	8.94e-09
carbohydrate metabolic process	ancient	1.40e-05
carbon-carbon lyase activity	ancient	3.54e-06
carbon-oxygen lyase activity	ancient	2.21e-03
carboxylic acid biosynthetic process	ancient	4.16e-08
carboxylic acid metabolic process	ancient	1.03e-21
carboxy-lyase activity	ancient	4.24e-03
catabolic process	ancient	2.38e-13

Continued on next page

Table C1 – continued from previous page

GO term	Age	p-value
catalytic activity	ancient	7.25e-35
cellular amine metabolic process	ancient	3.17e-23
cellular amino acid biosynthetic process	ancient	1.26e-09
cellular amino acid metabolic process	ancient	1.16e-20
cellular aromatic compound metabolic process	ancient	1.97e-10
cellular biosynthetic process	ancient	1.27e-30
cellular carbohydrate biosynthetic process	ancient	1.93e-03
cellular catabolic process	ancient	4.69e-09
cellular component biogenesis at cellular level	ancient	4.89e-04
cellular ketone metabolic process	ancient	2.96e-23
cellular macromolecule biosynthetic process	ancient	8.81e-09
cellular macromolecule metabolic process	ancient	5.46e-06
cellular metabolic process	ancient	7.08e-39
cellular nitrogen compound biosynthetic process	ancient	5.57e-21
cellular nitrogen compound catabolic process	ancient	3.01e-05
cellular nitrogen compound metabolic process	ancient	1.43e-28
cellular polysaccharide biosynthetic process	ancient	4.24e-03
cellular polysaccharide metabolic process	ancient	8.83e-04
cellular process	ancient	1.56e-14
cellular respiration	ancient	8.10e-03
coenzyme biosynthetic process	ancient	2.77e-08
coenzyme metabolic process	ancient	7.84e-20
cofactor biosynthetic process	ancient	3.26e-08
cofactor catabolic process	ancient	1.93e-03
cofactor metabolic process	ancient	1.90e-17
cytoplasm	ancient	8.9e-09
cytoplasmic part	ancient	1.52e-10
cytosol	ancient	1.78e-04
cytosolic part	ancient	1.63e-06
dicarboxylic acid metabolic process	ancient	4.03e-04
energy derivation by oxidation of organic compounds	ancient	6.56e-06
gene expression	ancient	2.2e-08
generation of precursor metabolites and energy	ancient	6.03e-10
glucose catabolic process	ancient	1.83e-04
glucose metabolic process	ancient	3.8e-05
glutamine family amino acid metabolic process	ancient	3.82e-04
heterocycle biosynthetic process	ancient	9.02e-15
heterocycle catabolic process	ancient	6.34e-05
heterocycle metabolic process	ancient	5.49e-28
hexose catabolic process	ancient	1.72e-05
hexose metabolic process	ancient	2.90e-07
hydrolase activity	ancient	8.29e-07
hydrolase activity, acting on acid anhydrides	ancient	6.62e-04
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	ancient	6.62e-04
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	ancient	8.10e-03
intracellular	ancient	1.12e-03
intracellular part	ancient	1.12e-03
isomerase activity	ancient	3.08e-03
large ribosomal subunit	ancient	1.08e-03
lyase activity	ancient	7.91e-10
macromolecule biosynthetic process	ancient	9.97e-09
macromolecule metabolic process	ancient	2.17e-03
metabolic process	ancient	2.95e-35
microbody	ancient	9.28e-03
mitochondrial matrix	ancient	1.02e-07
mitochondrial part	ancient	2.73e-08

Continued on next page

Table C1 – continued from previous page

GO term	Age	p-value
mitochondrion	ancient	1.23e-14
molecular function	ancient	7.68e-08
monocarboxylic acid metabolic process	ancient	4.21e-03
monosaccharide catabolic process	ancient	3.54e-06
monosaccharide metabolic process	ancient	2.11e-07
ncRNA metabolic process	ancient	8.14e-07
ncRNA processing	ancient	8.43e-05
nicotinamide nucleotide metabolic process	ancient	9.26e-03
nitrogen compound metabolic process	ancient	1.37e-26
nucleobase-containing compound biosynthetic process	ancient	1.32e-08
nucleobase-containing compound catabolic process	ancient	7.16e-05
nucleobase-containing compound metabolic process	ancient	1.56e-13
nucleobase-containing small molecule metabolic process	ancient	3.05e-23
nucleoside metabolic process	ancient	1.24e-04
nucleoside phosphate metabolic process	ancient	1.56e-20
nucleoside triphosphate catabolic process	ancient	2.16e-03
nucleoside triphosphate metabolic process	ancient	2.37e-05
nucleotide binding	ancient	3.10e-06
nucleotide biosynthetic process	ancient	4.32e-05
nucleotide catabolic process	ancient	3.17e-04
nucleotide metabolic process	ancient	3.32e-20
organic acid biosynthetic process	ancient	4.16e-08
organic acid catabolic process	ancient	6.39e-03
organic acid metabolic process	ancient	1.23e-22
organic substance metabolic process	ancient	3.19e-08
oxidation-reduction process	ancient	1.16e-08
oxidoreductase activity	ancient	1.46e-06
oxidoreductase activity, acting on CH-OH group of donors	ancient	1.28e-04
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	ancient	1.93e-03
oxidoreduction coenzyme metabolic process	ancient	4.03e-04
oxoacid metabolic process	ancient	2.51e-22
plastid	ancient	1.32e-04
polysaccharide biosynthetic process	ancient	8.35e-05
primary metabolic process	ancient	3.35e-24
purine-containing compound biosynthetic process	ancient	6.18e-05
purine-containing compound metabolic process	ancient	6.56e-13
purine nucleoside metabolic process	ancient	3.13e-03
purine nucleoside triphosphate catabolic process	ancient	8.03e-03
purine nucleoside triphosphate metabolic process	ancient	3.25e-04
purine nucleotide binding	ancient	3.27e-04
purine nucleotide biosynthetic process	ancient	3.80e-03
purine nucleotide catabolic process	ancient	1.10e-03
purine nucleotide metabolic process	ancient	1.60e-11
purine ribonucleoside metabolic process	ancient	3.13e-03
purine ribonucleoside triphosphate metabolic process	ancient	6.19e-04
purine ribonucleotide binding	ancient	3.27e-04
purine ribonucleotide catabolic process	ancient	4.17e-03
purine ribonucleotide metabolic process	ancient	3.26e-09
pyridine-containing compound metabolic process	ancient	4.24e-03
pyridine nucleotide metabolic process	ancient	4.24e-03
pyrophosphatase activity	ancient	1.11e-03
ribonucleoprotein complex	ancient	5.65e-03
ribonucleoside metabolic process	ancient	3.82e-04
ribonucleoside triphosphate metabolic process	ancient	3.25e-04
ribonucleotide binding	ancient	3.27e-04
ribonucleotide catabolic process	ancient	4.17e-03

Continued on next page

Table C1 – continued from previous page

GO term	Age	p-value
ribonucleotide metabolic process	ancient	4.23e-09
ribosomal subunit	ancient	1.57e-09
ribosome	ancient	5.52e-09
RNA binding	ancient	3.21e-03
small molecule binding	ancient	2.74e-05
small molecule biosynthetic process	ancient	1.15e-09
small molecule catabolic process	ancient	1.90e-03
small molecule metabolic process	ancient	3.40e-34
structural constituent of ribosome	ancient	3.01e-08
sulfur compound biosynthetic process	ancient	2.21e-03
sulfur compound metabolic process	ancient	4.51e-05
transferase activity	ancient	2.35e-07
translation	ancient	8.11e-10
tRNA metabolic process	ancient	3.19e-07

Appendix D: Material supplementary to Chapter 5

Table D1: Inter-fold sequence links. The pairwise list of different folds with significant alignments between representative models. The superfamilies of their most significant alignments are also given along with the E-value of the alignment and the number of significant alignments between these two folds (nalign)

fold1	fold2	sf1	sf2	E-value	nalign
b.1	c.1	b.1.18	c.1.8	3.7e-18	1
b.67	b.70	b.67.1	b.70.1	2.9e-04	1
a.56	d.15	a.56.1	d.15.4	5.3e-09	1
c.41	c.8	c.41.1	c.8.4	1.9e-04	1
b.145	b.86	b.145.1	b.86.1	2.0e-04	1
b.1	b.95	b.1.18	b.95.1	6.6e-06	1
b.18	c.41	b.18.1	c.41.1	4.4e-11	1
a.39	a.48	a.39.1	a.48.1	5.5e-12	1
a.138	a.3	a.138.1	a.3.1	1.7e-05	4
a.137	c.96	a.137.4	c.96.1	6.2e-09	1
c.3	c.4	c.3.1	c.4.1	1.2e-15	3
c.2	c.79	c.2.1	c.79.1	1.8e-05	1
b.36	c.14	b.36.1	c.14.1	1.6e-12	2
c.2	c.3	c.2.1	c.3.1	2.1e-06	9
c.31	c.36	c.31.1	c.36.1	5.3e-12	1
c.2	c.4	c.2.1	c.4.1	1.1e-10	2
b.98	d.92	b.98.1	d.92.1	1.7e-17	1
c.37	c.91	c.37.1	c.91.1	2.6e-04	1
b.34	d.93	b.34.2	d.93.1	7.3e-05	1
c.1	d.54	c.1.11	d.54.1	2.3e-09	1
b.71	c.1	b.71.1	c.1.8	1.3e-15	1
b.68	b.70	b.68.4	b.70.3	3.8e-38	10
b.92	c.1	b.92.1	c.1.9	3.1e-11	3
b.68	b.69	b.68.11	b.69.1	2.3e-78	23
a.216	a.24	a.216.1	a.24.9	5.3e-07	1
b.26	d.144	b.26.1	d.144.1	1.9e-06	1
a.35	b.82	a.35.1	b.82.1	5.5e-06	1
c.96	d.58	c.96.1	d.58.1	4.1e-29	1

Continued on next page

Table D1 – continued from previous page

fold1	fold2	sf1	sf2	E-value	nalign
b.67	b.68	b.67.1	b.68.6	2.2e-05	1
b.69	b.70	b.69.2	b.70.1	4.8e-44	12
c.2	c.66	c.2.1	c.66.1	1.1e-07	8
c.56	d.58	c.56.5	d.58.19	1.9e-16	1
c.2	c.5	c.2.1	c.5.1	1.7e-04	1
d.58	d.90	d.58.1	d.90.1	1.4e-08	1
a.1	d.58	a.1.2	d.58.1	1.9e-15	5
a.278	d.344	a.278.1	d.344.1	4.2e-05	1