

Biological Investigations through Sequence Analysis

Corin Yeats

Submitted for Degree

Doctor of Philosophy, University of Cambridge

October 2004

Sidney Sussex College, University of Cambridge

& The Wellcome Trust Sanger Institute

Acknowledgements

Many people have contributed, both accidentally and intentionally, both positively and negatively, to helping me get to here. Finding myself at the Sanger Institute with three years and 250 pages behind me and on the desk in front of me, has been as much from the efforts of others as from my own actions. I have always tried to approach life with an open mind and with a constant desire for learning; and I have tried to take lessons from both the positive and negative. And for instilling this attitude, and for providing many lessons in both of the above, I would like to thank my parents. They have made all things possible and given me the security to explore freely. Thank you.

Next up, from school: Chas, Andy

-and much love to everyone I've ever met with the name Oury -

Nick, Hamish. What can I say? We were there and we left, and it could have been very different. One word: Excellent. And in the nearly ten years since, second word: Excellent.

Anna, thank you, you've been wonderful.

And of course there are the people who have contributed directly to my work and learning. First and foremost my supervisor Alex Bateman has been an inspiration, giving me enough room to learn and putting in far more hours into my education than I had any right to expect. The whole of the Pfam group are superb and I wish them all much future success (is one paper in the top ten most cited enough?!). And I'd like to thank everyone I've collaborated with -especially Steve Bentley.

And in no particular order: Ali M, Ali W, Amy, {Bob, Mike, Barney), Ben D-J, Ben M, Ben S, Big Al, Billy, Buttercuts, Cath, Charlie C, Charlie T, Chris, Dan B, Dave U, Doug, Iffy, Jim, Jude, Matt & Anne (the antithesis of nuisance neighbours), Mike C, music & musicians everywhere, Nicola, Nikki, Tim, Waseem, Wee Al, Will, *et al.*

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Summary

The examination of three dimensional protein structures has revealed that most proteins are made up from modular building blocks. These blocks normally form stable globular structures, and carry particular functions - e.g. catalytic properties - and hence have been termed 'domains'. Domains can be considered both the functional and evolutionary units from which proteins are formed. It has also been demonstrated that if two protein amino acid sequences show significant similarity, then their structures also display similarity.

I have sought to take advantage of the huge amount of sequence data that is being generated by the current wave of genome sequencing projects to identify novel domains and build alignments of homologous sequences. These alignments provide a powerful means to integrate multiple sources of data and hence enable the derivation of novel biological knowledge without recourse to further laboratory experimentation.

Novel domains identified include: the PASTA domain, a β -lactam antibiotic binding domain, with various roles in eubacterial cell wall growth and maintenance; the eubacterial BON domain, a probable phospholipid membrane binding domain, with roles in osmotic shock protection and mechanosensitive channel function; the PepSY domain, which is likely to inhibit eubacterial M4 peptidases but is also found in archaea, and is possibly important in microbe-microbe interactions as well as self-protection and Bacillales sporulation.

Contents Listing

Acknowledgements	ii
Summary	iv
1 Overview	1
1.1 Aim	1
1.2 Background	1
1.3 Protein Domains, Repeats, Motifs and Families	6
1.4 Characteristic Properties of a Protein Domain	9
1.5 The Limitations and Difficulties of Domain Hunting	13
1.5.1 Domain Boundary Identification	13
1.5.2 The Stepping Stone Phenomenon	15
1.5.3 Replication of Experiments	18
1.6 Tools	21
1.6.1 Search Software	21
1.6.2 Alignment Software	25
1.6.3 Databases	31
1.6.4 Structural Collections and Classifications	37
1.6.5 Presenting Domain Architectures and Alignments	40
2 Identifying Novel Domains	44
2.1 Domain Hunt Methods	44
2.1.1 Introduction	44
2.1.2 Details of Methods	45
2.2 Domain Hunting in <i>Streptomyces coelicolor</i>	52
2.2.1 Introduction to <i>Streptomyces coelicolor</i> - a Complex Prokaryote	52
2.2.2 Methods	54
2.2.3 Summary of Results	54
2.2.4 Notes on Table of All Identified Novel Domains	57

2.3 Descriptions of Novel Domains	57
<i>HA</i>	57
<i>BTAD</i>	62
<i>ALF</i>	65
<i>SPDY</i>	68
<i>PASTA</i>	70
<i>HHE</i>	73
<i>PPC</i>	77
<i>FMN_bind</i>	81
<i>MbtH</i>	84
2.4 Significantly Extended Families	86
<i>SCP</i>	86
<i>FG-GAP</i>	92
2.5 Concluding Comments	92
3 Multi-genome Domain hunting	95
3.1 Rationale	95
3.2 Results	97
3.2.1 Summary of Results	97
3.2.2 Notes on Table of All Identified Novel Domains	98
3.3 Descriptions of Novel Domains	98
3.3.1 Domains Identified From Repeats	100
<i>PepSY</i>	100
<i>Gate</i>	100
<i>STN</i>	102
<i>Secretin_N</i>	105
<i>Secretin_N_2</i>	108
<i>Reg_prop</i>	110
<i>Y_Y_Y</i>	113
<i>DUF1533</i>	115
<i>Coat_X</i>	115
<i>Cleaved_adhesin</i>	118
<i>FIVAR</i>	118
<i>FlaE</i>	127
<i>Glug</i>	132
3.3.2 Domains Found Through Small Protein Clustering	132
<i>Coat_F</i>	132
<i>CTnDOT_TraJ</i>	132
<i>Dabb</i>	135

<i>Nif11</i>	142
3.4 Other Potential Uses	142
4 Detailed Investigations of Individual Domains	145
4.1 The PASTA Domain: A β-lactam-Binding Domain	145
4.1.1 Background	145
4.1.2 Searching for PASTA	147
4.1.3 Structure of PASTA	150
4.1.4 Roles of PASTA	151
4.1.5 PASTA and Cell Morphology	153
4.1.6 The PASTA Domain as an Antibiotic Target	154
4.1.7 Subsequent Research	156
4.2 The BON Domain: A Putative Membrane Binding Domain	157
4.2.1 Identification of the Conserved Regions	158
4.2.2 OsmY Comprises Two BON Domains	159
4.2.3 Other BON-containing Proteins	161
4.2.4 Phylectic Distribution	163
4.3 The PepSY Domain: A Putative Regulator of Peptidase Activity	164
4.3.1 Background to the M4 Peptidases	165
4.3.2 PepSY Domain Identification	166
4.3.3 Description of the PepSY Domain	166
4.3.4 Domain Architecture of the M4 Propeptide	170
4.3.5 Species Distribution of PepSY	170
4.3.6 PepSY Family Characteristics	171
4.3.7 PepSY Domains are Likely to be Inhibitors	173

4.3.8 The Biological Role of PepSY	174
4.4 Peptidase_A24 - the Prepilin Peptidase	175
5 Contributions to Genome Annotation Projects	180
5.1 Tropheryma whipplei	180
5.1.1 Background	180
5.1.2 The WiSP Protein Family	181
5.1.3 The WiSP Domains	183
<i>WND</i>	183
<i>CCD</i>	183
<i>He_PIG</i>	187
5.1.4 Implications for the Immune System	194
5.2 Burkholderia pseudomallei	194
5.2.1 Background	194
5.2.2 Novel Domains	195
<i>SCPU</i>	195
<i>BTP</i>	197
<i>PHB_acc</i>	199
<i>The Repetitive β-helix Surface Structure Superfamily</i>	199
5.3 Chlamydomonada abortus	205
5.3.1 Background	205
5.3.2 The Chlamydial Polymorphic Membrane Protein	206
<i>ChlamPMP_M</i>	206
5.4 Theileria annulata	211
5.4.1 Background	211
5.4.2 FAINT	212
5.4.3 The TASR Repeat Families	217
6 Conclusions	223

Bibliography	227
---------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----

Appendix A: List of All Domains in this Thesis	255
---	-----	-----	-----	-----	-----	-----	-----	-----	-----

Figure Listing

Figure 1.1: Growth of the UniProt database	5
Figure 1.2: Examples of different protein structure types	8
Figure 1.3: Graph displaying the average size of domains recognised by Pfam	11
Figure 1.4: A common protein clustering error caused by multidomain proteins	14
Figure 1.5: Graph of the average percent identity for each Pfam family	16
Figure 1.6: The iterative search methodology	18
Figure 1.7: Example of the Dotter output	26
Figure 1.8: A simple example of a "bad" alignment compared to a "good" alignment.	31
Figure 1.9: Example Pfam family page - the PASTA domain	33
Figure 1.10: Key to the architecture figures and some example architectures	43
Figure 2.1: General method for identifying novel domains	49
Figure 2.2: HA example alignment	58
Figure 2.3: HA example architectures	59

Figure 2.4: BTAD example alignment	63
Figure 2.5: BTAD example architectures	64
Figure 2.6: ALF alignment, architectures, and genome context	67
Figure 2.7: SPDY alignment and architectures	69
Figure 2.8: Evidence for the presence of a mobile DNA element	71
Figure 2.9: Alignment of the original HHE domains and predicted secondary structure against a Hemerythrin domain and known structure	74
Figure 2.10: HHE architectures	75
Figure 2.11: PPC alignment along with predicted and known secondary structure	78
Figure 2.12: PPC domain architectures	79
Figure 2.13: FMN_Bind alignment	82
Figure 2.14: FMN_bind domain architectures	83
Figure 2.15: Alignment and architectures for the MbtH domain	85
Figure 2.16: SCP domain alignment	88
Figure 2.17: SCP domain architectures	89
Figure 3.1: Simplified taxonomic tree of bacteria investigated in the multigenome hunt	96
Figure 3.2: Alignment and architectures for the Gate domain	101
Figure 3.3: STN alignment and architectures	103
Figure 3.4: Secretin_N example alignment	106
Figure 3.5: Secretin_N example architectures	107
Figure 3.6: Secretin_N_2 alignment and architectures	109
Figure 3.7: Example Reg_prop alignment	111

Figure 3.8: Example Reg_prop architectures	112
Figure 3.9: Y_Y_Y alignment and architectures	114
Figure 3.10: DUF1533 alignment and architectures	116
Figure 3.11: Coat_X alignment and architectures	117
Figure 3.12: Cleaved_adhesin alignment and architectures	119-120
Figure 3.13: Example FIVAR alignment	122
Figure 3.14: Example FIVAR architectures	123-125
Figure 3.15: Example FlaE alignment and architectures	129
Figure 3.16: Example Glug repeat alignment	130
Figure 3.17: Example Glug repeat architectures	131
Figure 3.18: Example Coat_F alignment and architectures	133
Figure 3.19: Example CTnDOT_TraJ alignment and architectures	134
Figure 3.20: Example Dabb alignment and architectures	136-137
Figure 3.21: Structure of the Dabb barrel	139
Figure 3.22: Example Nif11 architectures and alignment	142
Figure 4.1: Example PASTA alignment	148
Figure 4.2: Example PASTA domain architectures	149
Figure 4.3: Stereo view of the two PASTA domains of <i>Streptococcus pneumoniae</i> PBP2X	150
Figure 4.4: Distribution of resistance mutations in the PASTA domains of PBP2X	155
Figure 4.5: Example BON domain alignment	160
Figure 4.6: Example BON domain architectures	162
Figure 4.7: Example PepSY domain alignment	167
Figure 4.8: FTP motif example alignment	168

Figure 4.9: PepSY_TM example alignment	169
Figure 4.10: Example PepSY domain architectures	172
Figure 4.11: Peptidase_A24 example alignment and architectures	177
Figure 5.1: Example WiSP family architectures	184
Figure 5.2: Example WND alignment	185-186
Figure 5.3: CCD example alignment	188
Figure 5.4: Example He_PIG architectures	189
Figure 5.5: He_PIG example alignment	190
Figure 5.6: N-J Tree of all He_PIG domains from <i>Tropheryma whipplei</i>	193
Figure 5.7: SCPU example alignment and architectures	196
Figure 5.8: BTP example alignment and architectures	198
Figure 5.9: PHB_acc example alignment and architectures	200
Figure 5.10: Example Chlam_PMP alignment with related sequences	202
Figure 5.11: Fil_haemagg and Ice_nucleation example architectures	203
Figure 5.12: The <i>Chlamydomonas reinhardtii</i> polymorphic membrane protein family	207
Figure 5.13: ChlamPMP_M example alignment	209-210
Figure 5.14: FAINT example alignment	213-214
Figure 5.15: FAINT example architectures	216
Figure 5.16: Neighbour-Joining tree of the choline kinases of <i>Theileria parva</i> and <i>Theileria annulata</i>	218
Figure 5.17: Example alignments of the TASR short repeats of <i>T. annulata</i> (A) and <i>T. parva</i> (B)	219
Figure 5.18: <i>Theileria annulata</i> TASR_1 example architectures	220

Table Listing

Table 1.1: Results from Edgar's (2004) comparison of MAFFT, T-Coffee and Clustalw 30
Table 1.2: The ClustalX colouring scheme 42
Table 2.1: Table of all novel domains identified in <i>S. coelicolor</i> 56
Table 3.1: Table of all novel domains identified in the multigenome hunt 99
Table 5.1: Statistics testing whether the overlap between the <i>T. parva</i> TASR_2-containing proteins and the <i>T. annulata</i> TASR_1 proteins is by chance 221

1 Overview

1.1 Aim

The focus of the research in this Thesis is to generate novel biological knowledge through the transfer of information between related protein sequences. Currently there are around 1 million unique protein sequences available in public databases. Around a third of these proteins do not belong to any recognised and characterised family, and the majority contain regions that have not been described. Within these regions remains a huge amount of important biological information – and clustering them into sequence families allows both the synthesis of information from each family member and global analyses of family characteristics. The work carried out in this thesis aims to identify novel families of high interest, to refine known families and to correctly establish the homology borders within the member proteins. Statistical methods are used to identify potential new families in a high throughput manner, which are then manually investigated. Functional predictions are provided through the use of sequence analysis software and through the analysis of associated literature.

1.2 Background

As more protein structures have been solved, using X-ray crystallography and NMR, several trends and constraints of protein structure have become apparent. Of these, the most striking observation was that proteins are usually made up from several independently folding units, with the overall function of the protein being a composite of these substructures' functions. Furthermore, these substructures have been found to

have shuffled during evolution to create novel proteins with new emergent functions. The discrete and modular nature of these elements has led to them being termed domains; this also makes understanding protein domains a powerful way of understanding proteins.

It is also of note that there are already over 1 million proteins in public sequence databases, whereas it is estimated that there are between 1000 and 5,000 folds – a fold being the three dimensional structure a protein assumes in its native state – that exist in nature, with about 50% of proteins belonging to one of 800 folds (reviewed in Grant, Lee *et al.*, 2004; first estimated by Chothia, 1992). Therefore grouping these sequences into fold families and subfamilies makes the data much more manageable. Solving structures is expensive, time-consuming and labour intensive at best, and at worst is currently impossible – particularly with the extremely biologically interesting cell membrane-associated proteins. So while three dimensional structural analyses are highly informative, comparative methods of protein sequence and structure analysis are essential.

Certain observations from sequence analyses have led to the development of powerful tools for protein comparison and structure determination. First and foremost is that protein amino acid sequences divide up into discrete units, which can be found in differing contexts. Mapping these to the corresponding structures has shown that a “sequence domain” almost certainly maps directly to a “structural domain”. There are of course exceptions and qualifiers – for instance β -propellers are typically made up from between 6-8 sequence repeats and form a fold made up from 6-8 “blades”

(Murzin, 1992) - though of note haemopexin forms a four-bladed propeller (Gomis-Ruth, Gohlke *et al.*, 1996). All the blades are required to form the propeller, and hence all should be included as part of a single fold, but at a sequence level it would be seen as a series of homologous, and possibly gapped, repeats.

The second key observation is that if two protein sequences are shown to be evolutionarily related (homologous) then they will have the same tertiary structure – though again there can be exceptions (Grishin, 2001). There are now several powerful statistical tools for determining the likelihood that two sequences are related, some of which are described in chapter 1.6.

It is a common maxim in structural biology that function is encapsulated within the structure. If, through sequence analysis, we are able to demonstrate that set of sequences or sub-sequences are homologous, then we can transfer functional information associated with these regions. These two observations imply that if we can describe a family of related protein sequences and we know the physical structure of one of the proteins, then we can describe the function of all of them. This is because we should be able to construct comparative models based on the known structure and identify changes to the biochemistry of the protein. Developing comparative analysis technologies is currently the main approach in protein analysis as the cost of sequencing the gene is several orders of magnitude less than solving the structure of the protein, and *ab initio* structural prediction methods are still prone to significant inaccuracies (Aloy, Stark *et al.*, 2003).

In general the volume of publicly available protein sequence data has been expanding rapidly, driven by the current wave of genome sequencing projects, as indicated by the growth of the sequence repository, UniProt (see Figure 1.1). In turn sequence analysis has become part of the standard repertoire of biological research methods, and is now carried out on desktop computers by lab bench researchers and *en masse* on supercomputers by trained informaticians. A subfield of protein sequence analysis is domain hunting – the identification of novel protein domains from sequence data.

The concept of protein domains became apparent soon after the first structures were solved, and by the mid-1970s they were being considered in both sequence and structural terms (e.g. Wetlaufer, 1973; Edelman and Gall, 1969; Rossman and Liljas, 1974), with the first defined domain being the Ig domain (Edelman and Gall, 1969). The principle that they could be considered as mobile genetic units was put forward by Rossman and Liljas (1974) after analysing the similarity of nucleotide-binding domains in different structures.

Led by researchers like Eugene Koonin, Peer Bork, Chris Ponting & Kay Hoffman (to name a few) the *de novo* identification of domains has become a field in its own right. Approaches range from the purely automated (e.g. ProDom, see chapter 1.6.3; Servant, Bru *et al.*, 2002) to manually intensive (e.g. the BRCT domain; Bork, Hofmann *et al.*, 1997), and encompass combinatorial approaches (e.g. Ponting, Mott *et al.*, 2001). Several databases now collect and curate descriptions of these domains (see chapter 1.6.3), and provide tools for identification of known domains in new sequences. These use a variety of statistical methods and design philosophies. Others

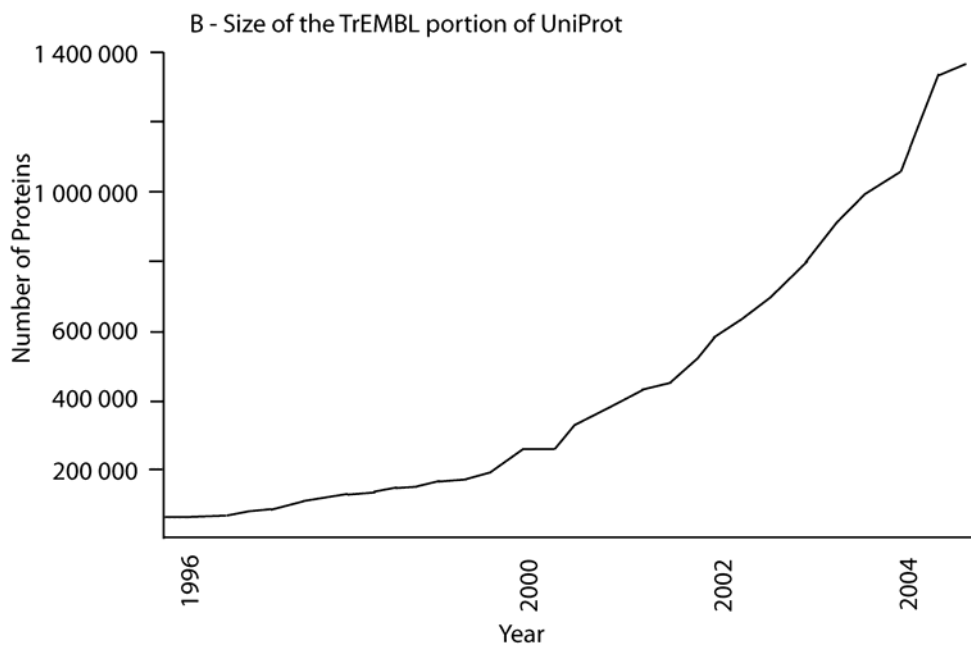
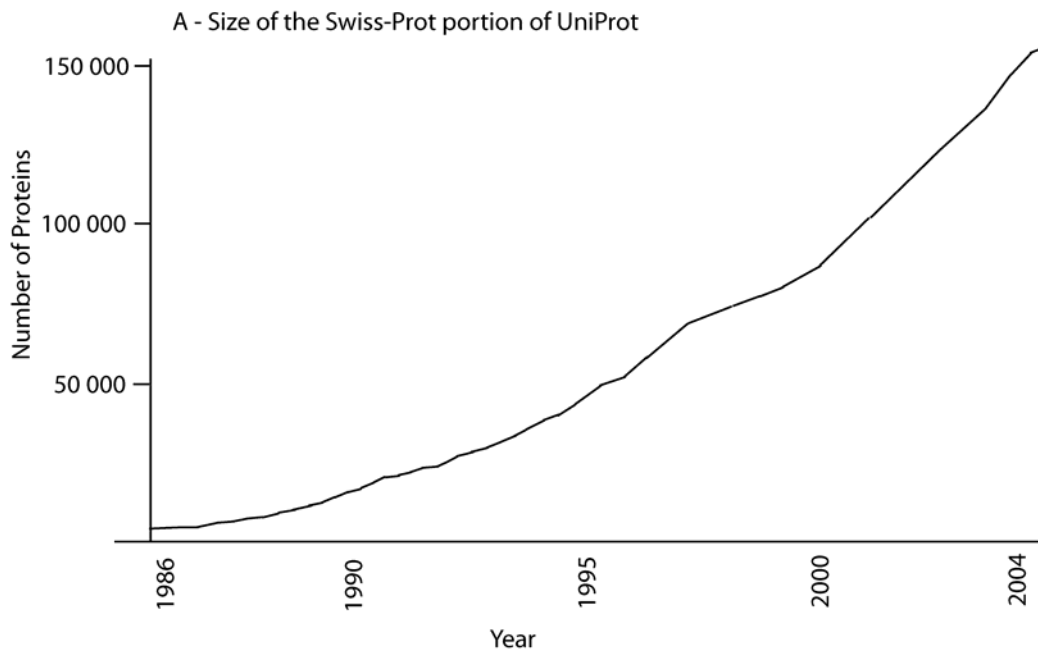


Figure 1.1: Growth of the UniProt database

(A) Swiss-Prot began in 1986 and has grown in a fairly steady fashion to the 150,000 proteins it now contains.

(B) The automatically generated supplement was begun in 1996 to keep up with sequence being generated by the large-scale sequencing projects begun in the 1990s. TrEMBL now contains more than 1,400,000 protein sequences - though there may be a high degree of redundancy relative to Swiss-Prot. Its growth appears to be exponential.

Both of these graphs are based on those presented at <http://us.expasy.org/sprot/>.

present the results for automatic domain detection, and update them when new sequences become available – for instance ProDom and the derivative Pfam-B (Bateman, Coin *et al.*, 2004), or ProtoMap (Yona, Linial *et al.*, 1999).

The growth in the power of domain identification from sequence has been driven by the large-scale sequencing projects of the last ten years. Previously the protein sequence databases were small and highly biased towards specific proteins or families of interest. Genome sequencing has led to a much wider range of proteins being sequenced, hence increasing the diversity of domains contained within the sequence database and the diversity of contexts these domains are found in. The increased diversity of sequences found in the protein databases can also allow subtler relationships to be derived, by the introduction of "Stepping Stone Sequences" - see chapter 1.5.2 for an explanation. As a result, not only is it possible to detect recently deposited novel domains, but also it is becoming easier to detect domains that were already present.

1.3 Protein Domains, Repeats, Motifs and Families

Proteins exhibit modular structures, with their overall function or fold being emergent from the modular components they are constructed from. The specific arrangement of modules is called the "domain architecture". All these components can be grouped into three classes - domain, structural repeat, and motif. When it is not possible to assign a component to a particular category, it can be classified as a family. These four types are the same as used by the Pfam (see chapter 1.6.2) database, around which the work in this thesis is based. While there is much discussion on what

constitutes a protein domain, the definition mostly depends on perspective; for a more detailed discussion of the precise differences see the review by Kong and Ranganathan (2004). Since much of the work presented has been done purely on protein sequence, without any available structure models, the most used definition of domain is the second one given below, but it should be noted that all three definitions largely overlap. The effective difference between them is on deciding where to position the edges of the domain within a protein. For instance a functional domain may be equivalent to an evolutionary domain and lie within a structural domain. Figure 1.2 shows examples of domains, motifs and repeats.

Three Common Definitions of a Protein Domain

- Structural**: An independently folding unit in a polypeptide chain, which forms its own hydrophobic core.
- Evolutionary**: A segment of amino acid sequence that is conserved in differing surrounding sequence contexts.
- Functional**: The minimum sequence required to encode a function in a protein, as determined by experimentation.

Definition of a Structural Repeat

A repeat is a conserved sequence that only forms a stable structure when present in more than one copy. Each repeat is not independently stable but all contribute to a final stable structure. Examples are the WD40 repeats (Neer, Schmidt *et al.*, 1994) and TPR repeats (Goebel and Yanagida, 1991). The number of repeats that make up the final structure may or may not be restricted to a range: WD40 repeats occur in sets of 6-8 and form a single propeller-like structure; the approximately 35 residue TPR repeats can occur anywhere between 2 and 50 times and form a solenoid structure.

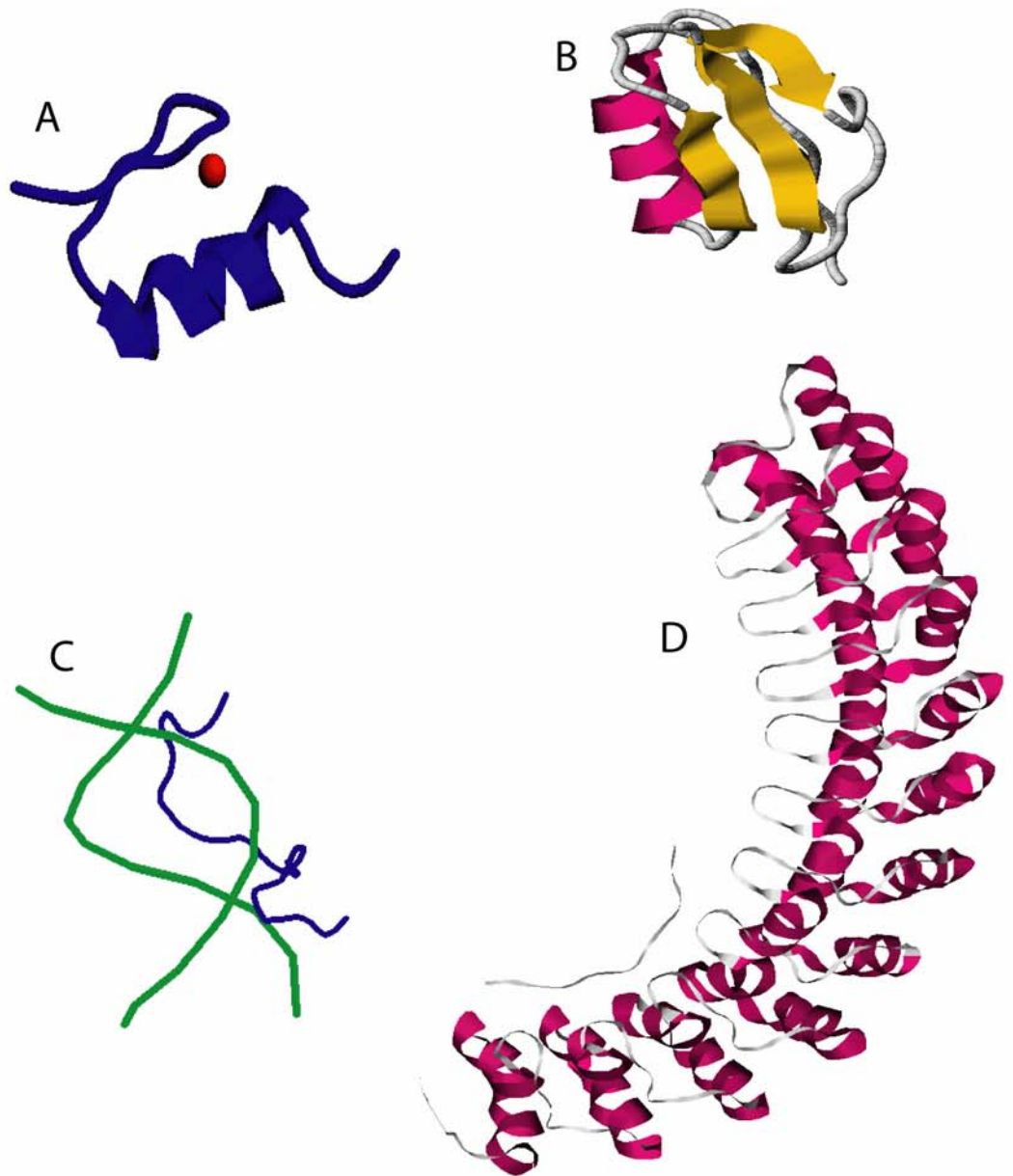


Figure 1.2: Examples of different protein structure types.

(A) **Zinc Finger domain:** Zinc fingers are able to form a stable tertiary structure in the presence of a stabilising zinc cation (red sphere). Image derived from PDB:1KLR.

(B) **PASTA domain:** PASTA domains form independently stable tertiary structures and can be considered as a classic structural domain. Image derived from PDB:1QMF.

(C) **AT-hook motif:** The AT-hook motif (blue) is a protein sequence that preferentially binds AT rich DNA (green) but has no stable structure. Image derived from PDB:2EZE.

(D) **Ankyrin repeats:** Several ankyrin repeats group together to form a higher order structure. Image derived from PDB:1N11

Definition of a Sequence Motif

A motif is an amino acid sequence that does not form an independently stable globular structure but has a specific function that is conserved between related sequences. For example the AT-hook motif is a short motif of around 13 residues that binds AT rich DNA (Nissen, Langan *et al.*, 1991). Although it is believed to form a particular secondary structure (Huth, Bewley *et al.*, 1997) its short size and lack of stabilising ligands means that it can not form a stable tertiary structure itself.

Definition of a Sequence Family

A family is a group of sequences that have been shown to be related using sequence comparison, but may consist of more than one domain, motif, repeat or combination thereof.

1.4 Characteristic Properties of a Protein Domain

As described above there are several ways of defining a protein domain, with the definition used being the one appropriate to the type of investigation. However, no matter the definition there are several common characteristics that typify what would be considered a domain. As discussed above, domains are the modular units of proteins, and so modularity would be expected. This can be expressed in several ways: Ideally the domain will be found in multiple architectures, as this demonstrates that it is independent of the surrounding sequence. Experimental evidence can also indicate modularity. For instance proteolytic degradation of the PulD protein (Nouwen, Stahlberg *et al.*, 2000) revealed the same N-terminus for the Secretin domain as the sequence based prediction made in chapter 3.3 (Secretin_N domain). Of course, there are exceptions to this apparently straight forward rule. In some proteins a domain may be dependant on another for correct folding. An example is the

strand swapping between homologous TOBE domains from different *Escherichia coli* ModE proteins (Hall, Gourley *et al.*, 1999; Koonin, Wolf *et al.*, 2000).

The second and possibly simplest property is that domains almost always measure between 50 and 400 residues in length (see Figure 1.3). The lower limit probably reflects the minimum number of residues required to form a stable structure. Stable structures usually are generally globular with a hydrophobic core. There are some exceptions in which strong stabilising interactions have allowed the formation of smaller stable structures. An example is the Zinc finger family, in which a Zn^{2+} ion stabilises a 22 amino acid structure (Miller, McLachlan *et al.*, 1985; depicted in Figure 1.2). Other interactions may include disulphide bridges and hydrogen bonding. If a region has been experimentally determined to be a functional domain, then it may be disordered – it has no stable tertiary structure – and maybe provides an electrostatic charge or some flexibility to the structure (i.e. SMC_hinge). Also transmembrane domains may not fold correctly until inserted into the membrane (i.e. Voltage-dependent K^+ channels; Jiang, Lee *et al.*, 2003). At the other end of the spectrum there are some giant domains - for instance the lipoxygenase domain is apparently a non-dividable structure of over 500 residues (Boyington, Gaffney *et al.*, 1993).

The reason for the lack of folds found that are larger than a few hundred residues in length is not clear. It is possibly due to several reasons rather than any particular one. For a start there may be a lack of unique structures beyond this threshold, with most possible stable forms being a composite of several smaller domains. Also larger

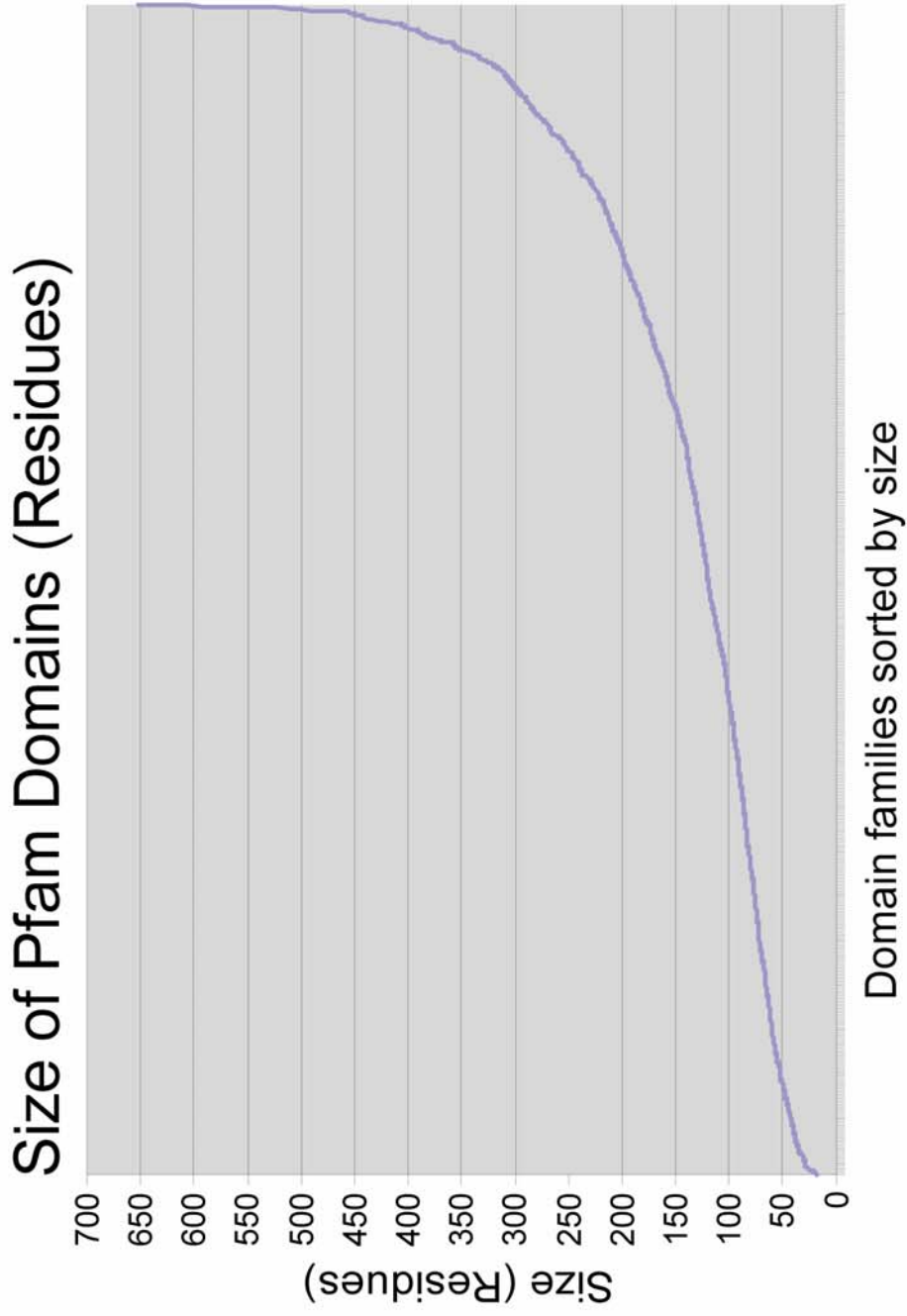


Figure 1.3: Graph displaying the average size of domains recognised by Pfam. For each domain family in Pfam 14 the average length was calculated, after discarding all fragment matches. Displayed are the results sorted by size along the X-axis. As can be seen the large majority of sequence domains are between 50 and 400 residues in length.

domains require more sequence to encode; to a certain extent natural selection minimises genome size, as evidenced by the general lack of intergenic space in bacterial genomes (average gene density = 86%, data from Genome Atlas; Pedersen, Jensen *et al.*, 2000). Furthermore an analysis by Lipman and co-workers (2002) found evidence for significant selection of shorter proteins. It is possible that longer domains would be selected against if there is a smaller domain that can carry out the same function, though this has not been observed.

The third property is that two related domains will also share function. So any member of, for instance, of the Transpeptidase domain family can be predicted to be a transpeptidase provided the catalytic residues are present. However, the extent to which this information transfer can take place varies for different domains and the form it will take can be subtle. The different transpeptidases may have slightly variant substrate specificity, but the basic reaction can be easily described for any. In contrast, the Ig domain shows a huge range of functions, and variants of the domain are able to bind nearly any chemical – one of their biological roles is forming the recognition sites in the immune system immunoglobulins. In this case the domain acts as a scaffold upon which functional motifs, which determine the specific function, can be hung. This mechanism of creating functional diversity is also commonly seen with structural repeats – for instance β -propellers (Murzin, 1992) and CASH repeat proteins (Ciccarelli, Copley *et al.*, 2002) show a similar range of functional diversity as the Ig domain.

So, although the basic concept of a domain is clear and straightforward, as always with biological systems, there are caveats that must always be borne in mind.

1.5 The Limitations and Difficulties of Domain Hunting

1.5.1 Domain Boundary Identification

Between 60 and 80% of proteins in a genome can be expected to consist of more than one domain (Teichmann, Park *et al.*, 1998; Gerstein, 1998). Hence when presented with a single amino acid sequence, the first problem in identifying novel domains is identifying the edges. Correctly identifying the edges of a domain can significantly alter the power of a predictive domain model (e.g. a profile HMM, see chapter 1.6), and lead to large expansions in the number of identified family members. An example from within this thesis is the PASTA domain. The PASTA model is similar to a previous model called PBP_C built by R. Finn, which correctly identified homologous penicillin-binding protein (PBP) regions, but failed to detect significant similarity to the PknB-like serine/threonine kinases (PSTKs). Subsequent to the creation of the PBP_C model, the crystal structure of PBP2X from *Streptococcus pneumoniae* was determined (Gordon, Mouz *et al.*, 2000). From this it was clear that the carboxyl-terminus (C-terminus) consisted of two identical domains and that the model covered the first domain and extended ten residues to the amino-terminus (N-terminus). The boundaries of the PASTA model were found in the sequence using the 'Repeat Hunt Method' described in chapter 2.1.2; it exactly covers one domain and is able to identify many novel homologies. Only a small correction to the model had a dramatic effect on its sensitivity.

Beyond having effects on the sensitivity of the model, correct boundary determination can also affect the quality of information transfer, the effectiveness of structure prediction software, and making crystals for structural analysis (Kong and Ranganathan, 2004). It also can be informative in the evaluation of automated clustering algorithms. A common flaw in many approaches for the automated clustering of protein families is that proteins that are only related by a single domain can be clustered, even though there is no overall functional link and the two proteins are not evolved from a single common ancestor. This type of error can be seen in the genome paper of *Streptomyces coelicolor* (Bentley, Chater *et al.*, 2002), in which the prediction of 44 PSTKs was reported on the basis of single linkage clustering. Using HMMs to predict the domain content shows that there are in fact 34 PSTKs. The discrepancy is caused by single-linkage clustering linking unrelated proteins through domains that they share. This is explained graphically in Figure 1.4.

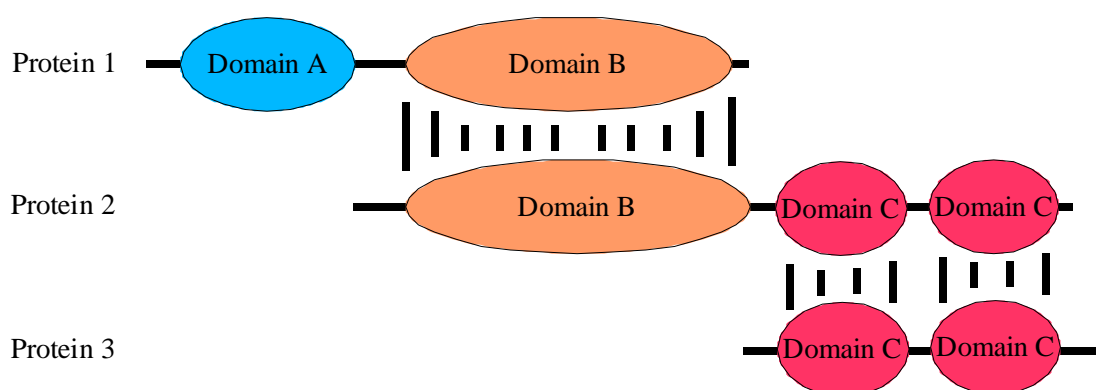


Figure 1.4: A common protein clustering error caused by multidomain proteins. Not knowing the domain structure of a protein under investigation can lead to missannotation. In this case a sequence comparison programme (i.e. BLAST) has identified significant sequence similarity between Protein 1 and Protein 2, as well as between Protein 2 and Protein 3. Naïve interpretation of this result would allow the transfer of information between Protein 1 and Protein 3; However, aided by the knowledge of the domain architectures we can see that there is not likely to be any functional similarity between Protein 1 and Protein 3.

Various different methods have been employed for the recognition of domain boundaries from sequence; I have mostly used manual or semi-automated approaches. These are described in more detail in chapter 2.1. Also there are many researchers developing automated approaches, with two main aims. One is for use over large data sets; the other is for predicting domains from sequence that have no obvious homologues in other proteins. Comparative approaches include mkdom2 (Gouzy, Corpet *et al.*, 1999) - the basis of the ProDom database and an evolution of the original Domainer script (Sonnhammer and Kahn, 1994) - and Gracy and Argos's (1998) pairwise comparison method that underlies DOMO.

Although these methods can be useful for large sets, they have yet to produce the accuracy of results that can be achieved through manual boundary determination – as discussed by Kong and Ranganathan (2004). Recent approaches, such as the combinatorial method developed by Nagarajan and Yona (2004) and the neural network-based method by Liu and Rost (2004) show some promise, and are starting to approach the accuracy of manual detection. The second method also has the advantage that it can take a single sequence and rapidly make a prediction, which can then be refined manually. The current state-of-the-art is reflected in Pfam-A's much higher coverage than Pfam-B despite only consisting of approximately 7,500 families compared to around 100,000 for Pfam-B.

1.5.2 The Stepping Stone Phenomenon

A general rule of thumb in pairwise biological protein sequence comparison is that if two homologous sequences show less than 30% identity (using any measure; May

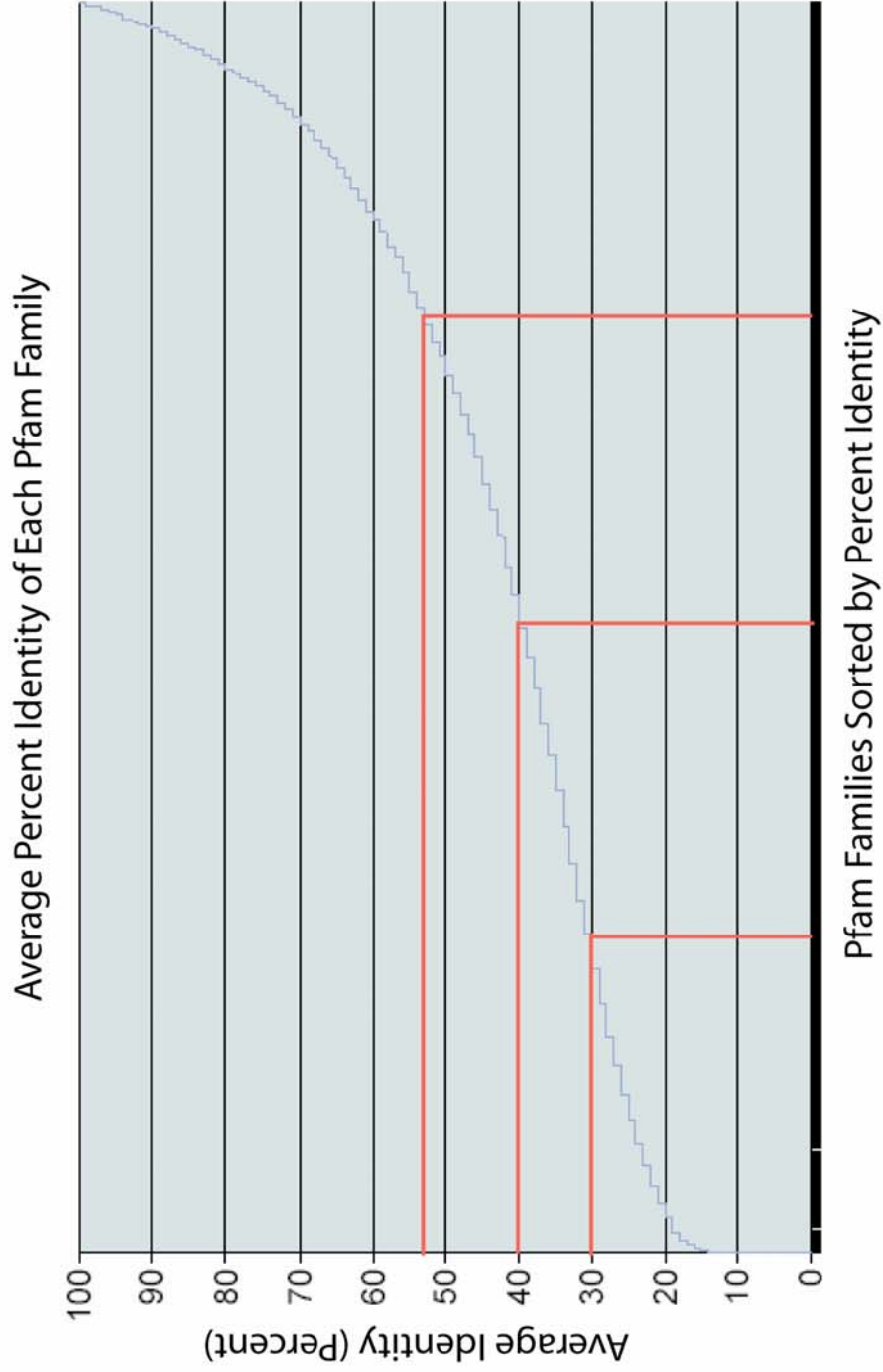


Figure 1.5: Graph of the average percent identity for each Pfam family
 The percent identity of each Pfam family was calculated using Sean Eddy's alifat software. The families are sorted on the X-axis according to their average percent identity. The red lines mark the quartile divisions. As can be seen half of Pfam families have an average sequence identity below 40%. This illustrates the power of iterative searching.

2004) then identifying the relationship is unlikely. However, 50% of Pfam families exhibit less than 40% average identity and 25% have less than 30% average identity (see Figure 1.5, statistics calculated using S. Eddy's "alifold" software). These distant relationships can be most easily detected by identifying an intermediate or, as they are also known, stepping stone sequence. This is a sequence that shows significant similarity to both distantly related sequences, and so can be used to infer a relationship.

This principle essentially underlies iterative searching: Newly identified homologues are included into the model and hence even more divergent homologues are detected (see Figure 1.6 for a graphical explanation). Prior to the genome sequencing projects, protein sequence databases were often biased towards specific proteins, species or sequence families of interest and so the necessary stepping stones were not present. As this is corrected subtle relationships are becoming apparent, but it also means that searches need regular repetition. As an example the HHE domain was identified in early 2002, and formed a cohesive and internally consistent family (Yeats, Bentley *et al.*, 2003). Repeating the searches in 2004 led to the merging of this family with the Hemerythrin family, which had been deposited in Pfam in late 1999. Until recently there was no obvious link between the two because the necessary sequences were not there - such as *Methanosarcina mazei* MM1985 (UniProt:Q8PVB8), a *Streptomyces parvulus* hypothetical protein (UniProt:Q70HY1) and *Shewanella oneidensis* SO3549 (UniProt:Q8EBG9).

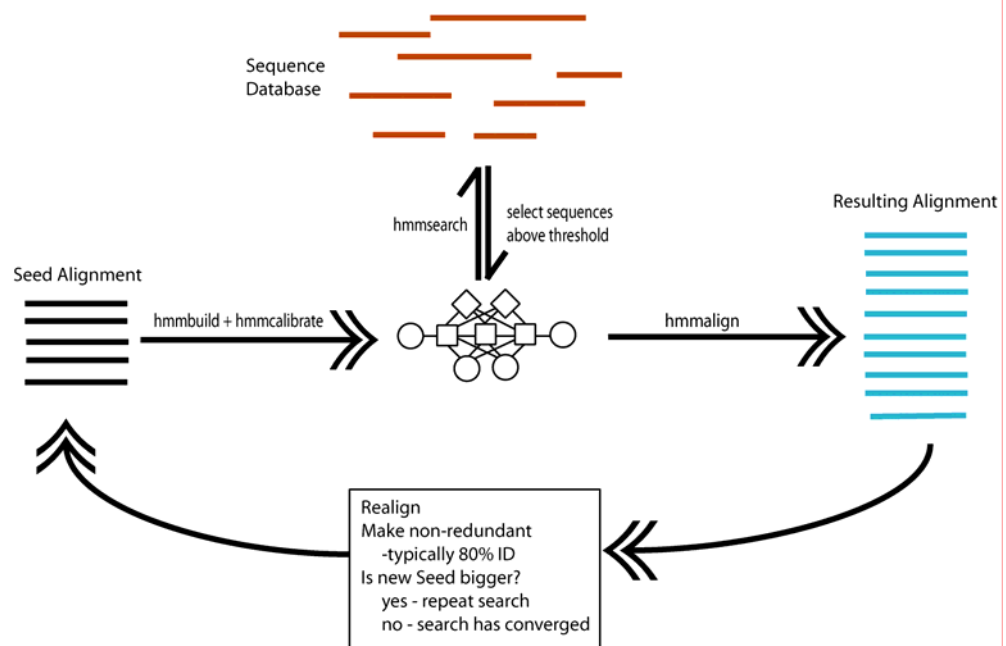


Figure 1.6: The iterative search methodology
 This process allows the researcher to start with a small number of sequences that are known to be related and detect homologues with a high degree of sensitivity and selectivity. By the arrows are the components of Sean Eddy's HMMER software that are used in each step - hmmbuild, hmmlcalibrate, hmmsearch, hmmalign. At the centre of the diagram is a simplified version of HMMER's Plan 7 HMM architecture (discussed further in Chapter 1.6.1).

1.5.3 Replication of Experiments

Despite the statistical basis and computational nature of domain identification from sequence, assigning a confidence score – a level of certainty that there are no false positives included – to a family is not simple. After each round of searching there are two questions: Are there any false positives included? Are there any members missing? Whilst stepping stone sequences allow iterative searching, as has been mentioned, sometimes the required sequences are not present in a database or may not even exist in nature. In this case it may be necessary to relax the inclusion threshold and incorporate sequences with a low similarity in order to identify distant homologues; concomitantly this increases the risk of including false positives. In this

case it is necessary to use reciprocal searches and other evidence to ensure that the relaxation of the threshold is valid.

Another technique for finding distant homologues is to create a smaller sequence database that is believed to be particularly enriched in the target domain (as discussed in chapter 2.3.4). Commonly used estimates of significance (E-values), including those used by BLAST and Prospero, for evaluating the significance of protein similarity scores are functions of database size: The larger a sequence database is, the more chance you would see an apparent match by chance. So by applying a knowledge-based filter, it is possible to reduce the database size while retaining all the copies of a domain, and hence increase the significance of any potential matches.

Using these different techniques, so as to build up a diverse domain family with a low level of conservation, makes statistical validation difficult. A solution that would have provided internal consistency to this thesis would have been to use a fixed release of UniProt (or Swiss-Prot/TrEMBL, see chapter 1.6.4) so that all searches were equivalent. However, the major protein sequence databases have new releases every couple of weeks, with a size doubling period of around 18 months; by not regularly updating, a vast amount of available information is being ignored and valuable stepping-stones may be missing - as was the case with the HHE/Hemerythrin domain.

There is also heterogeneity in the search tools, with some tools able to find more distant homologues in some families than the other tools. So given a starting sequence or alignment, several different results may be arrived at depending on the search tool

and the sequence database. There are also, of course, many different parameters and weighting systems that can be varied for each search tool, adding an extra layer of complexity. Various search tools are explained and discussed in chapter 1.6.1.

Rather than use a strict system of family building in which specific E-values are rigidly adhered to, the approach I have used in this thesis is to carry out controls that vary databases, search tools, starting points and also depositing the results in a public repository (the Pfam database) for further review.

Sequence Search Controls

- (a)** Reciprocal searches - varying the search start point.
- (b)** Vary the N and C-termini of the seed subsequence.
- (c)** Take sequences falling just below the inclusion thresholds as seeds.
- (d)** Use a different search tool – PSI-BLAST/BLAST/HMMER.
- (e)** Vary the sequence database – UniProt/GenBank/Selected sequences.
- (f)** Publish the family, either in the literature or in a public database, for peer review.
- (g)** Use different inclusion thresholds.
- (h)** Careful visual examination of the final alignment to identify inconsistent sequences.
 - can be aided by building a Neighbour-Joining Tree to group potential false positives.

The final decision as to whether the identified domain family was genuine, and that as many true members had been identified as possible with few (preferably none) false members included, is subjective but achieved through the consensus of several experiments. It is also important to be conservative in decision making until further

tests support the inclusion of more divergent sequences. Although these tests are not described in detail and are not, in some considerations, complete it is my belief that the families presented are correct. More importantly they are all readily available to the general public via the Pfam database for review and correction. This form of open peer review is probably the best way to ensure that models are as accurate as possible; indeed this open review allowed the realisation that two predicted PPC domains (see chapter 2.2.5) were false positives and they were removed from the alignment.

1.6 Tools

1.6.1 Search Software

HMMER (S. Eddy) and SAM (Hughey and Krogh, 1996)

Over the last decade the applications of Hidden Markov Models (HMMs) have proliferated in biological research. Uses include protein sequence comparison, splice-site prediction (i.e. Henderson, Salzberg *et al.*, 1997), transmembrane helix prediction (i.e. Krogh, Larsson *et al.*, 2001), signal peptide prediction (i.e. Nielsen and Krogh, 1998), and gene finding (i.e. Burge and Karlin, 1997; Meyer and Durbin, 2002). Their primary relevance to my work is that they underlie the search software I have mostly used - HMMER. HMMER also underpins the Pfam database (see 1.6.3) – around which much the work undertaken is based. In essence HMMER reads in a seed alignment and constructs a profile HMM. The architecture of the HMMER HMM, called 'Plan 7', has a core that consists of a node for each column of the alignment, each node consisting of three states - M, D, I (match, deletion, insert). The core is flanked by a B and an E (begin, end) state. The remaining five states control

algorithm-dependent features of the model, and can be varied to alter the type of model produced (see below).

The emission probabilities for the M state and the transition probabilities of the D state are generated from the multiple sequence alignment. In each column of the multiple sequence alignment the frequency of each amino acid is counted, and hence the emission probability of a particular amino acid appearing at each position can be derived. The transition probabilities of the insert states (I) are based on an internal evolutionary model. Since each node is considered separately, the probabilities assigned at node are independent of the other nodes, and hence higher order information can be lost. However, this seems to be not much of a problem in protein sequences as this type of approach has been successful.

By controlling the algorithm states HMMER can be used to construct two types of HMM – one is known global or 'ls' and the other is the local or 'fs' model'; both are local with respect to the protein sequence. The ls model will only find significant matches that extend over the whole model and will allow multiple non-overlapping hits per sequence. The fs model will report significant alignments that may not extend along the whole HMM, and also will allow multiple hits per sequence. This has an advantage over other methods in that the model itself encodes the fragment or global nature rather than using a different algorithm for searching the same model. One use is that specialised models can be built that capture detailed aspects of specific domains – e.g. a highly variable N-terminus but an absolute requirement for the C-

terminal 10 residues – and then searched against a sequence database using the same algorithm.

Searching the HMM returns a list of bit-scores for each sequence. From the bit score an E-value is calculated. This estimates the number of sequences one would expect to achieve at least that score that would exist by chance in the database or, the number of false positives. This is achieved by best fitting a histogram of scores generated from searching 5,000 random amino acid sequences which approximately reflect the composition and length of UniProt fitted to an extreme value distribution (EVD).

The mathematics that underlie the use of HMMs for sequence searching are well established and are described in detail by Durbin, Eddy *et al.* (1998) and so I do not propose to describe them in detail here. It is enough to know that they work and that the software has been rigorously constructed; HMMER is simple enough to use as a 'black-box' process.

HMMER is just one example of an HMM-based search package. Also popular is the SAM package created by Richard Hughey, Kevin Karplus and Anders Krogh. SAM also includes methods for secondary structure prediction and built-in iterative searching. Comparisons between HMMER and SAM show that at the near zero or zero error rate required for this project there is little difference in the performance of either package - the sequence composition of the seed alignment has a far greater effect on the sensitivity and specificity of the model - and that HMMER is also marginally faster on large sequence databases (Madera, Vogel *et al.*, 2004). The main

reasons for using HMMER were to allow easy interaction with the Pfam database and because it is well understood and supported within the lab.

BLAST/PSI-BLAST (Altschul, Madden *et al.*, 1997)

BLAST is a heuristic method for similarity searching that in essence simplifies the Smith-Waterman algorithm. It uses a significant amount of pre-processing and two key assumptions (listed below) so as to reduce the running time. The Smith-Waterman algorithm is derived from the Needleman-Wunsch algorithm for comparing two sequences. The key difference is Needleman-Wunsch compares the entire length of both strings - a global alignment – whereas Smith-Waterman can compare the sub-string of one sequence against any substring in another sequence – local alignment. The BLAST heuristic makes two assumptions:

(1) Most high-scoring local alignments contain one or more high scoring pairs of three letter substrings called 'words'. These locations can be quickly identified and used to grow a longer high-scoring alignment.

(2) Homologous proteins show extensive regions of similarity with no gaps in the sequence. This facilitates extending the words into local alignments.

BLAST is the most widely-used and possibly fundamentally important tool in bioinformatics. It has a very fast running time, which allows it to be used with genome sized datasets. For instance, searching a 65 letter query sequence against a protein database of 1,998,366 sequences (670,625,123 letters) using the NCBI default

gap penalties at the NCBI server, took less than 30 seconds. In contrast HMM-based methods are much slower and, without powerful compute farms, are impractical for large-scale analyses. It is also very adaptable and can be optimised for different types of search fairly easily. BLAST has been reviewed extensively and its uses well documented – for instance Korf, Yandell and co-authors' (2003) book "BLAST".

PSI-BLAST stands for Position Specific Iterated BLAST. It is a development of BLAST that has some similarities to HMMER and SAM in that it creates a profile of the family that it uses to search a sequence database. After starting with a standard BLAST search, the returned alignments are used to generate a Position Specific Score Matrix (PSSM) that is used to search again. This process can be repeated for a set number of rounds or until 'convergence' – when the searches identify the same number of sequences as in the previous round. It essentially uses the BLAST heuristic but is able to take a PSSM as input. It is not as sensitive as SAM or HMMER (Madera and Gough, 2002) and it deals with low complexity sequence less successfully – for a practical example see 2.3, the ALF repeat, and also noted by Chen (2003). On the upside it is much faster; this makes it an ideal tool for carrying out positive controls, or rapidly generating large numbers of seed alignments for refinement using HMMER.

1.6.2 Alignment Software

Dotter (Sonnhammer and Durbin, 1995)

Dotter is a tool for visualising protein to protein comparisons. It compares every amino acid in one sequence with every amino acid in a second. From this it produces

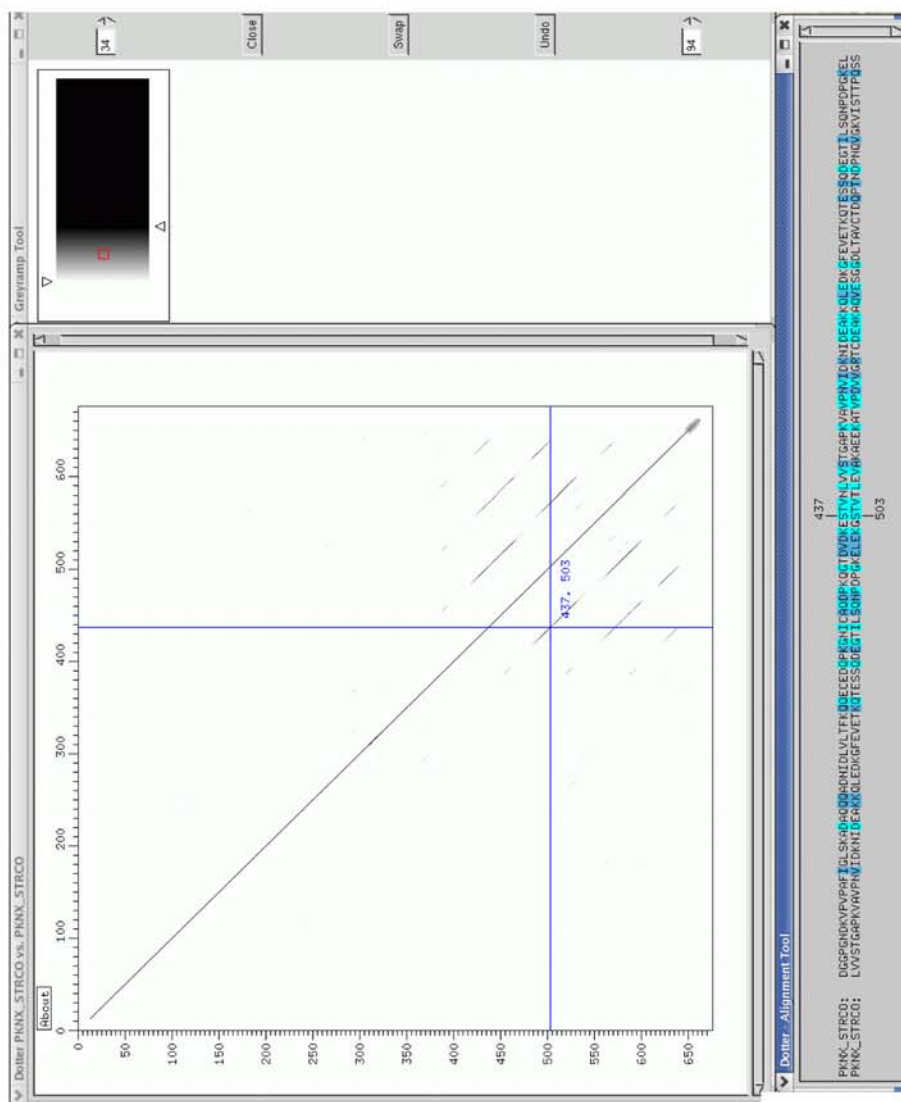


Figure 1.7: Example of the Dotter output
 Dotter is a useful tool for identifying repetitive regions within proteins. In this case a self-comparison of UniProt:Q9XA16 reveals four tandem repeats in the region 400-650; these repeats are four copies of the PASTA domain. The output consists of three parts - the dotplot, a grey ramp tool that alters the similarity threshold for displaying in the window, and an alignment viewer.

a dot plot with one sequence on the X-axis and the other on the Y-axis (see Figure 1.7 for an example). For easier visualisation the scores are averaged over a window that runs along the diagonal; work by E. Sonnhammer has found that 25 residues appears to be the most sensitive window size for identifying repeats and is used as default. This tool was used extensively during the work for this thesis, primarily for self-self comparisons, in order to identify novel repeated regions and to aid in interpretation of the results from Prospero (see below).

Prospero (<http://www.well.ox.ac.uk/rmott/ARIADNE/prospero.shtml>)

Prospero is part of the Ariadne software created by R. Mott. Prospero generates local alignments using the Smith-Waterman algorithm and then assigns accurate P-values (to within 5%, 95% of the time; Mott, 2000). The P-values are then multiplied by the database size, converting them into E-values. As discussed for HMMER, an E-value represents the expected number of false-positives occurring at that score in a database the size of the one searched. As implicated, the larger the database the greater the number of false-positives one would expect. Therefore, self-self comparison will return the lowest E-value for a particular score and will be more sensitive than searching against a sequence database. This principle underlies the approach used in many of the domain hunts undertaken. A second benefit of Prospero is that the output is easy to parse using computers compared to the graphical output of Dotter. This makes it very simple to carry out very large numbers of self-self comparisons and identify significant alignments, which can then be further processed and used to seed iterative profile-based searches (see chapter 2.1.3.2).

Multiple Sequence Alignment: - ClustalW, T-Coffee and MAFFT

Most of the sequence search software and processes described so far use or produce Multiple Sequence Alignments (MSAs). The sensitivity and specificity of HMMER can be significantly affected by the seed alignment from which it generates the HMM. Furthermore interpreting the patterns of similarity and identifying conserved residues is made much easier when the alignment is accurate. An accurate alignment has all structurally equivalent residues in the same column.

Given the size and number of alignments examined manual alignment is impractical, so three multiple sequence alignment programmes were used - ClustalW, T-Coffee and MAFFT. ClustalW is probably the oldest and most well known of the three (Thompson, Higgins *et al.*, 1994). It has the advantages of being fast and reasonably accurate. It is based on the progressive approach proposed by Hogeweg and Hesper (1984) and Feng and Doolittle (1987). To describe the process simply, pair-wise scores are determined for all the sequences by means of a substitution matrix, and are used to grow a Neighbour-Joining (N-J) tree. A series of pairwise alignments are carried out, starting with the most related sequences, then progressing to more distant sequences, and then aligning each of the sub-alignments so as to progressively build up an MSA. ClustalW includes some refinements to this process, which primarily focus on reducing errors in the pair-wise alignments. This type of algorithm is described as a greedy algorithm, and if an error is introduced early in the process its effects will be amplified and may disrupt the overall alignment. Also the global nature of ClustalW means that if one tries to align multidomain proteins that contain unrelated domains there can be deceptive misalignments.

T-Coffee is more recent, and uses a more complex alignment algorithm (Notredame, Holm *et al.*, 1998). Instead of using a substitution matrix, as used by ClustalW, it uses a PSSM, termed an "extended library", where the score for each pair of residues depends on their compatibility with the PSSM. The "primary library" is a collection of pairwise global alignments generated using ClustalW and local alignments generated by Lalign (Huang and Miller, 1991). The local alignments are used to create a consistency check, allowing the minimisation of potential errors during the build up of the progressive pairwise alignments. It is also possible to customise the extended library to improve its performance for specific families, or for ensuring that catalytic residues align. In comparison to ClustalW it performs better in general, though is much slower and impractical for alignments more than 200 sequences of length greater than 200 residues (personal observation).

MAFFT is the most recent of the three methods (Kato, Misawa *et al.*, 2002). Although the overall mechanism is similar to ClustalW it transforms the amino acid sequence into a sequence of polarity and volume values; these are aligned using a fast Fourier transformation and a novel scoring scheme. There are two implementations of MAFFT - a progressive method (FFT-NS-2) and an interactive refinement method (FFT-NS-i). I have exclusively used the FFT-NS-i implementation; it is much faster than the other tree programmes described, and also is as accurate.

Comparisons of the three methods have been carried out by various researchers. Presented below in Table 1.1 are the results of a recent test carried out by Edgar (2004), which was used for a comparison with his new sequence alignment

programme MUSCLE. The results from the test against BaliBASE (Thompson, Plewniak *et al.*, 1999) are presented below. Three other databases of alignments were also tested against, and similar results were found - PREFAB (Edgar 2004); SABmark (van Walle *et al.*, unpublished); and SMART (see below).

In practice all three alignment methods were used. MAFFT was typically used as the default; however, alignments were visually examined and if they did not appear satisfactory the other methods were tried. "Good" alignments are considered to have a minimal number of gaps - especially within secondary structural elements, and conserved motifs are immediately apparent. Bad alignments have unnecessary inserts, e.g. 'gappy', and do not line-up conserved motifs and secondary structural elements. For a trivial example of the difference see Figure 1.8.

<i>Method</i>	<i>Q</i>	<i>TC</i>	<i>CPU</i>
T-Coffee	0.882	0.731	1500
ClustalW	0.860	0.690	170
FFT-NS-i	0.844	0.646	16

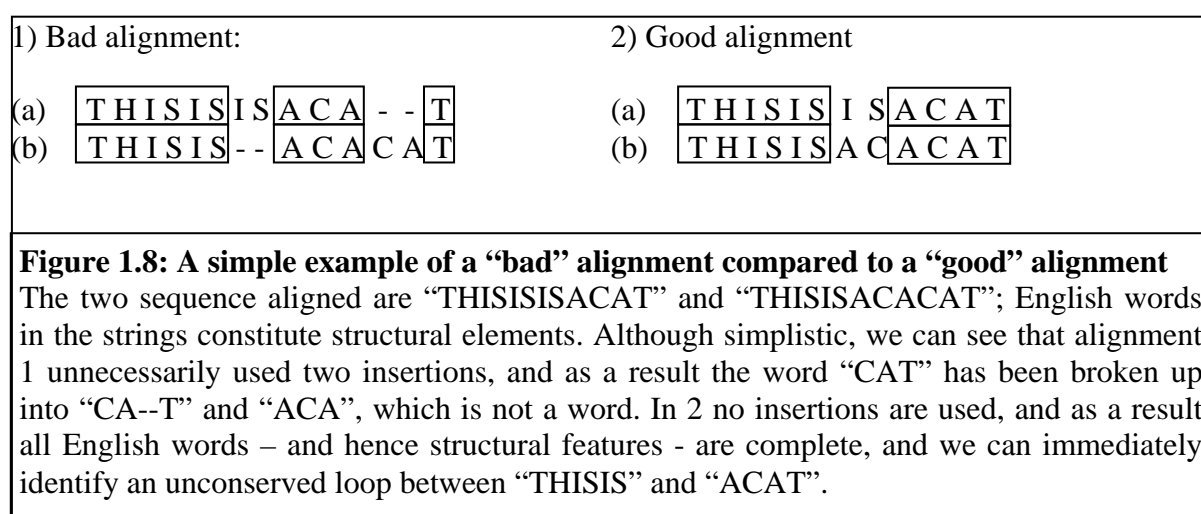
Q is the number of correctly aligned residue pairs divided by the number of residues pairs in the reference alignment.

TC is the number of correctly aligned columns divided by the number of columns in the reference alignment.

CPU is total CPU time in seconds.

Table 1.1: Results from Edgar's (2004) comparison of MAFFT, T-Coffee and ClustalW.

Whilst I generally found MAFFT and T-Coffee to be the most accurate, they do tend to push sequences to the ends of the alignment and leave gaps in the centre – MAFFT in particular. This is normally fine, but with short families composed of highly divergent sequences, some very poor alignments were produced (e.g. the FTP motif). I found that ClustalW performed the best with this type of family; T-Coffee was somewhere between the two. The accuracy of the alignment (with regards to the integrity of structural elements) does not overly affect the sensitivity of the HMM (Griffiths-Jones and Bateman, 2002) but it does make identification of conserved regions or residues harder and hence make analysis of the family more difficult. For a good review of the different methods of aligning multiple sequences see (Notredame, 2002).



1.6.3 Databases

Pfam (Bateman, Coin *et al.*, 2004)

Pfam is a two tier database for describing proteins. The aim of Pfam is to provide a comprehensive description of the domain content of the protein world, and to provide

tools for querying this data freely to the general research community. Pfam 13 (released April 2004) contains 7426 Pfam-A sequence families, which hit 74% of UniProt at least once (see Figure 1.9 for an example family page from the website). Pfam-A is a searchable database of manually curated sequence families. Each family consists of four primary elements:

- (1) A manually inspected SEED alignment of trusted sequences.
- (2) A global (ls) and a local (fs) HMM built from the SEED.
- (3) A description, including relevant literature.
- (4) An ALIGN file created by searching the HMMs against UniProt.


There are four family types in Pfam (see Figure 1.2 for examples). These fit into the definitions given for Domain, Repeat, Motif and Family in chapter 1.2. In Pfam 13 there are 5688 Families, 1464 Domains, 126 Repeats and 38 Motifs. Many of the families may actually represent domains, but a conservative judgement has been taken.

Pfam-B is an automatically generated supplement derived on ProDom (Servant, Bru *et al.*, 2002). ProDom is an automatically generated database of predicted domains - an outline of the method is provided in the description of ProDom below. ProDom regions that overlap Pfam-A domains are split or removed, depending on the type of overlap, hence creating an automatic description of homologies not detected by Pfam-A (the process is described by Bateman, Birney *et al.*, 2000). Pfam-B contains around 100 000 small families, which hit about 23% of UniProt.

Pfam Protein families database of alignments and HMMs

PASTA

Home Search by Browse by ftp iPfam Help



Accession number: PF03793

PASTA domain [Add Annotation](#)

This domain is found at the C termini of several Penicillin-binding proteins and bacterial serine/threonine kinases [1]. It binds the beta-lactam stem, which implicates it in sensing D-alanyl-D-alanine - the PBP transpeptidase substrate. It is a small globular fold consisting of 3 beta-sheets and an alpha-helix. The name PASTA is derived from PBP and Serine/Threonine kinase Associated domain.

NEW! This family forms **interactions** with other Pfam families, to view them click [here](#)

INTERPRO description (entry IPR005543)

The PASTA domain is found at the C-termini of several Penicillin-binding proteins (PBP) and bacterial serine/threonine kinases. It binds the β -lactam stem, which implicates it in sensing D-alanyl-D-alanine - the PBP transpeptidase substrate. In PknB of *Mycobacterium tuberculosis* ([SWISSPROT:P71584](#)), all of the extracellular portion is predicted to be made up of four PASTA domains, which strongly suggests that it is a signal-binding sensor domain. The domain has also been found in proteins involved in cell wall biosynthesis, where it is implicated in localizing the biosynthesis complex to unlinked peptidoglycan.

PASTA is a small globular fold consisting of 3 β -sheets and an α -helix, with a loop region of variable length between the first and second β -strands. The name PASTA is derived from PBP and Serine/Threonine kinase Associated domain [PUBMED:12217513](#).

QuickGO

FUNCTION : penicillin binding ([GO:0008658](#))

Figure 1: 1qmf
Peptidoglycan synthesis
 Penicillin-binding protein 2x (pbp-2x) acyl-enzyme complex

Key:

Domain	Chain	Start Residue	End Residue
Transpeptidase	A	289	609
PASTA	A	634	691
PASTA	A	692	750
PBP dimer	A	71	234

The Swissprot/PDB mapping was provided by [MSD](#)

1k25

Figure 1.9: Example Pfam Family Page - the PASTA Domain.

Each Pfam family has an automatically generated family page that displays a variety of information about the family. Some of this information is manually entered, while some is imported from other databases (i.e. InterPro), and some is calculated. The links to various tools make Pfam a useful workbench for domain family investigations. In this image the top half of the page is captured, showing annotation and structures. Below are links to graphical representations of the domain architectures, coloured alignments, HMM building information, other databases and cited articles.

InterPro (Mulder, Apweiler *et al.*, 2003)

InterPro is a front-end to a collection of databases. InterPro 7.2 (released March 2004) included Pfam (see above), SMART (see below), PROSITE (see below), PRINTS (Attwood, Bradley *et al.*, 2003), ProDom (see below), UniProt (see below), TIGRfam (see below), PIR superfamily (Huang and Miller, 1991), SUPERFAMILY (Madera, Vogel *et al.*, 2004), CATH (see below), SCOP (see below) and MSD (Golovin, Oldfield *et al.*, 2004). It provides facilities for both browsing the data and for searching sequences. The major benefit of InterPro is that it allows you to directly compare the predictions from different domain collections, and also compare these domains against a structural classification from SCOP (if available). Not all these databases were used in the work carried out, so a short description of the relevant ones is given in the section below.

SMART (Letunic, Copley *et al.*, 2004)

SMART is similar in form and function to Pfam (see 1.6.3) in its use of HMMs and in its construction of families – though it does not provide the full “ALIGN” files as constructed by Pfam. It is particularly focussed on modelling and describing domains found in signalling, extracellular and chromatin-associated proteins, whereas in other functional categories it is far less comprehensive. As of SMART 4.0 (released March 2004) it contained 667 domains.

PROSITE (Hulo, Sigrist *et al.*, 2004)

PROSITE is one of the original collections of sequence patterns (release 1 appeared in 1989). As of release 18.0 it contained “1,639 different patterns, rules or

profiles/matrices” and 1200 documentation entries. This diversity of model types reflects the history of sequence searching during the 1990s. Initially much sequence analysis was carried out using pattern matching techniques such as 'regular expressions'. These patterns tended to take the form “G-x(8,10)-[FYW]-x-G-[LIVM]-x-[LIVMFY]-x(4)-G-K-[NH]-x-G-[STAR]-x(2)-G-x(2)-[LY]-F” (in this case PS00845; CAP_GLY_1). However, profile methods subsequently have come to dominate sequence analysis due to their superior sensitivity, specificity and broader application; as a result PROSITE's earlier models are patterns and their later ones are generalised profiles (Bucher, Karplus *et al.*, 1996). PROSITE has detailed documentation for each of its families.

TIGRfam (Haft, Selengut *et al.*, 2003)

Release 3 (October 2003) had 1976 families, of which 1004 are "equivalogs", 330 are "other equivalogs" (proposed equivalogs for which the function is not known) and 642 are "other" (families for which it is not known if the function is conserved). Equivalogs are proposed to be families of functional equivalence. The difference in definition to an orthologue is worth noting: orthologues are homologous proteins that have separated due to a speciation event, but the function is not necessarily conserved; in contrast equivocals may be separated by any evolutionary process - such as lateral gene transfer, but the function is conserved. It is a rapidly growing resource – 350 new families were added between release 2.1 and release 3 (about 1 year). The families are more functionally specific than Pfam, allowing for greater confidence in the functional description that accompanies a match, but it is not yet as comprehensive.

ProDom (Servant, Bru *et al.*, 2002)

As mentioned above ProDom is an automatically generated domain database, from which Pfam-B is derived. Although automated methods are not as accurate, either in terms of defining the correct domain boundaries or in completeness of the families, ProDom does effectively capture genuine homologies and so can provide a useful starting point for a researcher looking for interesting sub-regions within a protein. The algorithm for its construction is also of interest, as the same principles are behind a method used in this thesis (see chapter 2.1.2). The assumption is made that the shortest amino acid sequence is representative of a domain. This sequence is then searched against UniProt (see below) using PSI-BLAST. Any matching regions and the query sequence are removed from the database and assigned a family number. This process is iterated using the shortest sequence remaining until no sequences with detectable homologies are left. Three filters are applied to the sequence database first; all sequences marked as 'fragment' are removed, low complexity regions are masked using 'seg' (Wootton and Federhen, 1993), and regions shorter than 20 amino acids are excluded.

Swiss-Prot/TrEMBL or UniProt or 'sptr' (Apweiler, Bairoch *et al.*, 2004)

The work in this thesis is mostly based on searching HMMs against a sequence database. The sequence database of choice was Swiss-Prot and its supplement TrEMBL. Founded in 1986, Swiss-Prot is a manually curated sequence database, with various functional and structural annotations attached. The increasing rate of DNA sequence production meant that a large volume of data was unavailable between releases, so an automated supplement was created – TrEMBL (Translated EMBL).

Figure 1.1 shows how UniProt has grown between October 2001 and July 2004. It should be noted that much of TrEMBL is redundant; new entries can often already be represented in Swiss-Prot or TrEMBL, or be a fragment of a larger protein. As of 2003 Swiss-Prot merged with the Protein Information Resource (PIR) to form the Universal Protein Knowledgebase (UniProt). This wasn't so much a combining of sequence data, but a merging of resources and infrastructure so as to produce a single high quality database that was able to keep up with the generation of sequence data. As can be seen from the slower growth of Swiss-Prot as compared to the near exponential growth of TrEMBL (see Figure 1.1) this was becoming a problem. As of May 2004 it still consisted of Swiss-Prot and TrEMBL; hence although the later work is done against UniProt rather than Swiss-Prot/TrEMBL, from the researcher's point of view they can be considered interchangeable.

1.6.4 Structural Collections and Classifications

wwPDB (Berman, Battistuz *et al.*, 2002)

The Worldwide Protein Data Bank (wwPDB) was established in 1971 as the Protein Data Bank (Bernstein, Koetzle *et al.*, 1977) to be “the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data” for the public. It is currently the product of the collaboration between the Japanese PDBj group, the European MSD group, and the American RCSB PDB group - hence "wwPDB" (Berman, Battistuz *et al.*, 2002). The other major structural databases – i.e. CATH, SCOP – are all built on top of it. A website provides querying services and an FTP site provides the underlying data freely for download. As of the 4th May 2004 it

contained 25,343 three dimensional structures, including 22,936 proteins, peptides and viruses and representing just under 4000 folds.

PDBSum (Laskowski, 2001)

PDBSum is a web-based interface to summary information contained within the PDB files and from structural analysis software, as well as linking to some relevant structural and sequence data in other databases. The information is presented in a pictorial manner, making it very easy to understand and interpret. It also shows the position of structural domains, as determined by CATH (see below) against the sequence allowing for easy cross-comparison with Pfam.

CATH (Pearl, Bennett *et al.*, 2003)

CATH is hierarchical system of protein structure classification based on a combination of automated approaches and manual validation. Proteins are split into domains and the structures characterised. The domains are then described in accordance with eight groups of criteria, which are:

Class – derived according to the secondary structure content: all α ; all β ; α/β ; and "few secondary structures". For example the PASTA domain is α/β (see Figure 4.1), whereas the Hemerythrin domain is all α (see Figure 2.9).

Architecture – describes the structure in terms of the orientation of secondary structure elements without reference to their connectivity.

Topology – determined by the order and type of secondary structure elements.

Homologous Superfamily – proteins that are thought to be evolutionarily related and hence homologous.

Sequence Family - groups of structures that show at least 35% sequence identity - as structure is highly conserved at this level.

Non-identical - groups structures that are at least 95% identical; useful for creating non-redundant datasets.

Identical - groups structures that are 100% identical in sequence terms.

Domain - the leaf of the CATH tree; this refers to structural domains as discussed in chapter 1.2.

SCOP (Murzin, Brenner *et al.*, 1995)

SCOP (Structural Classification Of Proteins) is another hierarchical system of protein structure classification that categorises domains in terms of their structural elements. The assignments are made based on a variety of evidence, including automated and manual interpretation of the data. The final assignments are determined by expert knowledge; and hence this system is probably the most accurate. There is some delay between a structure being deposited in the PDB and its classification in SCOP - e.g. as of July.9.2004 there were 25977 protein-containing PDB structures, and 20169 classified in SCOP. The classifications are:

Class – The same as CATH's 'Class' (see above), except that SCOP separates the α/β class into two types: α/β , in which the different types of secondary structure are mixed together in the fold; and $\alpha+\beta$, in which the

different types of secondary structure are largely segregated. Also SCOP has a "multidomain protein" class for proteins that consist of several different folds that have no obvious homologues, as well as a membrane protein class and a small protein class; it does not have the "few secondary structures" class.

Fold – groups of structures that have the same major secondary structure elements and topology (same as CATH's 'Topology' above) but show little or no overarching sequence similarity.

Superfamily – groups of structures that are likely to have evolved from a common ancestor, but have significantly diverged in sequence and function.

Family – groups of sequences that can be shown to have evolved from a single ancestor. This is defined by a sequence identity of greater than 30% or high structural and functional conservation.

National Center for Biotechnology Information (NCBI)

The NCBI website provides a simple front end to a range of bioinformatic tools and data resources (Wheeler, Church *et al.*, 2004). Of particular relevance is the PSI-BLAST server which searches the "nr" database - a mostly non-redundant composite peptide database made up from compilation of several resources. This provides an analogous system to the HMMER searching of UniProt used in this work, and so is a very useful positive control for the searches carried out. The NCBI also hosts a searchable biological/biomedical literature abstracts database (PubMed), a genetic disease mutations database (OMIM), authoritative taxonomy listings, a BLAST server for partially complete microbial genome sequencing projects and a range of other services.

1.6.5 Presenting Domain Architectures and Alignments

For all the novel families presented in this thesis, three pieces of information are supplied. These are an architecture figure, an alignment figure and a secondary structure prediction. These all conform to the same style discussed here - where there are specific variations these will be noted in the relevant figure caption. The domain architectures are presented in a 'Beads-on-a-String' style of representation. This view represents the protein sequence as a line with features depicted as coloured boxes. The features shown are Pfam-A families, signal peptides (SignalP; Bendtsen, Nielsen *et al.*, 2004), transmembrane helices (TMHMM; Krogh, Larsson *et al.*, 2001), low complexity regions (seg; Wootton and Federhen, 1993), and coiled-coils (ncoils; Lupas, Vandyke *et al.*, 1991). The key to the domain figures is shown below in Figure 1.10, along with a few example architectures. Unless indicated all the images are taken directly, and without alteration, from the Pfam website; this is to ensure that the data shown is publicly available, reviewed and consistent. In general most or all of the different architectures for a family will be shown.

Associated with each protein shown are its UniProt accession, its common name, and the species it is found in. It should also be noted that where possible all the proteins in a figure have been shown on the same scale. However, in some cases members of a domain family can diverge in length by an order of magnitude; in these cases scaled depiction is not realistic. To compensate the lengths are marked by each protein.

The alignments have been drawn in Jalview (Clamp, Cuff *et al.*, 2004), using the ClustalX (Thompson, Higgins *et al.*, 1994) colouring schema for different amino acid

groups (given below in Table 1.2). The sequences shown are essentially arbitrarily selected but have been picked in order to show the variety in the family as well as its typical form. The colours for each amino acid group are shown in Table 1.2, the colour being chosen according to the residue type and most conserved property in the column. Each sequence is shown with its UniProt accession number and the start/end coordinates of the domain. Another sequence alignment viewer I have commonly used is Belvu by Erik Sonnhammer; however, it does not include the ClustalX colouring scheme and so is not used to create the alignment figure images.

Residue Type	Frequency in Column	Colour	Description
ACFHILMVWY	>60%	Blue	Hydrophobic
DE	>50%	Magenta	Negatively Charged
KR	>60%	Red	Positively Charged
STQN	>50%	Green	Polar Charged
C	>85%	Pink	Cysteine
G	>85%	Orange	Glycine
P	>85%	Yellow	Proline
FYW	>50%	Cyan	Aromatic

Table 1.2: The ClustalX colouring scheme.

This scheme is the one used for the alignment figures shown in this Thesis unless otherwise indicated.

Under each sequence alignment is a secondary structure prediction, unless there is a known three dimensional structure. α -helices are indicated by red cylinders, whereas β -strands are indicated by yellow arrows. These predictions have been made using three programmes: JPred (Cuff and Barton, 2000), PHDsec (Rost, 1996) and PROF (also by B. Rost, but unpublished). Most of the older predictions have been made using JPred, whereas the more recent predictions are made using PROF and PHDsec. The reason for this change is more to do with the development of the servers supplying the service than improvements in accuracy. Whilst in the text for each family it may name either PROF or PHDsec, in reality both methods were run for

each family and it was checked that the results were largely in agreement. The exact output chosen for representation was dependant on how well it agreed with the shape of the alignment. If the two methods showed significant disagreement then the sample alignment was altered and further predictions run. In some cases a transmembrane helix prediction (blue box) takes the place of the secondary structure prediction. The predictions were made using TMHMM (Krogh, Larsson *et al.*, 2001).

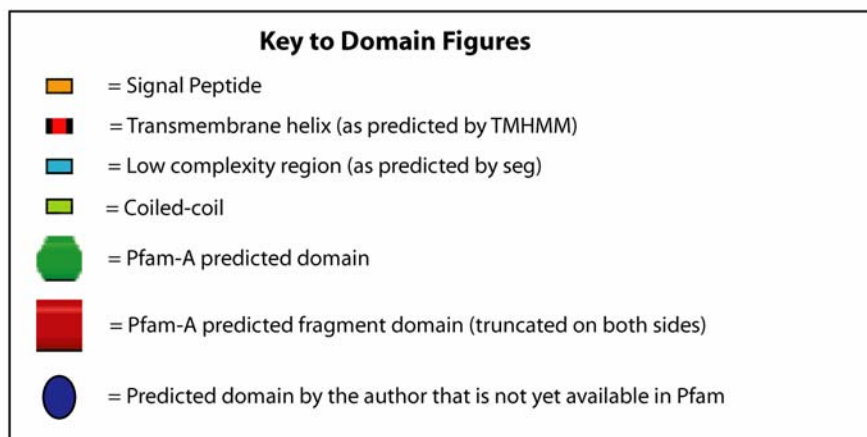
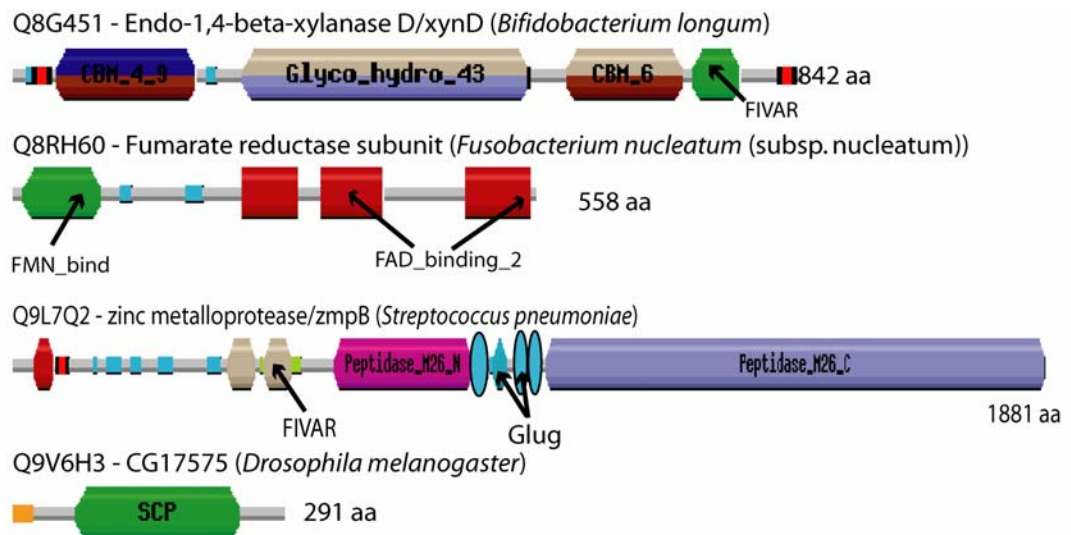


Figure 1.10: Key to the architecture figures and some example architectures
 Seven types of features are shown for each protein depicted in the architecture diagrams in this thesis. The methods for predicting each type of feature are noted in the main body of the text. Above each protein is its UniProt accession code, a common gene or protein name and the species in italics. Also near the C-terminus of each protein is its length in amino acids, so as to help relate the size of the proteins. Where possible they have been shown on the same scale.

2 Identifying Novel Domains

2.1 Domain Hunt Methods

2.1.1 Introduction

As discussed in chapter 1 the work in this thesis is mostly based around semi-automatic methods of domain detection. The idea behind these approaches is to take a large set of proteins and to generate a number of targets that potentially represent a domain or other interesting feature. Each of the targets is then analysed by hand and their validity assessed. The two methods I have most employed - small protein clustering and internal duplication identification - are described below. To a certain extent it is not the method that is the determinant of the success of a novel domain search - or "hunt" as they are also termed - but the starting dataset (Altschul, Boguski *et al.*, 1994). Most of the work in this thesis is based on using complete proteome data sets; this is to increase the chance that I will find domains that are of general biological relevance, rather than finding rare or uninteresting domains due to an unnatural bias in the starting data set. Sometimes very restricted sets are used - such as the *Chlamydomonas reinhardtii* polymorphic membrane protein family discussed in the chapter 5.3 - in order to identify domains involved in specific processes.

In general I have tended to tailor the parameters used in these methods so as to produce targets that have a high chance of being a domain, rather than producing large numbers of targets. This was done for the following reason. Since domain copy number in the tree of life follows a power law (Qian, Luscombe *et al.*, 2001) it can be assumed that most of these domains are of relatively low general interest. Also approximately 50% of the total sequenced amino acids do not yet belong to any

family – Pfam 14 cover 53.1% of all residues in UniProt 43.2/26.2. So by attempting to identify the most represented domains, there is a good chance that many high interest but novel domains should be found. General descriptions of the methods used are in chapter 2.1.2 and specific details of how they are applied are found the relevant sections.

As well as the primary high throughput techniques, two other techniques for working on small numbers of proteins are also presented.

2.1.2 Details of Methods

Small Protein Clustering (SPC)

This is a very simple method that can rapidly generate potential domain families. The main assumption made is that a protein of less than 100 amino acids is likely to be composed of a single structural domain. This assumption can be considered reasonable since domains are rarely less than 50 residues in length. A second assumption is if a small protein is important to universal cellular biology then it will be represented at least once in most genomes, but it may only be represented once in any particular genome. By investigating multiple genomes simultaneously these proteins should become easier to detect. Small protein clustering also drives the ProDom algorithm, the automated approach mentioned in chapter 1.6.3. I developed a four step process for identifying potential new families; the principles and details of this approach are given below.

Step 1: A set of proteins of less than 101 residues in length was assembled. An all-against-all BLAST was carried out and the proteins clustered using single-linkage

clustering according to a score threshold. A conservative clustering threshold was used so as to prevent the clustering of unrelated sequences. I determined the cut-off by trying a range of values and finding the region in which changing the threshold caused little variance in the composition of the clusters around this mark. Since the datasets that I used this method on never contained more than 6000 proteins, the separation of signal and noise was clear. Further confirmation was obtained from visual analysis of the alignments and from alignments found to be related to known Pfam-A families. The threshold was typically about 50-70 bits.

Step 2: All clusters that corresponded to Pfam-A families and singlet proteins with no homologues were now removed from the set. Comparing the excised cluster to the Pfam-A family also provides a useful check on the stringency of the clustering cut-off score. If the clustering scores were stringent enough there should be no sequences that the Pfam-A family does not identify. The clustered sequences were then aligned using T-Coffee or MAFFT.

At this stage it can become apparent that some of the proteins are significantly shorter than the rest. Predicting the start and ends of proteins purely from DNA sequence is still imprecise, and so can lead to the prediction of truncated proteins. However, this can be confirmed by initially discarding these sequences and then searching the final global HMM against the original DNA sequence from which this protein was predicted. If a significant match is found along the length of the HMM then it can be assumed that the protein was mispredicted and the amino or carboxyl terminus should be extended. If the match is still only partial then either the predicted

protein could be a pseudogene or the predicted domain is incorrect and should be truncated.

Step 3: The aligned clusters were then used as seeds for an automatic search using HMMER. At convergence the families were realigned with T-Coffee or MAFFT and a single round of searching carried out. If any new family members are identified then the iterative search process was repeated.

Step 4: The final stage consists of the manual analysis. This involves improving the MSA, which can make the searches more sensitive and makes it easier to identify important residues, and trying to detect remote homologues. Further analyses included structural prediction, literature searching, feature prediction (e.g. transmembrane regions, disulphide bridges), genome context investigation and phylogenetic tree building. The principle is to use the MSA to correlate as much information as possible together and interpret it. Most of these tools are discussed in chapter 1.

Repeat Identification (RI)

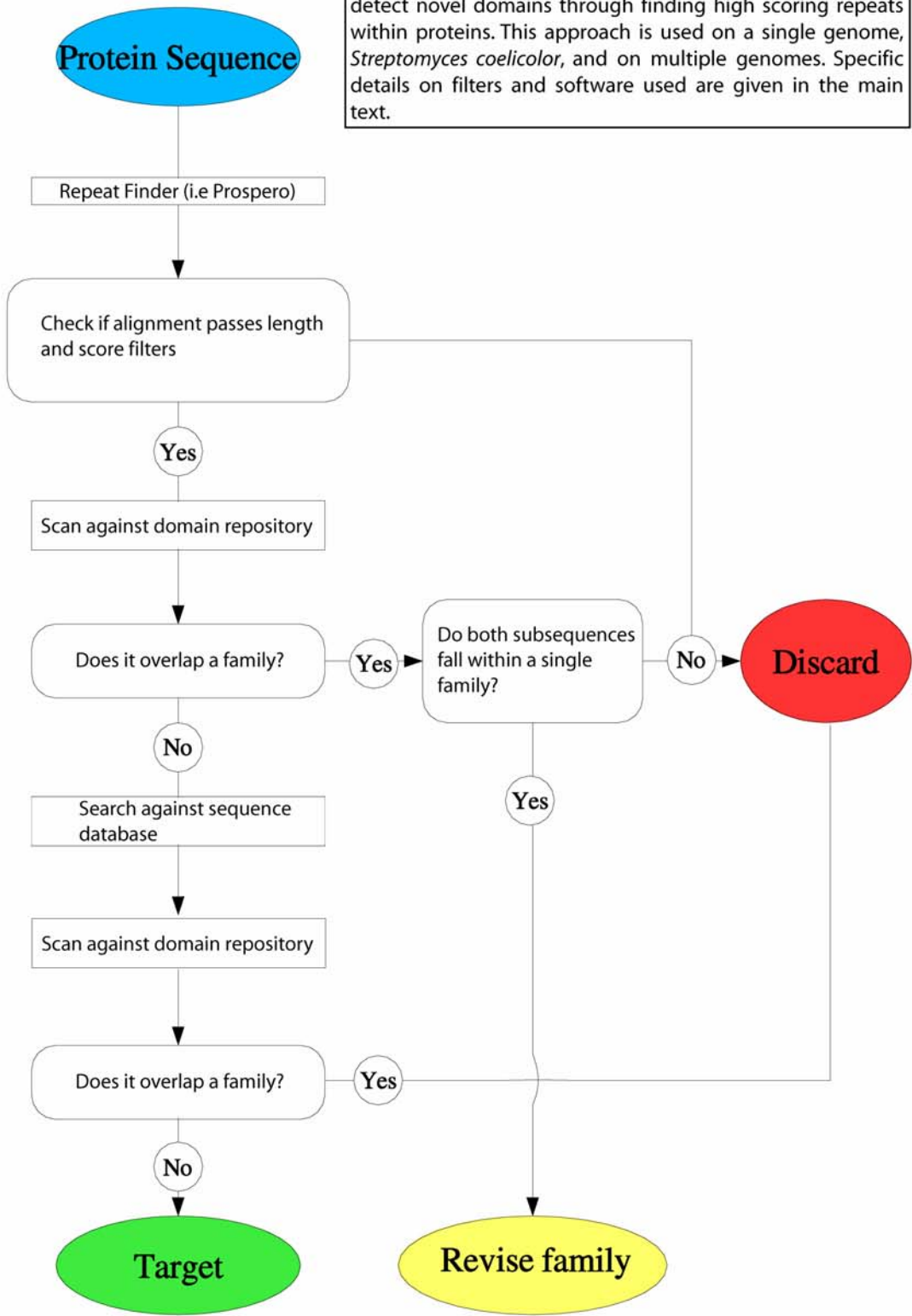
Repeat identification refers to the process of identifying domains through finding repeats within a protein, and is the method I've most used. In the past the discovery of internal repeats within a proteins sequence has led to the discovery of novel domains (e.g. Fong, Hurley *et al.*, 1986 and Haslam, Koide *et al.*, 1993). In 2001, Ponting, Mott *et al.* codified a procedure to take advantage of the apparent frequent occurrence of internal duplications in proteins, and successfully applied it to *Drosophila melanogaster*. A slightly modified version of this process is described below and depicted as a general method in Figure 2.1.

Its main advantage over other *ab initio* domain prediction methods is that it is very quick to do - for instance all the target generation searches for *Streptomyces coelicolor* (see chapter 2.2) were carried out within a day - and the targets produced have a high conversion rate into novel domains. Interestingly virtually no catalytic domains were detected using this method in this thesis, whereas many structural and substrate-binding domains were. Binding domains are frequently duplicated so as to increase substrate affinity; however, there are many instances of catalytic domains also being duplicated. Whether this bias in the results reflects that the majority of domains are not catalytic or that possibly the wide-spread catalytic domains have already been detected through laboratory-based experimental work, is not clear.

Step 1: A set of protein sequence data was assembled, such as the complete proteome of an organism. Low complexity regions were masked using 'seg' (Wootton and Federhen, 1993). Each protein was searched against itself using Prospero.

Step 2: Highest scoring matches were retained for each sequence and a series of filters applied to remove matches that were unlikely to be novel domains. Firstly, all matches which have an E-value greater than 0.001 were discarded. With genome sized datasets (1,000 – 40,000 proteins) this gives a very low chance of producing a false positive. For instance, if we expect one false-positive in a thousand Prospero predictions, then within the 124 targets that were generated (see chapter 2.2.3) we

Figure 2.1: General method for identifying novel domains
 The method shown here provides an outline on how to detect novel domains through finding high scoring repeats within proteins. This approach is used on a single genome, *Streptomyces coelicolor*, and on multiple genomes. Specific details on filters and software used are given in the main text.



would expect that there was an approximately 10% chance that one of the predictions was a false-positive.

Secondly, alignments with a length of less than 30 residues were removed. Such short duplications are unlikely to be genuine domains. Thirdly alignments where the start points of each subsequence are separated by less than 45 residues ('shift') were discarded. These are more likely to be structural repeats that are not stable in isolation (e.g. the β -propeller forming WD40 repeats, as discussed in chapter 1.3).

The fourth filter was scanning the potential targets against Pfam-A in order to determine if they were already part of a Pfam-A family. If an overlap is found the target is discarded – unless both subsequences fell within a single Pfam-A. This implies that the family represented more than one sequence domain or repeat and so needed rebuilding. An overlap is defined as there being a protein with residues that were found in both the Pfam-A family and the target alignment.

Step 3: The alignments generated by Prospero were used as an initial alignment to make profile-HMMs using the HMMER 2.2 software. If the pair of sequences in the Prospero alignment overlapped each other, these overlap regions were removed from the alignment. Profile HMMs are built in local (fs) and global (ls) mode. The resulting profile HMMs were scanned against UniProt and an alignment constructed from significant matches, using an inclusion threshold of 0.01. This alignment was then compared again to the Pfam-A database to see if the search had detected any similarities to known families. This step removes targets that are distant homologues

of previously described families. In some cases the missing members were subsequently added to the Pfam SEED alignments.

Step 4: The previous three steps help to narrow down the number of potential domains to analyse. The remaining targets were validated and investigated as described in Step 4 of the short protein clustering method (above).

Sequence Fragmentation

The principle here is to take a likely multidomain protein (e.g. longer than 200 amino acids) and to split it up into 50 or 100 amino acid blocks. Potentially one of these blocks may fall within the boundaries of a domain, and hence be able to identify homologues. Then the alignment can be extended at the amino and carboxyl termini so as to cover the whole domain. However, this approach is manually intensive and is very unreliable for a range of reasons – i.e. the domain may only have two highly similar regions that are spaced more than 100 residues apart.

“Blocky Alignments”

When proteins with similar domain architectures, but with one or two inserted or deleted domains, are aligned the alignments can take on a blocky appearance. This is because the shared domains are aligned together, but large inserts are required between them or at the amine and carboxyl termini of the protein to achieve this. So seeing a blocky alignment can provide the viewer with a good clue that it contains multiple domains. This approach was successful in determining the correct edges of the peptidase unit of the type IV signal peptidase, Peptidase_A24 (see chapter 4.4).

There have been attempts to automate the identification of blocky alignments and hence then derive domains, but they have been of limited success when compared to the accuracy of manually identified domain boundaries. For instance this is the basis for Domination (George and Heringa, 2002).

2.2 Domain Hunting in *Streptomyces coelicolor*

2.2.1 Introduction to *Streptomyces coelicolor* – a Complex Prokaryote

Streptomyces coelicolor is a representative of a group of high G+C (72.1%) Gram-positive bacteria whose successful adaptation is demonstrated by their almost ubiquitous presence in soil (Hodgson, 2000). This is largely accounted for by their broad metabolic capacity allowing them to cope with the many variables in their environment. They are able to utilise a wide range of food sources including the debris from plants, insects and fungi. Streptomycetes are also famed for their production of a range of secondary metabolites including antibiotics and other chemotherapeutic compounds. Unusually for bacteria, streptomycetes exhibit complex multicellular development, with branching, filamentous mycelia giving rise to aerial hyphae which in turn bear long chains of reproductive spores. These three developmental stages also display differential 'tissue-specific' gene expression (Hopwood, 1988).

Also unusual is the size and structure of streptomycete chromosomes. *Streptomyces coelicolor* has a linear chromosome, which at 8,667,507 base pairs was the largest complete bacterial genome sequence available in 2002 (Bentley, Chater *et al.*, 2002). At each end of the chromosome there are telomeric-like structures that contain repetitive DNA, including several palindromic sequences that may form stable

secondary structures (Huang, Lin *et al.*, 1998); they nearly identical to each other and are known as the Terminal Inverse Repeats (TIRs). Unusually the streptomycete plasmid SCP1 is also linear and has similar, though not identical, repetitive telomeric structures; the smaller SCP2 plasmid is circular. The genome is predicted to encode 7825 proteins – around twice as many as most sequenced bacterial genomes, more than the eukaryote *Saccharomyces cerevisiae*, and still the largest sequenced eubacterial proteome. This plethora of proteins reflects both a multiplicity of novel protein families and an expansion within known families when compared to other bacteria and thus is a good resource in the search for novel protein domains

Thus *S. coelicolor* provides a good proof of principal test-bed for domain hunting as an investigative tool. The rich variety of domains and metabolic paths encoded increases the probability that novel domains will be identified and that novel systems will also be delineated. The complete sequence also allows the domains to be investigated in the genome context, which can provide functional insights through identifying the function of proteins in the same operon or close proximity. Its acquisition of genes from a wide variety of sources also may increase the probability that identified domains will be found in other organisms. As an example it contains a type of collagenase (Peptidase_M9) that is only found in small group of mammalian pathogens in the Proteobacter and in the Firmicutes – both groups being unrelated to the Streptomycetes.

2.2.2 Methods

Both the RI and the SPC methods (see Section 2.1.2) were used for investigating the *S. coelicolor* proteome. The specific thresholds used and the results of each are presented below.

<i>Proteome Size</i>	<i>Short Proteins</i>	<i>UniProt Release</i>	<i>Pfam Release</i>	<i>Date</i>
7846	597	40/18	7.4	Dec 2001- March 2002

Repeat Identification

As discussed in chapter 2.1.3.2 identification of repeated sequence within a protein is a powerful and sensitive method of identifying novel domains, and has been previously successful. The method was applied to all 7846 proteins and the resulting targets manually investigated.

Short Protein Clustering

The short protein method was also applied to *S. coelicolor* in order to determine if the assumption that important small proteins may be represented multiple times was valid. The four step process described in 2.1.3.3. was applied to 597 proteins with a length of less than 101 amino acids. A BLAST clustering threshold of 50 bits was used.

2.2.3 Summary of Results

Repeat Identification:

From an initial set of 124 possible domain targets 31 novel domains were identified, giving a 25% success rate. Sixteen targets were removed due to overlaps with Pfam-A

families. Of the targets that lay within Pfam families, most related to the same set of overlapping families: Patched (PF02460), SecD_SecF (PF02355), and MMPL (PF03176). These targets probably identify a highly divergent transmembrane domain that occurs in pairs, and is found within these families. Table 2.1 lists and briefly describes all novel domains identified in the domain hunt processes. There were also significant extensions to two Pfam-A families – the SCP domain and FG-GAP repeats.

Small Protein Clustering

From an initial set of 597 short proteins 35 clusters were derived, accounting for a total of 102 proteins. There were 26 size two (two proteins) clusters, 4 size three clusters, 2 size five's, a size six, a size seven, and a size 15 cluster. All the clusters above size three were part of Pfam-A families - DUF397 (PF04149), CSD (PF00313), Whib (PF02467) and DUF320 (PF03777). DUF397 accounted for the size fifteen and the size six clusters. DUF320 was found by both hunt processes. As a positive control the iterative search steps were carried out on the annotated clusters. These were all simple to develop in to good approximations of the Pfam-A families. When the remaining clusters were iteratively searched only one family significantly extended — the MbtH family (see below). Three small families of less than 10 sequences – GvpG (PF05120), GvpK (PF05121) and spdb (PF05122) – were also produced.

<i>Pfam</i> Accession No	Family Name	<i>Pfam</i> Type	Basic Function	No of copies in <i>S. coelicolor</i>	Antibiotic biosynthesis	Cell Wall Biosynth	Cell Wall/ Periplasm	Replication	Secreted
A) Novel Families									
PF03457	HA	Domain	Putative RNA binding domain	21				X	
PF03621	MbBH	Domain	Possibly involved in antibiotic biosynthesis	2	X				
PF03625	DUF302	Domain	Unknown function	3			X		X
PF03640	Lipoprotein 15	Repeat	Unknown function	6			X		X
PF03703	DUF304	Domain	Unknown function	4			X		X
PF03704	BTAD	Family	Bacterial transcriptional activator domain	12	X				
PF03710	GlnE	Domain	Glutamate-ammomia ligase adenylyltransferase	2					
PF03713	DUF305	Domain	Unknown function	6			X		X
PF03714	PUD	Domain	Putative carbohydrate binding domain	2			X		X
PF03724	DUF306	Domain	Unknown function	2			X		X
PF03729	DUF308	Repeat	Unknown function	6			X		X
PF03733	DUF307	Domain	Unknown function	2			X		X
PF03752	ALF	Repeat	Putative signal transduction domains	16					X
PF03756	AlfA repeat	Repeat	A-factor biosynthesis	2	X				
PF03771	SPDB	Domain	(Probably) mobile element replication	16					
PF03777	DUF320	Domain	Unknown function	11			X		X
PF03779	SPW	Repeat	Unknown function	2			X		X
PF03793	PASTA	Domain	Cell wall peptidoglycan sensor domain	9		X	X	X	X
PF03794	HHE	Domain	Unknown function	7		X			
PF03795	YCII	Domain	Probably enzymatic domain	3					
PF03860	DUF326	Domain	Unknown function	6			X		
PF03984	DUF346	Repeat	Unknown function (β -propeller)	7			X		X
PF03988	DUF347	Repeat	Unknown function	4			X		
PF03990	DUF348	Domain	Unknown function	3			X		X
PF03992	ABM	Domain	Antibiotic biosynthesis monooxygenase	3	X				
PF03993	DUF349	Domain	Unknown function	3					
PF03994	DUF350	Domain	Unknown function	2			X		X
PF03995	DUF351	Domain	Unknown function	4					X
PF04151	PPC	Domain	PKD-like peptidase C-terminal domain	3					X
PF04205	FMN_bind	Domain	FMN-binding domain	2			X		X
PF05120	GvpG	Domain	Gas vesicle protein G	2					X
PF05121	GvpK	Domain	Gas vesicle protein K	2					
PF05122	SpdB	Domain	Mobile element transfer proteins	2					
B) Previously Described New Pfam Families									
PF03458	UPF0126	Domain	Unknown function	4			X		X
PF03459	TOBE	Domain	Transport-associated OB fold domain	9			X		
PF03707	MHYT	Repeat	Putative ligand receptor	6			X		X
PF03989	DNA_gvraseA_C	Repeat	DNA-binding β -propeller	8				X	
C) Significantly Extended Families									
PF00188	SCP	Domain	Unknown function	4			X		X
PF01839	EG-GAP	Repeat	Putative β -propeller	57			X		X

Table 2.1: Novel domains found in *Streptomyces coelicolor*

2.2.4 Notes on Table of All Novel Domains Identified

Table 2.1 lists all the domains identified in this project. Part A shows entirely novel families. Part B shows families not in Pfam, but described elsewhere. References are: UPF0126 – Swiss-Prot; TOBE (Koonin, Wolf *et al.* 2000); MHYT (Galperin, Gaidenko *et al.*, 2001); DNA_gyrase_C (Qi, Pei *et al.*, 2002). Part C lists significantly extended families. Domains highlighted in blue are discussed below. Some basic functional information, whether it cell wall associated for instance, is provided for each family

2.3 Descriptions of Novel Domain

HA (Helicase Associated domain; PF03457)

See Figure 2.2 for an example alignment and architectures. The domain is typically seventy residues in length and is predicted by JPred to have an α -helix fold. It appears to mostly only be found in the streptomycetes, though an HA-containing helicase is found in *Chlamydia muridarum*, and a protein consisting of three copies of the domain (UniProt:Q98RX4) is found in the eukaryotic algae *Guillardia theta* and *Giffithsia japonica*. Investigation into the *C. muridarum* genome identified an extensive region of laterally transferred genes (LTGs) - the "plasticity zone" - and also three genes outside of this region were determined to be LTGs on the basis of comparison with *Chlamydophila pneumoniae* (Read, Brunham *et al.*, 2000). The HA helicase was one of these three LTGs; whether it is functional or expressed is not known.

Examination of the position of the HA domain-containing proteins, using Artemis (Rutherford, Parkhill *et al.*, 2000), on the *Streptomyces coelicolor* genome gives some

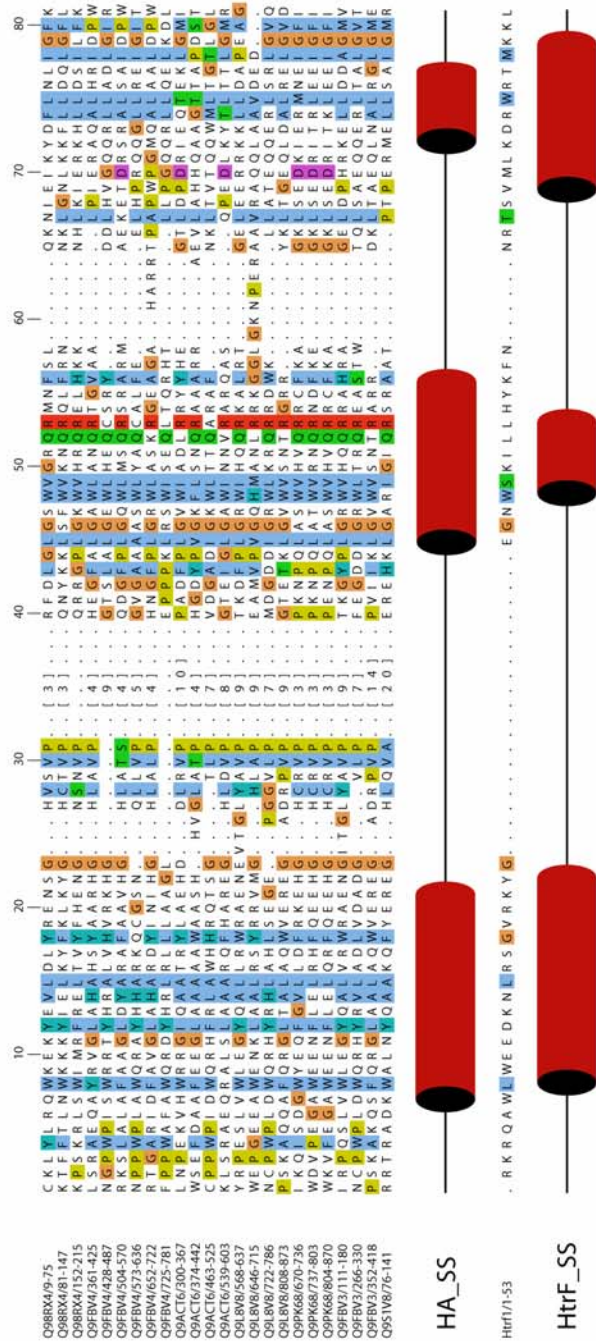


Figure 2.2: HA example alignment
 As well as an example alignment of HA domains the DNA-binding domain of HtrF is shown below, along with its known secondary structure (PDB:1BA5). This sequence was aligned to the HA domains with T-Coffee. As can be seen there are some intriguing similarities between the two.

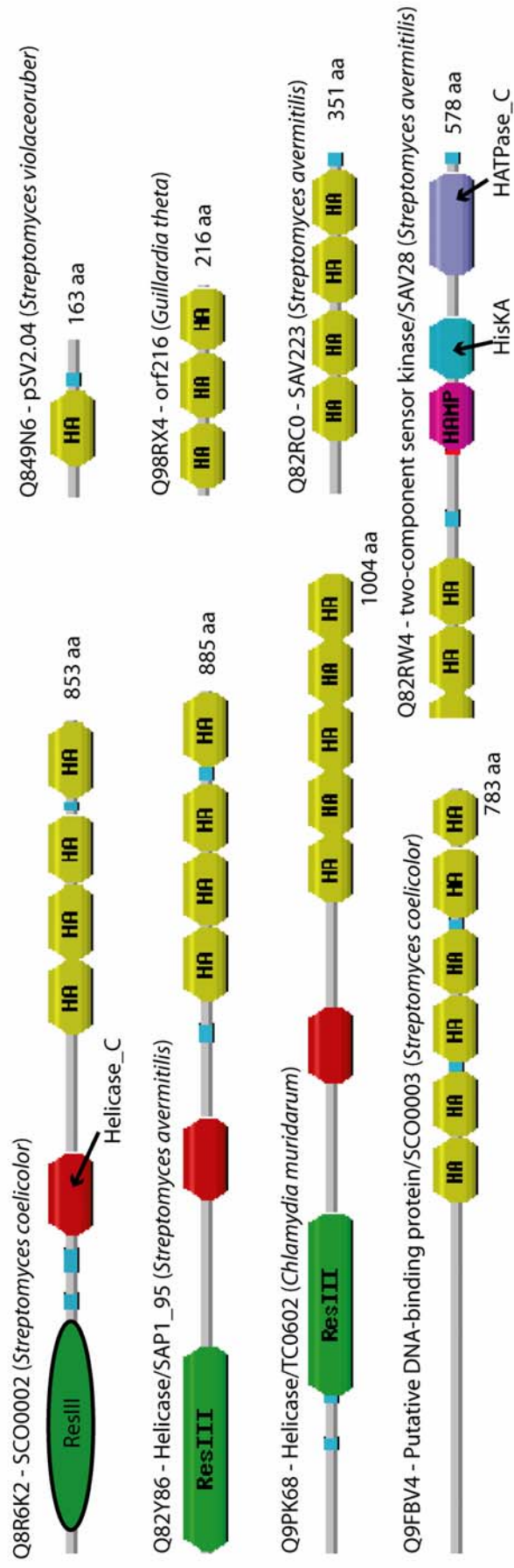


Figure 2.3: HA example domain architectures

suggestion of the HA-protein's function. The second and third ORFs from each end of the chromosome lie between 1.2 Kb and 6.2 Kb from the ends - well inside the 22 Kb TIRs. The second gene from each end is identical to the other (SCO0002 and SCO7845; UniProt:Q8R6K2) as are the HA-containing genes third from each end (SCO003 and SCO7844; UniProt:Q9FBV4). SCO0002 and SCO7845 have an N-terminal helicase (ResIII) domain, a central Helicase_C domain and 4 C-terminal HA repeats. SCO003 and SCO7844 have 6 C-terminal HA repeats and N-terminal region of unknown function, though it may contain a helix-turn-helix DNA-binding motif (score = 3.12, 50% probability as predicted at http://npsapbil.ibcp.fr/cgi-bin/primanal_hth.pl). One more gene encoding a single HA domain, SCO0034 (UniProt:Q9S1V8), is found at one end of the core region, about 7 Kb upstream from the nearby TIR. The origin of replication is centrally located on the chromosome, so this would make it one of the last genes duplicated during replication.

Specific complexes are required for maintaining the ends of the linear streptomycete chromosomes (Hinnebusch and Tilly, 1993), and the appearance of the genes encoding these domains in the TIRs suggests that the proteins may be involved in forming these complexes. This is further evidenced by the observation (Bey, Tsou *et al.*, 2000) that similar helicases appeared at the end of several of the streptomycete chromosomes investigated as well as the linear plasmids. A knockout mutation experiment they carried out was inconclusive; chromosome linearity was maintained, but the region of protein substituted did not include the ResIII domain or two of the HA domains, so it is possible that the helicases still retained enough functionality. This may be an example where an experiment has failed to knock out all of a protein's

function due to not considering the domain structure, especially if the core function of these proteins resides in the HA domains.

If this domain is involved in maintaining the linear TIRs then we would also expect to find a HA-containing helicase on the streptomycete linear plasmid SCP1, as plasmids contain all the proteins necessary for their reproduction. In fact there appears to be two HA-containing helicases on the SCP1 plasmid; however, only one is complete – SCP1.216 – whereas SCP1.136 is missing the N-terminal ResIII domain. It is possible SCP1.136 does not encode a functional protein. In contrast the circular SCP2 does not encode an HA helicase.

There are no clear conserved catalytic residues in the alignment, such as the polar residues (R, N, D, C, E, Q, H, K, S, T), suggesting that these domains have a binding function. The secondary structure prediction of the HA domain as a three-helical bundle is also suggestive of the Myb-like domain – a general DNA-binding domain. Aligning the sequence of the DNA-binding domain of Htrf1 (UniProt:P54274; human telomeric protein) against the HA domain alignment with T-Coffee showed interesting similarities between them (see Figure 2.2).

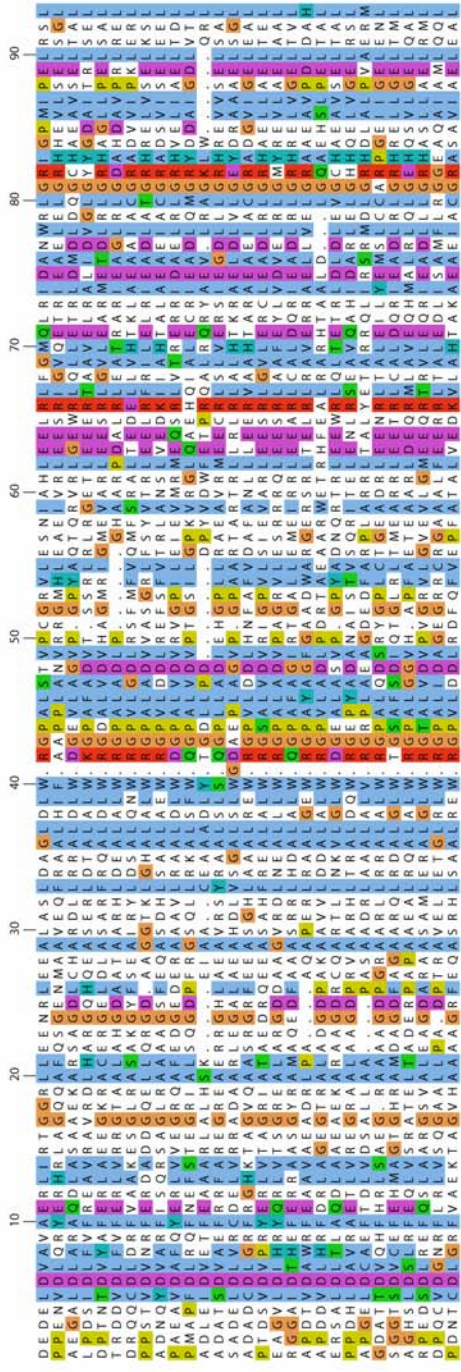
One of the three key tryptophan residues in Myb-like DNA binding domain aligns to a tryptophan residue in HA, another lies adjacent to a tryptophan, and the third aligns with a structurally similar leucine. The first helix appears to align well, but the second is longer in HA and the third is shorter. As to whether there is a true evolutionary or functional relationship between the HA domain and the Myb-like domain, the evidence is not conclusive but the number of similarities is at least striking.

Eukaryotic and Streptomyces telomeres are significantly different in structure, but the Myb-like domain may provide a plausible structure model for determining if and how the HA domains interact with DNA.

HA domains are also found at the N-terminus a two-component regulatory histidine kinase in *Streptomyces avermilitis*. From the organisation of the domains it would seem that HA domains fulfil the role of the sensor (see Figure 2.3); this fits with the prediction from the conservation pattern that HA is a binding domain. It is hence probable that this protein is involved in the maintenance or biogenesis of the telomeres; however, *S. coelicolor* does not have this regulator.

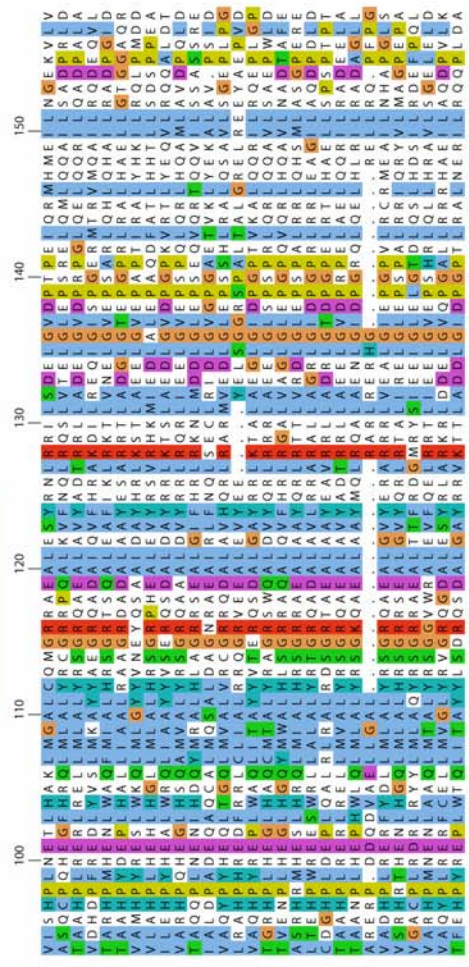
BTAD (Bacterial transcriptional activator domain; PF03704)

This domain was not directly derived from the initial target. Although a repeat was detected (residues 790-896:870-975) with an E-value of 4.73×10^{-4} using Prospero on the masked sequence of SCO4426 (UniProt:P25941), the validity of the repeat could not be verified by other means. However, I noticed that an undescribed amino terminal region (residues: 119-263) was related to a number of other bacterial proteins and investigated further; see Figures 2.4 and 2.5 for alignment and architectures. This region had been briefly mentioned as an uncharacterized domain (Aravind, Dixit *et al.*, 1999). In fact, subsequent work by David Studholme (personal communication) has shown that the C-terminus of this protein is made of highly divergent TPR repeats, and also that the BTAD region may be as well, but he was unable to confirm this hypothesis.



- AC24_STRCO/09-255
- Q9HLL3/98-254
- P970C0/95-250
- DNRL_STRPE/101-257
- Q9KTY4/103-253
- Q05797/97-253
- O54494/121-276
- O53145/101-257
- Q9KWX4/101-257
- O68896/98-254
- Q98H59/97-231
- O68913/102-258
- Q9L096/914-1065
- P71486/101-257
- Q9L096/914-1065
- Q9L545/101-257
- Q9L8V5/102-258
- Q98D13/100-252
- Q98K99/103-256
- O59076/115-270
- Q9X501/112-268
- Q98883/112-235
- Q9KCC3/192-347
- Q9KCC4/105-261
- Q92389/122-277
- Q92A48/114-270
- REDD_STRCO/175-330
- YC07_JMYCTU/106-262

BTAD_SS



- AC24_STRCO/09-255
- Q9HLL3/98-254
- AF58_STRCO/115-270
- P97060/95-250
- DNRL_STRPE/101-257
- Q9KTY4/103-253
- O05797/97-253
- O54494/121-276
- O53145/101-257
- Q9KWX4/101-257
- Q68896/98-254
- Q98H59/97-231
- Q9L096/914-1065
- Q9L545/101-257
- P71486/101-257
- Q9L096/914-1065
- Q9L545/101-257
- Q9L8V5/102-258
- Q98D13/100-252
- Q98K99/103-256
- O59076/115-270
- Q9X501/112-268
- Q98883/112-235
- Q9KCC3/192-347
- Q9KCC4/105-261
- Q92389/122-277
- Q92A48/114-270
- REDD_STRCO/175-330
- YC07_JMYCTU/106-262

BTAD_SS

Figure 2-4: BTAD example alignment

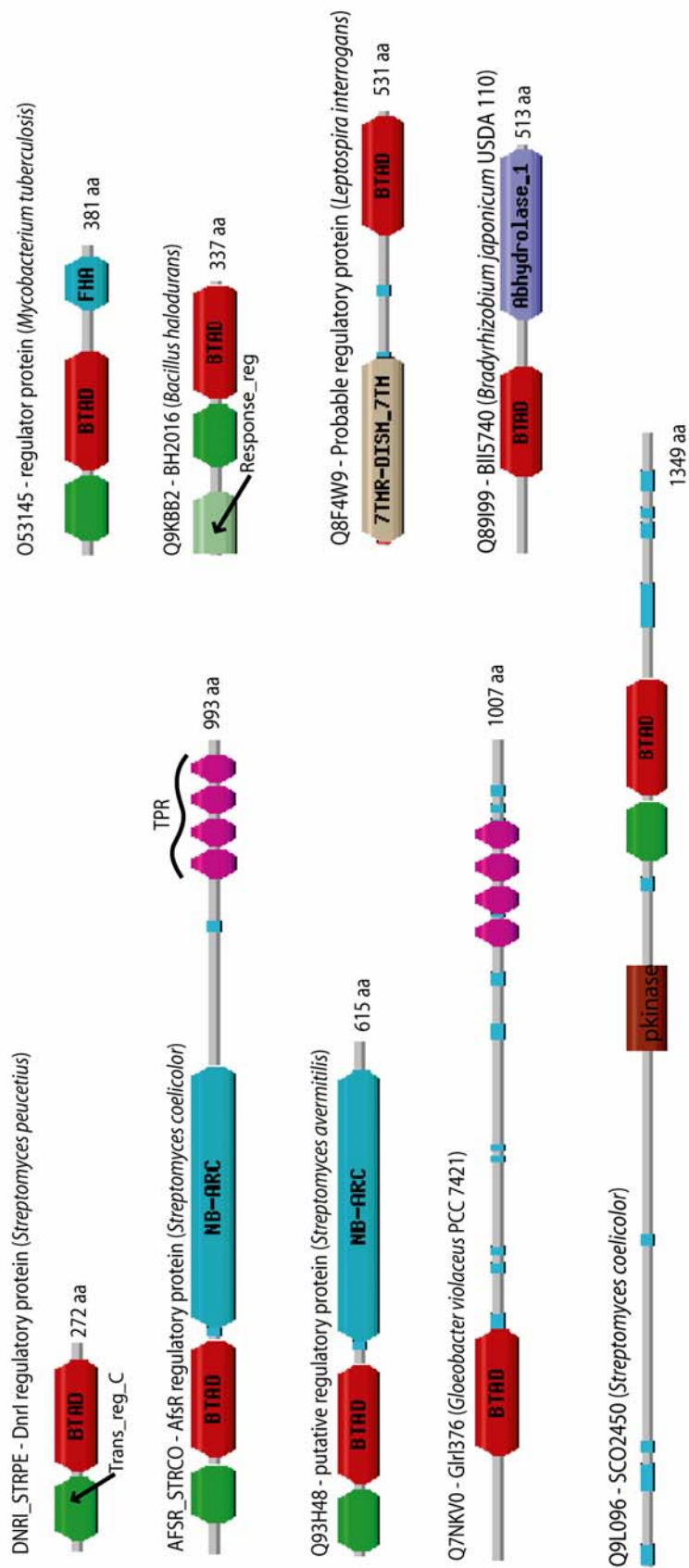


Figure 2.5: BTAD example architectures

The BTAD domain is disparately distributed across bacteria, though wide-spread. One of the proteins it is found in – AfsR – is a global secondary metabolite regulator of *S. coelicolor* (Floriano and Bibb, 1996). This protein has two basic functions – binding DNA and recruiting RNA polymerase. The first of these is carried out by the OmpR-like DNA-binding domain (Trans_reg_C; PF00486), whereas the second is carried out by the region C-terminal to the BTAD domain. This region includes the ATP-binding NB-ARC domain (PF00931) and three TPR repeats (PF00515). AfsR's DNA-binding activity is modulated by serine/threonine phosphorylation (Umeyama, Lee *et al.*, 2002); of note, there are no conserved serines or threonines in the BTAD domain so the phosphorylated residues probably occur elsewhere in the protein.

A mutation analysis by (Sheldon, Busarow *et al.*, 2002) on DnrI of *Streptomyces peucetius* suggests that the BTAD domain is essential to its function. A possible explanation is that it mediates oligomerisation with other transcription complex proteins, or even that it mediates interactions between DnrI monomers binding tandem repeats in a promoter region. There are eleven pathway-specific regulatory proteins in *S. coelicolor* that contain this domain, including a DnrI homologue and RedD, five of which are found in antibiotic synthesis clusters. It is possible that the BTAD domain mediates interactions between the global regulator AfsR and the downstream pathway-specific regulators.

ALF (Adenine-Leucine-rich conserved (F)phenylalanine; PF03752)

This family occurs as two sets of four forty-five residue tandem repeats in three *S. coelicolor* proteins and as three tandem repeats in an *S. avermilitis* secreted protein. The repeats have a predicted secondary structure of three α -helices (See Figure 2.6).

When the work on this domain was originally carried out (February 2002) these proteins were all described being involved in chemotaxis sensory transduction in UniProt; however this annotation was incorrect and probably came about for the following reason. To the C-terminus of each set of repeats is a low complexity and coiled-coil region. For all three proteins InterProScan found a chemotaxis sensory transducer region (IPR:004089; PS50111) between the two ALF-repeat regions. In contrast, searching these regions with HMMER 2.2g against SWISS-PROT and TrEMBL found no significant similarity to other chemotaxis proteins; similarly using PSI-BLAST at the NCBI found several false-positives – proteins that were unrelated to each other – but no chemotaxis signal transduction proteins. The sequence in this stretch is very alanine rich, and so could lead to significantly high-scoring matches on the basis of the apparent conservation of the alanines despite a lack of conservation in other positions. So it seems likely that the apparent homology is incorrect. This result is no longer reported by InterPro, but the example does illustrate the dangers of naively trusting automatically assigned annotation. One of the proteins, SCP1.201 (UniProt:Q9ACV2), also contained a Hint domain (N-terminus: SM00306, IPR003587; C-terminus: PS50818, IPR002203) at its C-terminus, which is the first identified in *S. coelicolor* (discussed more below).

In bacteria, genes with related functions – i.e. part of the same metabolic pathway or signalling pathway – are typically found to be near each other in the genome and the genomic neighbourhood of the ALF proteins does give some clues as to their possible function (see Figure 2.6 for a depiction). Two of the proteins, SCO6198 (UniProt:Q9Z5A4) and SCO6593 (UniProt:O87848), are located on the chromosome

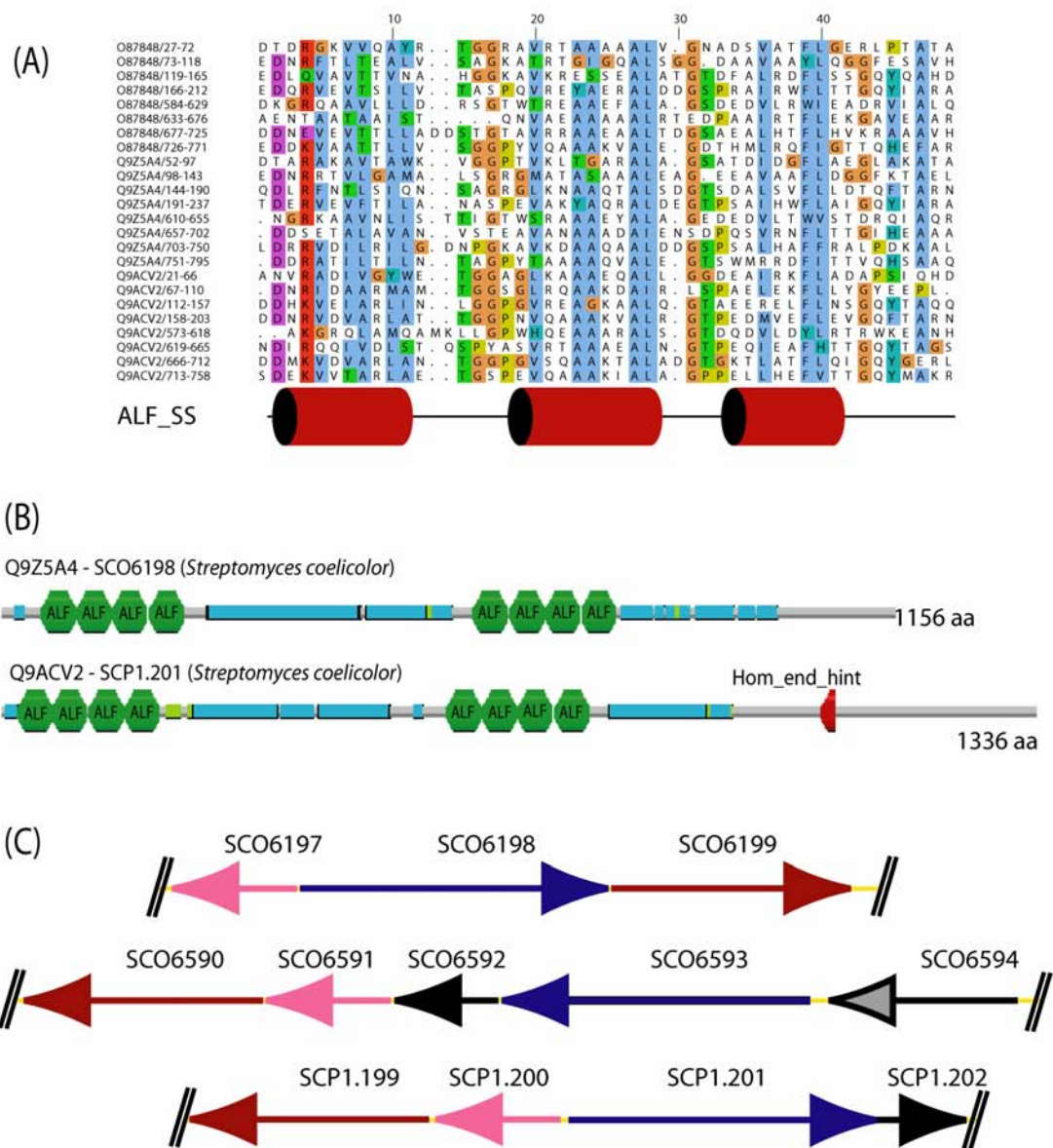


Figure 2.6: ALF alignment, architectures, and genome context
 Parts A and B depict an example alignment and architectures for the ALF repeat as standard. Part C shows the ALF-containing proteins of *Streptomyces coelicolor*, along with their immediate gene neighbourhood. The arrows indicate the direction of transcription. Homology between the proteins is indicated by colour; black indicates no homologues were found in *S. coelicolor*; black and grey indicates that homology was found to a protein in *S. coelicolor* but not from within these three regions. As can be seen, although direction and order are not conserved, the local gene neighbourhood consists of related proteins in all three cases.

adjacent or close to secreted esterases (SCO6199 and SCO6590) and several other probable secreted proteins of unknown function (SCO6197; SCO6592, SCO6591, SCO6594). SCP1.201 is located on the SCP1 plasmid. Again this gene is located near a secreted esterase (SCP1.199) and a secreted protein of unknown function (SCP1.200). Homology searches showed that SCO6197, SCO6591 and SCP1.200 are all homologues, though no other homologues were found. No relationships were found for SCO6592, while SCO6594 was found to be homologous to the C-terminal portion of SCO0545. SCO0545 does not have a known function but there are several catabolic enzymes in the same region. Given the conservation of the associated genes it seems possible that they represent a conserved system and that the ALF regions act as a substrate or product recognition domain that passes a signal to or from the secreted esterases.

The Hint module does not contain the homing endonuclease, and so is probably no longer an active mobile genetic element; this concurs with the apparent lack of other inteins in the *S. coelicolor* genome. This implies that the plasmid has passed through another species that has mobile intein elements. It may still fulfil a functional role as most of the bacterial Hint domains are found in secreted and cell wall associated proteins (personal communication: S. Petrovsky).

SPDY (Serine-Proline-Aspartate-Tyrosine motif; PF03771)

This domain typically occurs in pairs, is approximately 90 residues in length and has two conserved tryptophans and a proline (See Figure 2.7). It is only found in a region of the *S. coelicolor* that is believed to be an integrated genetic element, e.g. a plasmid or transposon (Bentley, Chater *et al.*, 2002). The edges of the mobile element can be

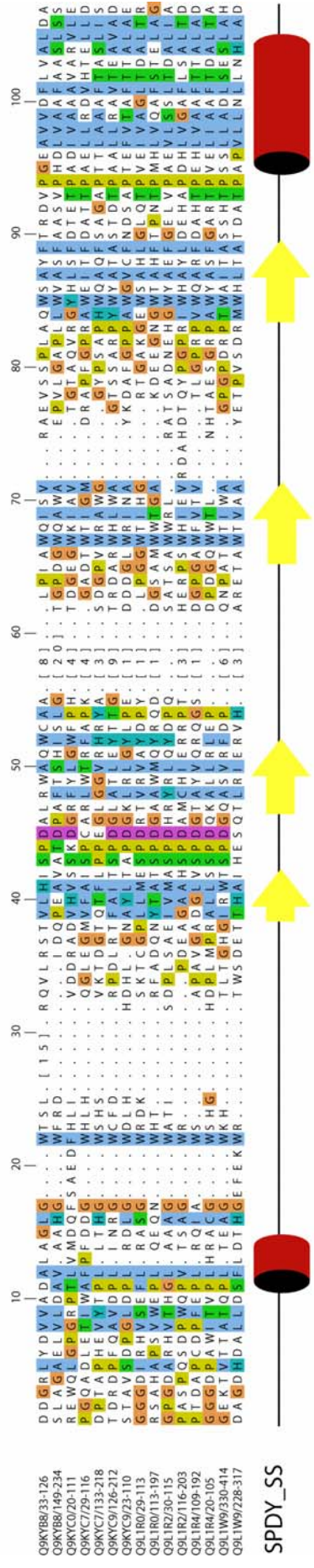


Figure 2.7: SPDY alignment and architectures

detected by viewing a plot of the composition of the DNA. A fairly recently introduced element would be expected to have G+C content and G-C ratio that is markedly different from the genomic background. These graphs are shown in Figure 2.8. The region appears to consist of two sections: a 'core' mobile element region with the essential replication genes and a flanking region containing arsenic resistance genes and a polyketide synthase (personal communication: S. Bentley; see Figure 2.8). So this element may be important in mobilising these loci between strains. All of the SPDY domains occur in the core region, indicating that they are important in the replication of the element – though it is not possible to assign them a precise role. The lack of occurrences of this domain in any other known proteins indicates that this region of the genome represents a previously undescribed type of mobile genetic element.

PASTA (Pbp And Serine/Threonine kinase Associated; PF03793)

The PASTA domain is discussed in greater detail in chapter 4.1; In this section I will discuss its relevance to *Streptomyces coelicolor*. It is a small, approximately 70 amino acids, globular α/β domain that binds cell wall peptidoglycan. Typically organisms that have PASTA domains have two PASTA-containing proteins. One is a PASTA-containing serine/threonine protein kinase (pPSTK), which is thought to be a key regulator of cell wall peptidoglycan cross-linking and hence essential to growth and development. The other is a PASTA-containing penicillin-binding protein (pPBP), which is one of essential peptidoglycan cross-linking enzymes. For a type example see *Streptococcus pneumoniae* PBP2X (UniProt:PBPX_STRPN).

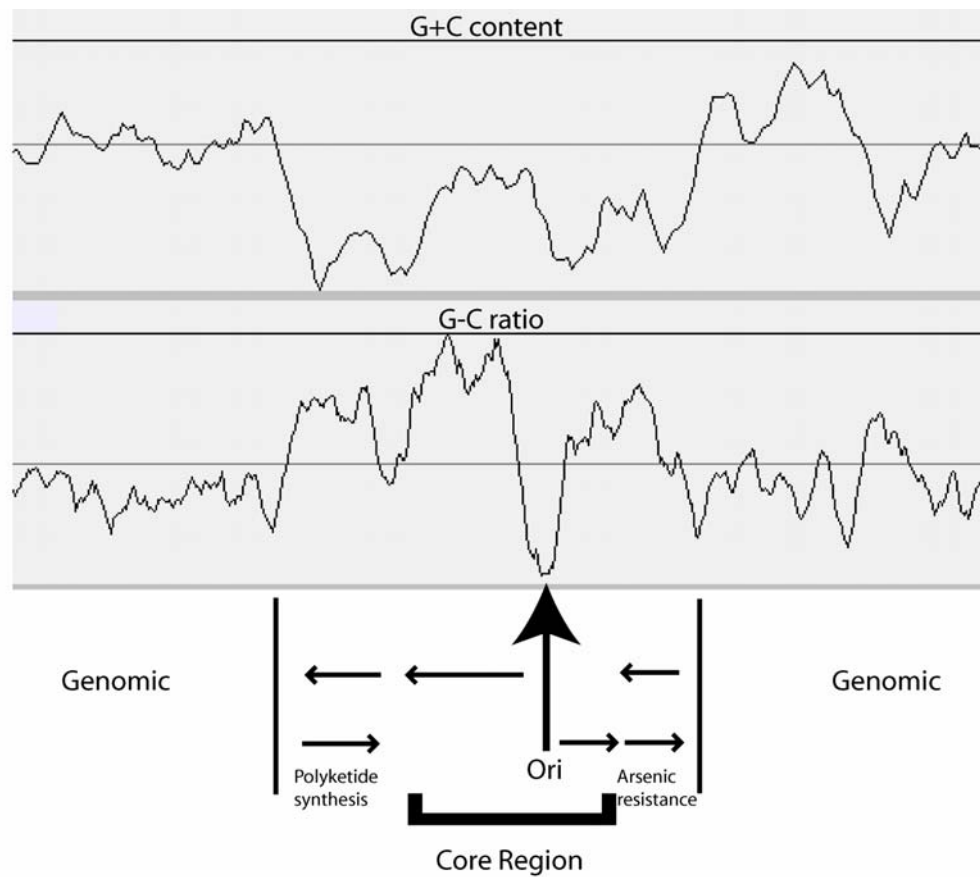


Figure 2.8: Evidence for the presence of a mobile DNA element
 The SPDY domains are all found in a region of the genome believed to be a mobile element. As can be seen the G-C ratio and G+C content show a marked difference to the background genome. This element appears to carry genes for arsenic resistance and synthesis of a polyketide.

However, uniquely amongst the sequenced microbial genomes, *S. coelicolor* has three pPSTKs and no pPBP. The PASTA domains show very little identity to each other in each PSTK. The simplest explanation is that each pPSTK regulates different stages of growth and division, each of which uses different peptidoglycans (as reported in Kalakoutskii and Agre, 1976). Since each stage of the *S. coelicolor* developmental cycle has a slightly different environment and growth requirement, different biochemical properties are needed for each type of cell wall. This also fits with there being no pPBP as it would be specific to a single peptidoglycan structure; so I propose it uses an alternative localisation system, perhaps similar to that used by

Deinococcus radiodurans or Gram negative bacteria – both of which do not have pPBPs. There is a protein containing a single PASTA domain, SCO4557, but whether it is involved in localisation of the PBPs is not known.

The identification of the protein containing a single PASTA domain does illustrate the stepping stone sequence phenomenon. Despite extensive searching when I first identified this family I did not find this match as it was too divergent from the rest of the family. Subsequent iterative searching against UniProt 44.0/27.0 rather than 40.0/18.0 allowed expansion of the family and its subsequent inclusion.

Intriguingly *S. coelicolor* has three principle cell morphologies and it may be that each pPSTK regulates the development of each type. The correlation between an organism having several distinct cell wall morphologies and having more than one pPSTK or pPBP is discussed further in chapter 4.1.

As for the relatives of *S. coelicolor*, the pPSTK StoPK-1 of *Streptomyces toyacaensis* have been shown to be involved in growth and resistance to antibiotics, and disrupting it causes changes in its mycelial morphology (Neu, MacMillan *et al.*, 2002). Also PSTK inhibitors block sporulation and slow the induction of antibiotic resistance (Neu and Wright, 2001). *Streptomyces avermilitis* has the same set of PASTA proteins as *S. coelicolor*.

HHE (Histidine-Histidine-Glutamate motif; PF03794)

This domain provides a good example of the "stepping stone" phenomena discussed in chapter 1.5 and mentioned above in the PASTA domain report. When first

identified (Yeats, Bentley *et al.*, 2003), this family was iteratively searched until convergence; when the searches were repeated 18 months later (UniProt Release 43.2/26.2 rather than 40/18) significant similarity to another Pfam family Hemerythrin was detected. This had two effects: it is possible to test the predictions made about HHE, and secondly our understanding of the Hemerythrin domain can be refined and expanded.

The HHE domain was predicted to be a 60 amino acid two α -helical cation-binding domain (see Figures 2.9 and 2.10 for examples). It was mostly found in prokaryotes, though some plant and fungal homologues were also identified (e.g. UniProt: Q9LJQ1). Noticeably the HHE domain mostly occurs in pairs, though there are apparently some examples of singlets (e.g. UniProt: Q92Z80). The MSA highlighted two conserved histidine residues, both of which reside within the predicted helices, and a conserved glutamate; combined with the occurrence of the HHE domain in a predicted cation-transporting ATPase, this is suggestive of a cation binding site. For instance two histidines and a glutamate are used to coordinate Zn^{2+} ions in Carboxypeptidase A.

Hemerythrin has previously been described as a 120 residue, four or five helical domain and has been best studied in a sandworm system analogous to haemoglobin (for a review see Kurtz, 1999). It binds two Fe^{2+} ions through four histidines and two glutamates (Kurtz, 1997), though it has also been shown to bind other cations,

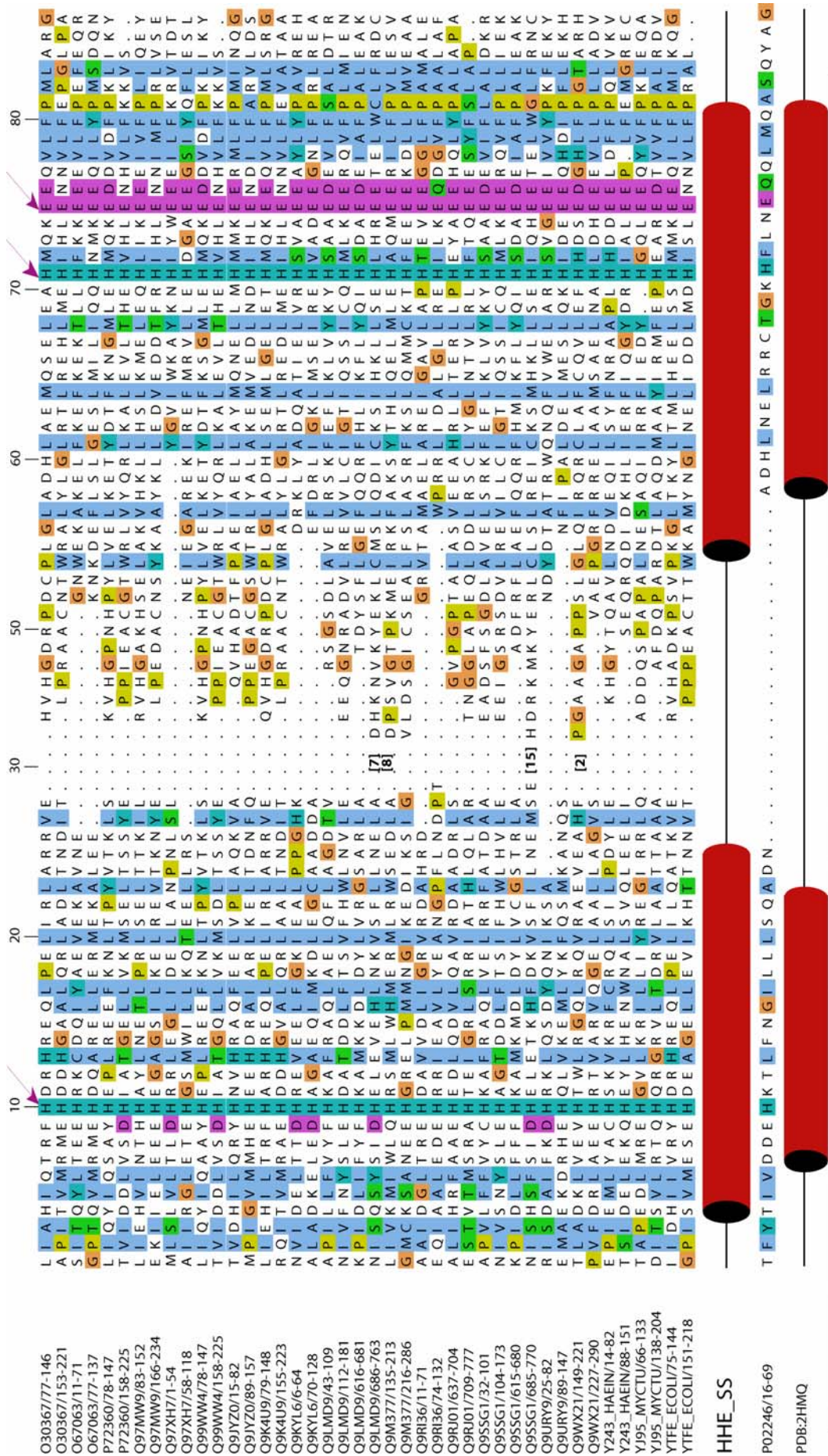


Figure 2.9: Alignment of original HHE domains and predicted secondary structure against a Hemerythrin domain and known structure
 The ligand-binding residues are marked by the purple arrows.

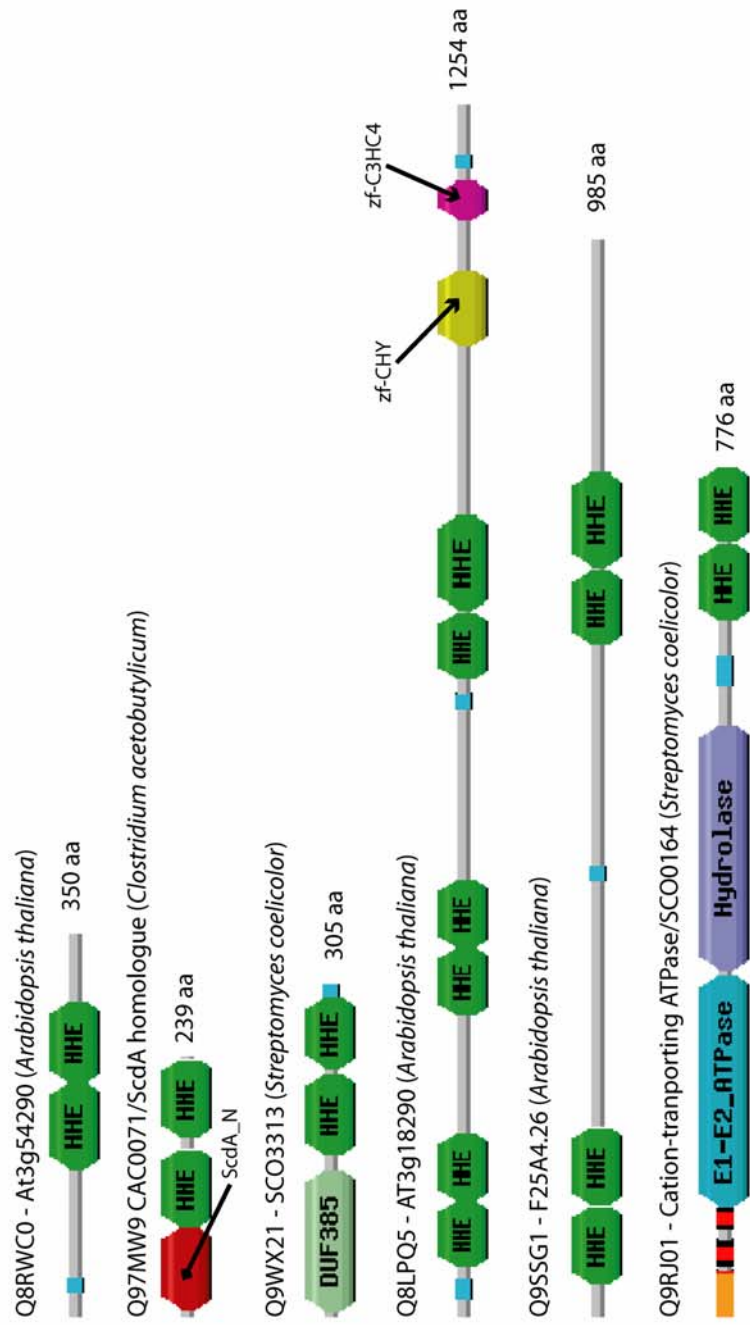


Figure 2.10: HHE architectures

including mercury (Clarke, Sieker *et al.*, 1979). The Fe^{2+} ions then typically coordinate an oxygen atom. This description can now be refined to state that the hemerythrin structure consists of two homologous domains of around 60 residues, each of which binds a Fe^{2+} ion. A family of proteins related to the hemerythrins, called myohemerythrin and also found in sandworms, has been found to bind both Cd^{2+} and Fe^{2+} (various refs, including Deloffre, Salzet *et al.*, 2003). In sandworms both families form homomeric complexes of HHE domains (for hemerythrin see PDB:1A7D, and myohemerythrin see PDB:1I4Z).

There are also members of this family, e.g. NorA and DnrN, that were initially discovered as part of the HHE family and not hemerythrin and that are described as being involved in the regulation of NO response in denitrifying bacteria (Pohlmann, Cramm *et al.*, 2000; Vollack and Zumft, 2001). For instance if *Pseudomonas stutzeri* DnrN is deleted then its nirSTB operon responds more slowly to nitrate. Given the conservation of ion-chelating function in the Hemerythrins it is possible that these HHE domains also bind a cation, which is then used to sequester NO. Also a low cytoplasmic oxygen concentration is essential for denitrification. So alternatively it may be involved in maintaining the anoxic environment of the cell during denitrification through scavenging free cytoplasmic oxygen, or up-regulating anoxia maintenance systems after sensing free molecular oxygen in the cell.

It has been noted that a deletion mutant of the *Staphylococcus aureus* homologue of DnrN, ScdA, exhibits defects in the cell wall, growth and development (Brunskill, deJonge *et al.*, 1997). Subsequent work has shown that it is regulated by SrhSR (Throup, Zappacosta *et al.*, 2001), which is the global regulator that allows *S. aureus*

to switch its metabolism from aerobic to anaerobic. While (Brunskill, deJonge *et al.*, 1997) suggest that ScdA is a regulator of development, the evidence of its domain structure combined with its involvement in *S. aureus* survival when moving into anoxic environments suggests that its specific role may be either to scavenge O₂ from outside the cell or to provide an intracellular store. Alternatively it may function as a positive regulator and if there is no oxygen bound to the HHE domains it will up-regulate self-protective systems. The defects identified by Brunskill and colleagues, and noted above, may be ascribable to damage caused by oxidative stress. These proteins have another domain - ScdA_N - at the N-terminus, which does not have an identifiable function, but may transduce the signal from the sensor HHE domains to the next downstream element. Determination of the function of the ScdA_N domain should help to resolve how these proteins function.

The domain is now identified to be wide-spread in the web of life, with instances occurring in humans, plants, worms, fungi, bacteria, archaea and elsewhere. It appears to be a successful alternative to haemoglobin for chelating cations and binding molecular oxygen.

PPC (Bacterial Prepeptidase C-terminal domain; PF04151)

These domains are typically ninety residues in length and found at the C-termini of secreted peptidases (See Figures 2.11 and 2.12). These domains are found in at least four different classes of peptidases, the metallopeptidase families M4, M9 and M28, and the serine peptidase family S8 (as defined by Rawlings, Tolle *et al.*, 2004). They are also found in the plant Ubiquitin Fusion Degredation proteins (UFD1 domain) and tyrosinase. In *Pyrococcus furiosus* pyrolysin the PPC domains are cleaved off

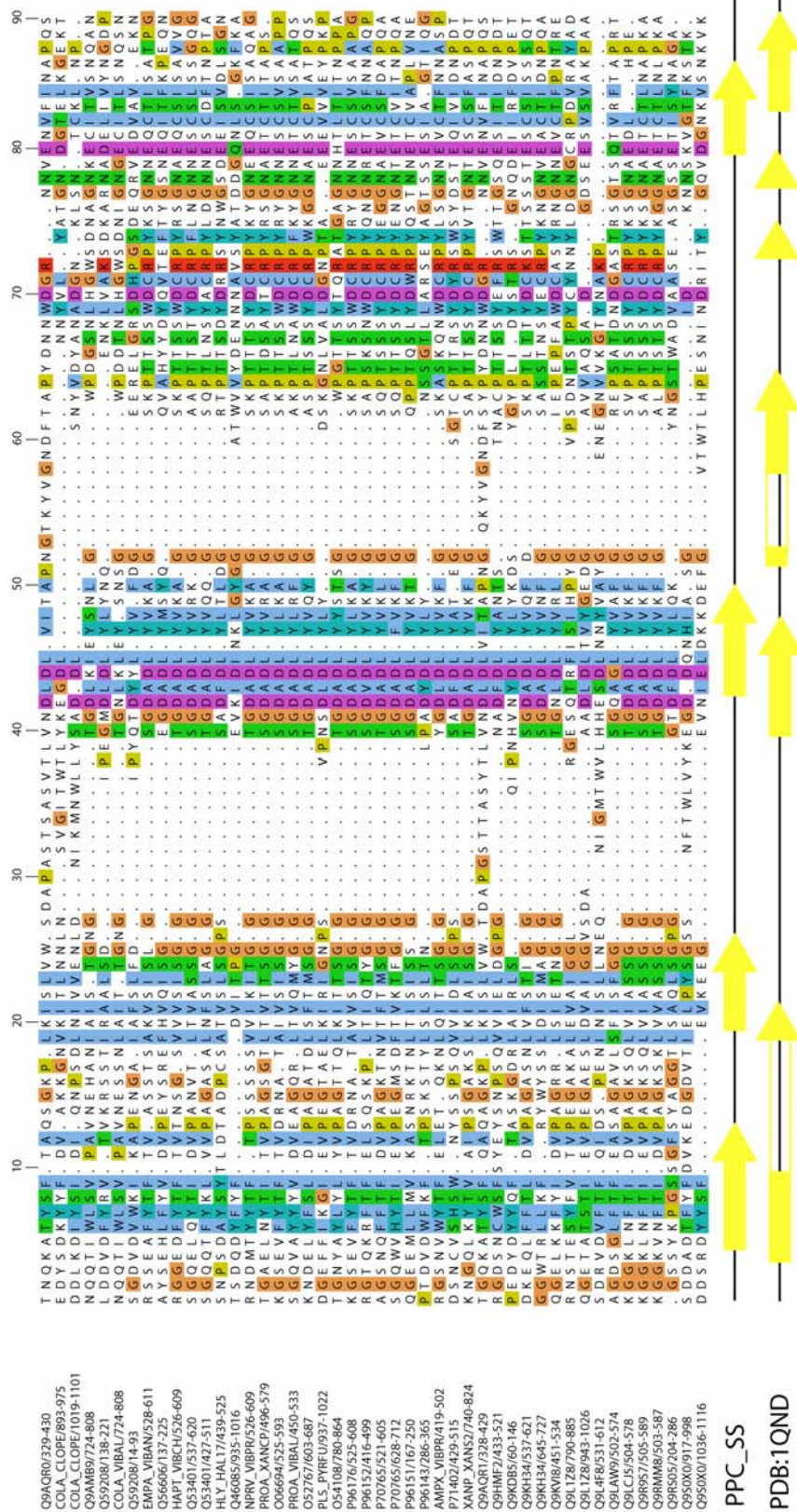


Figure 2.11: PPC alignment along with predicted and known secondary structure
 Since the initial prediction of the structure of this domain the C-terminal PPC domain of Q9S0X0 (residues: 1035-1116) has been solved. The structure is largely in agreement with the predicted structure for the family. Where a secondary structure element overlaps a gap in the alignment it has been extended to cover the gap.

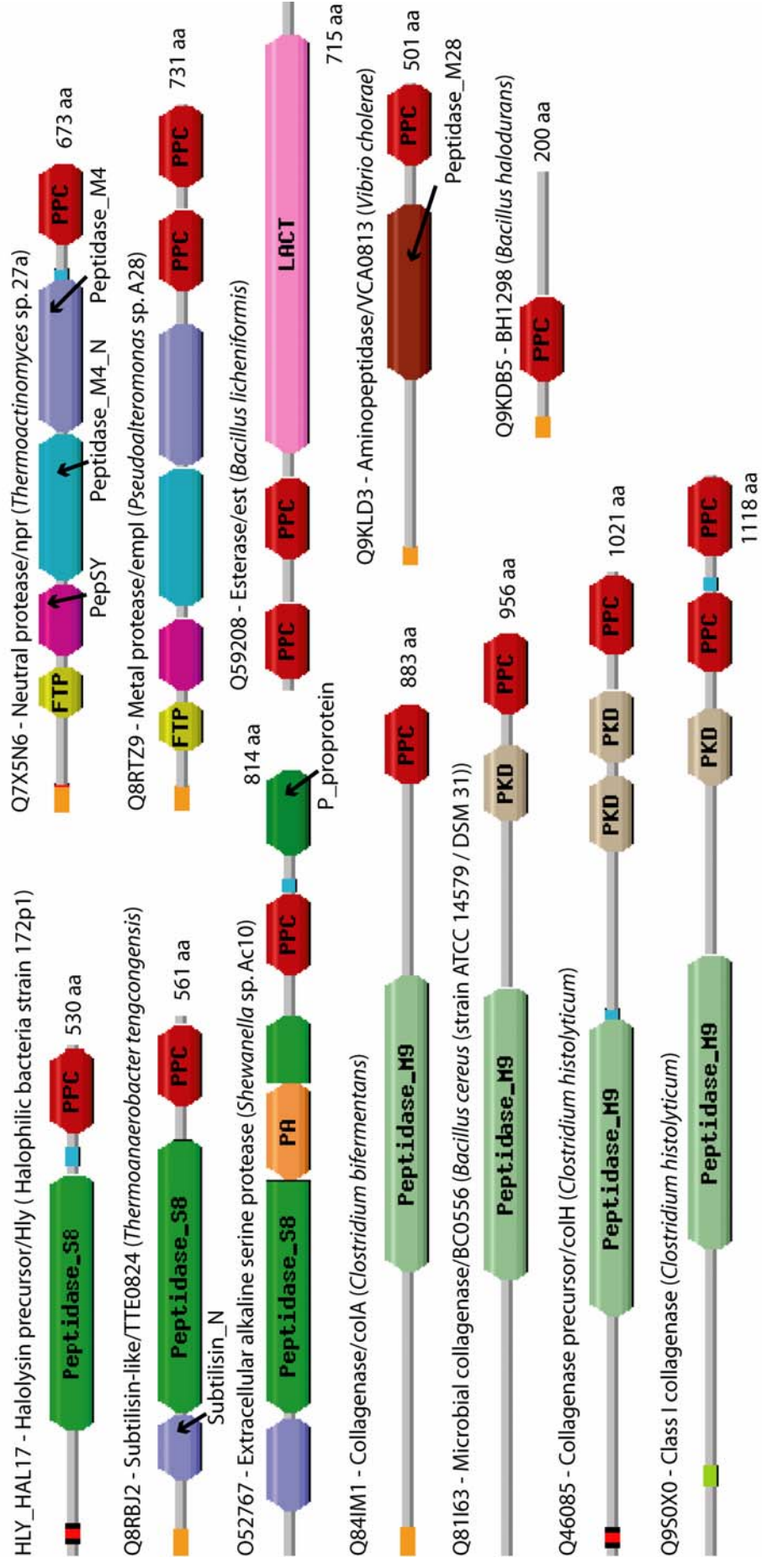


Figure 2.12: PPC domain architectures

subsequent to secretion, but prior to activation of the peptidase (Voorhorst, Eggen *et al.*, 1996). Although termed the prepeptidase C-terminal domain it is also found at the N-terminus of a couple of proteins (e.g. UniProt: Q59208).

The original publication of this domain (Yeats, Bentley *et al.*, 2003) suggested that it was likely to belong to the Ig fold, based on the MSA (personal communication: A Bateman) and its apparent interchangeability with the PKD domain in various architectures (e.g. compare the architectures of Q81I63, Q899Y1, Q9S0X0 and Q46085). Furthermore it was predicted to either be involved in localisation of the enzyme or in acquisition of the substrate. Subsequent to these predictions, the crystal structure of a PPC domain was determined by Wilson, Matsushita *et al.* (2003), which showed that it actually to belong to the jelly roll fold. Their work has shown that this domain binds the triple-helix of collagen in a reaction mediated by calcium ions; however, the Ca²⁺-binding site lay in a linker that does not fall within the PPC domain, so possibly the involvement of Ca²⁺ is restricted and not true of all PPC domains. A similar conclusion is reached by Wilson, Matsushita *et al.* (2003) on the basis of site-directed mutagenesis. It should also be noted that not all PPC domains necessarily bind collagen; further direct experimentation is needed to clarify their overarching function.

Comparing the resolved secondary structure to my previous predictions shows that most of the predicted strands were roughly in the correct positions. However, the alignment reveals that the PPC domain crystallised is atypical compared to most of the rest. The second predicted strand appears to have been deleted whereas another has been inserted between the third and fourth predicted strands. As for the two very

short strands, this region of the alignment is not well conserved and may not form them in the homologues. Still, this result and the HHE result suggest that most of the secondary structure predictions can be taken with confidence.

FMN bind (Flavin MonoNucleotide-binding; PF04205)

This domain represents a sixty residue region that includes an FMN-binding site (see Figures 2.13 and 2.14), as determined in the NqrC proteins of *Vibrio cholerae* (Barquera, Hase *et al.*, 2001) and *Vibrio alginolyticus* (Hayashi, Nakayama *et al.*, 2001). The NqrB proteins, which also bind FMN through a threonine residue and are part of the same complex, do not show any obvious similarity. The region is found in several electron transport chain proteins; for example the RnfG electron transport protein is part of a chain that supplies electrons to both nitrogen fixation and DNP reduction in *Rhodobacter capsulatus* (Jouanneau, Jeong *et al.*, 1998). Other examples include the NosR/NirI nitrous oxide reduction regulatory proteins. The FMN_bind proteins appear to form a few distinct groups; for instance the NqrC homologues are about 250 amino acids in length and contain one domain. The NosR-related proteins are around 800 residues, and also have several transmembrane helices towards the C-terminus. The ProSite 4Fe-4S model (PS00198) detected possible matches in the NosR proteins. These were confirmed by iteratively searching from these start points; within two rounds of searching the family overlapped with the Pfam-A families NIR_SIR and Fer4. This suggests that the regulatory mechanism of the NosR proteins involves charge movement. FMN_bind also occurs in fumarate reductases in association with the FAD_binding_2 domain.

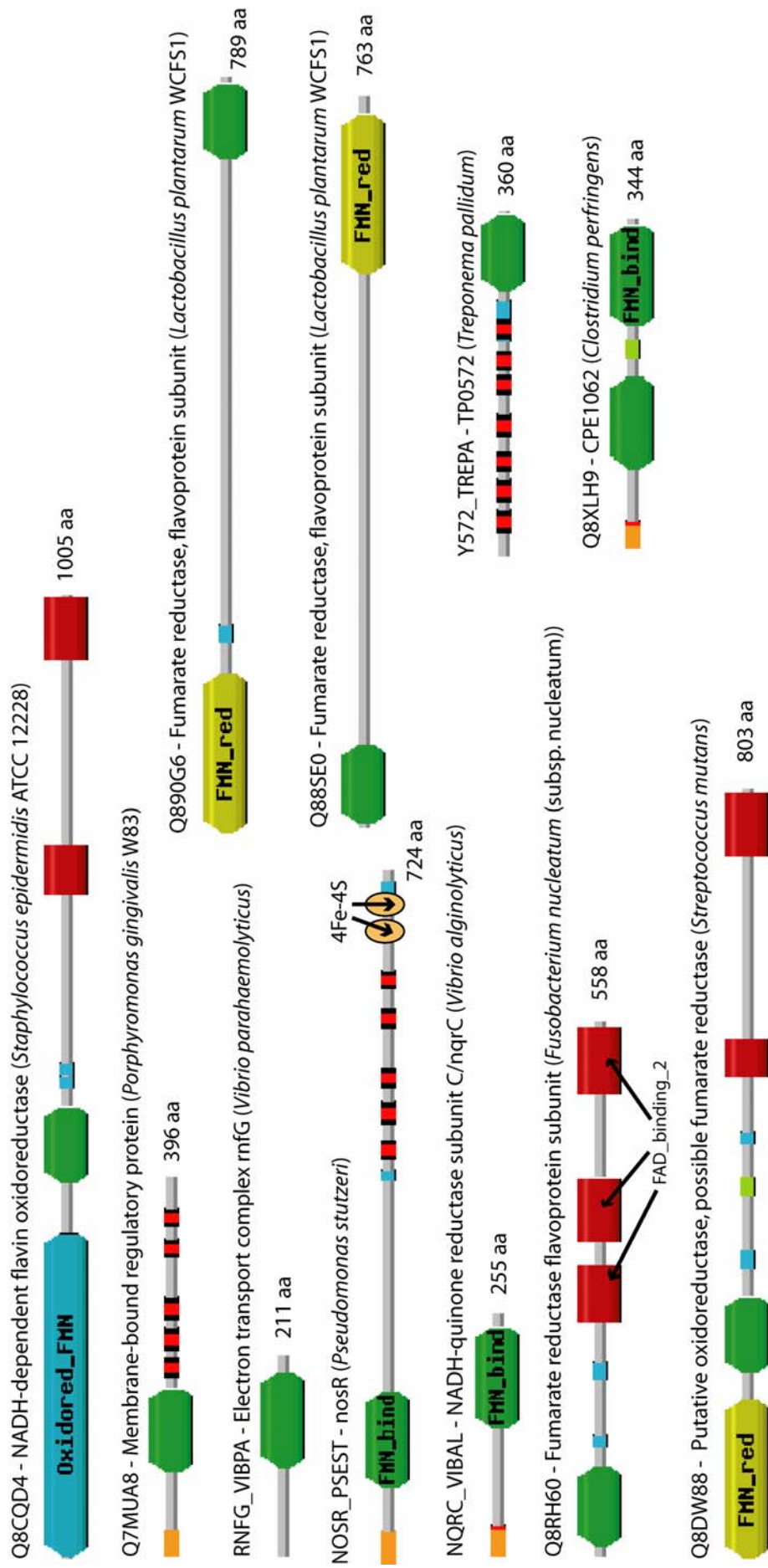


Figure 2.14: FMN_bind domain architectures

MbtH (MbtH-like proteins; PF03621)

This domain is named after the MbtH protein from *Mycobacterium tuberculosis* (UniProt:O05821). The domain is typically 70 residues in length and covers the full length of the protein, though NikP1 from *Streptomyces tendae* (UniProt:Q9F2E7) also contains two domains common to antibiotic synthesis proteins: an AMP-binding domain (PF00501) and a Phosphopantetheine attachment site domain (PF00550). It is found in the Actinomycetes, the Proteobacteria gamma subdivision and in the Rhizobium/Agrobacterium group. Several of these proteins have been implicated in antibiotic biosynthesis in streptomycetes (for instance nikkomycins: Lauer, Russwurm *et al.* (2001); simocyclinone: Galm, Schimana *et al.* (2002); coumermycin A1: Wang, Li *et al.* (2000), and the formation of siderophores such as *E. coli* enterobactin or *M. tuberculosis* mycobactin (reviewed by Crosa and Walsh, 2002). In the biosynthesis of siderophores they do not seem to have a direct role, as a complete synthetic pathway can be built up of mycobactin without assigning to a role to MbtH (and similarly with enterobactin and the MbtH-like YbdZ); so it is likely that it is involved in either regulation of production or an accessory role, with a similar function in antibiotic synthesis. There are several conserved residues, including three tryptophans that may have functional importance (See alignment and architectures in Figure 2.15).

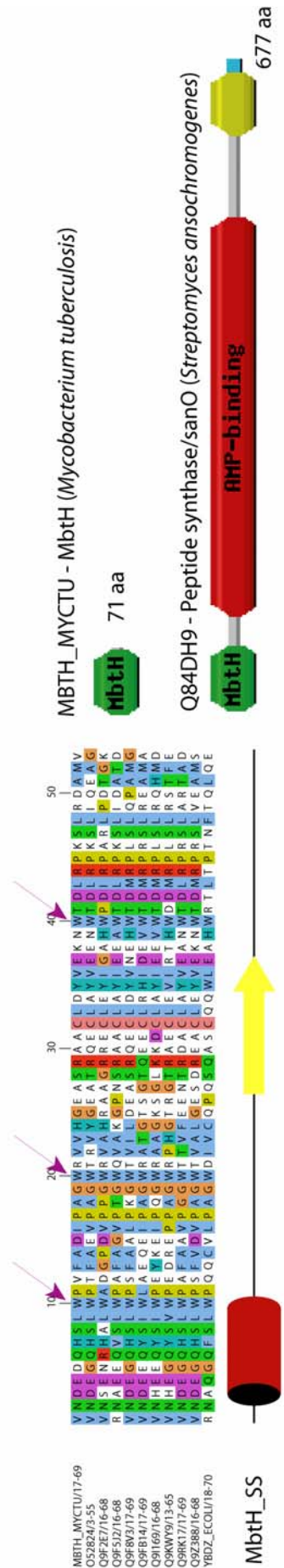


Figure 2.15: Alignment and architectures for the MbthH domain
 The purple arrows above the alignment mark three invariant tryptophan residues that may have functional importance.

2.4 Significantly Extended Pfam Families

SCP (Secreted Cysteine-Rich Proteins; PF00188)

The SCP domain was initially identified as a eukaryote-only domain (Szyperski, Fernandez *et al.*, 1998). Members of the family have been found to be involved in a wide variety of biological processes. For instance they are involved in several mammalian developmental processes, most notably sperm maturation (Maeda, Nishida *et al.*, 1999) and sperm-egg fusion (Roberts, Ensrud *et al.*, 2002), and are up-regulated in several tumours (Yamakawa, Miyata *et al.*, 1998; Asmann, Kosari *et al.*, 2002). Clear evidence has been found of *Xenopus* sperm following the concentration of 'Allurin' – an SCP-containing protein (Olson, Xiang *et al.*, 2001). They are also commonly used by insects and reptiles as mammalian toxins - as an example pseudochetoxin (from king brown snake) appears to bind the extracellular portion of cyclicnucleotide gated ion channels (CNG channels) blocking their function (Brown, Haley *et al.*, 1999). The eukaryotic branch of the family is characterised by all its members being secreted and the domains being rich in cysteines – which are thought mostly to form stabilising disulphide bridges.

The first report of this domain in bacteria is by Ponting, Aravind *et al.* (1999). However, recent evidence allows the expansion of their results and the formation of a hypothesis of the molecular function of this domain, and so it was discussed in detail in Yeats, Bentley *et al.* (2003); also a model was created and deposited in Pfam (see Figure 2.16 and 2.17 for alignment and architectures). The most obvious difference between the bacterial and eukaryotic copies is the absence of the disulphide bridges in the bacterial SCP proteins. It has been suggested that there is an active site, based on analysis of the 3D NMR image of plant PR14a and comparison with human GliPR

(Szyperski, Fernandez *et al.*, 1998). Alignment with the prokaryotic versions allows us to determine that three of the four residues predicted to make up the site are conserved between the eukaryotic and prokaryotic subfamilies (See Figure 2.16). This reveals the site to consist of two histidines and a glutamate - similar to the Hemerythrin/HHE domain above.

Review of the data available for this family had previously led to the conclusion that it was somehow involved in extracellular signalling; however, the protein is very large for a signalling molecule and, even though there is evidence for an active site, there is little evidence for the generation of a smaller signalling molecule. A recent paper by (Milne, Abbenante *et al.*, 2003) suggested that Tex31, which contains a single SCP domain, is a Ca²⁺-dependant protease. While the evidence for it being a protease is not definitive, mostly due to possible left-over impurities (personal communication: N Rawlings), the evidence for Ca²⁺-binding is quite strong. This fits with the identification of the conserved histidines and glutamate (see HHE domain above), and also fits with the involvement of SCP-domain proteins in many diverse processes. For instance cell polarity has been well-established as being of fundamental importance in determining growth directions of pollen tubes and fungal hyphae.

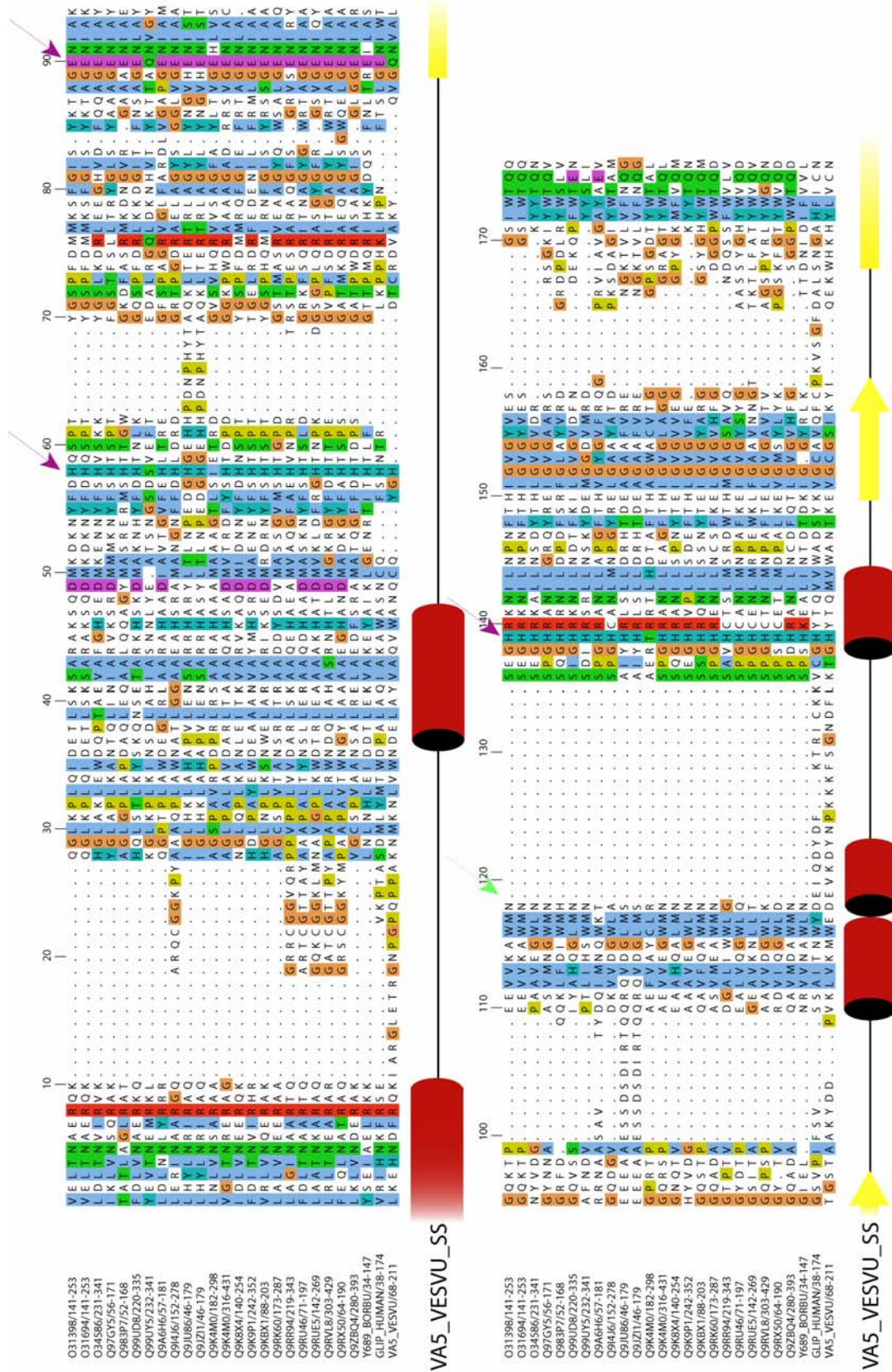


Figure 2.16: SCP domain alignment

The secondary structure provided in this figure is based on the three-dimensional structure of VA5_VESVU (PDB:1QNX). The arrows above the alignment mark the residues predicted by Szyperski et al (1998) to form an active site. One of these, marked by the green arrow, is not conserved between the bacterial and eukaryotic versions, ruling it out from being an important part of the active site.

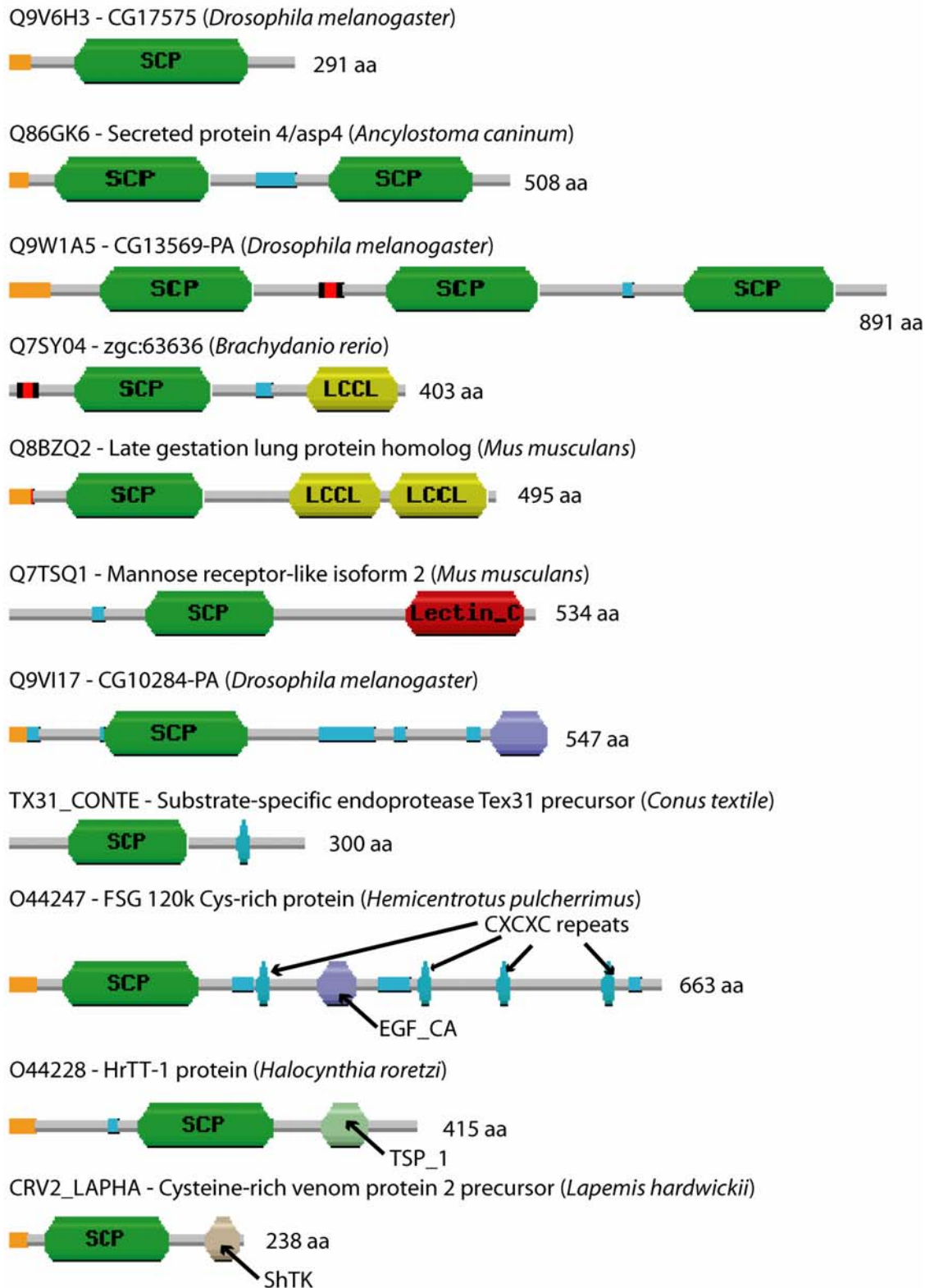


Figure 2.17: SCP domain architectures

To understand how SCP-containing proteins may interact with the known cell polarity mechanisms it is necessary to understand the role of Ca^{2+} concentrations. In the case of fungi, establishment of the gradient causes actin polymers to align down the length of the cell and then to transport cell wall components and polymerases to the growth tip (reviewed by Sheu and Snyder, 2001). This allows expansion at the tip and so osmotic pressure can drive growth. Possibly important to this process is the export of Ca^{2+} ions from one end of the cell (Silverman-Gavrila and Lew, 2003), and import towards the posterior, as has been shown in pollen (Malho, Read *et al.*, 1995).

Such processes have been shown to underlie development in animals and plants. Furthermore the concepts that sperm may follow a Ca^{2+} -gradient in *Xenopus*, and that cell polarity may be one of the first factors established during sperm-egg fusion do not seem implausible. For a start waves of calcium have been seen to emanate from *Xenopus* oocytes (Eidne, Zabavnik *et al.*, 1994). Indeed it is not impossible that sperm-egg fusion uses a conceptually similar process as hyphal growth or pollen tube growth. It is already known that actin filaments in the sperm acrosome polymerise and push the acrosomal membrane into the egg cell. Then their cytoplasm merge and the sperm nucleus transfers. Creation of a Ca^{2+} -gradient could serve as the trigger for this process as it would polarise the sperm cell, causing the actin filaments to rearrange and drive the membrane. It may also be found that the egg cell polarises - this would allow it to transport lytic factors to the correct place in the cell membrane to facilitate sperm entry.

In bacteria cell polarity has also been shown to be important in the establishment of specialised organs at different locations in the cell and in replication (Shapiro,

McAdams *et al.*, 2002). For instance, *Caulobacter crescentus* has an assymetrical life cycle, which at one division produces two different types of cells. This is achieved through establishing a clear cell polarity in a process involving an actin-like cytoskeletal element, MreB, which has an innate polarity (Gitai, Dye *et al.*, 2004). *C. crescentus* does also have an SCP-containing protein – CC2118 (UniProt:Q9A6H6).

I propose that SCP-containing proteins are going to be important to the establishment of cell polarity, and effect local Ca^{2+} concentrations in the extracellular medium so as to amplify any charge imbalance. This could happen in three ways. SCP domains may sequester calcium ions, hence reducing the extracellular concentration. They could carry out a more sophisticated version of this activity by carrying the ions through the extracellular medium and depositing them for import into the cell. A third possibility is that they could cap ion channels – as pseudochetoxin apparently does. Whatever the mechanism by which SCP domains function, it would then be logical for charged cytoskeletal elements (e.g. MreB) to lie along the polarity gradient and carry proteins to their target. This model would complement the known pathways of establishing cell polarity, but it is entirely hypothetical and requires experimental testing.

Collation of the processes that SCP-domain proteins are involved in suggests that they may be involved in many of the early developmental pathways in eukaryotes, and in the localised differentiation of bacterial, and possibly archaeal, cells. If the predictions made above are correct than SCP proteins form part of a remarkable universal system for patterning individual cells and modifying their behaviour at specific localities. Conversely their basic function has been co-opted by various organisms for

application as a toxin (i.e. king brown snake) and to modulate the host immune system (dog hookworm's neutrophil inhibitory factor; Moyle, Foster *et al.*, 1994).

FG-GAP (PF01839)

Several *S. coelicolor* proteins were identified that were found to be related to FG-GAP repeats. The Pfam family from version 7.4 contained only 5 bacterial members. By merging in the *S. coelicolor* proteins it was possible to expand the family, and in Pfam 7.5 there were thirty nine bacterial members – including fourteen in *S. coelicolor*. An extra thirty-four eukaryotic family members were also identified (Pfam 7.5), An archaeal protein (UniProt:O28333) was also identified, and it now appears that the euryarchaea in general contain them. FG-GAP repeats form a β -propellor (Springer, 1997). FG-GAP domains can now be regarded as near universal domains that are likely to have an important role and are an ancient β -propeller family.

2.5 Concluding Comments

The primary purpose of this research was to identify novel protein domains for which information could be easily derived, and that were of biological significance to *Streptomyces coelicolor*. The hunt methods employed were optimised to produce a short list of targets with a good chance of being a domain. This was achieved through restricting the search to only looking for repeated domains and by using strict length filters and overlap filters. Whilst many potential novel domains were missed, detecting them would have involved developing a more complex process for delineating domain boundaries, searching proportionally more targets and having to carry out many more searches to identify distant homologues. To underline the speed of this approach there are 204 copies of the novel domains listed in Table 2.2 in S.

coelicolor alone, not including the SCP and FG-GAP families. In order to discover this many domains in *S. coelicolor* it was only necessary to investigate 145 potential families, most of which could be discarded quickly. The primary reason for this was that no matches were found to other proteins. This suggests that once a sufficient number of genomes have been sequenced comparative scans like this one will be even more useful. The BTAD domain is the only domain not derived directly from a target, but rather the region was highlighted by the investigation.

Examples, such as the PASTA domain (see chapter 4.1), also demonstrate that reasonably large gains in biological knowledge could be made through the delineation of the domain structures of these proteins and the taxonomical distribution of the domains. Similarly with SCO0002 and SCO0003 a strong functional link can be made between them due to the occurrence of HA domains in the C-termini of both of them. Given the location in the telomeres of the chromosomes and the associated helicase domain, we hypothesise that the HA domains bind DNA; we also note that predicted structural similarities to the Myb-like DNA-binding domain may provide a model for its function. Previously such a hypothesis could only be made based solely on their close proximity within the telomeres of the chromosome. Not all the predictions made lead to the identification of novel domains but rather to the expansion of known domain families. Most of these are not reported as they do not particularly enhance our understanding of the domains or *S. coelicolor*; however, a couple – SCP and FG-GAP – show large information gains. This demonstrates that the approach employed by Ponting, Mott *et al.* (2001) also works well in bacteria and has helped elucidate information specific to the species (e.g. the HA domain), to bacteria (e.g. the PASTA domain), and general biology (e.g. SCP).

Also once one member of a family is described information can be transferred to its relations. This is enhanced by the deposition of the families into Pfam; any further investigations into the streptomycetes using Pfam will automatically annotate these domains, increasing the knowledge and understanding of these remarkable organisms.

3 Multi-genome Domain Hunting

3.1 Rationale

Although the approach used in chapter 2.1 was successful in finding new biological information about *S. coelicolor* specifically and bacteria in general, further studies in other bacteria – including *Deinococcus radiodurans* and *Mycoplasma genitalium* – failed to uncover as many novel domains (only the BON domain reported – see chapter 4.2). Partly this was because some of the families that occur as repeats had already been identified, but may also have been because of the nature of bacterial genome structure. Bacterial genomes often appear to consist of a general core genome – the housekeeping genes and other essential metabolic or biosynthetic processes – and then a set of niche specific genes. As a caveat this generalisation does not extend to symbionts as core functions can be shared between the partners. The niche-specific genes are typically less characterised than the more wide-spread core genes, and so represent a better source of novel domains; and *S. coelicolor* has an enormous number of niche-specific genes compared to most other bacteria. In essence the bigger the genome the more chance of success. Another problem was that these investigations tended to generate information that was very specific to a species and not of general application to bacteria. So a more general approach was developed.

In principle, the more genes surveyed the more chance a rare duplication event may be identified, leading to the delineation of a domain's boundaries. Also domains of interest are likely to occur in several genomes. So 13 genomes (see Figure 3.1 for list) were processed as in chapter 2.2 and then the repeat pairs clustered using single-linkage clustering, in the same manner as in the small protein clustering method. The

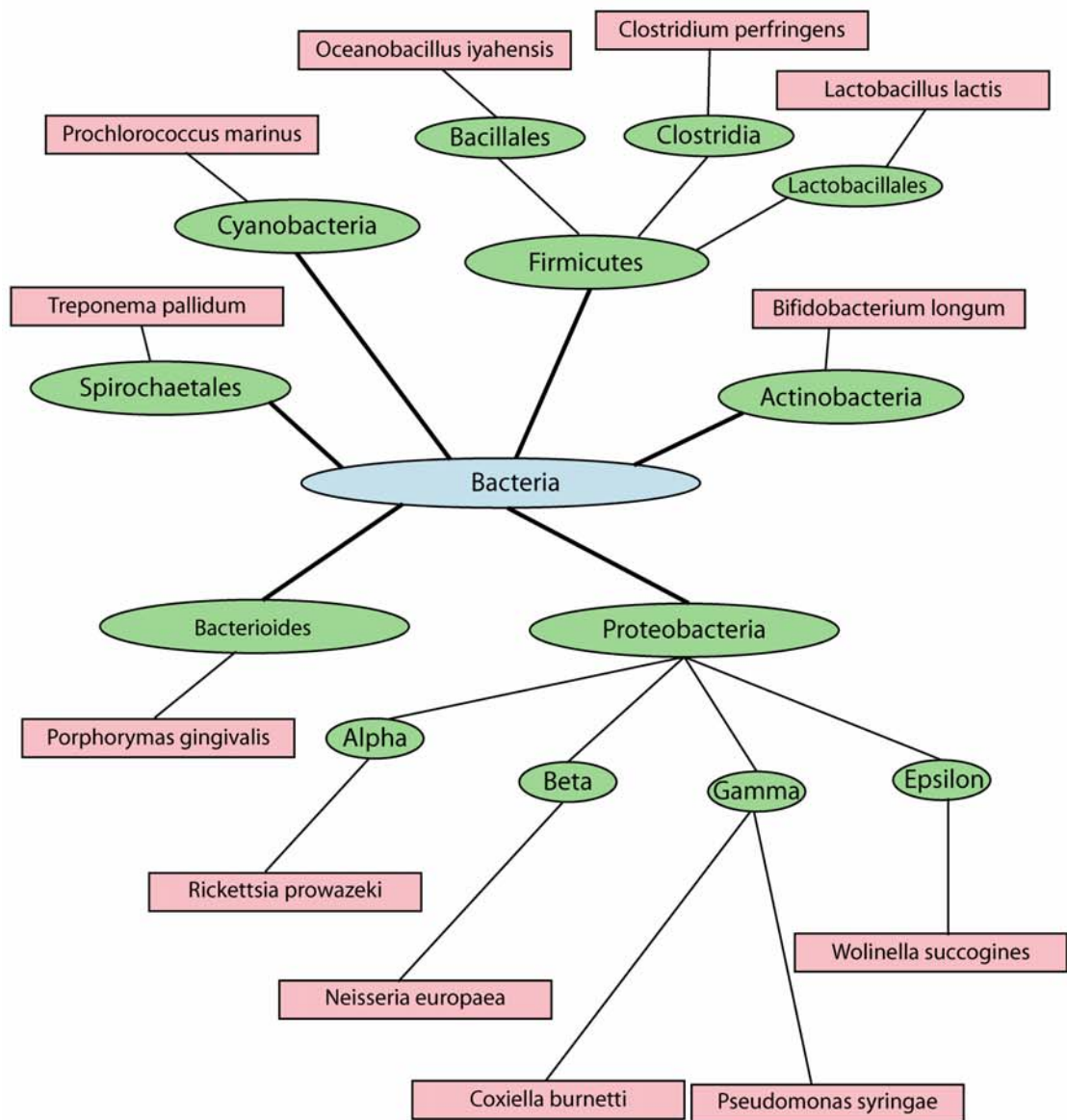


Figure 3.1: Simplified taxonomic tree of bacteria investigated in the multigenome hunt
 The bacterial species investigated in the multigenome hunt are indicated in the pink boxes on the leaves of the tree; the species groups are shown in the green ovals. When selecting these species an attempt was made to have a selection of species from all over the eubacterial kingdom. The taxonomy was taken from the NCBI's taxonomic database.

two advantages of this approach are greater sensitivity and that targets are ranked in likely importance – more widespread domains will produce larger clusters. The disadvantage is that the number of genomes prevents detailed contextual analysis within a reasonable time frame.

In this project the criteria for selecting a genome for investigation was simply that it wasn't too related (based on taxonomy) to one already chosen (see Figure 3.1), and that it was fully sequenced. The result was a “metagenome” of 29,173 proteins that gave reasonable coverage of the taxonomic tree.

3.2 Results

<i>Total Proteins</i>	<i>Short Proteins</i>	<i>UniProt Release</i>	<i>Pfam Release</i>	<i>Date</i>
29173	3091	41.25/25.14 & 42.5/25.6	11 & 12	Oct 2003- April 2004

3.2.1 Summary of Results

Repeat Identification

A total of 96 clusters that passed through the filters (see chapter 2.1.2) were found. The clusters that failed the second (overlap) filter were also investigated, as this filter proved to be overly restrictive. If one sequence had a single residue overlap the cluster failed. So the overlaps were manually checked and if the overlap only represented only a small portion of the alignment they were added to the list of targets – this led to the additions of an extra six targets. In total 4190 novel domains, repeats and motifs in 30 families were identified in UniProt 42.5/25.6. The families are listed in Table 3.1.

Summary of Small Protein Clustering Results

In total, 3091 proteins of less than 101 residues in length were clustered into 243 clusters, using a BLAST score threshold of 50 bits. Of the 243 clusters 124 had more than two proteins in them. 140 of these clusters significantly overlapped with Pfam families and so were discarded. After iterative searching, 17 new families were identified. In fact the actual number was slightly higher, but several of these families were only found in specific regions of *Lactobacillus lactis* that corresponded to mobile elements (Bolotin, Wincker *et al.*, 2001). None of these families had homologues from outside *L. lactis* and were not investigated further. In total 363 new domains, repeats and motifs were identified in UniProt 44.0/27.0.

3.2.2 Table of All Novel Domains and Families Identified

Table 3.1 lists all the new families identified during this investigation as well as some basic functional information. A similar set of accessory information is supplied as in Table 2.1. Domains reviewed in this Thesis are highlighted in blue.

3.3 Descriptions of Novel Domains

In this section, the novel domains produced from the multigenome hunt are described in a similar manner as chapter 2. In chapter 3.3.1 I describe the novel domains identified by the repeat identification hunt, while in chapter 3.3.2 I describe some domains identified by small protein clustering. The PepSY domain is noted here, but it is discussed in detail in chapter 4.3.

Pfam Accession No	Family Name	Pfam Type	Basic Function	No of copies in UniProt 44/27	Antibiotic biosynthesis	Spore Coat Formation	Cell Wall / Periplasm	Replication	Secreted
A) Novel Families									
PF03413	PepSY	Domain	M4 Peptidase Inhibitor			X	X		X
PF03958	Secretin N	Domain	Secretin N-terminal Domain				X		
PF07494	Reg_prop	Repeat	Regulatory β -propeller				X		
PF07495	Y_Y_Y	Motif	Unknown Regulatory Function				X		
PF07503	zH-HYPF	Domain	HypF-type Zinc Finger Domain						X
PF07550	DUF1533	Family	Unknown function						
PF07551	DUF1534	Family	Unknown function						
PF07552	Coat_X	Domain	Bacillales Coat X Domain			X	X		
PF07553	DUF1535	Domain	Unknown function				X		X
PF07554	FIVAR	Domain	NAG-binding Domain				X		X
PF07556	DUF1538	Family	Unknown function						
PF07559	FlaE	Domain	Flagellar Hook Protein Domain				X		X
PF07560	DUF1539	Family	Unknown function				X		
PF07561	DUF1540	Family	Unknown function						
PF07563	DUF1541	Family	Unknown function						
PF07577	DUF1547	Family	Unknown function						
PF07578	LAB_N	Family	Lipid A Biosynthesis N-terminal Domain				X		X
PF07581	Glug	Repeat	Short Cell Surface Repeat				X		X
PF07613	DUF1576	Family	Unknown function				X		X
PF07615	Ykof	Family	Unknown function						
PF07634	RtxA	Repeat	RTX toxin Repeat				X		
PF07655	Secretin_N_2	Domain	Secretin N-terminal Domain				X		
PF07660	STN	Domain	Secretin N-terminal Domain				X		
PF07670	Gate	Domain	Nucleoside Recognition				X		
PF07671	DUF1601	Family	Unknown function						
PF07675	Cleaved adhesin	Family	RepA-Ksp complex component				X		X
PF07853	CTnDOT_TraJ	Domain	Conjugative Transfer Protein J Domain				X		X
PF07862	NifH1	Domain	NifH1 Nitrogen Fixation Domain						
PF07865	DUF1652	Family	Unknown function						
PF07866	DUF1653	Family	Unknown function				X		
PF07867	DUF1654	Family	Unknown function						
PF07868	DUF1655	Family	Unknown function						
PF07869	DUF1656	Family	Unknown function						X
PF07870	DUF1657	Family	Unknown function				X		X
PF07871	DUF1658	Family	Unknown function						
PF07872	DUF1659	Family	Unknown function						
PF07873	YabP	Family	Unknown function						
PF07874	DUF1660	Family	Prophage Protein of Unknown Function						
PF07875	Coat_F	Domain	Bacillales Coat F Domain				X		
PF07876	Dabb	Domain	Stress-responsive Dimeric α/β Barrel				X		X
PF07877	DUF1661	Family	Unknown function						
PF07878	DUF1662	Family	Unknown function						

Table 3.1: Novel domains found in the multigenome hunt

3.3.1 Domains Identified Through Repeats

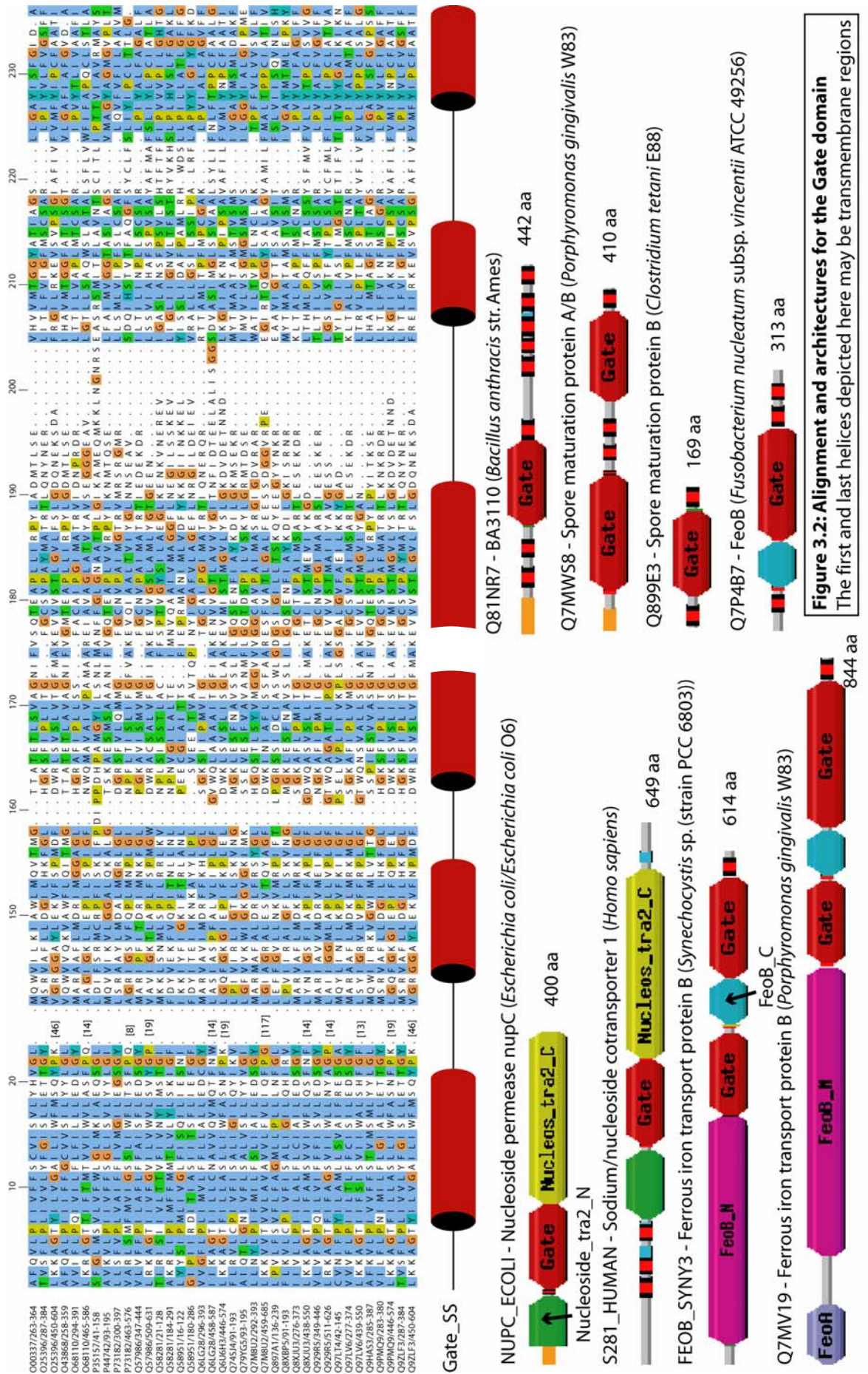
PepSY (PF03413 – M4 Peptidase inhibitory domain)

This domain is discussed in greater detail in chapter 4.3.

Gate (PF07670 – potential membrane channel specificity domain)

This domain is apparently ubiquitous, with copies found in all domains of life, and in most species it occurs between 1 (*Neurospora crassa*) and 11 (*Bacillus anthracis*) times. It shows a variety of architectures (see Figure 3.2) and is predicted to be an $\alpha+\beta$ fold, though most of the structure is made up from six α -helices (using PROF). The helices at the N and C-termini are both mostly predicted to be transmembrane regions by TMHMM (see below). Species that have the most copies include the enteric pathogens *Escherichia coli* and *Shigella flexneri*. It is also found in the human concentrative nucleoside transporter proteins (hCNT) 1, 2 and 3. These proteins are the Na^+ -dependant active transporter channels for the uptake of nucleosides from the cellular environment in the recovery pathways of many cells. Hence they are important physiological proteins, but they also have an important pharmaceutical role in the determining the uptake of nucleoside-based drugs, as used in the treatment of chronic lymphocytic leukaemia for instance.

Loewen, Ng *et al.* (1999) carried out mutagenesis assays which located the nucleoside specificity function of these proteins to the region now delineated as the Gate domain. However, this domain is found as two copies in the eubacterial FeoB proteins; these proteins are active GTP-dependant Fe^{2+} transporters and there is no current evidence that they can transport nucleosides - indeed *E. coli* also has an Gate-



containing hCNT homologue (NupC) that can transport nucleotides (Loewen, Yao *et al.*, 2004). So it seems that the Gate domain's function is more than to determine the specificity of the channel. One role that fits the available data is that is that Gate may be a nucleoside-binding domain, and in the FeoB proteins Gate recruits GTP. A second possibility is that the Gate domain either forms the channel or the opening to the channel, and hence that small changes to its structure can allow it to transport specific substrates from a wide range of possible substrates. Whilst the first one seems more likely, there is no evidence directly supporting either role. TMHMM, in general, predicts that the Gate domain is found on the cytosolic side of the membrane; this supports both suggested functions. Of note, the eukaryotic nucleoside transporters have an N-terminal extension, which contains three transmembrane helices, that is not present in the prokaryotic versions.

In terms of structure, the Gate domain normally seems to contain two transmembrane helices at its amino- and carboxyl-termini, though in some cases another pair of helices seems to be inserted in the centre. This may be a misprediction by TMHMM, which is used to make the Pfam transmembrane helices predictions, but certainly several of the copies have substantial insertions.

STN (Secretin and TonB N-terminus domain; PF07660)

The STN domain is around 50 residues long and is predicted to form an α/β fold (see Figure 3.3). It is one of a number of domains that I have identified at the N-terminus of secretin proteins (e.g. Secretin_N, BON, and Secretin_N_2). The bacterial secretins are membrane channels involved in the Type II/III secretion systems. The family are defined by a Secretin domain that forms the physical channel.

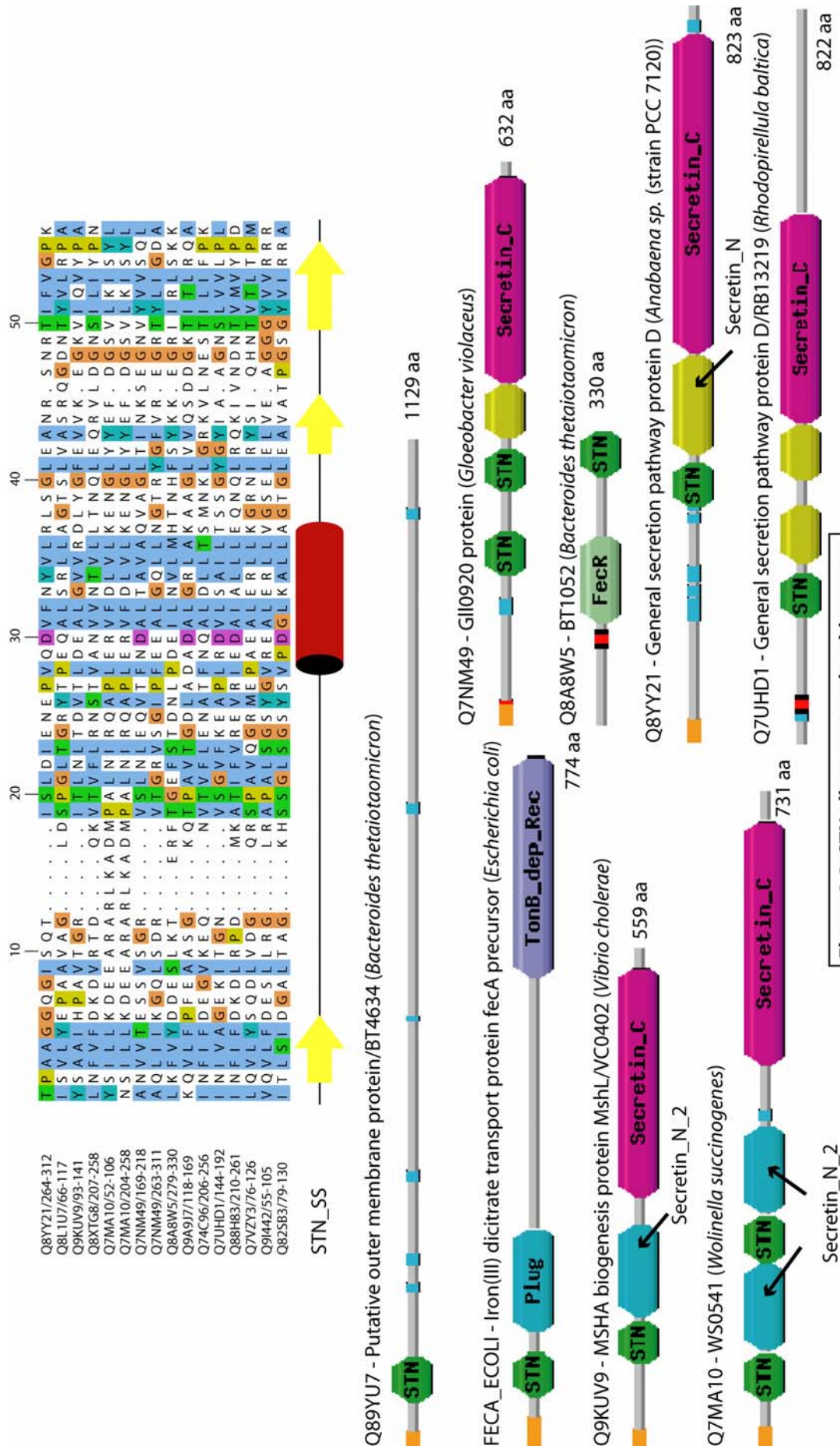


Figure 3.3: STN alignment and architectures

The various domains found to the N-terminus of the Secretin domain probably modulate the function, specificity and localisation of the channel. These channels are discussed more below and in chapter 4.2 (the BON domain). However, it is worth noting at this stage that elucidation of the various N-terminal domains will allow biologists to correctly classify the secretin subfamilies and allow much more accurate information transfer; based on architectures, Pfam 14 recognises 16 different secretin families. The secretins are a vastly important bacterial family due to their involvement in many processes, such as bacterial mating, iron sequestration, pathogenesis, niche adaptation and pilus formation. The STN domain is found adjacent to the Secretin_N (see below), the Secretin_N_2 (see below) and TPR families.

STN also occurs at the N-terminus of several TonB-dependant receptors (TDRs). The domains that form the channel (TonB_dep_rec) and the entrance (Plug) have both been delineated already so STN domains must carry out an alternative function. Another domain found at the N-terminus of secretins, investigated in Chapter 4.2 and called the BON domain, is believed to bind phospholipid membranes and hence may aid in localisation and stabilisation of the channel. Since STN is found in a similar context it may carry out a similar role. Some TDRs have N-terminal TPR repeats (Pfam:PF07719; UniProt:Q88H83) as well, which suggests that some of these channels have complexes recruited to the cytosolic side. Indeed, it may do neither role but be involved in modulating the channel response to an unknown signal. So, to confidently assign a functional role, direct experimentation is likely to be necessary.

Secretin_N (Secretin N-terminal domain; PF03958)

This domain occurs at the N-terminus of 70% of the Secretins in Pfam 14. It is normally 60-90 residues in length, though some copies contain large fifty residue insertions, and is predicted to have a mixed α/β fold (see Figure 3.4). The original Secretin_N model (called GSPII_III_N) actually contained one and a half repeats, resulting in many unusual fragment matches being reported. By resolving the correct boundaries it was possible merge all the fragments into a cohesive family and merge in the NolW-like family. It also allowed confirmation of the Secretin domain boundaries, since commonly this domain follows directly after a Secretin_N domain.

Experimental support for the Secretin N-terminal boundary was provided by a limited N-terminal proteolytic degradation experiment carried out by Nouwen, Stahlberg *et al.* (2000). They identified a peptidase resistant C-terminal domain in *Klebsiella oxytoca* PulD protein that began just before the region they described as conserved in all secretins. This correlates with the sequence evidence and subsequently the refined model has been confirmed by its lack of overlaps with STN, Secretin_N_2 and the BON domain - all of which occur adjacent to it.

It has not been specifically described or tested before, but from context it is possible to make some educated guesses as to its function. The Secretins either have a BON domain, one or more Secretin_N, or one or more Secretin_N_2 domains. The BON domain has been deduced to be a lipid membrane binding domain; therefore its substitution by Secretin_N suggests that Secretin_N may also fulfil a similar function. This may not be specifically binding the phospholipid membrane, but may be the similar but more general role of anchoring the internal end of these channels.

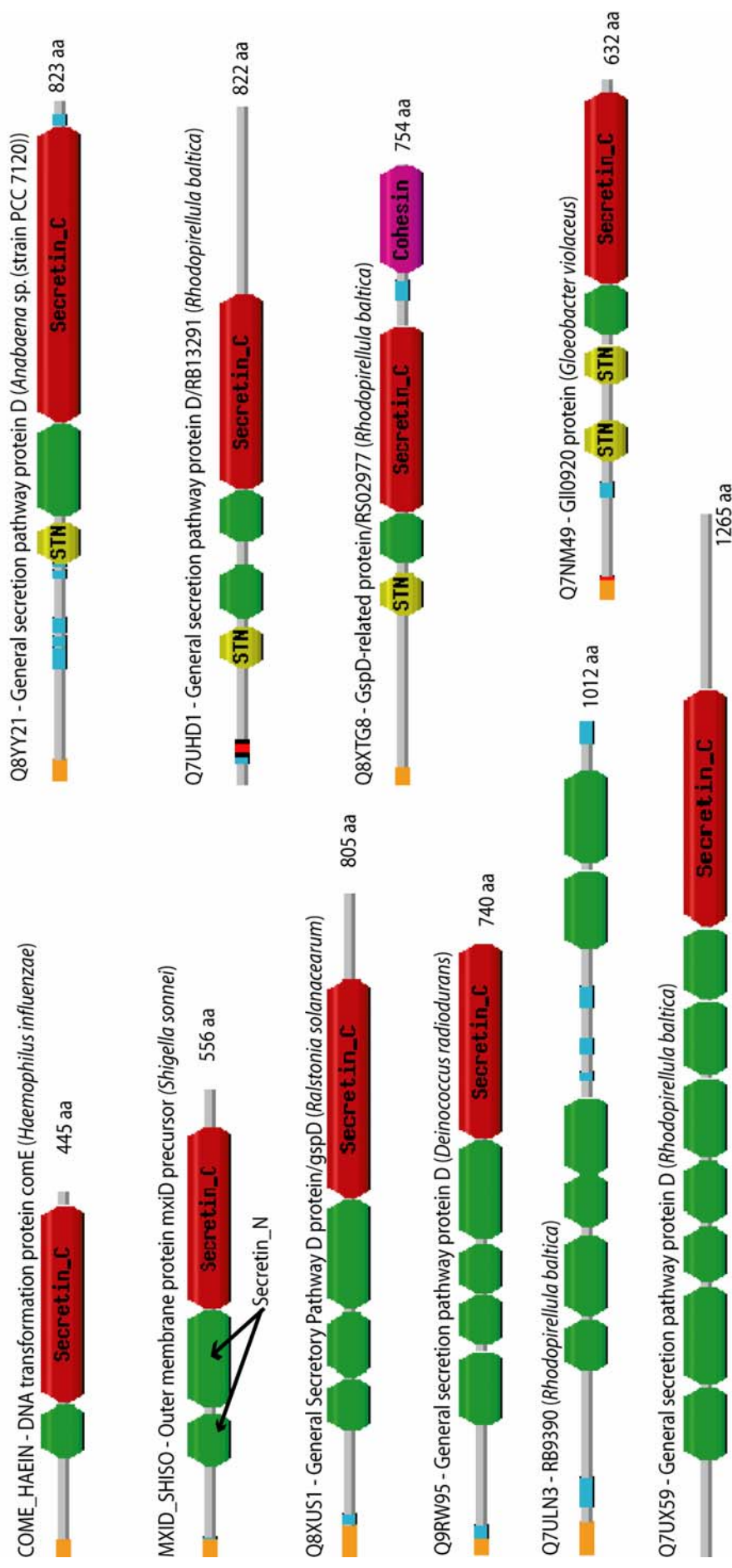


Figure 3.5: Secretin_N example architectures

There are some proteins that only contain Secretin_N domains (see Figure 3.5). Firstly *Rhodopirellula baltica* contains three proteins (UniProt:Q7ULN3, UniProt:Q7UET9, UniProt:Q7UU67) that contain Secretin_N domains and no other known domains. The function of these proteins is not obvious but *R. baltica* has an unusual proteinaceous cell wall structure, which contains no peptidoglycan. The *R. baltica* secretins also have an unusual six copies of Secretin_N. Binding domains also often appear in multiple copies in order to increase affinity for the binding partner.

Secondly, the NolW proteins of Rhizobium species also consist of a single Secretin_N domain. The Nol proteins make up the complex that defines host specificity in Rhizobial-plant interactions. Inactivation of NolW extends the host range of Rhizobium fredii strain USDA257 (Meinhardt, Krishnan *et al.*, 1993). These proteins are of considerable interest as the nitrogen-supplying nodules formed in the plant roots are critical importance in agricultural systems.

Although there is no direct experimental evidence, I postulate that the Secretin_N domains are critical to the correct and stable localisation of the Secretin channel in cell membrane, either through interactions with other membrane-associated proteins or through directly contacting the membrane.

Secretin N 2 (Secretin N-terminal domain; PF07655)

Another of the Secretin N-terminal domains, Secretin_N_2 is around 80 residues in length, contains a variable length serine rich region and is predicted to have an α/β fold (predicted using PROF). It is only found in a small number of Secretins in the Epsilon- and Gamma-proteobacteria that are involved in secretion of Mannose

Q7MA10/108-203
 Q7MA10/260-355
 Q7MHC4/148-242
 Q7WXX5/184-272
 Q9F533/193-274
 Q9K2G7/24-105
 Q9KUV9/145-248

T K T F K I N Y V G M D R S G V S N T E V S I S R D D G I N S S S S A L G S S Q G S S G S S F Q R S S V S G S K S G I N
 T K T F K I N Y V G M D R S G V S N T E V S I S R D D G I N S S S S A L G S S Q G S S G S S F Q R S S V S G S K S G I N
 T V T I P V D Y I Q F Q R S G R S L T S I V T G S V T S T G S S G S S A L G S S Q G S S N S N S G D N T T A S G G T R
 T R T F R M Y A . . F D D V N T V D S T V R S G M T T A A G I S G D G S G S T G Q N G S S G I S G D S G S K Q T
 T R S F P I T Y . . M D S N V A Y N S K V S G T M S S G S T G S S G G M T G D A S N T Q T
 T R T F Q F T F . . L N T N I T S N A S V T S G S T S M G T S G G S T N S S V S G D S S S S Q Q
 T V T I P V D Y L Q F K R T G R S L T S I T T G T I T N T D I N N S S S S I S S N S S D G S S S N S N S R R S D A R G G T E

Secretin_N_2_SS
 Secretin_N_2_SS

Q7MA10/108-203
 Q7MA10/260-355
 Q7MHC4/148-242
 Q7WXX5/184-272
 Q9F533/193-274
 Q9K2G7/24-105
 Q9KUV9/145-248

I K A E D G F R F W E G I Q A E I L A I L N R P G D S Y T L . . P Q A
 I K A E D G F R F W E G I Q A E I L A I L N R P G D S Y T L . . P Q A
 I E I T E S D F W P L L Q Q A V A G L I G S G K G Q S V V V T P Q A
 T S S E L K T S I L S D I E N S I N S M L T P S M G R M S L S R A T G
 T T V E M K S L Y N D L K S E V S M L T P G T G R M Y L S T G S L
 T T V Q S R N V Y D D M K K T L E T M V T P Q K G R F W L S A S T G
 I E T T N E S D F W P L L E K A V A Q L L G G S G G Q T V I V N P Q A

Secretin_N_2_SS
 Secretin_N_2_SS

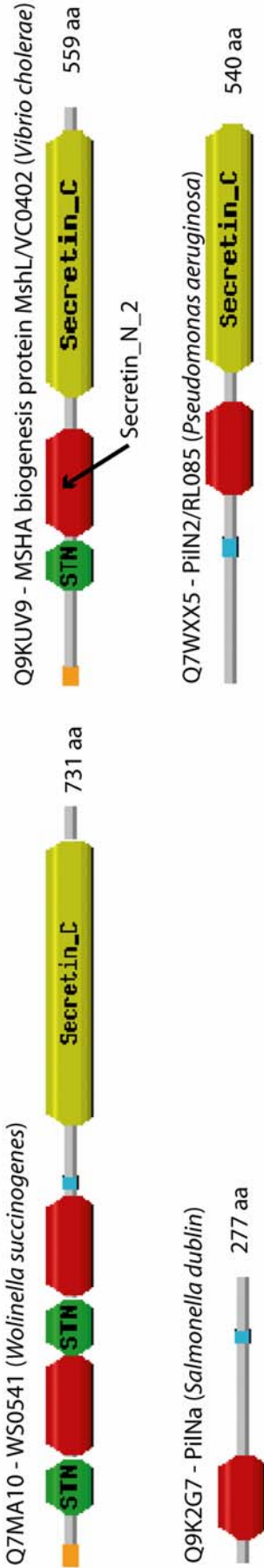


Figure 3.6: Secretin_N_2 alignment and architectures

Sensitive Haemagglutinin Type IV pili (MSHA). See Figure 3.6 for example architectures and an alignment.

The difference in function between Secretin_N and Secretin_N_2 domains is not clear. However, similarly to Secretin_N, there are proteins that consist of only the Secretin_N_2 domain; e.g. the *Salmonella dublin/typhi* PilNa protein (UniProt:Q9K2G7). These may associate with a Secretin to form a functional channel or to recruit specific complexes.

Reg_prop (Regulatory Protein Propeller; PF07494)

The conserved core of this repeat is around 25 residues long and is predicted to be a β -strand (using PROF), though the actual length of the structural element is probably longer. Between all the identified repeats there are large gaps – around 25-30 residues. This pattern of conservation is similar to that seen for the Ig-like He_PIG domains found in the WISP proteins of *Tropheryma whipplei* (see chapter 4.1); these also only have a short conserved core, even though in some individual proteins the repeat can be clearly identified as being about 100 residues long. In the Reg_prop repeats even the conservation of the core is fairly weak, with only a single residue showing any consistency – an aromatic residue near the end of the alignment (see Figure 3.7). These repeats normally occur in multiples of seven (see Figure 3.8) and show significant sequence similarity to β -propeller families, such as WD40 and PPQ, which often also consist of seven blades. β -propellers are involved in mediating a large variety of interactions (Pons, Gomez *et al.*, 2003). Reg_prop repeats are all found in regulators, mostly variants of two main architectures. Some of these regulators are hybrid two component regulators – or 'one-component regulators'; they

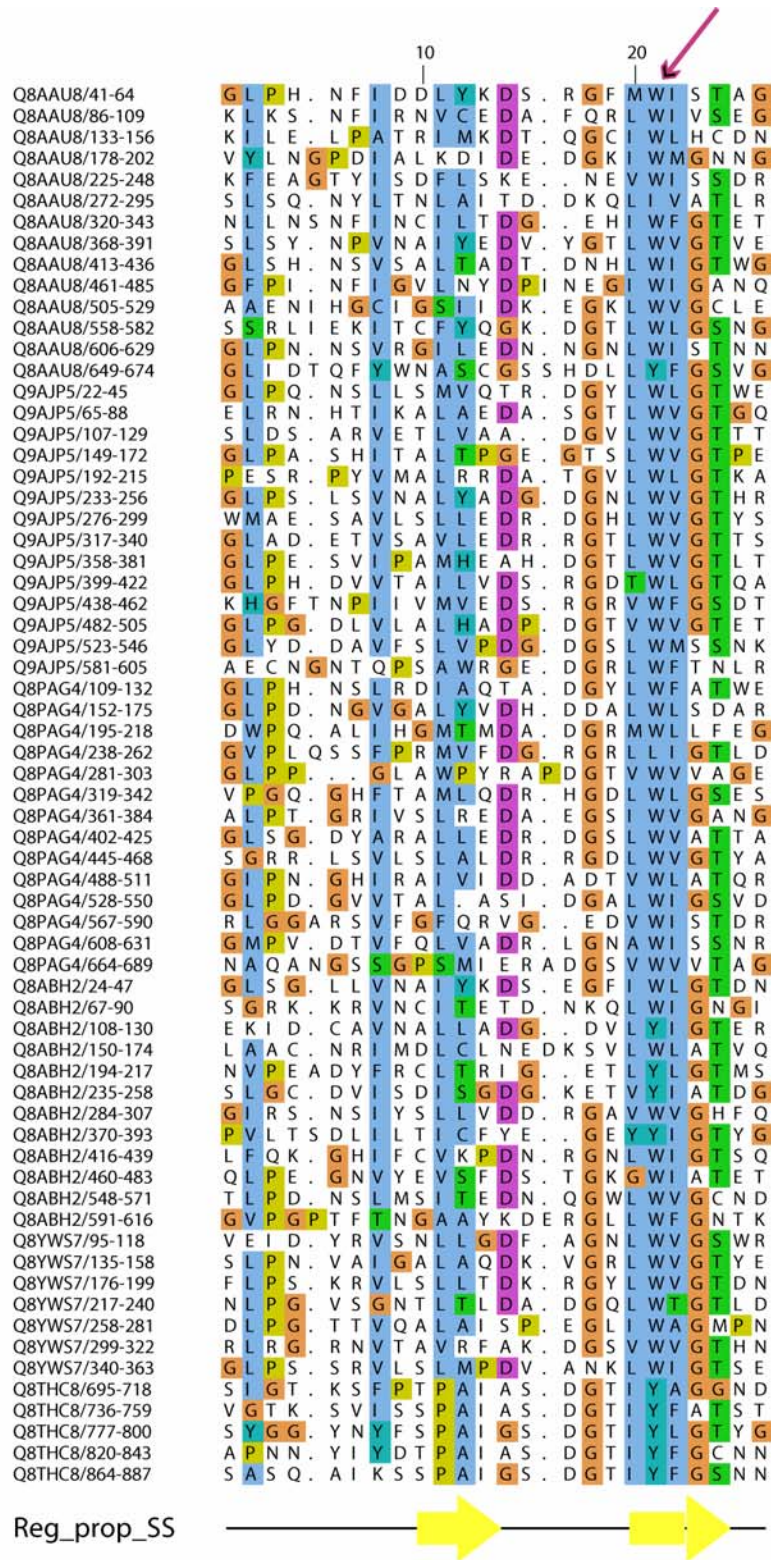


Figure 3.7: Example Reg_prop alignment
The purple arrow above the alignment marks a mostly invariant aromatic residue.

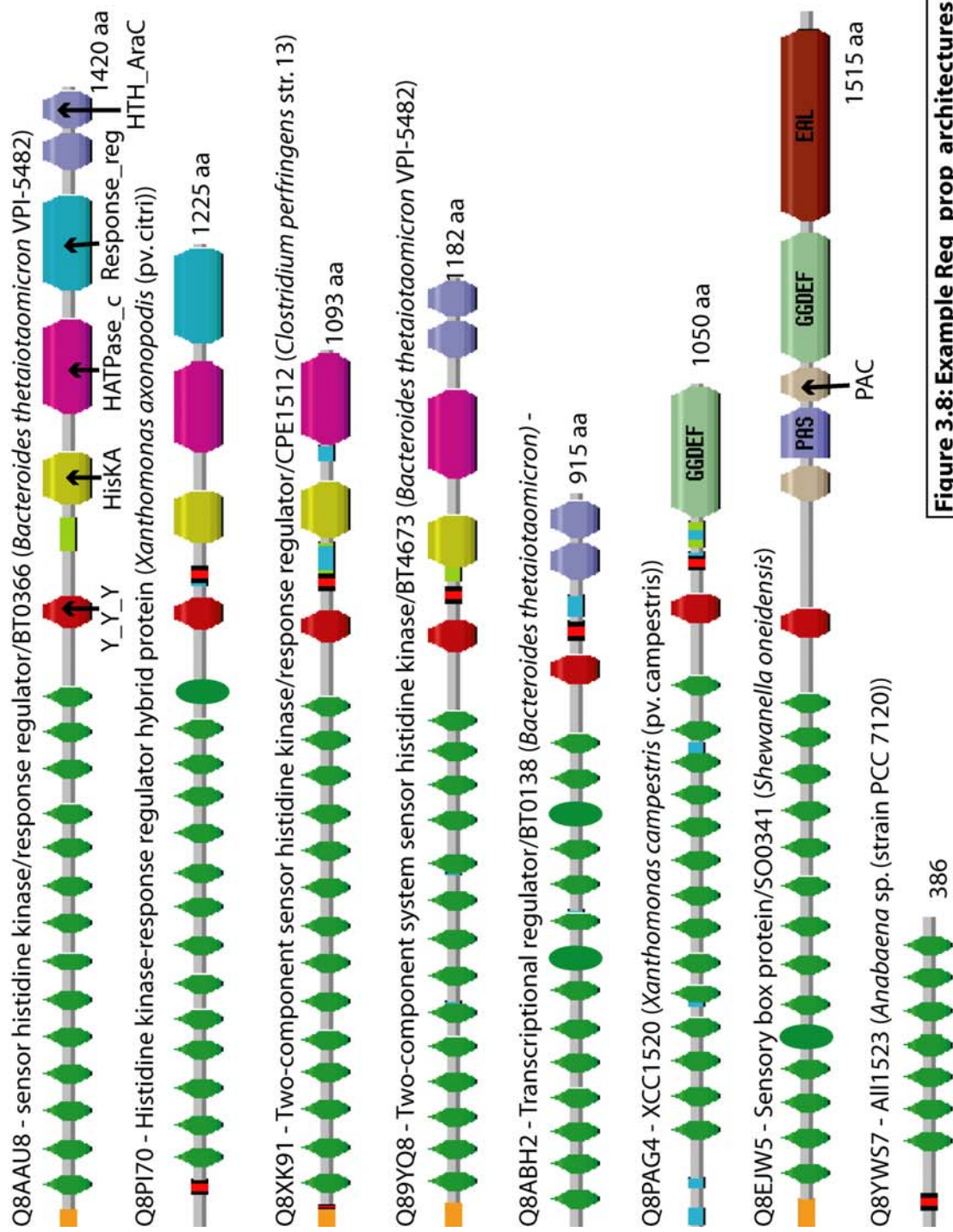


Figure 3.8: Example Reg_prop architectures

have both the signal receiver domains and the DNA-binding AraC-like Helix-Turn-Helix (AraC_HTH) domain. The others do not have the AraC_HTH domains. Both these types of regulator protein have the signal receiver domains at the N-terminus and the response modulator domains towards the C-terminus. So it is likely that these propellers bind a particular substrate and allosterically signal to the response modulator domains.

Y Y Y (Conserved Tyrosine Motif; PF07495)

This motif typically occurs in the hybrid 'one component regulators' that also contain the Reg_prop propellers (see above), at the C-terminus of the cytosolic portion of the regulator. It does also occur in a few proteins in multiple copies, sometimes by itself (e.g. UniProt:Q891H4) and sometimes with the peptidoglycan binding domain PG_binding_1 (PF01571; e.g. UniProt:Q97G63). The alignment (see Figure 3.9) highlights three conserved tyrosine residues and a glycine residue, which are likely to be the functionally important residues; it is not clear what this function is.

Its appearance as a single copy and as tandem repeats, suggests that it may form an independent stable structure, but its short length (40 residues) would seem to suggest otherwise. As discussed in chapter 1.3 this is right on the limit of the minimum domain size, unless there are some significant stabilising interactions. Visual examination of the alignment suggests that these do not include disulphide bonds. It is possible this domain may show a similar pattern of conservation to the Reg_prop repeats, with the structural domain being larger than the sequence domain.

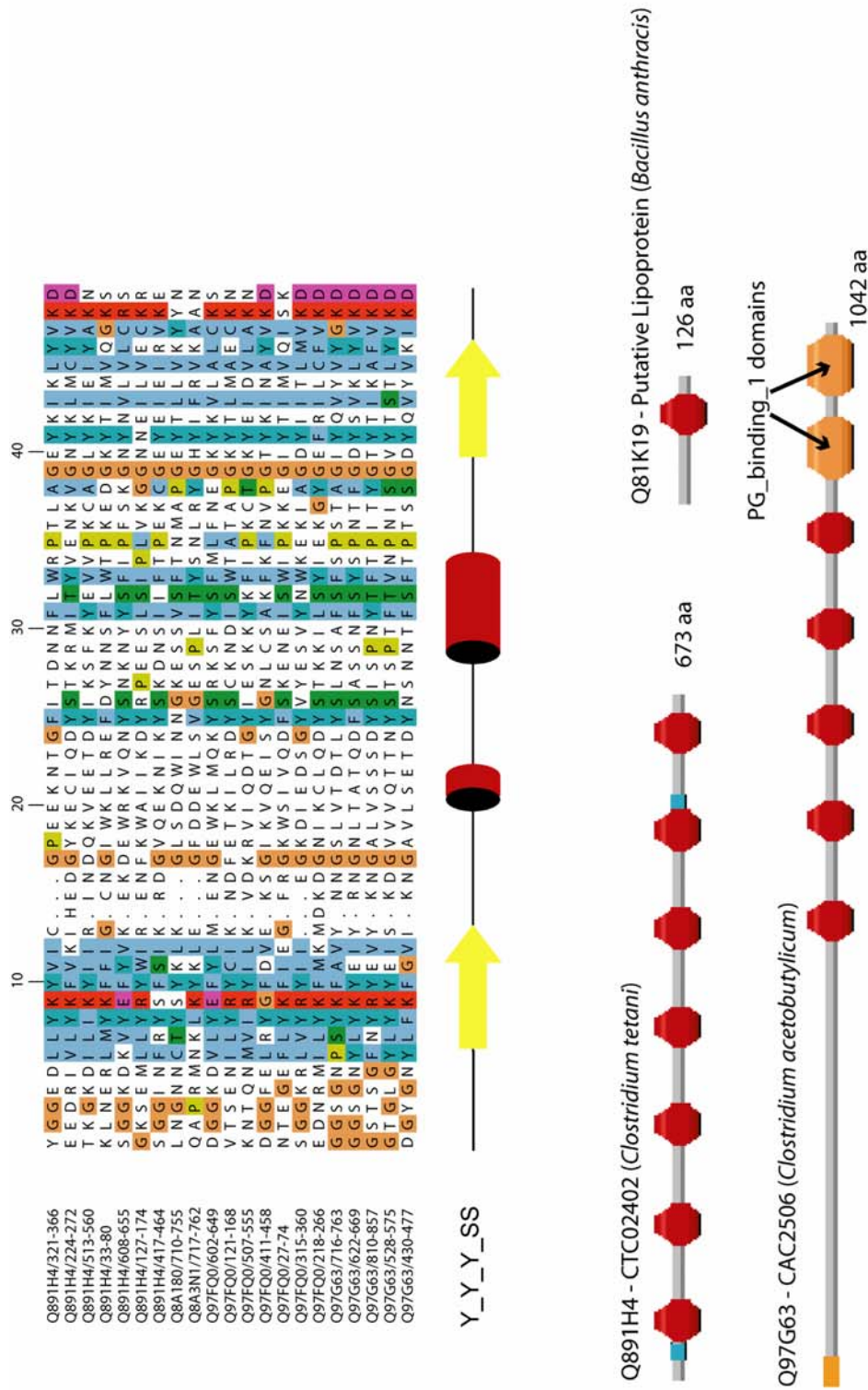


Figure 3.9: Y_Y_Y alignment and architectures

Secondary structure predictions are partially conflicting, but do confirm the existence of β -strands at the amino and carboxyl termini. In a couple of proteins it appears to overlap the SMART model of PKD (e.g. UniProt:Q8A241); however, I was not able to confirm this relationship and suggest that the SMART matches are spurious.

DUF1533 (Domain of Unknown Function 1533; PF07550)

This 60-70 residue predicted α/β (mostly β) domain is found in a small number of Firmicute proteins (see Figure 3.10). It is not obvious what the function of this domain might be, but it is found in conjunction with the NEAT domain (Andrade, Ciccarelli *et al.*, 2002), which is involved in iron siderophore import (see Figure 3.9 for architectures). This process is of critical importance in many pathogens, such as the human pathogenic Firmicutes.

Coat_X (Bacillus Coat Protein X domain; PF07552)

The Bacillales spore coats include two insoluble proteins CotX and CotV. CotV is composed of a single copy of this domain, whereas CotX contains two tandem repeats (see Figure 3.11). CotX appears to contribute around 30% of the insoluble fraction of the *Bacillus subtilis* coat, and so is likely to be a major component of the structure (Zhang, Fitzjames *et al.*, 1993). It does seem likely that CotX and CotV interact as they share domains, expression and cellular location, and combined together they may fulfil a structural role. The domain is around 60 residues in length and is predicted to form an α/β fold. Elucidation of the domain boundaries should aid structural studies of the Bacillales spore coat.

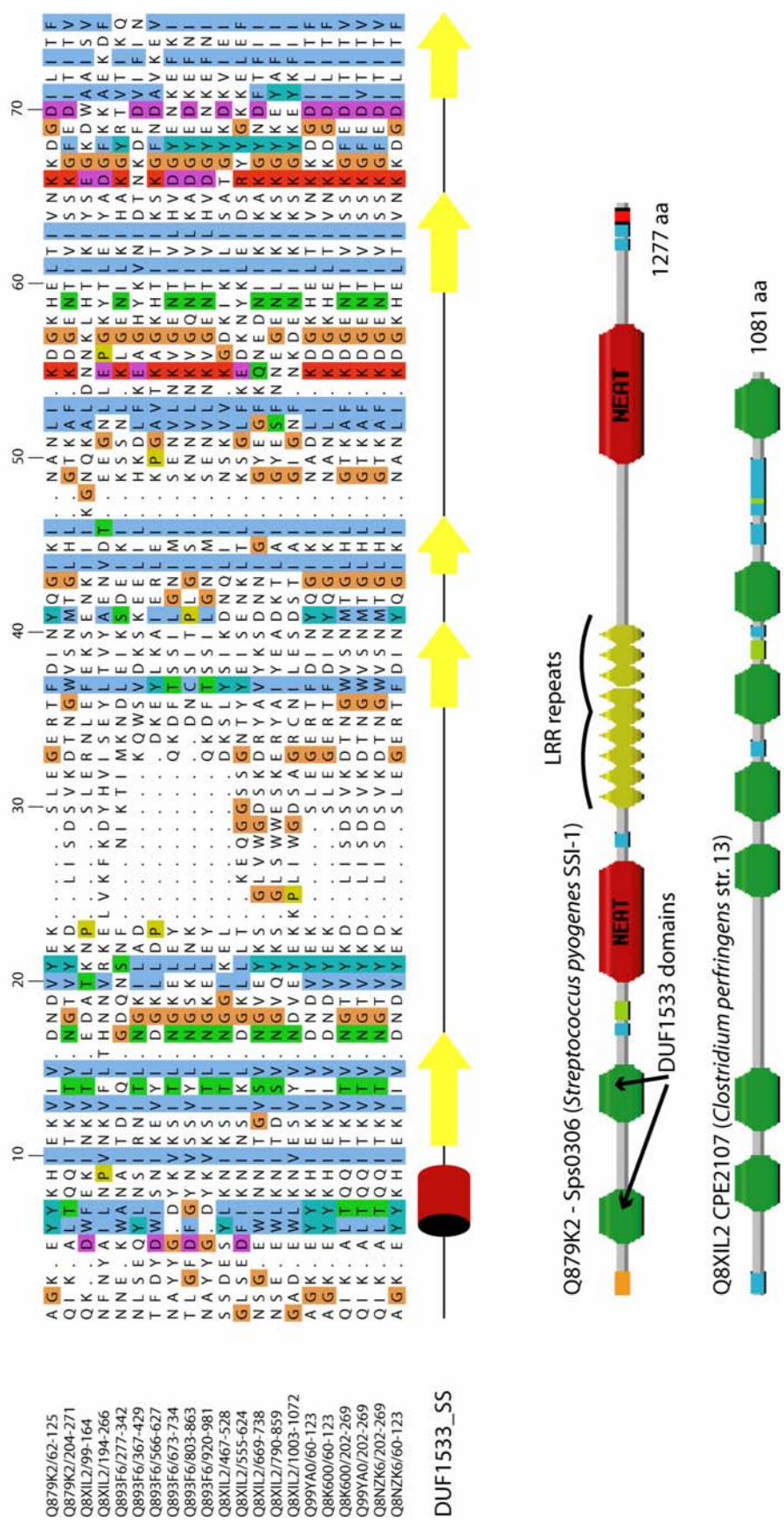


Figure 3.10: DUF1533 alignment and architectures

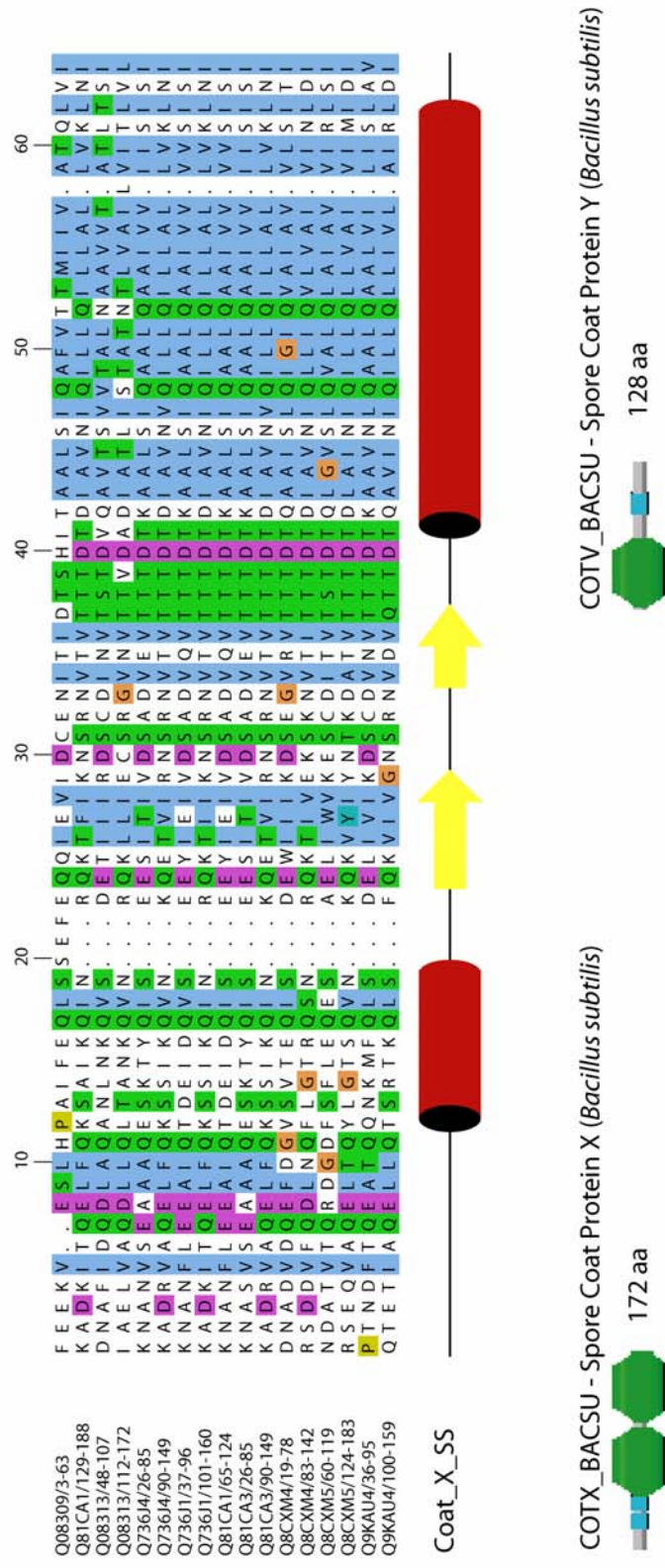


Figure 3.11: Coat_X alignment and architectures

Cleaved Adhesin (PF07675)

This domain seems to be limited to the periodontal pathogen *Porphyromonas gingivalis*, and more specifically, to a group of proteins that form the extracellular RgpA-Kgp virulence complex (Slakeski, Cleal *et al.*, 1999). These domains are cleaved from the precursor proteins and form part of the adhesins. Other domains found in these proteins include Peptidase_C25, a Plug domain, Formyl_trans_N and possibly FN3 (as predicted by SMART). It is possible that these domains are related to FN3, but the relationship is not clear beyond the overlap of SMART's FN3 (Fibronectin Type III domain) model and Pfam's Cleaved_Adhesin model. Secondary structure predictions suggest that Cleaved_Adhesin is mostly composed of β -strands, with a single α -helix – which does not contradict the possibility that these are divergent FN3 domains.

The occurrence of a Plug domain (Oke, Sarra *et al.*, 2004) is particularly surprising, as these domains are almost always found at the N-terminus of the TonB-dependant receptor channels, where they act as the plug or gate. See Figure 12 for architectures and alignment. These domains may form the scaffold for the virulence complex or they may recognise the host cell – which would correspond with the FN3 possibility. A third role is that they form part of the secretion apparatus for the formation of the RgpA-Kgp complex; this may account for the occurrence of the Plug domain.

FIVAR (likely NAG-binding motif; PF07554)

As of Pfam 14.0 this domain was found in 43 different architectures – hence its name “Found In a Variety of ARchitectures” (See Figure 3.12). The domain itself is around 60 residues in length, and highly divergent (see Figure 3.13); structurally it is

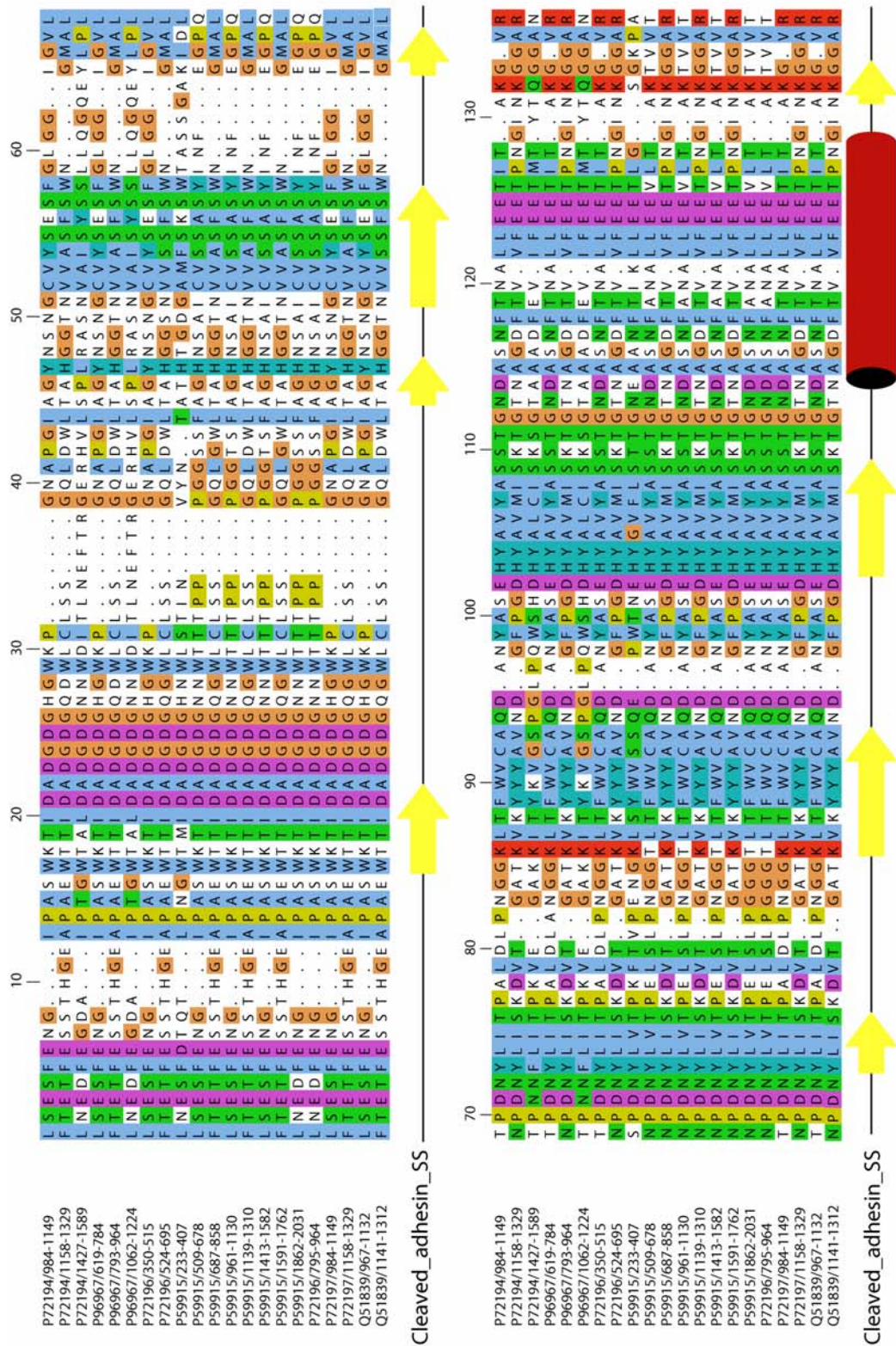


Figure 3.12: Cleaved_Adhesin alignment and architectures (Page 1)

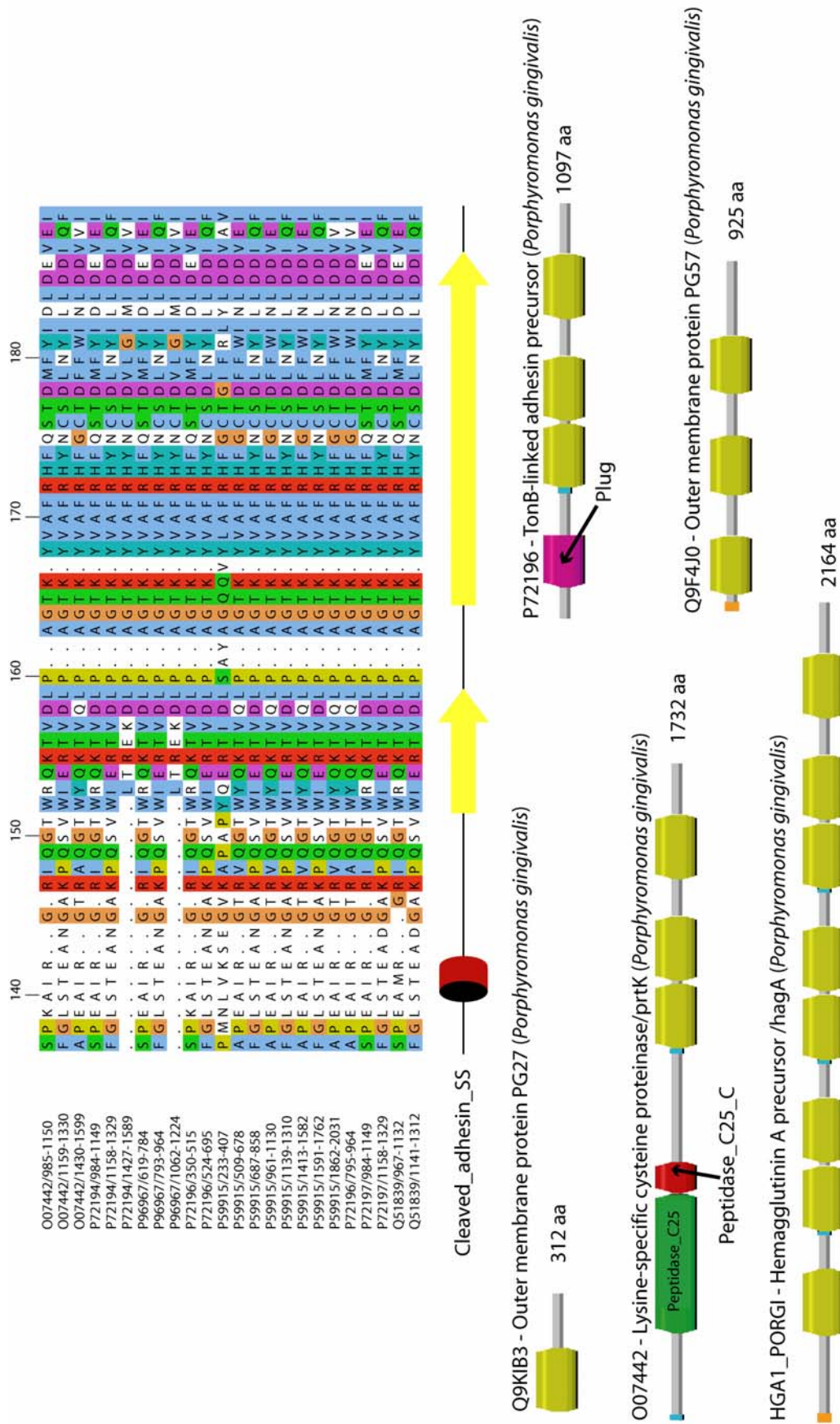
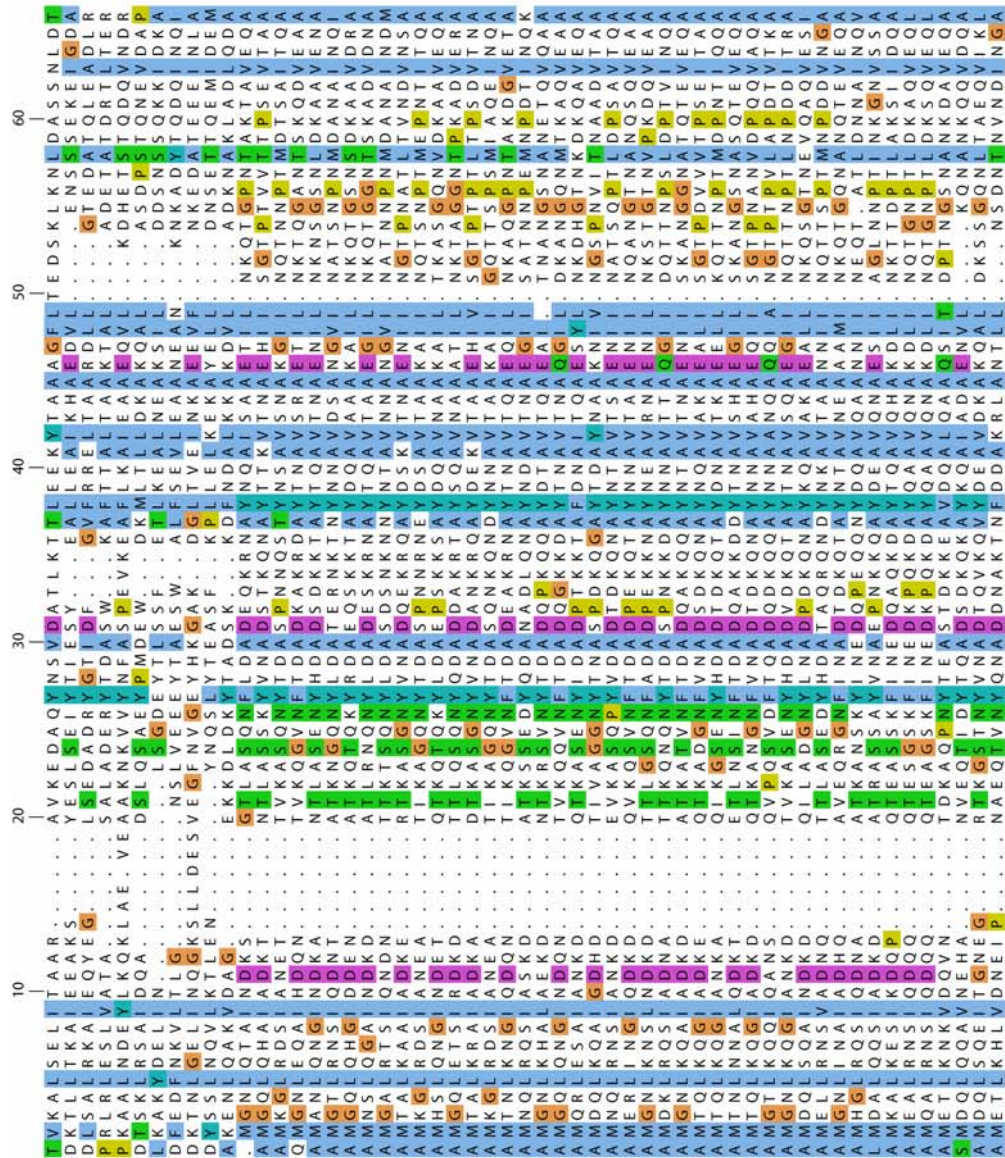


Figure 3.12: Cleaved Adhesin example alignment and architectures (Page 2)

predicted to be composed of several α -helices (using PHDsec). It occurs both as a single copy and as up to ten tandem repeats, with varying domains at the amino and carboxyl termini, which strongly suggests it is an independently folding unit. All the copies identified are found in the Actinobacteria and the Firmicutes except a few in the archaeal Thermococcaceae species.

Despite the enormous range of contexts there are some clear conserved themes that enable us to guess at its function. Most of the proteins are cell surface proteins – as evidenced by the N-terminal signal peptides and also from the occurrence of haemagglutinin domains such as Myco_haema and Big_2, Big_3 and Big_4. There are also many sugar binding and hydrolysis domains associated – for instance CBM_6 (Carbohydrate-Binding Module 6), G5, Glyco_hydro_43 (glycosyl hydrolase family 43), Hyaluronidase, Sialidase, and Peptidase_S8 (subtilase). As an example it is found in EndoD of *Streptococcus pneumoniae*, an endo-beta-N-acetylglucosaminidase that acts on complex asparagine-linked oligosaccharides (see Figure 3.14).

Bacterial cell-cell interactions are often mediated by polysaccharides, with cell surface proteins recognising and attaching to the sugars and also processing them. This type of activity is especially important in biofilm formation (reviewed in (O'Toole, Kaplan *et al.*, 2000)). FIVAR occurs in the same context as other polysaccharide processing and recognition modules – as mentioned above – and so it



- 005121/112-171
- 08XL5/1565-1617
- 069822/1459-1512
- 069822/1527-1578
- Q8C752/1227-1288
- Q8C450/816-867
- P26831/1363-1410
- P26831/1427-1479
- P26831/1498-1557
- 09A1R8/987-1038
- 0954K2/1603-1654
- Q931R6/1-38
- Q931R6/126-184
- Q931R6/252-310
- Q931R6/378-436
- Q931R6/504-562
- Q931R6/630-688
- Q931R6/756-814
- Q931R6/882-940
- Q931R6/1008-1066
- Q931R6/1134-1192
- Q931R6/1260-1318
- Q931R6/1386-1444
- Q931R6/1512-1570
- Q931R6/1638-1696
- Q931R6/1764-1822
- Q931R6/1890-1948
- Q931R6/2142-2200
- Q931R6/2268-2325
- Q931R6/2393-2451
- Q931R6/2519-2577
- Q931R6/2645-2703
- Q931R6/2771-2829
- Q931R6/2897-2955
- Q931R6/3023-3081
- Q931R6/3149-3207
- Q931R6/3275-3333
- Q931R6/3401-3459
- Q931R6/3527-3585
- Q931R6/3663-3711
- Q931R6/3779-3837
- Q931R6/3905-3963
- Q931R6/4031-4089
- Q931R6/4157-4215
- Q931R6/4283-4341
- Q931R6/4409-4467
- Q931R6/4535-4592
- Q931R6/4660-4718
- Q931R6/4786-4844
- Q931R6/4912-4970
- Q931R6/5038-5096
- Q931R6/5164-5222
- Q931R6/5290-5348
- Q931R6/5412-5471
- Q931R6/5666-5722

FIVAR_SS

Figure 3.13: FIVAR alignment

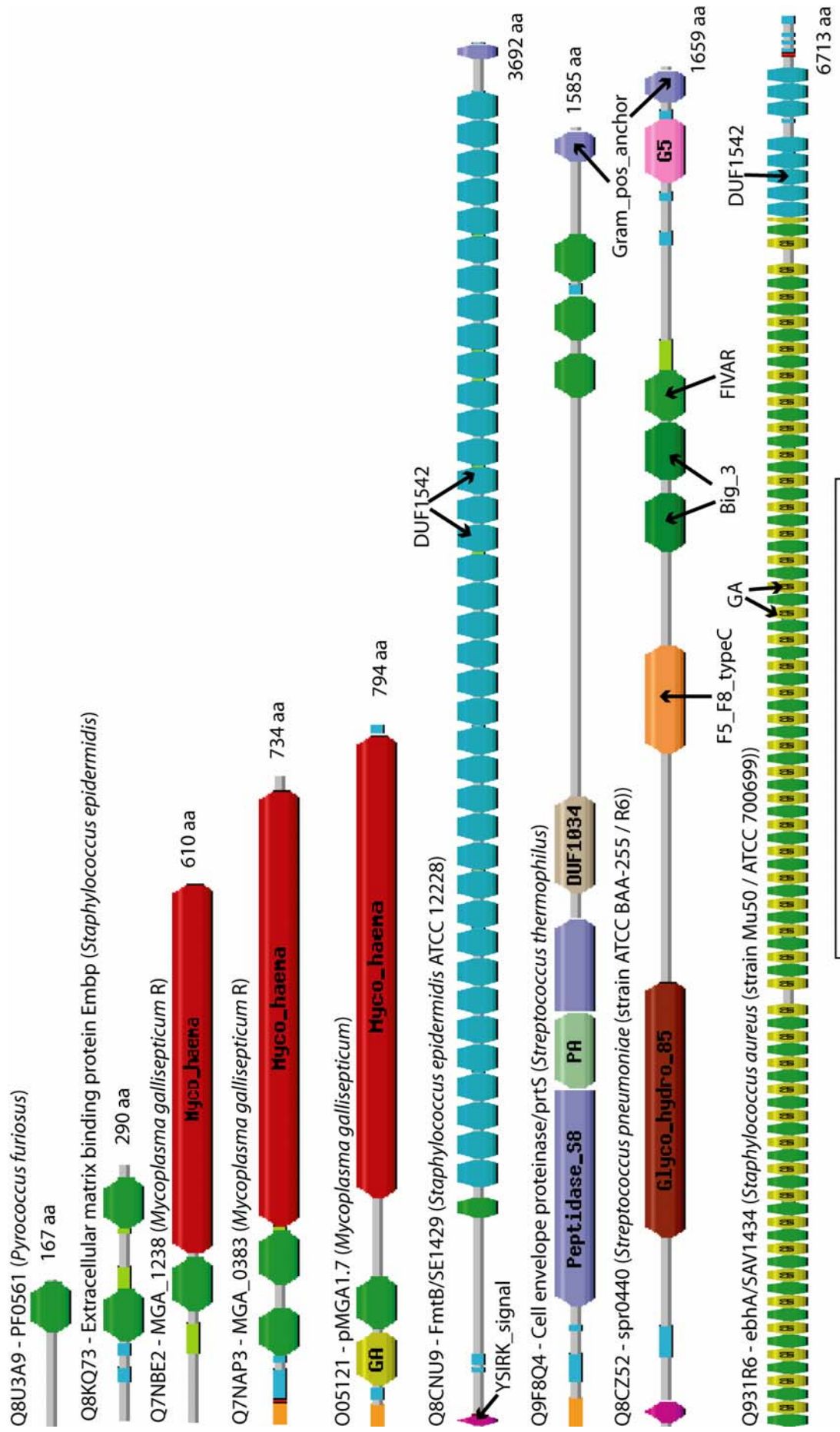


Figure 3.14: Example FIVAR Architectures (Page 1)

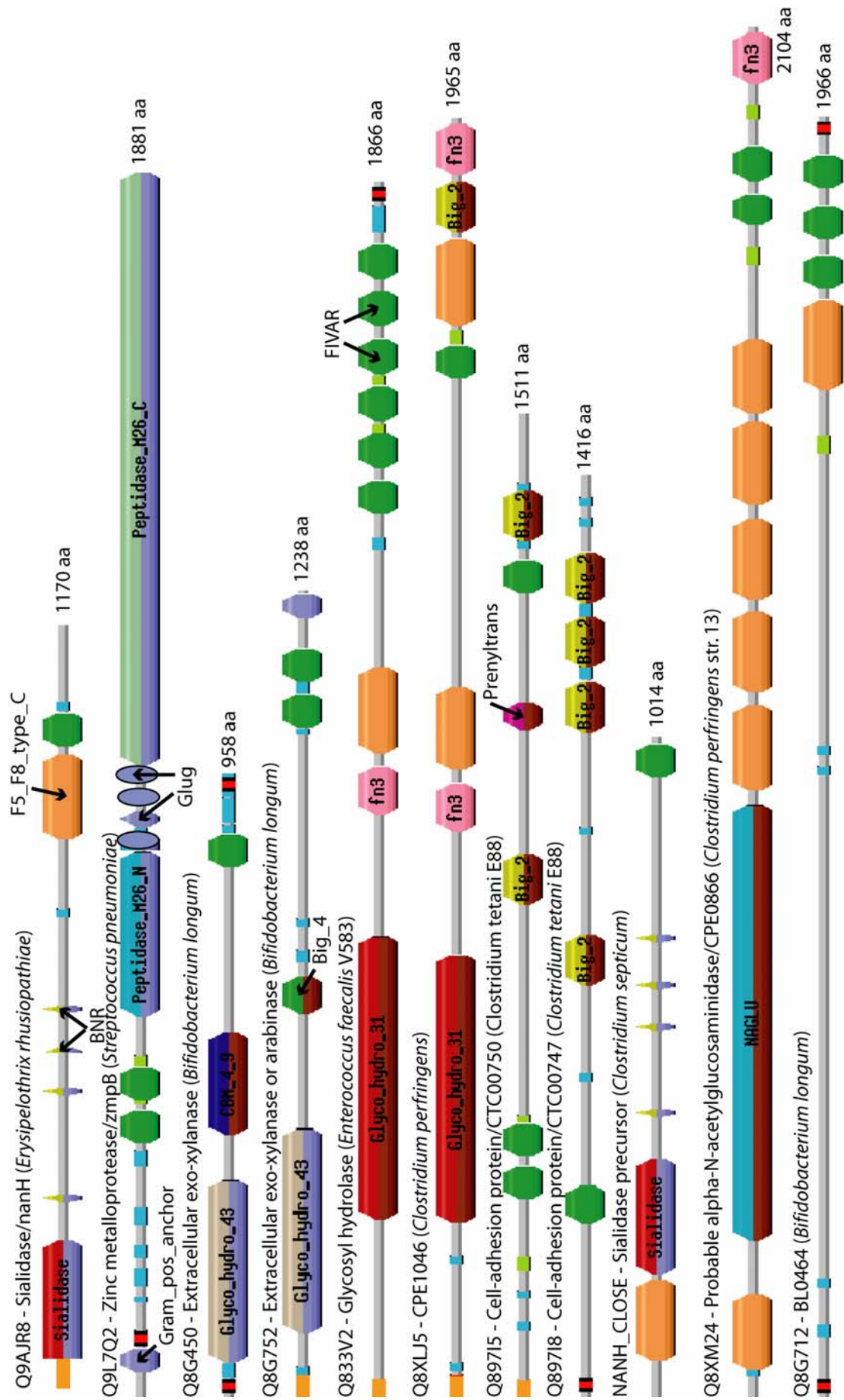


Figure 3.14: Example FIVAR Architectures (Page 2)

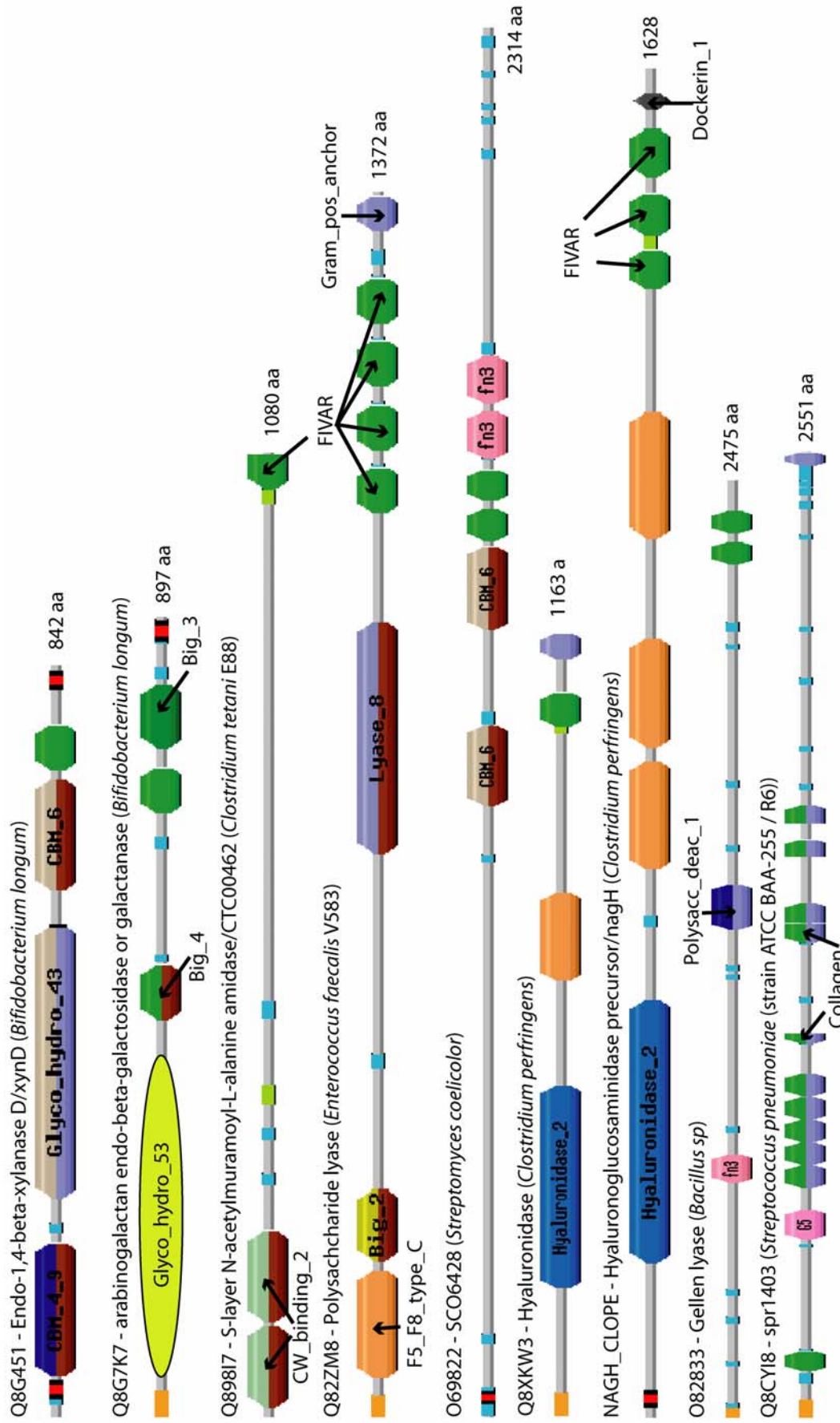


Figure 3.14: Example FIVAR architectures (Page 3)

seems likely that it also is a sugar recognition motif. Most of the processing enzymes identified metabolise N-acetylglucosamine (NAG), which the G5 domain is also thought to bind (unpublished observations: A. Bateman, S. Bentley & C. Yeats), and so it may be that FIVAR does as well. The need for a different module may come from recognising different bonds or slightly different polymer structures. Some of the proteins it is found in are noted for being methicillin-resistance factors - for instance *Staphylococcus aureus* FmtB (UniProt:Q99QR6). Disruption in this protein causes *S. aureus* both to produce an altered cell wall peptidoglycan and also lowers its resistance to methicillin (Komatsuzawa, Ohta *et al.*, 2000). Adding NAG to the growth media restores resistance; this implies that FmtB is in some way involved in the acquisition or synthesis of NAG.

Another line of evidence in support of it binding to a form of N-acetylglucosamine comes from the architectures of the IgA1-specific metallo-endopeptidase M26. IgA1 prevents the adhesion of bacterial cells to mucosae and subsequent colonization; to counter this, the Streptococcae encode an IgA1-specific peptidase. Normally these proteins have a G5 domain near the N-terminus of the M26 peptidase unit. In one instance the G5 domain has been substituted by two FIVAR domains, suggesting functional equivalence.

Although further structural and mutagenesis studies are required to fully understand the function of the FIVAR domain, it is clear from the huge range of architectures and consistent themes that it is an important contributor to the cell wall structure and cell-cell interactions in this group of Gram-positive organisms. It gains particular interest from many of these bacteria being animal pathogens.

FlaE (Flagella hook domain; PF07559)

This region is generally around 100 residues in length and contains several conserved aromatic and glycine residues. It is predicted to be composed of β -strands, and perhaps forms a β -helix (see Figure 3.15). It is found in flagella hook proteins (FlgE), which form the filamentous rod that extends from the bacterial cell surface. Although found in a few contexts it is not clear what the nature of this family is. It could either fulfil a structural role or recruit other factors.

Subsequent to the identification of this domain and its analysis, a portion of the *E. coli* FlgE protein was crystallised (Samatey, Matsunami *et al.*, 2004). This portion included the predicted FlaE domain (residues 169-282). Analysis of the structure revealed a domain (named 'D2') that extended from 145-284 – very similar to the predicted domain boundaries, and all the secondary structure elements were found in the predicted domain. This domain was found to be an eight stranded β -barrel. The secondary structure of the domain has been included beneath the sequence alignment (Figure 3.15) for comparison to the predicted structure. Hence, this family provides another blind test as to the accuracy of the domain boundary predictions and shows that they probably approximately correct.

Glug (Short G-G-L-hyd-G repeat; PF07581)

The Glug repeat is disparately distributed across the eubacterial kingdom, except for a small family in the eukaryote *Giardia lamblia*, a protein in the archaea *Methanosarcina acetivorans*, and one in the algal virus *Ectocarpus siliculosus*. The repeat is about 25 residues long and contains a conserved hyd-G-G-L-hyd-G motif

(where 'hyd' is hydrophobic), from which the name is derived (see Figure 3.16). It is found in secreted and cell surface associated proteins, in association with the IgA1-specific metallo-endopeptidase M26, haemagg_act (PF05860), and the nickel chelating CbiX domain (see Figure 3.17). Secondary structure prediction suggests that it forms an all- β fold. The repetitive and short nature of Glug is reminiscent of the Fil_haemagg repeat (See Chapter 5.2), which forms adhesive regular filaments that coat the cell. Similarly Fil_haemagg is also composed of β -strands and it forms a β -helix; hence, by analogy Glug may also form a helix. However, this is certainly not definitive.

O30404/601-729
P16322/169-283
P35806/324-471
P50610/472-598
P57422/179-285
P75937/169-282
Q7VJR0/493-615
Q83RT3/410-523
Q83WMS/588-716
Q884Z9/176-292
Q884Z9/324-436
Q885A0/176-322
Q914P9/174-343
Q9K010/178-315
Q9RQB6/194-343

SEQLKLSAF...SAGLEIYDS
FVSDADSYNKG...GTVTVYDSDS
SKGVKPDFE...VQIPLSYDSDS
ASGKFTHTAT...HATSI...DVIYDSDS
SKPEYMTY...ISINYK
FASNAADSYNKG...GVSIVFSDS
SDFIQMPSY...QTNVEIYDSDS
FASNAADSYNKG...GVSIVFSDS
SEQLFLSAF...SAGLEIYDSDS
FDPAATYTRAI...GRTIQDA...GSGTGA...GAVP...AADHELKQYFVK
FDSLDTKYSN...DMFANAI...FDSDS
FADDTKYSN...MKYS...TPIYDSDS
FAPSDAATY...NSSSSLG...IYDSDS
FDITDPE...TYNRTTSS...TIYDSDS
AQILES...TWSTEFKVYDSDS

FlaE_SS

O30404/601-729
P16322/169-283
P35806/324-471
P50610/472-598
P57422/179-285
P75937/169-282
Q7VJR0/493-615
Q83RT3/410-523
Q83WMS/588-716
Q884Z9/176-292
Q884Z9/324-436
Q885A0/176-322
Q914P9/174-343
Q9K010/178-315
Q9RQB6/194-343

...GEP...TNI...IVGT...ARF...NND...GSL...AS...YT...P...RT...INF...S...PIN...
...AAT...PT...AST...L...K...F...NE...NG...L...L...E...S...GG...T...V...N...I...T...T...GT...
...N...NG...Q...I...S...T...G...I...E...F...T...D...G...K...L...K...N...T...G...S...
...A...A...P...N...V...F...E...G...G...R...L...H...F...N...D...G...S...L...A...G...M...N...P...L...L...Q...F...D...P...K...
...D...D...K...E...T...I...K...N...S...F...D...L...T...F...N...D...D...G...E...L...T...S...D...N...V...F...N...I...T...S...K...D...
...I...M...V...S...P...D...Y...V...V...H...T...I...E...F...D...A...D...G...N...P...V...A...E...P...V...L...A...F...G...D...K...
...N...S...I...A...K...T...A...T...L...E...F...N...A...G...T...L...V...D...G...A...M...A...N...I...A...T...G...A...
...G...E...P...N...N...I...V...G...T...A...R...E...F...N...D...G...S...L...A...N...Y...S...P...K...N...I...N...F...S...P...N...
...N...S...V...A...P...L...H...V...I...L...E...Q...K...P...N...G...S...L...S...Y...S...G...N...S...Q...H...
...D...S...T...E...P...L...T...A...N...I...V...F...D...S...G...S...V...R...L...T...G...S...G...
...T...T...T...A...A...D...K...N...D...L...T...F...D...S...G...N...L...V...A...A...P...G...T...V...
...S...N...K...T...P...M...S...F...N...V...T...F...D...A...S...G...O...M...T...S...V...R...A...P...
...V...V...G...D...A...A...S...P...T...G...V...G...H...T...M...R...F...N...D...G...T...L...S...L...N...N...G...Q...P...I...V...T...E...P...
...R...V...G...I...G...T...T...D...G...V...Q...N...S...F...I...V...R...E...D...N...N...G...H...L...A...S...V...T...D...A...G...N...V...T...S...P...[18]A...G...A...P...T...R...H...T...E...D...V...N...[6]S...K...N...T...I...Q...F...S...D...K...S...T...T...K...A...Y...E...Q...D...D...G...Y

FlaE_SS

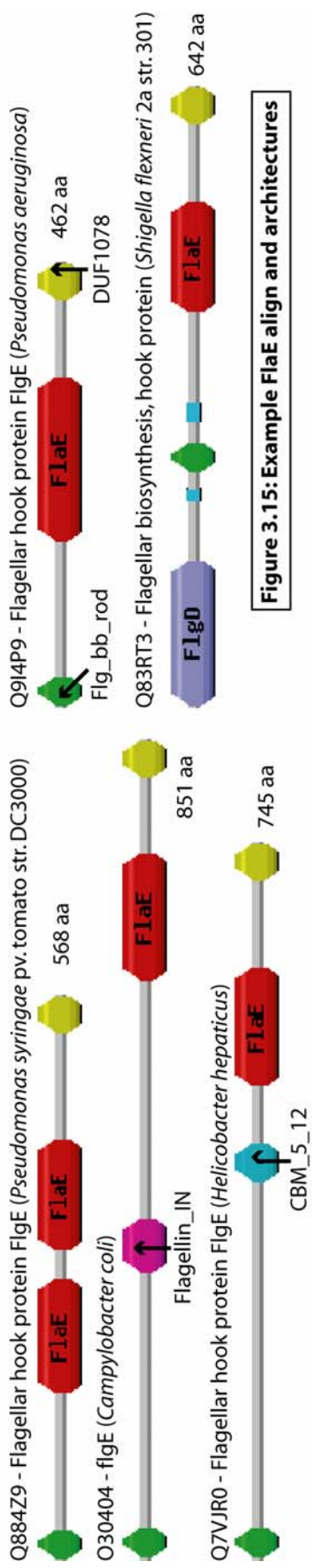
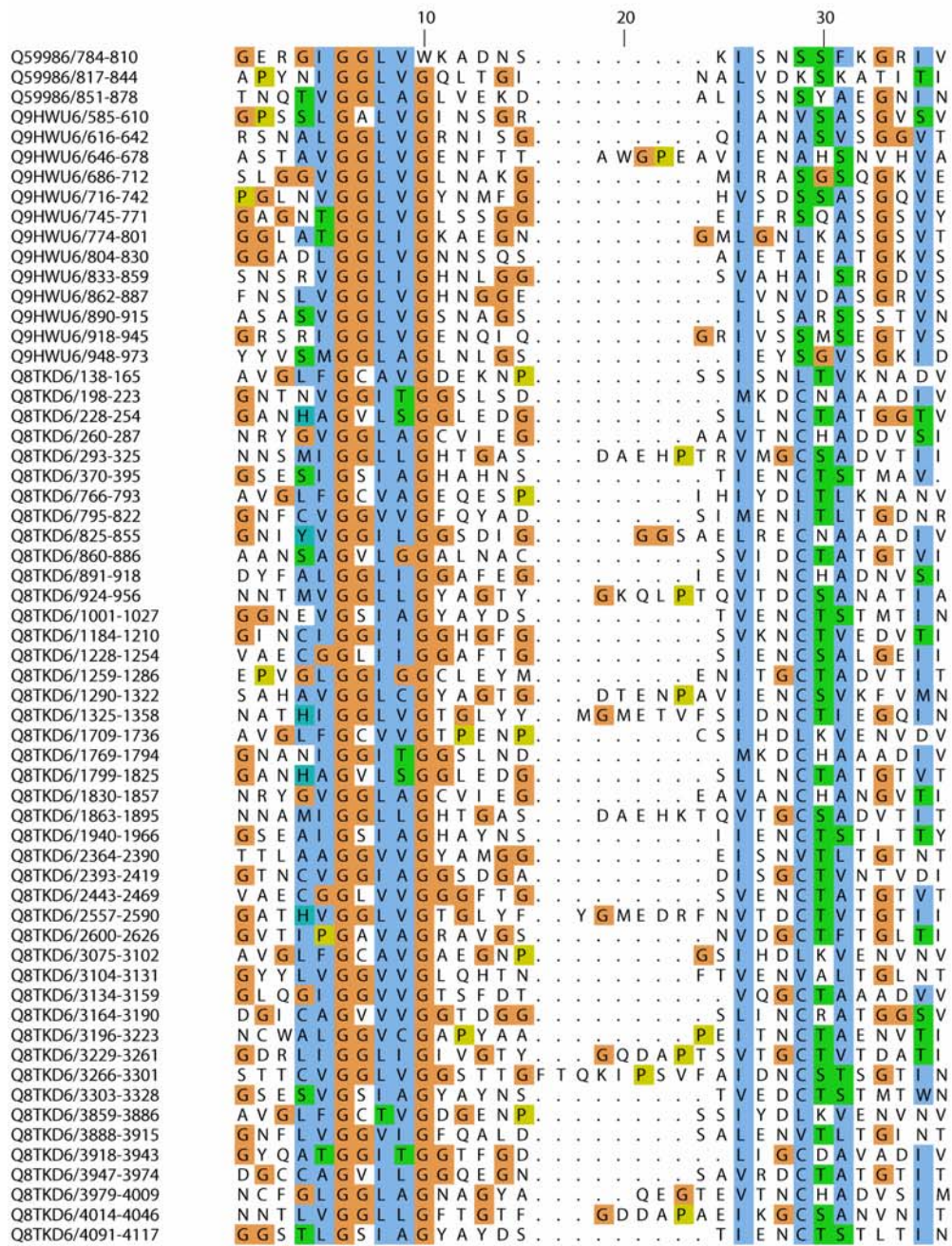


Figure 3.15: Example FlaE align and architectures



Glug_{SS}



Figure 3.16: Example Glug repeat alignment

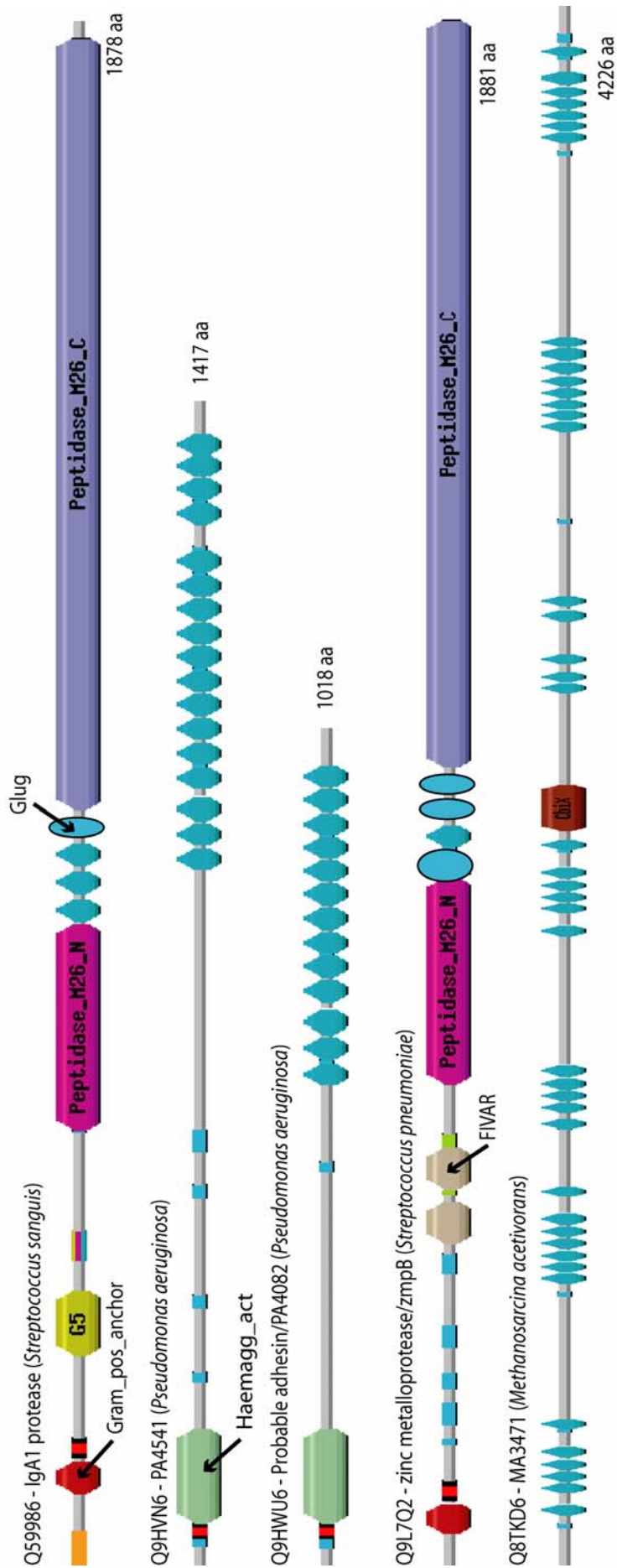


Figure 3.17: Example Glug repeat architectures

3.3.2 Domains Identified Through Protein Clustering

Coat_F (Coat protein F; PF07875)

This domain is mostly found in the Firmicutes, though the Proteobacterium *Ralstonia eutropha* has it as well, and occurs in multiple copies in the Bacillales genomes. Most of the species it is found in appear to have single copy of a two domain Coat_F protein and several copies of a single domain Coat_F protein (see Figure 3.18). Between related species the Coat_X gene copy number can be highly variable; for instance all of the Clostridiaceae have only a single Coat_F protein, except *Clostridium acetobutylicum*, which has nine. The variety of architectures, particularly within a single genome is reminiscent of the Coat_X proteins identified in the *S. coelicolor* hunt. Like Coat_X, Coat_F proteins contribute the spore wall. It is approximately 60 residues in length and is predicted to form an α -helical fold (PROF). The alignment shows that there is very little sequence conservation, and no residues are entirely conserved. There is a short motif in the centre that may be functionally important, possibly an interaction or attachment site. Like Coat_X, I would suggest that Coat_F forms a structural component of the spore coat; the variety in gene copy number may reflect some adaptability in the cell wall formation.

CTnDOT_TraJ (Conjugative Transfer Protein J; PF07863)

This family is currently only found in *Bacterioides thetaiotamicron* (5 proteins) and *Porphyromonas gingivalis* (9 proteins). It is an approximately 60 residues domain with a predicted α/β fold (see Figure 3.19). Somewhat surprisingly the CTnDOT_TraJ proteins in *B. thetaiotamicron* have a different architecture to the *P. gingivalis* proteins. All the former species' proteins are around 340 residues in length, have a long N-terminal region containing 5 transmembrane helices, and the

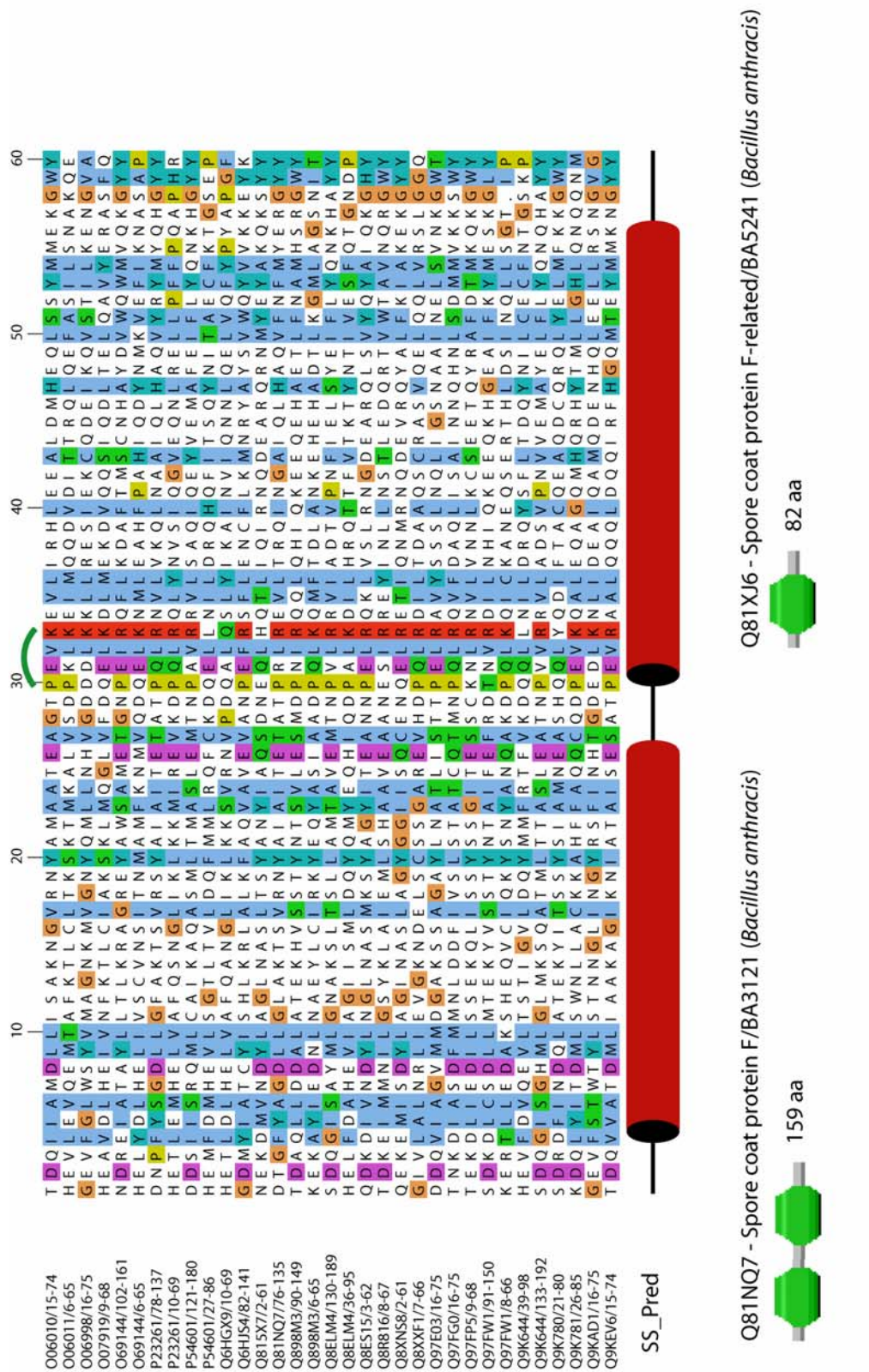


Figure 3.18: Example Coat_F alignment and architectures
 The green bracket above the alignment marks the most conserved region of the family

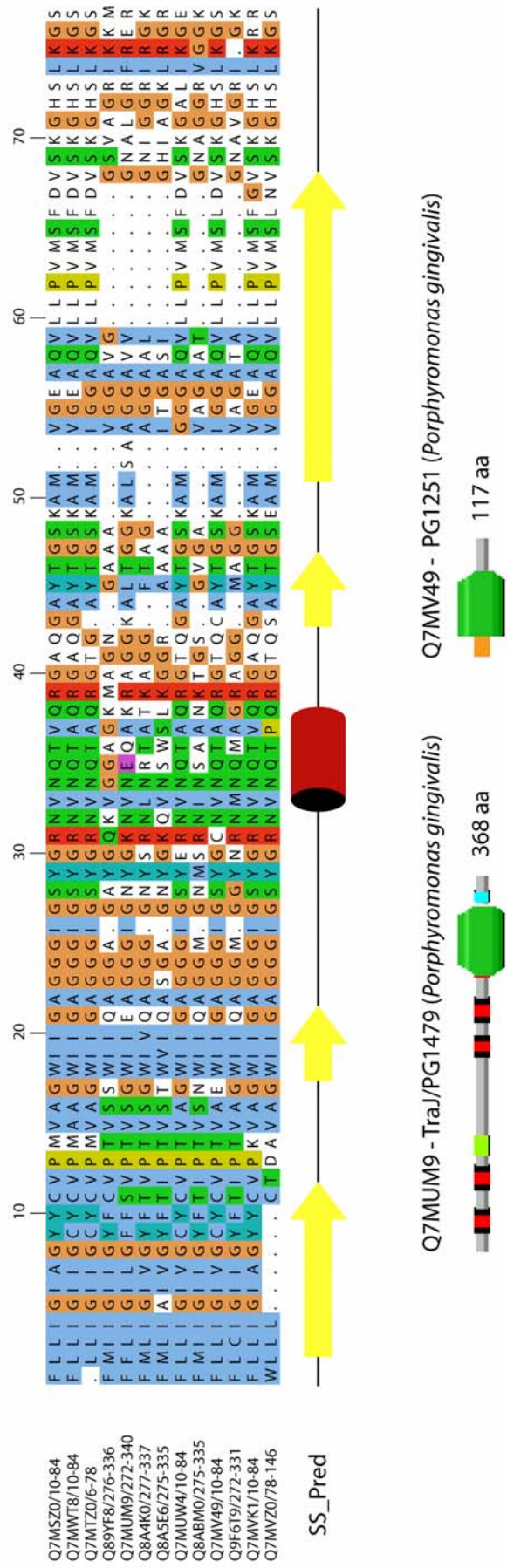


Figure 3.19: Example CTnDOT_TraJ alignment and architectures

CTnDOT_TraJ domain at the C-terminus; in contrast the latter has one protein of this type and 8 of around 100 residues in length with an N-terminal signal peptide (see Figure 3.19).

The conjugative transposons (CTns) of the Bacterioides are believed to be important in the distribution of antibiotic resistance between these species (Whittle, Shoemaker *et al.*, 2002); whilst normally they are commensal gut dwellers, some strains have pathogenic capabilities. *P. gingivalis*, as discussed in Cleaved_adhesin above, are periodontal pathogens and if they have the same type of conjugative transposons, they will be able to distribute antibiotic resistance genes by the same methods. Hence understanding the mechanisms and components of these transfer systems is important in preventing the wide distribution of antibiotic resistance amongst these pathogens.

Conjugative transposons use a specialised pilus to attach and transfer DNA to another bacterium. The consistent identification of signal peptides and transmembrane regions in the CtnDOT_TraJ proteins indicates that they are involved in this extracellular structure; the precise role is not clear, but they could form part of the pilus, perhaps for recognising another bacterium or perhaps as a structural component.

Dabb (Stress responsive dimeric α/β barrel; PF07876)

This family is disparately distributed across the kingdoms of life, with copies found in plants, fungi, most eubacteria, and the some of the euryarchaea; however, it is fairly divergent (average identity around 20%) and may be more widely represented but current searches could be limited by the composition of the sequence databases. Most of the proteins it is found in consist solely of a single Dabb domain, except a plant

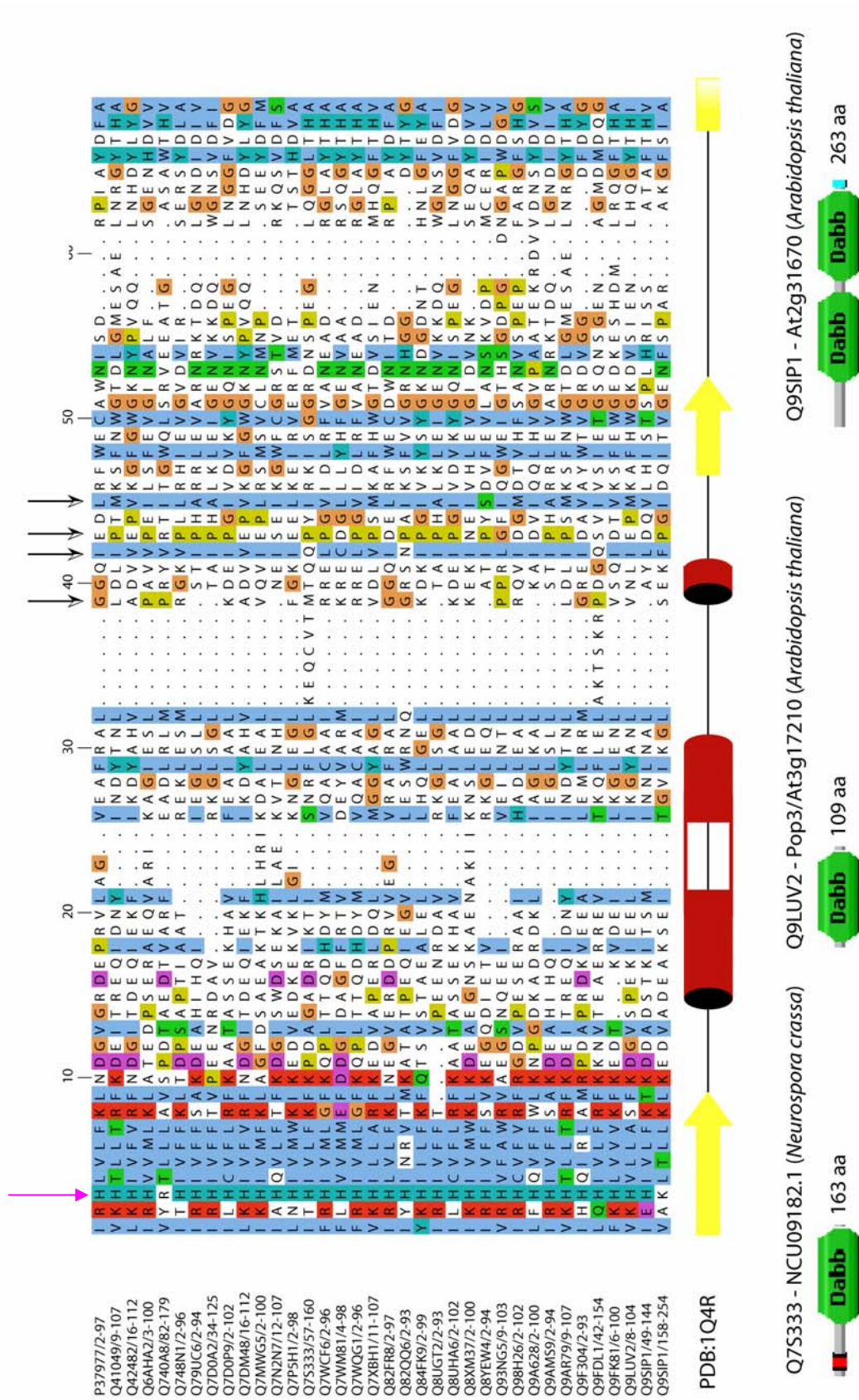


Figure 3-20: Example Dabb alignment and architectures (Page 1)
 The residues that contact a magnesium ion in PDB1Q4R are marked by the black arrows above the alignment. PDB:1Q4R corresponds to UniProt:Q9LUV2

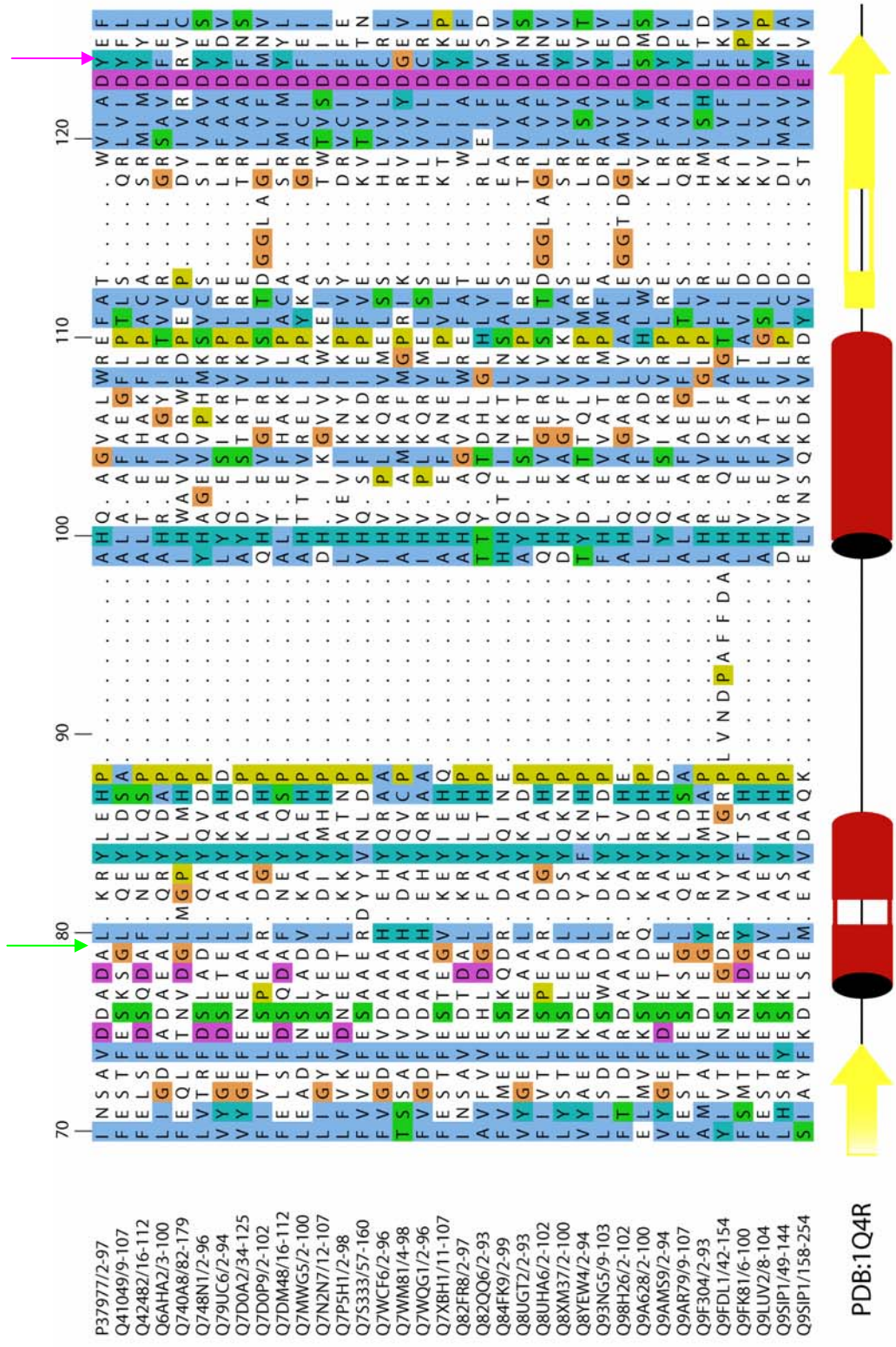


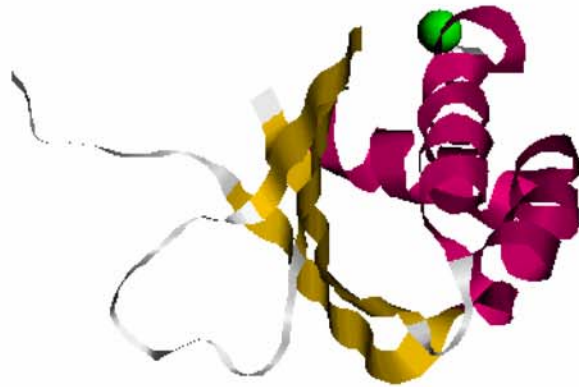
Figure 3.20: Example Dabb alignment and architectures (Page 2)

sub-family, which consist of two Dabb domains. There is also a single occurrence of this domain at the C-terminus of an F_bP_aldolase (fructose 1,6-bisphosphate aldolase) domain in *Hydrogenophilus thermoluteolus* (Swiss:Q9ZA13; see Figure 3.20). The domain forms half of an α/β barrel of approximately 200 residues (see Figure 3.20).

Mostly these proteins have not been well studied or characterised, despite the solution of the three dimensional structure. These proteins have been implicated in recovery from salt stress in plants – the Pop3 protein from *Populus balsamifera* (Gu, Fonseca *et al.*, 2004). The structure comes from one of the *Arabidopsis thaliana* POP3 homologues, but the molecular function of this protein is not specifically known (Lytle, Peterson *et al.*, 2004). Resolution of the structure of this protein found that it forms a homodimer that folds into an α/β barrel (see Figure 3.21). This fits with the discovery of the duplicated domains in some plant proteins – which may form a monomeric barrel.

To some extent it is surprising that this is not the norm as two copies are required to form the structure. Having to use two peptide chains to form a functional protein may allow the host cell to translate it without activating it, hence giving it a high degree of control over the function of the Dabb proteins. This fits with finding them in stress responses, when a plant may need to implement a rapid correction in the cytosolic conditions and may not have time to start up transcription. Structure 1Q4R shows two Mg^{2+} ions in complex with the structure, one in each half of the structure. The residues that coordinate the ions are marked in Figure 3.19. The function of these ions

A.



B.

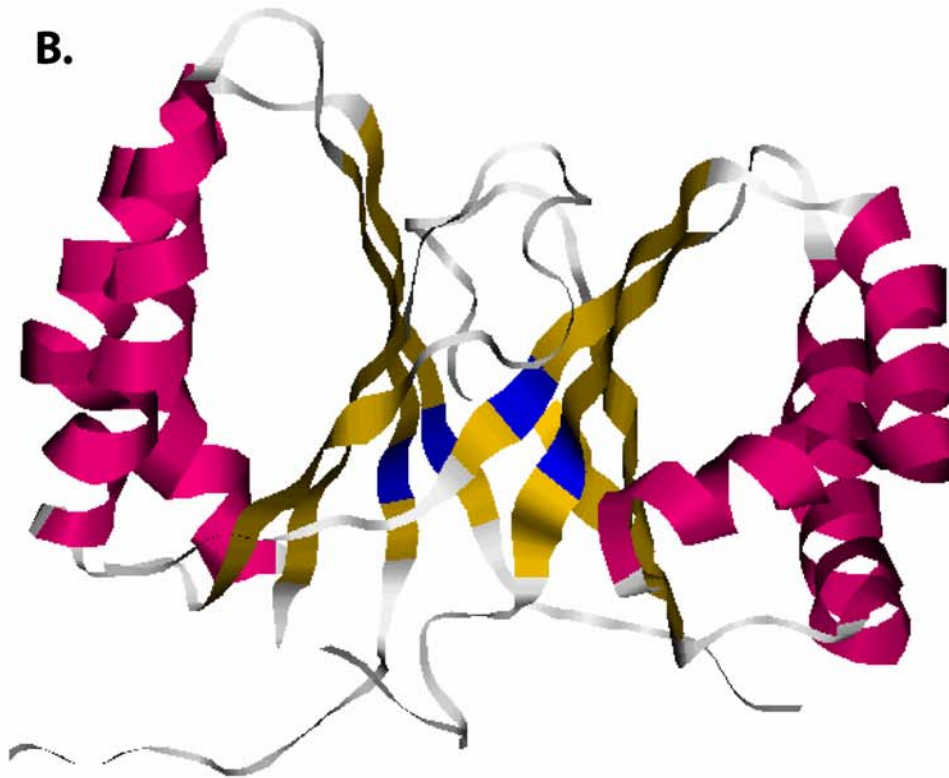


Figure 3.21: Structure of the Dabb dimeric barrel

Part (A) shows PDB:1Q4R. This image shows just a single Dabb protein, which only forms half the structure. The structure is coloured according to the secondary structure; the magnesium ion is coloured green.

Part (B)^{shows} PDB:1Q53. This image shows the complete dimeric barrel. The structure is coloured as for part B, but the magnesium ion is not shown; also the two nearly invariant residues in the alignment are coloured blue. As can be seen they lie in adjacent strands in the three dimensional structure.

may be to stabilise the structure, as they seem to sit in pockets on opposite sides of the overall structure rather than forming a central active-site.

It also raises the question as to whether this family represents a true domain. If we consider it in the terms of the three definitions given in chapter 1 then it only fulfils the requirements of the evolutionary domain. If, though this has not been tested, it can form an independently stable "half-barrel" then it could possibly be considered as a structural domain that has some of the properties of a structural repeat. In either case, the functional domain requires two copies of Dabb. These issues do not particularly affect characterisation or comprehension of this family, but they do blur the lines between the different types of structures, and raise questions about how they evolved.

Finding a copy of this domain at the C-terminus of a F_bP_aldolase domain (see Figure 3.20) suggests an involvement with sugar metabolism - fructose 1,6-bisphosphate aldolase catalyses the reversible condensation of dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate into fructose-1,6-bisphosphate. This family shows, currently, some weak similarity to the EthD Pfam family (Q89V09 is hit with an E-value of 1.6). This family is identified by Superfamily (Madera, Vogel *et al.*, 2004) as being a family of dimeric α/β barrels (Superfamily:SSF54909).

The annotation for the EthD family by Simon Moxon suggests that they are involved in the degradation of ethyl tert-butyl ether (ETBE) – a common pollutant. This is based on work by Chauvaux, Chevalier *et al.* (2001) who demonstrated that EthD is required for the degradation of ETBE in *Rhodococcus ruber*, but were unable to assign an exact function. So it is possible that these two families are related, but as

with the HHE and Hemerythrin families, the necessary stepping stone sequences are not yet present in the databases. Further evidence for this comes from the functional annotation. Of note, Chauvaux, Chevalier *et al.* (2001) suggest that only a few bacterial species are able to degrade ETBE.

Investigation of the structure and alignment in conjunction reveals some clues about which residues may be involved in function. At positions 3 and 123 of the alignments are two nearly invariant residues – a histidine and an aspartate respectively. These residues lie adjacent in the structure (marked in blue in Figure 3.21) and lie in the centre of the barrel. Examination of the side-chains is inconclusive as to whether these residues may form a hydrogen bond (personal communication: R. Finn). As for being catalytic, they also face away from the central channel, which appears to be the most likely active site. They may, however, re-orientate in the presence of the substrate. There is also a mostly conserved phenylalanine, but again initial examination is inconclusive as to what role it performs. The aspartate and phenylalanine are also present in the EthD family, whereas the histidine does not appear to be.

So, in summation, it is not obvious what the catalytic or binding behaviour of this domain might be, but its wide spread distribution suggests that it may be of some interest to biotechnology. Creation of an encompassing sequence family should help speed up research into these proteins.

Nif11 (Nitrogen fixation 11; PF07862)

Nif11 is an all- α fold domain of approximately 50 residues, found only in a few cyanobacterial species and the unrelated *Azotobacter vinelandii* (see Figure 3.22).

This family seems to be particularly expanded in *Prochlorococcus marinus* species, with strain mit9313 having 35 Nif11-containing proteins. The function of these proteins is unknown but it has been implicated in nitrogen fixation in *Azotobacter vinelandii* (Jacobson, Brigle *et al.*, 1989).

3.4 Other Potential Uses

As can be seen from both the multigenome hunt and the *Streptomyces coelicolor* hunt results, the Repeat Identification and Small Protein Clustering approaches are effective at identifying novel domains with a high success rate, and amenable to using many different data sets. Possible other hunts include:

(i) Focus on a particular system (e.g. the bacterial cell wall) by obtaining as many proteins and their homologues as possible (e.g. search every protein with an N-terminal signal peptide). This is analogous to previous investigations where domains involved in particular processes have been identified by using a dataset composed of functionally linked proteins. For instance, Mushegian and Koonin (1996) identified domains involved in development by constructing a database of proteins that were retrieved from NCBI non-redundant database using the key word "developmental".

(ii) Searching for novel systems in particular lineages, for instance by getting every 'hypothetical' protein in a particular group of species.

(iii) Finding environmental adaptations, e.g. by concatenating together the genomes of pelagic bacteria.

In all the searches the greater number of proteins the greater the chance of success, as the chance of including rare domain duplications increases. However, the greater number of proteins and genomes that the investigator looks at, the greater difficulty in fully exploring the data associated with each one. This is reflected in the results from the two different hunts discussed so far. In the *Streptomyces coelicolor* hunt it was easier to make functional predictions based on genome context and known physiological function of surrounding genes; this meant that I was able to make predictions for domains like ALF and SPDY. However, the PASTA domain was probably the only novel domain with a high level of general interest to biology. In the multigenome hunt annotation was much more difficult, but a more functionally interesting set of domains was identified. These included the various domains found at the N-termini of the secretins, the PepSY domain, the Dabb domain and the FIVAR domain.

4 Detailed Investigations of Individual Domains

In this chapter I will provide detailed descriptions of four domains. These domains appeared to be of higher interest than most and so were subjected to a lengthier and more detailed analysis than normally carried out. The first domain, PASTA, was identified during the *Streptomyces coelicolor* domain hunt; the second was found during a similar investigation of *Deinococcus radiodurans*, and the third was found during the multigenome hunt. The fourth domain discussed, Peptidase_A24, was identified by chance and not as part of a systematic hunt. It is included as it makes a useful point about how evolutionary and functional arguments can be resolved through detailed identification of homologies.

4.1 The PASTA Domain: a β -Lactam-Binding Domain

This work was part of the *Streptomyces coelicolor* domain hunt but this particular domain was examined in greater depth than the others (Yeats, Finn *et al.*, 2002). Several factors contributed to the decision to study it in greater detail. These included the existence of a three dimensional structure, its potential involvement in cell wall biosynthesis, and also finding it in the medically important Penicillin Binding Proteins. This work was carried out in collaboration with A. Bateman and R. Finn. Their specific contributions are as indicated in the text.

4.1.1 Background

While investigating the *Streptomyces coelicolor* homologue of *Mycobacterium tuberculosis* PknB, a serine/threonine protein kinase (a PSTK), I identified a novel domain that is found in its C-terminus and in the penicillin-binding proteins (PBPs).

This domain was termed PASTA (for Penicillin-binding protein And Serine/Threonine kinase Associated domain).

PSTKs are a relatively uncharacterised group of proteins in eubacteria. However, recent studies show that they are more widespread than previously thought (Av-Gay and Everett, 2000). PSTKs are typically signal transducers and are involved in bacterial growth, developmental regulation and pathogenesis. The extracellular portion normally consists of one or more sensor domains that upon binding induce alterations in the intracellular conformation. This in turn activates a signaling cascade.

There are two main types of PBP: low molecular weight and high molecular weight (Brenot, Trott *et al.*, 2001). The high molecular weight PBPs further subdivide into two groups based on the architecture of their domains: types I and II. The difference in function between these two groups is not fully understood. High molecular weight PBPs are the main architects and repairers of the bacterial cell wall, functioning through cross-linking of peptidoglycans on the surface of the cell wall. The transpeptidase domain recognises and nucleophilically attacks the penultimate D-alanine of the peptidoglycan precursor through the active-site serine residue (Ser337 in *Streptococcus pneumoniae* PBP2X). The resulting acyl enzyme intermediate then reacts with the side chain of another unlinked peptidoglycan (diaminopimelate, modified lysine or ornithine derivatives) to give the cross-linked product (Lee, McDonough *et al.*, 2001). This product forms a reinforcing mesh that envelops the cell, and makes up a substantial portion of the cell wall.

Penicillin-type (β -lactam) antibiotics function by being structurally analogous to the unlinked peptidoglycan. They acylate the active-site serine - blocking the function of the PBP. This process prevents the bacterium from replicating and maintaining the structural integrity of its cell wall. The β -lactam antibiotics are currently the most commonly used antibiotics worldwide (Lee, McDonough *et al.*, 2001; Gordon, Mouz *et al.*, 2000).

4.1.2 Searching for PASTA

Examination of the *S. coelicolor* PknB homologue (UniProt:Q9XA16) by Dotter and Prospero identified four tandem repeats of approximately 70 residues in the C-terminal half of the protein (joint observation with A. Bateman). An alignment of these repeats was used as a starting point for iterative searches using of HMMER 2.2g against the SWISS-PROT (release 40) and TrEMBL (release 18) databases. An inclusion E-value threshold of 0.01 was used. After two rounds of searching, homology to the C-terminus of high molecular weight PBPs was identified. This finding accorded with a previously noted similarity between the C-termini of PknB and PonA (Av-Gay and Everett, 2000). Further searching revealed PASTA domains in a group of uncharacterised proteins (e.g. *Borrelia burgdorferi* BB0063; UniProt:O51090) and archaeal peptidyl-prolyl isomerase (e.g. *Methanococcus kandleri* MK0796; UniProt:Y796_METKA).

These results were confirmed by use of PSI-BLAST at the NCBI server with default E-value cut-offs. Figure 4.1 shows an example alignment, and Figure 4.2 shows example domain architectures. The PASTA domain is distributed mainly in the Gram-positive bacteria, most notably among species of the genera *Bacillus* and *Clostridia*. *S.*

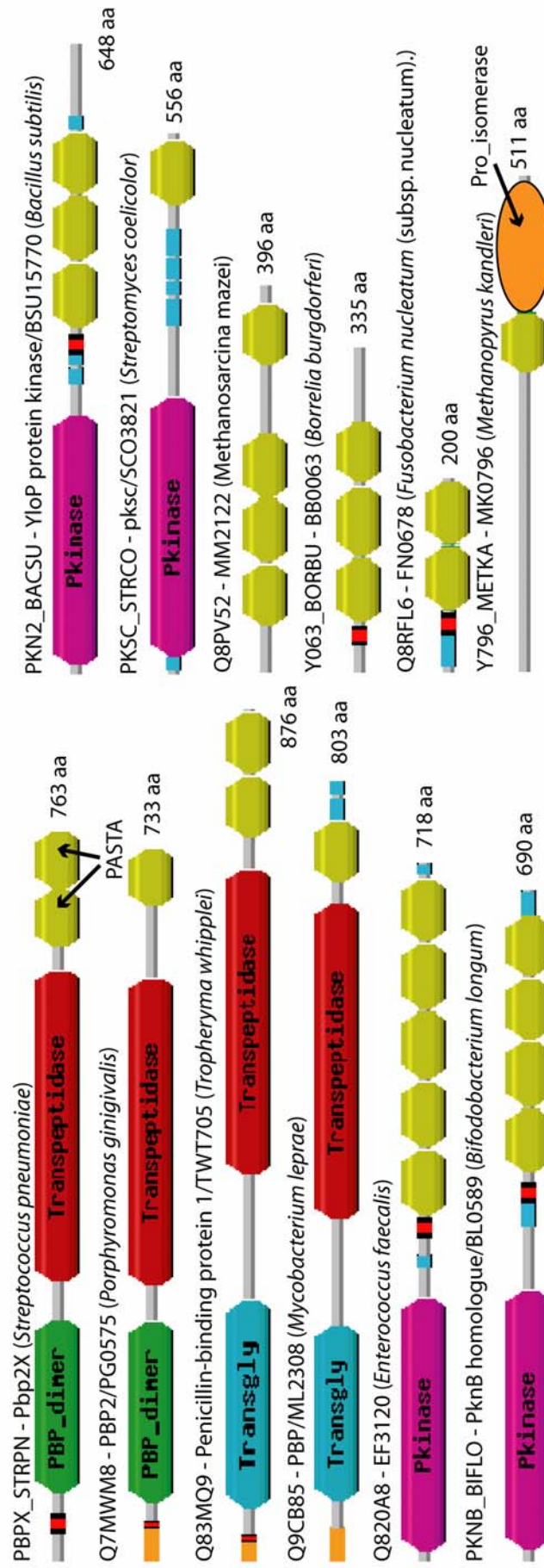


Figure 4.2: Example PASTA domain architectures

coelicolor has five PASTA-containing proteins, with eleven copies of the domain in total. Matches were also found amongst the Deinococci, Spirochaetes, Thermotogales, the Euryarchaea and others.

4.1.3 Structure of PASTA

The conserved region occurs both singly and in multiple copies, which suggests that it is a domain rather than a structural repeat (Figure 4.2). Confirmation of this notion came from the crystal structure of the soluble portion of PBP2X (PDB accession 1QMF) from *S. pneumoniae*, which contained two consecutive copies of the PASTA repeat (Gordon, Mouz *et al.*, 2000). The high molecular weight PBPs, PBP2X and PBP2B, are the primary resistance determinants in *S. pneumoniae* for several classes of β -lactam antibiotics (Grebe and Hakenbeck, 1996). Each repeat was a small globular fold consisting of three β -strands and an α -helix, with a variable length loop region between the first and second β -strands (see Figure 4.3).

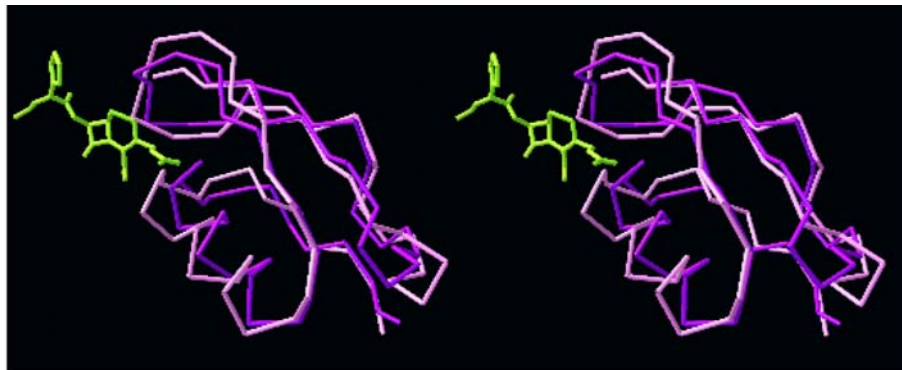


Figure 4.3: Stereo view of the two PASTA domains of *Streptococcus pneumoniae* PBP2X
The two PASTA domains of PBP2X are shown, one in purple (residues 636-691) and the second in pink (694-750), superposed on each other. As can be seen there is little variance in the structure between the two domains. Areas of difference between the two domains mainly occur close to the position of the bound cefuroxime (green). This image is derived from PDB:1QMF and was created by Robert Finn.

When the structures of the two PASTA domains were superimposed, a root mean square deviation of 1.4 Å was found (data supplied by R. Finn). This finding indicates a strong structural conservation of the PASTA domains, which contrasts with the sequence identity of only 10.5%. Of note is the unusual head-to-toe orientation of the two PASTA domains with respect to each other; this would seem to allow PASTA domains to form oligomers with substrate-binding pockets (see 4.1.4 below) facing in opposite directions, and also may allow polymerisation of PASTA domain-containing proteins. This is highly speculative, but in support Madec, Laszkiewicz *et al.* (2002) demonstrated that the PASTA-containing extracellular region of PrkC kinases readily dimerised.

4.1.4 Roles of PASTA

The PSTKs are essential for growth and development in *M. tuberculosis* (Drews, Hung *et al.*, 2001). Typical PSTKs have an extracellular sensor portion, which can be made up from more than one domain, and an intracellular kinase domain. In PknB, PASTA domains are predicted to make up the entire extracellular portion, which strongly suggests that it is a signal-binding sensor domain. Furthermore, there is a lack of obvious conserved catalytic residues in the sequence alignment, which rules out an enzymatic function, and binding domains commonly occur in multiple copies (e.g. CBM_3, PDZ).

In the structure of PBP2X, two bound cefuroxime molecules were observed. One was, as expected, bound to the active-site Ser337. The β -lactam ring of the second cefuroxime was associated with the first PASTA domain through van der Waal's interactions (Figure 4.1 shows contacting residues). This part of the antibiotic

molecule is the part that is analogous to the unlinked peptidoglycan. This feature suggests that the domain binds unlinked peptidoglycan, although probably with a low affinity because tight binding would block the activity of the transpeptidase domain.

To analyse the physiological role of PknB and its homologues further, I examined the genome context around the genes coding for these proteins. The surrounding genes were not highly conserved, but pPSTKs were generally in the vicinity of signalling and cell-wall-biosynthesis protein-encoding genes. For instance, the STRING server (von Mering, Huynen *et al.*, 2003) finds significant association between PknB and the PknA-like PSTK family, which has been shown to regulate the morphological changes involved in cell division (Chopra, Singh *et al.*, 2003). It also finds an association with the protein phosphatase 2C family (i.e. UniProt:Q8VKT2). If the PASTA domain binds unlinked peptidoglycan, PknB could act as a sensor for the concentration or presence of unlinked peptidoglycan. It then could, directly or indirectly, activate the downstream cell-wall-biosynthesis proteins, including the PBPs. Here, the PASTA domain has another role – localizing the biosynthesis complex to unlinked peptidoglycan.

The functions of the uncharacterised group of PASTA-containing proteins are not clear, but these proteins are generally found in bacteria that do not have a PASTA-containing PSTK. It is possible they act as a sensor for an alternative signalling system.

Peptidyl-prolyl isomerases catalyse one of the rate limiting steps of protein folding. Eukaryotes typically encode many versions and abundantly express them (Pahl, Brune

et al., 1997); archaea appear to use fewer. Archaeal homologues that have been previously studied do not contain a PASTA domain; perhaps this particular version is involved in refolding cell surface associated proteins or in the formation of specific cell surface complexes.

4.1.5 PASTA and Cell Morphology

As noted in the description of PASTA domain as it relates to *S. coelicolor* (chapter 2.3), it appears that each of the pPSTKs regulates the formation of different cell morphologies. Indeed, there seems to be a direct correlation between the number of possible cell types and the number of pPSTKs. In order to further substantiate this observation other species with more than one pPSTK or pPBP were also investigated.

The Corynebacteriaceae, Actinomycetes distantly related to the Streptomycetes, have two pPSTKs. They display irregular and variant morphologies; the two kinases may reflect a change in the cell wall composition allowing different cell wall morphologies.

Another Actinomycete, *Tropheryma whipplei*, also appears to confirm the hypothesis that having more than one pPSTK links to having more than one cell morphology. Detailed studies by Pahl, Brune *et al.* (1997) found that it had a distinct extracellular form to its better known intracellular form, which responded differently to staining – indicating a change in cell wall composition. It also contains two pPSTKs in its otherwise reduced and compact genome.

All of the sequenced Bacillae appear to have two pPBPs adjacent in their genome; one contains a single PASTA domain (SpoVD) and one contains two PASTA domains (FtsI or PBP2B). In *Bacillus subtilis*, SpoVD has been noted to only be involved in sporulation (Daniel, Drake *et al.*, 1994), which uses an alternative peptidoglycan (Atrih and Foster, 2001), whereas FtsI is specific for growth at the septum (Scheffers, Jones *et al.*, 2004). Given the conservation of the genomic arrangement, these functional assignments are likely to be true of the rest of the Bacillae. An exception to this is the *Bacillus cereus* group. Both *B. cereus* and *B. anthracis* have a third pPBP, containing two PASTA domains, at a different locus from the other two. The role of the extra pPBP in the *Bacillus cereus* group is not clear, but it would seem to imply that the *Bacillus cereus* group can utilise an alternative peptidoglycan in their cell wall, and possibly even form a new cell type.

4.1.6 The PASTA Domain as an Antibiotic Target

Having found this association between PASTA domains and β -lactams, it seemed that the PASTA domain itself might represent a viable antibiotic target. It certainly meets several of the criteria: it is an extracellular domain; it is found in essential proteins; it is not found in eukaryotes; and a known antibiotic binds to it. To examine its importance as an antibiotic-resistance determinant, I examined the distribution of mutations in 39 PBP2X sequences from resistant isolates of *S. pneumoniae*. The analysis concurred with the observation by Dessen and colleagues that the PASTA domains are indeed mutational hotspots (Dessen, Mouz *et al.*, 2001), and that the mutations consistently occur at the same sites (see Figure 4.4). It is possible that these mutations may have spread through transformation rather than independent mutation

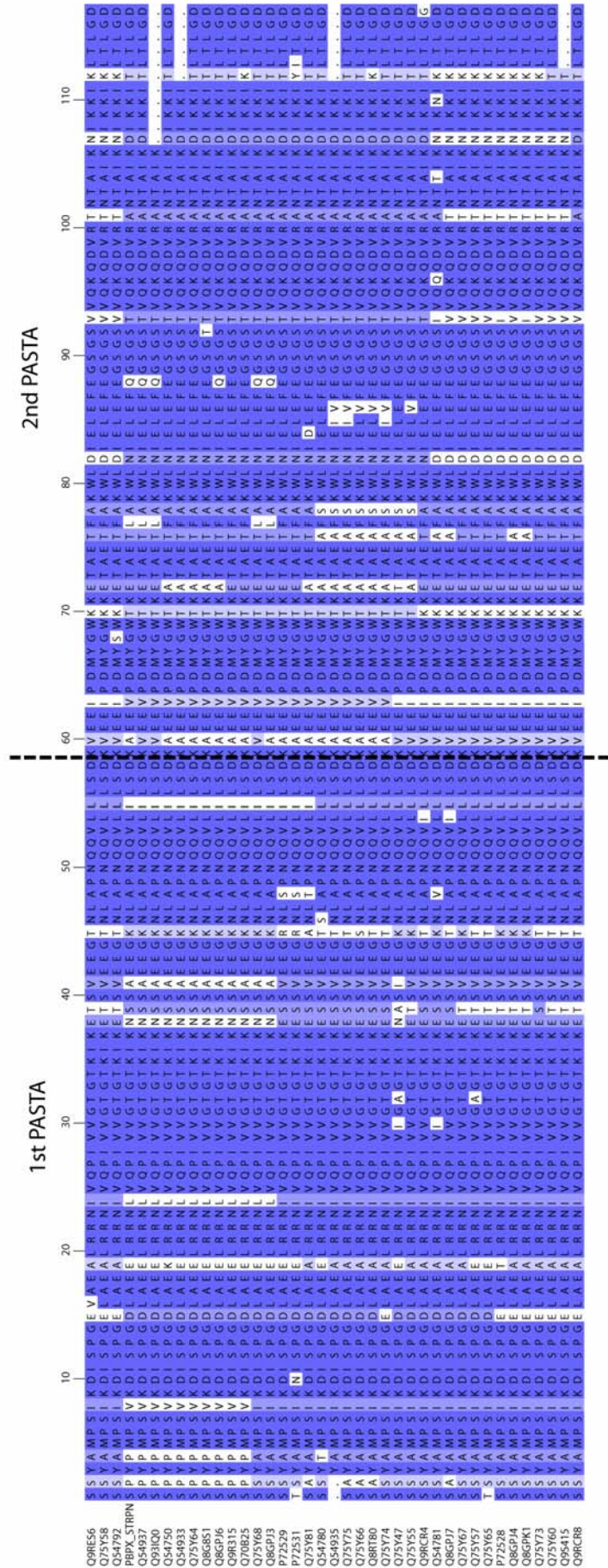


Figure 4.4: Distribution of resistance mutations in the PASTA domains of PBP2X
 This alignment was created by taking the 86 copies of *Streptococcus pneumoniae* PBP2X C-terminus that included both PASTA domains from UniProt, aligning them with MAFFT, and making it non-redundant (no identical sequences). The alignment is coloured according to conservation using Jalview. Whilst the source of each sequence has not been fully investigated, they are largely generated from the sequencing of β -lactam antibiotic resistant strains (e.g. Reichmann, Konig *et al.*, 1996 or Laible, Spratt, *et al.*, 1991). Despite the geographic and strain variation contained in this alignment the mutations are consistently in the same residues, suggesting that they are secondary resistance determinants. The dotted line indicates the end of the first PASTA domain and the start of the second.

events, but in either case it is clear that these particular mutations has been selected for in response to antibiotic challenge.

Biochemical assays show that resistant strains of *S. pneumoniae* have abnormal branched peptides in their cell walls (Garciabustos and Tomasz, 1990). Changes in the active site are required to maintain the efficiency of the PBP complex; so perhaps changes in the PASTA domain are also required to maintain the efficiency of localization of the PBPs to unlinked peptidoglycans. Therefore, the function of the domain is open to disruption from antibiotics. In the case of the PSTKs, PknB has already been put forward as a good antibiotic target (Drews, Hung *et al.*, 2001). Characterisation of the PASTA domain confirms that idea and suggests a possible class of compounds that could attack them.

4.1.7 Subsequent Research

The identification of the PASTA domain (Yeats, Finn *et al.*, 2002) has aided several research groups investigating the physiological role or structure of PASTA-containing PSTKs. Most of this work has focussed on the structure and biochemical action of the catalytic portion of the kinase, as these proteins are relatively uncharacterised in bacteria. For instance, Boitel, Ortiz-Lombardia *et al.* (2003) identify a conserved interaction between PknB and PstP – a protein phosphatase. The two proteins occur in the same operon in *M. tuberculosis*, and this seems to be conserved across the Actinobacteria, including *S. coelicolor*. In this operon are also a *pbpA* (an HMW PBP) and a *rodA* gene, both of which are involved in cell wall biosynthesis. Both the conservation of this system across several species and the inclusion of cell

morphology determinant genes, strongly reinforce the concept that the pPSTKs are an important part of the cell wall surveillance mechanism.

Strong, Graeber *et al.* (2003) constructed a genome wide functional linkage map in *Mycobacterium tuberculosis*, using information from gene order, phylogenetic profile and known protein function. They then used this map to assign novel annotation to uncharacterised proteins. One of the results to come out of the research was that *pknB*, *pknA*, *ppp*, *pbpA*, *rodA* and *Rv0019c* are functionally linked and that they are likely to be involved in cell wall biosynthesis.

Work by Echenique, Kadioglu *et al.* (2004) on the *Streptococcus pneumoniae* kinase *StpK* has shown that it is important in virulence and competence triggering during infection. *StpK* also helps prevent *LytA*-induced autolysis and resist low concentrations of cell wall directed antibiotics. The authors suggest that these functions are induced by stresses on the cell wall, which are detected by the PASTA domains.

These lines of work, while not providing a precise definition of the function of the PASTA domain, support the predictions made above.

4.2 The BON Domain: A Putative Membrane Binding Domain

As demonstrated in chapter 2.2 scanning genomes for novel domains is an effective method for elucidating both organism-specific information and more widespread biological processes. I decided to carry out such a scan on *Deinococcus radiodurans* because it is renowned for its ability to repair extensive DNA damage caused by

radiation and rehydration from a desiccated state (Englander, Klein *et al.*, 2004). As mentioned in chapter 3.1 the investigation did not recover many novel domains, and there was little functional information associated with them; however, one domain was examined further. This domain is involved in osmotic-shock protection and other cell-membrane-localized processes through its interactions with phospholipid membranes (Yeats and Bateman, 2003).

4.2.1 Identification of the Conserved Regions

A conserved repeat was identified in *D. radiodurans* protein DR0888 (UniProt:Q9RVY3) using Prospero and subsequent to masking low complexity regions using seg. Residues 9-75 aligned to 124-190 with an E-value of 1.2×10^{-6} . The aligned pair was then searched against SWISS-PROT (40.31) and SP-TrEMBL (22) using HMMER 2.2g. The initial search found a suggestive (E-value = 0.16) match to *Xanthomonas axonopodis* protein XAC0682 (UniProt:Q8PPK4), residues 53-116. This region was then used to initiate a set of iterative HMMER searches. Both global (ls) and fragment (fs) models were built and searched with, and results combined using E-value cut-offs of 0.1 and 0.01, respectively. Alignments were examined visually and potential false-positives removed between rounds. T-Coffee and manual editing were used to produce the final alignment (see Figure 4.5 for an example alignment). After 13 rounds the searches converged to identify 61 proteins, including both the regions from DR0888, confirming the validity of the initial suggestive match. To ratify the searches, an equivalent process was carried out at the NCBI PSI BLAST server using an E-value cut-off of 0.002. The occurrence of these regions as singlets in some proteins and in varying surrounding domain contexts implies that they are

structurally stable in isolation and are true domains. The domain was termed the BON (bacterial OsmY and nodulation) domain.

4.2.2 OsmY Comprises Two BON Domains

The BON domain is typically around 60 residues long and is predicted to have an α/β fold (shown in Figure 4.5). There is a conserved glycine residue and several hydrophobic regions. This pattern of conservation is more suggestive of a binding or structural function rather than a catalytic function. The OsmY protein is an *Escherichia coli* 20 kDa outer membrane or periplasmic protein (Yim and Villarejo, 1992) that has RpoS-controlled expression in the stationary phase under normal growth conditions (Weichart, Lange *et al.*, 1993). It is also expressed in response to a variety of stress conditions, in particular, helping to provide protection against osmotic shock (Yim and Villarejo, 1992; Bernstein, Bernstein *et al.*, 1999; Oh, Cajal *et al.*, 2000).

One hypothesis is that OsmY prevents shrinkage of the cytoplasmic compartment by contacting the phospholipid interfaces surrounding the periplasmic space (Oh, Cajal *et al.*, 2000; Liechty, Chen *et al.*, 2000). This would physically prevent the inner membrane from shrinking by attaching it to the more rigid outer membrane. The symmetrical domain architecture of two BON domains (see Figure 4.6) suggests that they contact the surfaces of phospholipids, with each domain contacting a membrane.

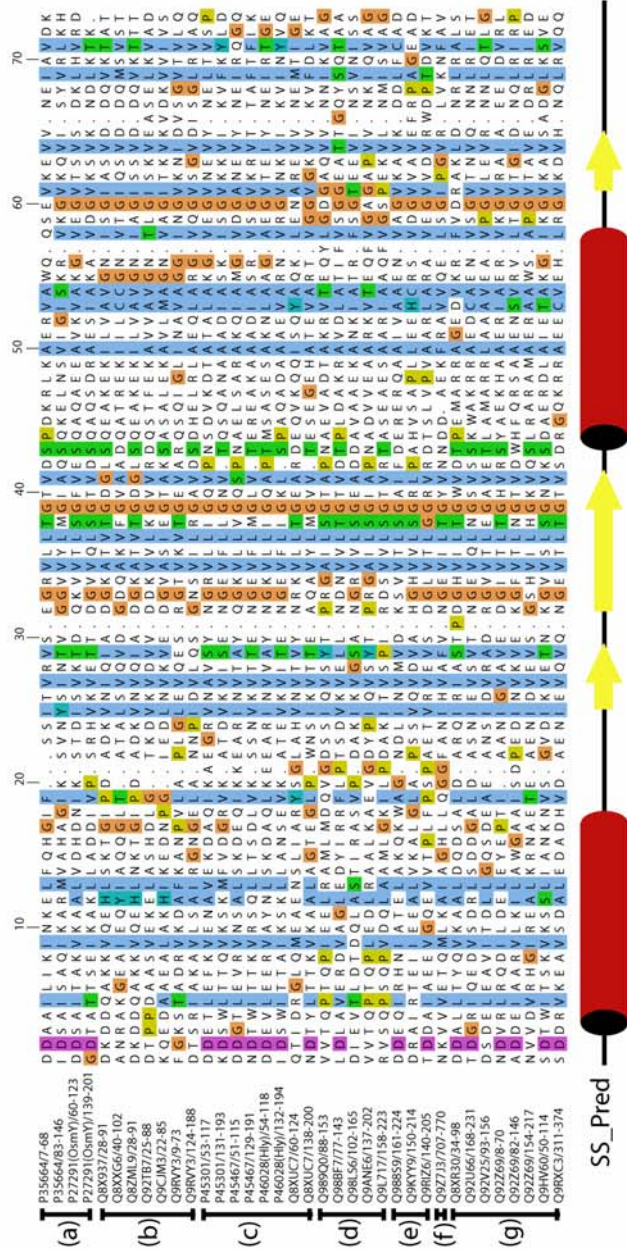


Figure 4.5: BON domain example alignment
 The sequence alignment is grouped according to protein function. The groupings are as such: (a) OsmY protein-related (Swiss: P27291); (b) LysM domain-associated (Pfam: PF01476); (c) HLY protein-related (Swiss: P46028); (d) modulation specificity; (e) CBS domain-associated (Pfam: PF00571); (f) regulatory; and (g) other.

Notably, a group of putative haemolysins also consist of two BON domains. The assignation of haemolytic activity is based on the conferment of haemolytic activity to *E. coli* after transformation with a plasmid that contained DNA sequence from *Actinobacillus pleuropneumoniae* (Ito, Uchida *et al.*, 1993). To my knowledge, no other work has been carried out to confirm that the encoded protein (HLY) is directly responsible for this activity; however, the ability to interact with cell membranes would be expected of a haemolysin.

4.2.3 Other BON-Containing Proteins

The other occurrences of BON further support the hypothesis of it associating with phospholipid membranes (see Figure 4.6). It occurs in association with two membrane-pore forming domains - Secretin (Bitter, Koster *et al.*, 1998; Drake and Koomey, 1995) and MS_channel (Kloda and Martinac, 2001). MS_channel proteins are mechanosensitive ion channels and have been implicated in osmotic regulation; some appear to function in response to membrane deformation (Perozo, Kloda *et al.*, 2002). None of the BON-containing MS_channel proteins have been specifically characterised; a possibility is that the BON domain reacts to deformations in the plasma membrane, and allosterically signals to the ion channel domain. The most characterised of the BON Secretins is CpaC of *Caulobacter crescentus*. CpaC forms a polar pilus that forms in a specific location in the cell, along with another pilus subcomponent CpaE (Skerker and Shapiro, 2000). This pilus is required for the progeny swarmer cell of a sessile stalk cell to move away from its mother. Homologues of the CpaC Secretin are also found in the Rhizobiales.

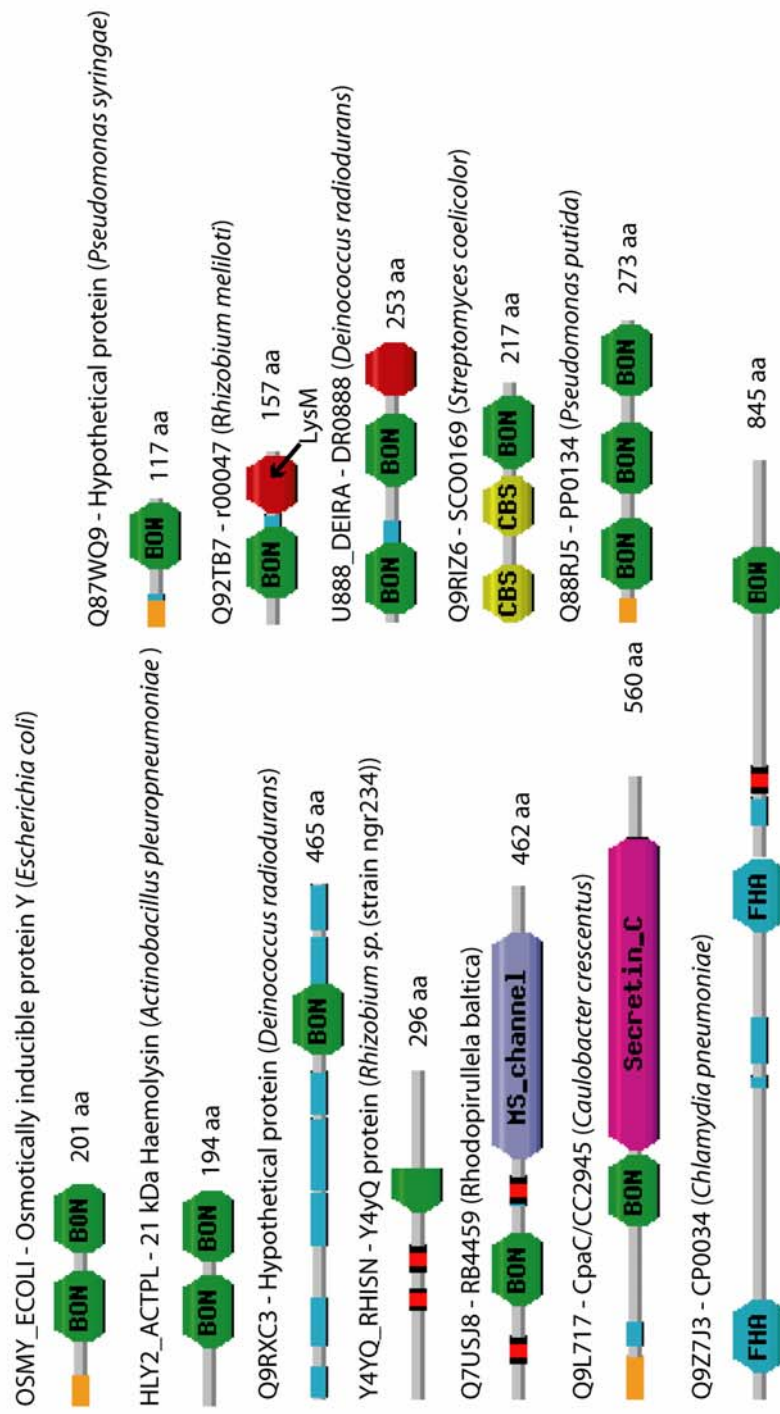


Figure 4.6: Example BON domain architectures

The BON domain also co-occurs with the cell-wall peptidoglycan-associating LysM domain (Bateman and Bycroft, 2000) in the Rhizobiales. Although this protein is not annotated in *Mesorhizobium loti* it is found between the nitrogen fixation regulators, FixL and FixJ, and the nitrogen fixation operon, FixS through to FixN; *Bradyrhizobium japonicum* has three such LysM and BON containing proteins.

Finally, it is found in a set of Chlamydia putative regulatory proteins (e.g. Q9Z7J3). These proteins are described by UniProt as having homology to an adenylate cyclase, but this is because they also contain a forkhead-associated (FHA) domain (Hofmann and Bucher, 1995; see Figure 4.6). They do not actually have any adenylate cyclase function. This is an example of the single-linkage mis-annotation as discussed in chapter 1.5.1. Nothing is known about the function of this family of regulators. If the hypothesis that the BON domain binds phospholipid membranes is correct, then these regulators may detect deformations in the cell membrane. Hence they could form part of a mechanism for regulating genetic responses to the cell being under osmotic or mechanical duress.

4.2.4 Phyletic Distribution

Most proteobacteria seem to possess one or two BON-containing proteins, typically of the OsmY-type proteins (data not shown); outside of this group the distribution is more disparate. The family is unusually expanded in Burkholderia, a genus containing several significant mammalian pathogens with varying host ranges. The number of BON-domain proteins varies between Burkholderia species, suggesting that they perform specific roles in their respective lifestyles. *B. pseudomallei* have eight BON-containing proteins (personal communication: M. Holden). Included in this set are a

protein with a single BON domain and two proteins with an unusual three consecutively repeated BON domains. Homologues of this protein are only found in some other Burkholderia and in the main symbiosis or pathogenicity plasmids of Ralstonia (plant pathogens) and Rhizobiales. The Rhizobiales show a similar variety in the number and type of BON domain proteins in their genomes as the Burkholderia. For instance *Sinorhizobium meliloti* has ten BON proteins, including many single BON-containing proteins, while *Mesorhizobium loti* has three. This distribution suggests that these proteins play a role in host invasion or host-cell interactions. Within the completed genomes, no clear operon structures associated with BON domains were found.

In conclusion, the BON domain is likely to be a phospholipid-binding domain that is involved in a variety of biological processes.

4.3 The PepSY Domain: A Putative Regulator of Peptidase Activity

During a search for novel protein domains in bacterial genomes a repeated region in TTE0861 of *Thermoanaerobacter tengcongensis* (sequenced by Bao, Tian *et al.*, 2002) was identified. Homology searches found this region to be spread throughout bacterial species, most significantly in the N-terminal propeptide of the M4 family of peptidases (as classified in Rawlings, Tolle *et al.*, 2004). This region is termed PepSY for Peptidase (M4) and subtilis' YpeB (Yeats, Rawlings *et al.*, 2004). The M4 family of metallopeptidases are a widespread family that are mostly found in both Gram-negative and Gram-positive eubacteria, but are also sporadically found in fungi (*Neurospora crassa*) and archaea (*Methanosarcina acetivorans*).

4.3.1. Background to the M4 Peptidases

Some members of the M4 peptidases, notably bacillolysin (EC 3.4.24.28) and thermolysin (EC 3.4.24.27), are among the most commonly used enzymes in industry. Their general biological role is not well understood, but it appears that they are often involved in the generation of nutrients in the local environment. However, several pathogens have adapted this function for the breakdown of host tissue. For instance, both *Vibrio vulnificus* (vibriolysin, EC 3.4.24.25) and *Pseudomonas aeruginosa* (pseudolysin, EC 3.4.24.26) use M4 peptidases to invade host tissue (Miyoshi, Nakazawa *et al.*, 1998; Heck, Morihara *et al.*, 1986).

Typically, a species has only one M4 peptidase, but the family is expanded in some; for example, *Bacillus subtilis* has two and *Streptomyces coelicolor* has five. M4 peptidases are typically translated as propeptidases, with a secretory signal sequence, N-terminal propeptide and a two-domain peptidase unit. Some examples (e.g. thermolysin) show broad substrate specificity, whereas some (e.g. vibriolysin) appear to show a far more limited range of substrates – although this might be attributable to a lack of characterization. In most cases, the propeptide is cleaved through full or partial auto-catalysis in the periplasm, but remains non-covalently attached (Kessler, Safrin *et al.*, 1998). Several studies have shown that the propeptide has inhibitory and chaperone activities (e.g. Marie-Claire, Roques *et al.*, 1998), and that – provided the sequence similarity is not too low (e.g. less than 20%) - the propeptide from one peptidase can substitute for the propeptide from another (Tang, Nirasawa *et al.*, 2003).

4.3.2 PepSY Domain Identification

Examination of TTE0861 (UniProt: Q8RBF9) using dotter identifies five repeats of 60-75 residues that are interspersed by regions of 15 or more residues (see Figure 4.7 for coordinates). Aligning these repeats with MAFFT enabled an iterative search against Swiss-Prot (release 41.25) and TrEMBL (24.14) with both fragment and global hidden Markov models generated by HMMER using the maximum entropy weighting. Hits with an E-value of less than 0.05 (fragment) and less than 0.1 (global) were included in subsequent rounds and the search repeated until convergence. The alignment was periodically realigned with MAFFT and manually adjusted. The separation of signal-to-noise was not distinct and so reciprocal searches were carried out from multiple starting points; these included aligning all identified M4 propeptide regions, excising the PepSY region and using this as an initial search seed, and PSI-BLAST searching at the NCBI. Eventually more than 270 copies were identified. An alignment of example sequences is given in Figure 4.7.

I used similar approaches to identify two associated families: the PepSY_TM transmembrane helix family (PF03929), and the FTP (for fungalysin/thermolysin propeptide; PF07504) motif (see chapter 4.3.3 and Figures 4.8 and 4.9).

4.3.3 Description of the PepSY Domain

The PepSY domain varies from 60-90 residues in length and is predicted to have an α/β fold (Figure 4.7). It often occurs as a single copy and in multiple domain architectures (see Figure 4.10); this suggests that it is stable in isolation and is a true domain. Similarity between some of the family members is low, and only a couple of regions show strong conservation. First, an aromatic residue is often found in the

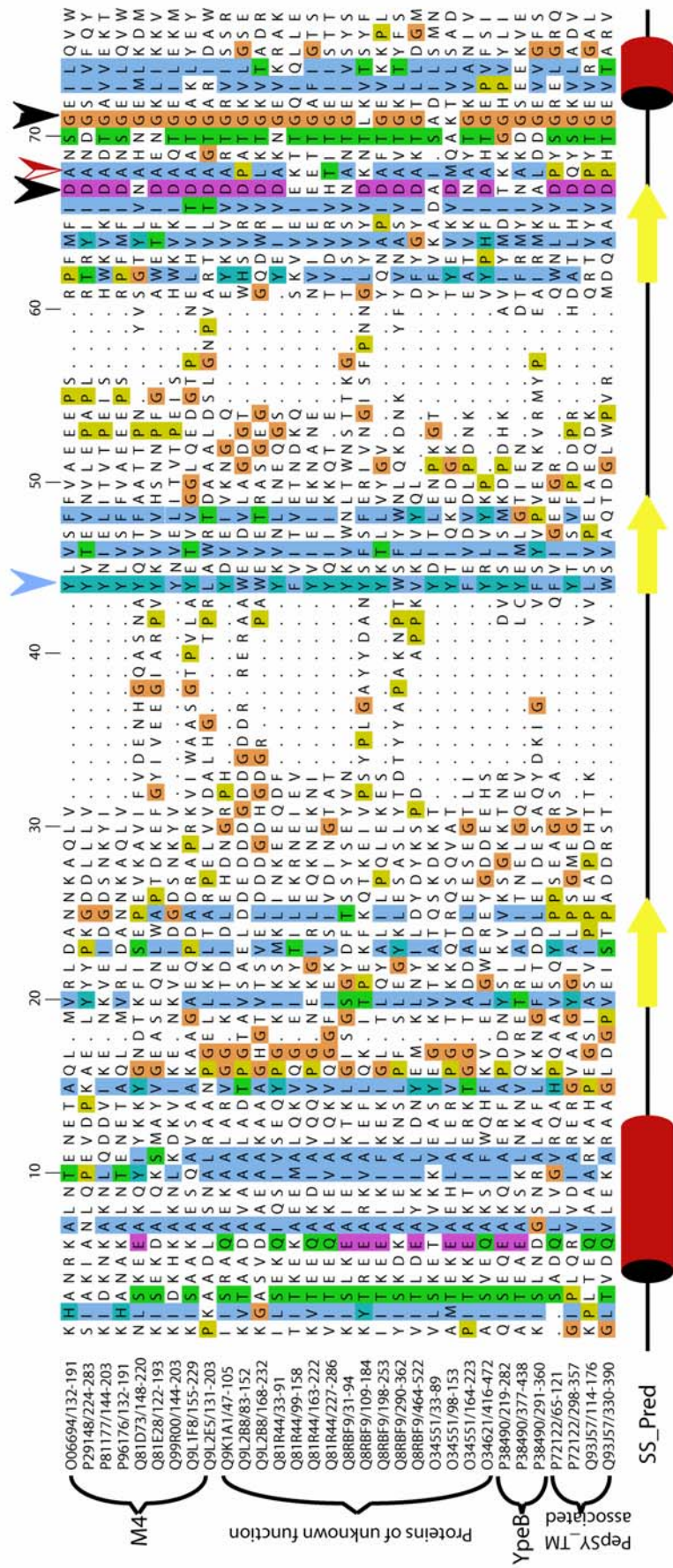


Figure 4.7: Example PepSY domain alignment

The particularly conserved region discussed in the main body of the text is marked by the two black arrows. The locus of the A183V mutation of pseudolysin (UniProt: P14756) described and characterised by Braun, Bitter, *et al.* (2000) is marked by the red arrow. The blue arrow marks the mostly conserved tyrosine or aromatic residues mentioned in the main body of the text.

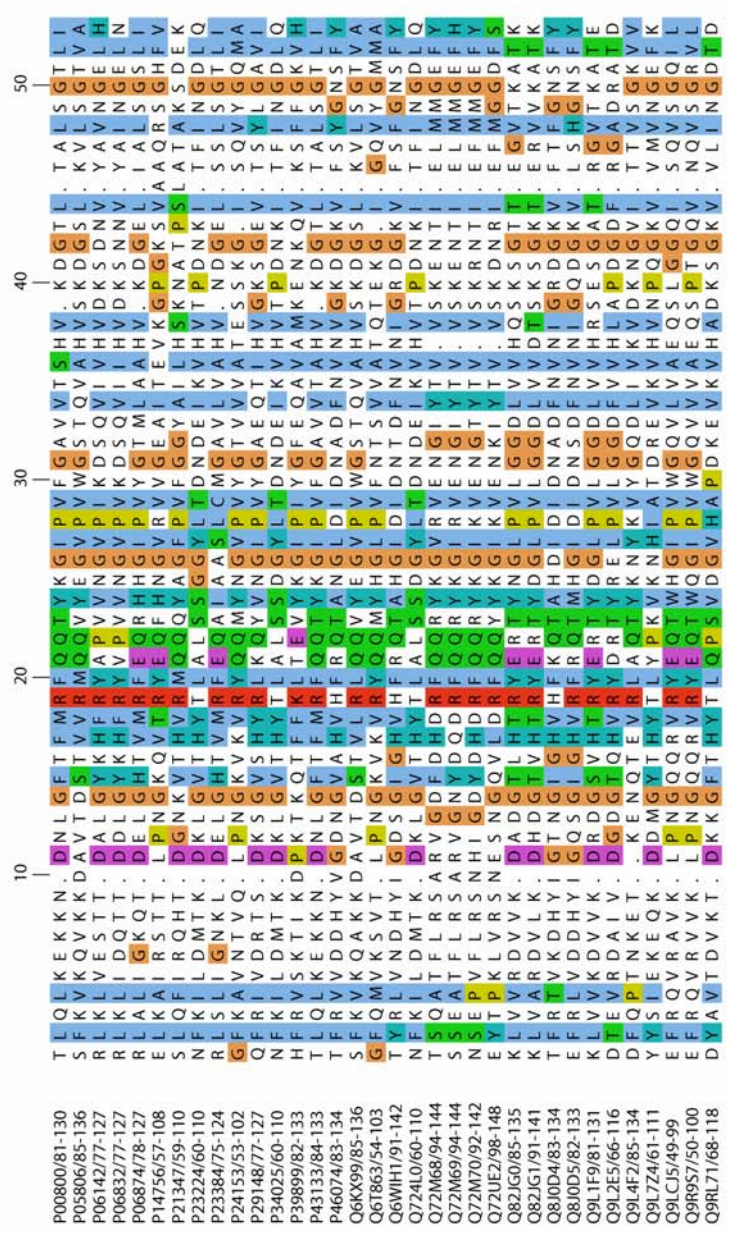


Figure 4.8: FTP motif example alignment

middle of the alignment; second, at the C-terminus, a hyd-Asp-hyd-Xaa-Xaa-Gly (where “hyd” is a hydrophobic residue and "Xaa" is any) motif is fairly conserved. In several proteins, the C-terminus of the PepSY domain coincides with the end of the protein or the start of another domain (e.g. UniProt: O34551 and UniProt: P29148). These observations give us confidence in the positioning of the domain boundaries.

4.3.4 Domain Architecture of the M4 Propeptide

Examination of an alignment of the propeptides of the M4 peptidases identified a second conserved region near the N-terminus of PepSY. Searching (in the same manner as described above) revealed that this region is also present in the eukaryotic M36 peptidases – but not in the bacterial group of uncharacterised PepSY-containing proteins. The M36 peptidases (the fungalysins) are believed to belong to the same structural fold as the M4 peptidases and have the same active site architecture. The PepSY domain does not appear to be present in the fungalysins. This suggests that this second region of conservation – named FTP (PF07504) – is separate from the PepSY domain and has a separate function. Computational and visual examination of the region between the FTP motif and the M36 peptidase unit did not reveal any similarity to the PepSY domain.

4.3.5 Species Distribution of PepSY

Most eubacterial species have one or two copies of PepSY – mainly in the M4 peptidase – but some have more: *B. anthracis* has 11 PepSY-containing proteins; *Staphylococcus aureus* has five but the closely related *S. epidermis* has only two. This suggests that the expansion of this family is specific to the biology of the organism, but is not specifically linked to pathogenicity for instance. PepSY domains are also

found in several archaeal species, although the only identified archaeal M4 peptidases are in *M. acetivorans*. Perhaps surprisingly, the M4 peptidase in *N. crassa* does not have a typical propeptide and, concordantly, no PepSY domains have been identified in fungi.

4.3.6 PepSY Family Characteristics

PepSY-containing proteins appear to fall into three main groups (Figure 4.10): (i) the M4 peptidases, (ii) those with no ascribed functions, and (iii) PepSY_TM associated. Most members of the second group normally have either one or two copies of the domain and no other domains, but some have three, four or five copies. TTE0861, as well as having five copies of PepSY, has three SLH (S-layer homology) domains at the C terminus. SLH domains anchor proteins to the S layer of the bacterial cell wall (Mesnage, Fontaine *et al.*, 2000). Some members of this group have predicted signal leader peptides.

In most cases, members of the third group normally have two PepSY domains, each flanked by a pair of conserved homologous transmembrane helices named PepSY_TM. The membrane topology of these proteins is - in most cases – predicted by TMHMM to hold the PepSY domains to the exterior of the cell.

Signal peptide and transmembrane helix predictions consistently suggest that group (ii) and (iii) proteins are either held on the cell surface or secreted. A notable exception to this rule is the YpeB homologue group [within group (ii)]. These form a small group of Bacillales proteins. The *ypeB* gene in *B. subtilis* is in a bi-cistronic operon with *sleB*. SleB is one of the primary cortex lytic enzymes and is essential for

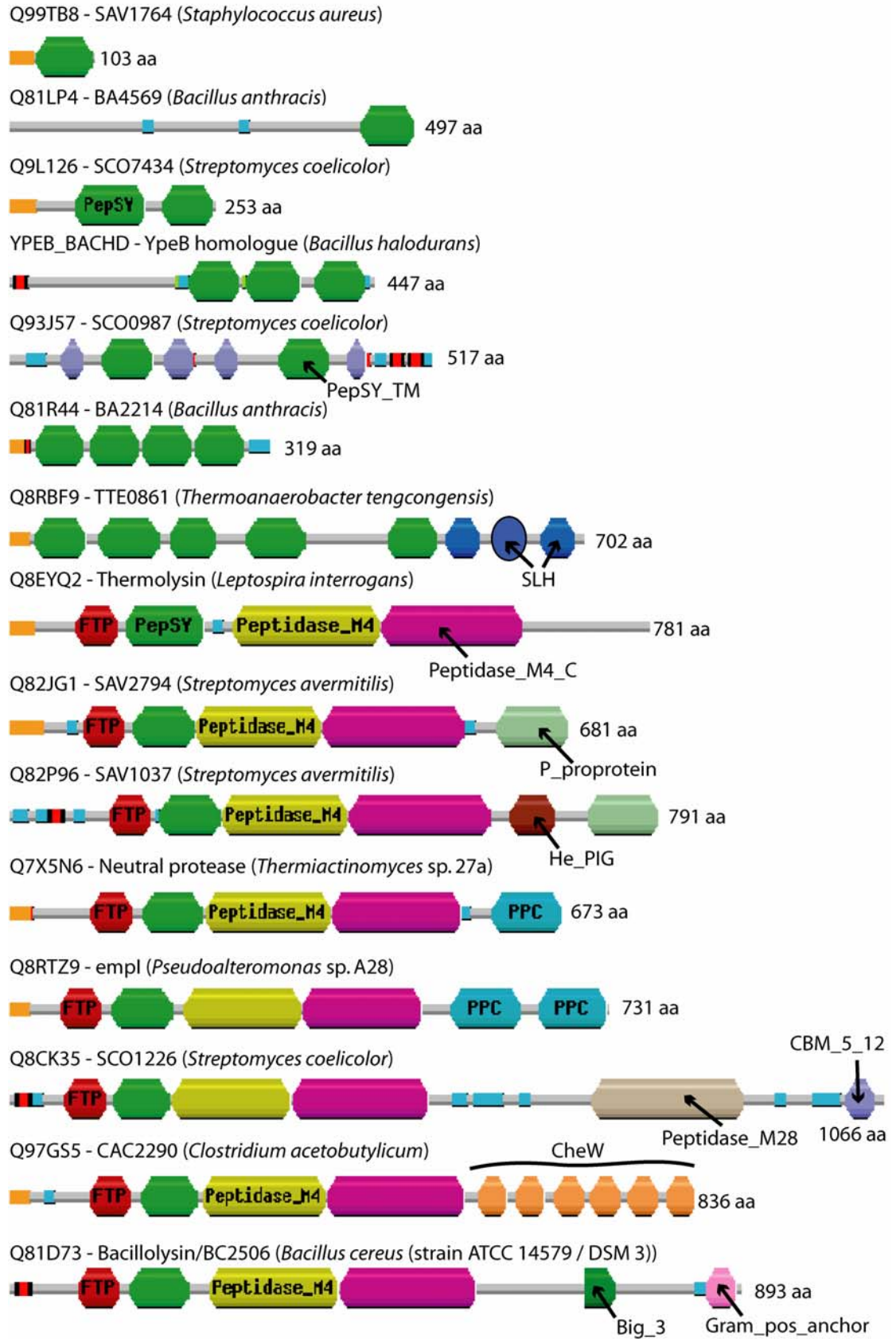


Figure 4.10: Example PepSY domain architectures

the germination of the spores. It has been shown that its expression, localisation and/or stabilisation, requires the co-expression of *ypeB* and that both proteins are co-localised to the inner membrane and integument. It has been hypothesised that SleB is either a peptidase or a lytic transglycosylase, but the reaction has not been described (Boland, Atrih *et al.*, 2000).

4.3.7 PepSY Domains Are Likely to be Inhibitors

Where examined, the propeptide shows strong inhibitory activity (e.g. Braun, Bitter *et al.*, 2000; Tang, Nirasawa *et al.*, 2003), and this function is likely to be conserved. Given that most of the M4 propeptides appear to be substitutable, the chaperone and inhibitory functions must lie within the conserved regions. Braun, Bitter *et al.* (2000) identified two mutants of *Pseudomonas aeruginosa* pseudolysin – one of alanine to valine at position 183 (A183V; within the PepSY domain) and the other of threonine to isoleucine at position 45 (T45I; outside of PepSY) – and examined their effects on the dissociation of the propeptide from the peptidase unit (see Figure 4.5 for position of Ala183). The T45I mutation was mildly disruptive to cell growth; the A183V mutation led to severe growth retardation, cell leakage and ultimately cell lysis. The interpretation of these results is that the A183V propeptide rapidly dissociates from the peptidase prior to export from the cell, and that the peptidase has folded correctly because it is active and proceeds to digest the cell from the inside (Braun, Bitter *et al.*, 2000). The T45I propeptide mostly remained associated and therefore this region is not involved in inhibition. This evidence appears to confirm that the inhibitory function resides principally in the PepSY domain, but the chaperone activity does not.

Since PepSY is only found in the M4 propeptide and is not associated with any other peptidase families, the inhibitory activity may have limited specificity. However, the tight co-expression of *ypeB* with *sleB* suggests that PepSY might have a broader range of inhibition. The transcriptional coupling of a lytic enzyme to its inhibitor to prevent premature or misplaced activation has been shown several times recently (e.g. Massimi, Park *et al.*, 2002; Rzychon, Sabat *et al.*, 2003). SleB shows no detectable sequence similarity to the M4 peptidases, and so for it to be inhibited by PepSY would imply that at least some instances of PepSY have a broad specificity of inhibition. Alternatively YpeB may be protecting SleB, but there are no reports on the processing of SleB during or just prior to sporulation.

4.3.8 Biological Role of PepSY

The PepSY domain has significant biological roles both in the control of M4 peptidases through their propeptide and in the germination of Bacillales spores. Furthermore, their presence in a diverse family of secreted and cell wall-associated proteins suggests that they might play a part in regulating protease activity in the cell's local environment. This might have a special significance in pathogenesis and the formation of microbial communities. If the bacterial population increases in density, whether through aggregation or reproduction, then individual cells must use mechanisms that prevent them from eating each other. One way would be to switch off secreted peptidase production, but there are clearly risks to this strategy - e.g. if production is suddenly triggered by an unusual event - and so it makes sense for a complementary self-protection method to be employed. A further intriguing idea is that PepSY domain-containing proteins could be used to block the progress of pathogens that use an M4 peptidase to invade tissue.

4.4 Peptidase_A24 - the Prepilin Peptidase

The prepilin peptidase is the aspartic acid peptidase (family A24 as defined by MEROPS) that cleaves the signal peptide required for secretion and assembly of bacterial pili. However, as archaeal genomes began to be sequenced, it was noted that they also contained signal peptide-like sequences at the N-termini of archaeal flagella components and other secreted proteins. Both the bacterial pili and archaeal flagella are essential for motility. Initial investigations were unable to identify homologues of the prepilin peptidases, so it was proposed that they used an alternative system (Jarrell, Correia *et al.*, 1999). The argument was essentially resolved by Albers, Szabo *et al.* (2003) through the identification of a "Cluster of Orthologous Proteins" (Tatusov, Fedorova *et al.*, 2003), COG1989, which contained bacterial prepilin peptidases and the archaeal *Sulfolobus solfataricus* sequence SSO0131 (now termed PibD). They assayed this protein for activity and found that it was capable of processing *S. solfataricus* signal sequences.

At about the same time as this work was carried out I was examining the Pfam 9.0 alignment of eubacterial Peptidase_A24 proteins. It was clear, upon visual examination, that the alignment was composed of multidomain proteins that had varying architectures. This produced a blocky alignment, as described in chapter 2.1.2. It was also clear that there was a region of approximately 100 residues that was found in all these proteins. I excised this region and used it to iteratively search against UniProt (14.25/24.14). Within a single iteration archaeal homologues were identified, and the searches converged after another two rounds. Some new Eubacterial A24 peptidases were also found, and these consisted of a single Peptidase_A24 domain, which covers the entire length. This gives confidence in the

accuracy of the deduced domain boundaries. Figure 4.11 shows an example alignment and example architectures.

This family is fairly poorly characterised, with the two active site aspartate residues only recently discovered (LaPointe & Taylor, 2000). These residues are both found in the 100 residue region I excised and are absolutely conserved (see Figure 4.11), suggesting that all the homologues found are active peptidases. The result from the sequence analysis allows the generalisation of the result of Albers, Szabo *et al.* (2003), and it is now possible to state that the signal peptidases are close to ubiquitous in Eubacteria, Crenarchaea and Euryarchaea.

The method used by Albers, Szabo *et al.* (2003) is essentially complementary to the approach I have used. Through their combinatorial approach they provide strong evidence for the existence of signal peptidases in archaea, and definitively in *S. solfataricus*. My approach is unable to demonstrate the conservation of function, but it is able to generalise, confirm and extend their data.

It is also now possible to further define variances in these proteins. As mentioned above the domain was identified because it was found in multidomain proteins that aligned poorly. So I also built sequence families for these other regions. This has allowed the identification of four different domain architectures, two of which are exclusive to the archaea. Whilst it is not clear what the function of these accessory domains are, it has been noted that *Pseudomonas aeruginosa* PilD N-methylates the precursor protein as well as processing it. In contrast PibD does not have this activity. PilD has a DiS-P-DiS (PF06750) domain at the N-terminus, while the PibD has the

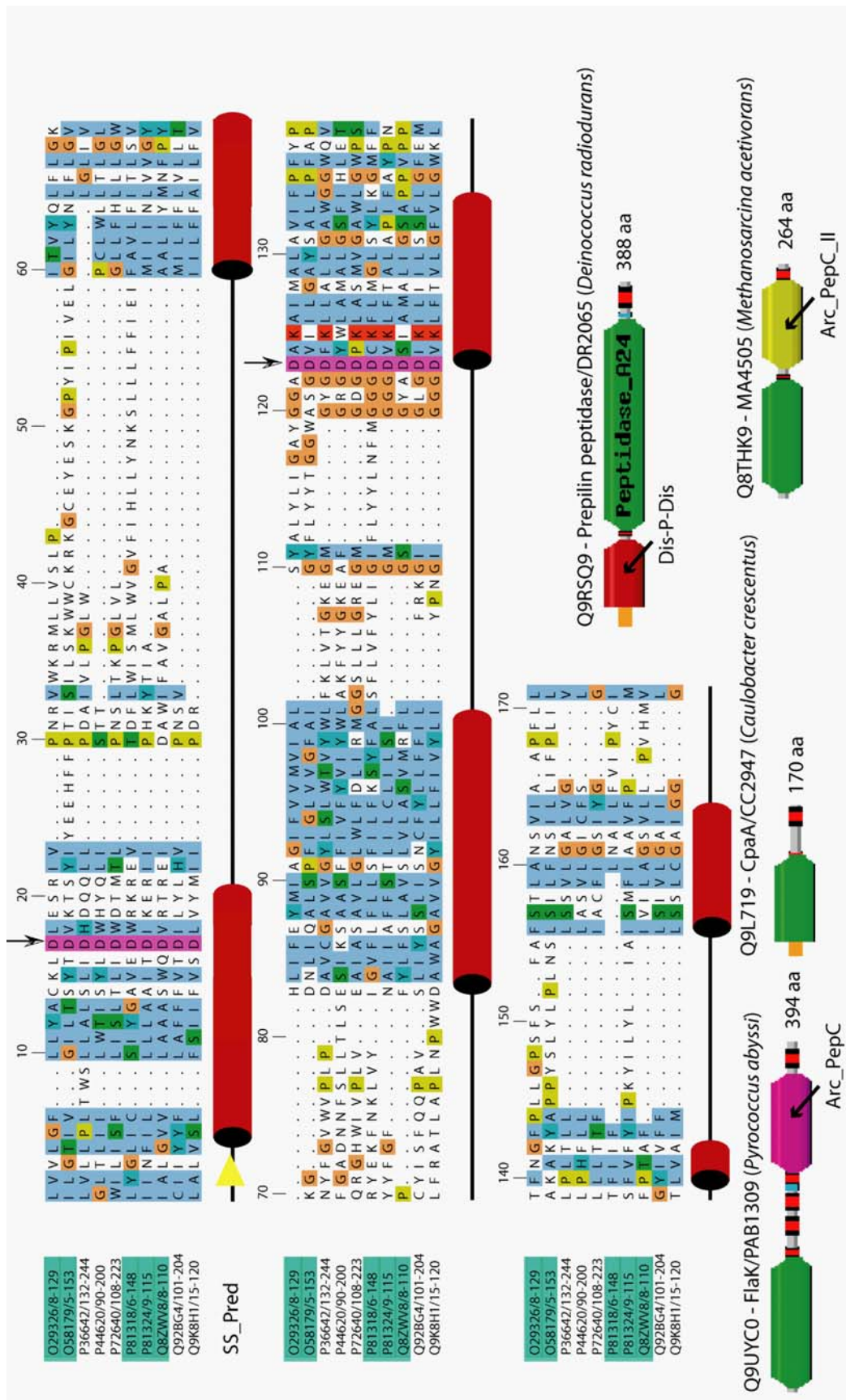


Figure 4.11: Peptidase_A24 example alignment and domain architectures

In the alignment the archaeal sequences are highlighted with green boxes. As can be seen they appear to be structurally similar to the eubacterial versions. The black arrows above the alignment indicate the two active site apertures

peptidase domain at the N-terminus and an Arc_pepC_II domain at the C-terminus. Hence it may be that the methylation activity is found in the DiS-P-DiS domain. Researchers can now begin to understand the differences in function between the various A24 peptidases through appreciation of the variances in domain architecture.

Expansion of the family also highlights some more questions to be resolved. Whilst most species have one Peptidase_A24 protein, it is now clear that some have more. For instance, *P. aeruginosa* has two, with the undescribed second consisting of only the Peptidase_A24 domain. It appears that most of the proteins with this architecture are described only as “hypothetical proteins”, presumably because the similarity to the other signal peptidases had not been found. However, some of them do have some annotation in the literature. *Actinobacillus actinomycetemcomitans* TadV has been implicated as being part of the Tight Adhesion operon. This operon is described by Kachlany, Planet *et al.* (2001) as encoding the assembly and release of a long bundled fimbrial leader pilus (Flp), and similar operons are found across bacteria and archaea.

It would seem that these previously unrecognised A24 peptidases indicate the existence of alternative secretory systems for specialised pili or flagella. This observation is supported by research into ApfD of *Actinobacillus pleuropneumonia*, which is architecturally similar to TadV. It is suggested by Boekama, Van Putton *et al.* (2004) that ApfD is the leader peptidase responsible for processing the pilus required for adhering to lung epithelia. This prediction is based mostly on *in silico* analysis though and so is not definitive.

Of the distribution of A24 peptidases, the most unusual species are the Vibrionaceae. While they commonly have more than one, *Vibrio vulnificus* strain cmp6 has one while *Vibrio vulnificus* strain YJ016 has three. What the differences between these two strains imply is not clear.

In conclusion, delineation of the correct domain boundaries of the A24 peptidases simplifies characterisation and enables the correct classification of the family members. As is now clear the signal peptidases are almost ubiquitous in the prokaryotes, and that archaeal flagella and bacterial pili are processed according to a similar secretory mechanism. It will be interesting to see how a species with more than one signal peptidase is able to target proteins to the correct pathway.

5 Contributions to Genome Annotation Projects

In this chapter I will describe contributions I have made to the annotation of newly sequenced genomes. A major part of the Sanger Institute's work is to sequence and annotate the genomes of pathogens and other microbes of economic or medical importance. These genomes can be an ideal source of interesting novel domains. By working with experts in a species' biology, interesting proteins can be rapidly identified and examined in detail. In turn they can also help provide a detailed understanding of a novel domain. The resulting observations provide a useful framework for future research. I have carried out investigations into four species, three prokaryotes and one eukaryote. These investigations have involved single protein families (the WiSP proteins of *Tropheryma whipplei* – chapter 5.1) and whole genomes (*Burkholderia pseudomallei* – chapter 5.2). Some of the work refines previous knowledge (Chlamydia polymorphic membrane protein family – chapter 5.3); while some is entirely novel. For instance, in *Theileria annulata* and *Theileria parva* I describe the initial characterisation of two correlated, but unrelated, short repeat families (chapter 5.4).

5.1 *Tropheryma whipplei* (Bentley, Maiwald *et al.*, 2003)

5.1.1 Background

Tropheryma whipplei is the causative agent of Whipple's disease, an extremely rare multisystemic chronic infection, with symptoms developing over several years. Currently Whipple's disease infects tens of people every year (Fenollar and Raoult, 2001). Primarily it reduces the body's ability to absorb carbohydrate and fat nutrients by destroying the microvilli on the surface of the small intestine, but it also has effects on the immune system. The organism had resisted characterisation due to difficulty in

culturing it, but in the year 2001 a method of growing it on human fibroblasts was developed (La Scola, Fenollar *et al.*, 2001) and in 2002 its genome was sequenced (Bentley, Maiwald *et al.*, 2003).

It is a small Gram positive rod-shaped bacterium that belongs to the Actinomycetes, though it is not closely related to any of the cultured relatives (Wilson, Blitchington *et al.*, 1991). It also appears to have a trilaminar appearance, with the outer membrane possibly being derived from the host (Silva, Macedo *et al.*, 1985). The genome consists of a single chromosome 925, 928 base pairs in length, which encodes 784 protein sequences, and has a low G+C content (46.3%) relative to the other Actinomycetes.

5.1.2 The WiSP Protein Family

Analysis carried out by the sequencing team – the "PSU" – found that it had a reduced genome size and was likely to be dependent on the host for several essential compounds - for instance it is missing genes required for amino acid biosynthesis and carbohydrate metabolism. However, the genome has an unusually low coding sequence density (84.4%), largely due to two non-coding DNA repeat clusters – RC1 and RC2. As noted in chapter 1.4 the average gene density for a bacterium is around 86% and this figure would be expected to be higher in an intracellular pathogen due to selective pressure for a minimum genome size (i.e. the Chlamydia typically have a coding density of 90%). Three proteins (TW157, TW161, and TW570) from a family denominated WiSP (for Whipplei Surface Proteins) were associated with these regions. The annotation team identified the WiSP proteins through clustering of the genome using a single-linkage clustering method developed by A. Bateman to reveal

14 related proteins (Bentley, Maiwald *et al.*, 2003). It is worth noting at this point that may be not all of these are expressed. Some of the domain architectures appear to be fragments of a complete protein (i.e. Q83N67) and so may be pseudogenes.

Several other lines of investigation by Bentley and co-workers highlighted this family. 10 of these 14 proteins have N-terminal signal peptides, and five appeared to have C-terminal transmembrane helices. Of the 17 genes in the genome that have pronounced nucleotide anomalies, WiSP proteins account for 11 of them and exhibit unusual dinucleotide content, codon usage and positional base preference. One of them, TW642, is one of five *T. whipplei* genes that appear to be under the control of a phase variable mechanism. Phase variation is a random process by which genes can be switched on and off between generations through length variation in short repeat tracts (reviewed in van der Woude and Baumler (2004).

Of the most unusual finds associated with this family, it was discovered that of all the 48 variable loci in the shotgun clones, all but one were located in one of two WiSP-encoding genes - TW157 and TW570. TW157 is located in RC1 and TW570 is located in RC2. Since the population from which the genome sequence was derived was clonal, this variation was not initially present in the culture but must have arisen during passaging. Further investigation then revealed that all the variable sequences found in these proteins were also found in the repetitive intergenic portions of RC1 and RC2. This implies that the variation in TW157 and TW570 was generated by a novel gene conversion mechanism, presumably involving recombination between coding and non-coding repeats, and so thereby generating novel alleles.

The WiSP proteins appear to be surface proteins. The implication of having such an intricate mechanism for rapidly varying them and of having such a large amount of the genome sequence dedicated to them is that they are major antigens. Hence I carried out a novel domain hunt in order to determine relationships to proteins in other species and characterise them further.

5.1.3 The WiSP Domains

Three domains were identified in the WiSP proteins - the WiSP N-terminal domain (WND), the C-terminal conserved domain family (CCD) and the WiSP β -stranded domain (He_PIG – for Haemagglutinin Putative Ig-fold). All the WiSP proteins contained the He_PIG domain except TW774, which only contained a CCD domain - see Figure 5.1 for full architecture diagram.

WND (WiSP N-terminal Domain; PF07861)

The WND domain is around 260 amino acids long and is highly conserved, with only a difference of a few residues between the copies. The domain sequence showed compositional bias, with a high proportion of serine and threonine residues (see Figure 5.2 for an alignment). It is predicted to be composed mostly of β -strands, and some α -helices. The function of this domain is not clear.

CCD (WiSP C-terminal Domain; PF07860)

This family is found at the C-terminus of TW113, and accounts for the whole length of TW774. TW776 and TW113 are very similar except for the absence of a signal peptide at the N-terminus of TW776 and the C-terminal domain is truncated. TW774 shows 94% identity to the C-terminus of TW113. The function of this region is

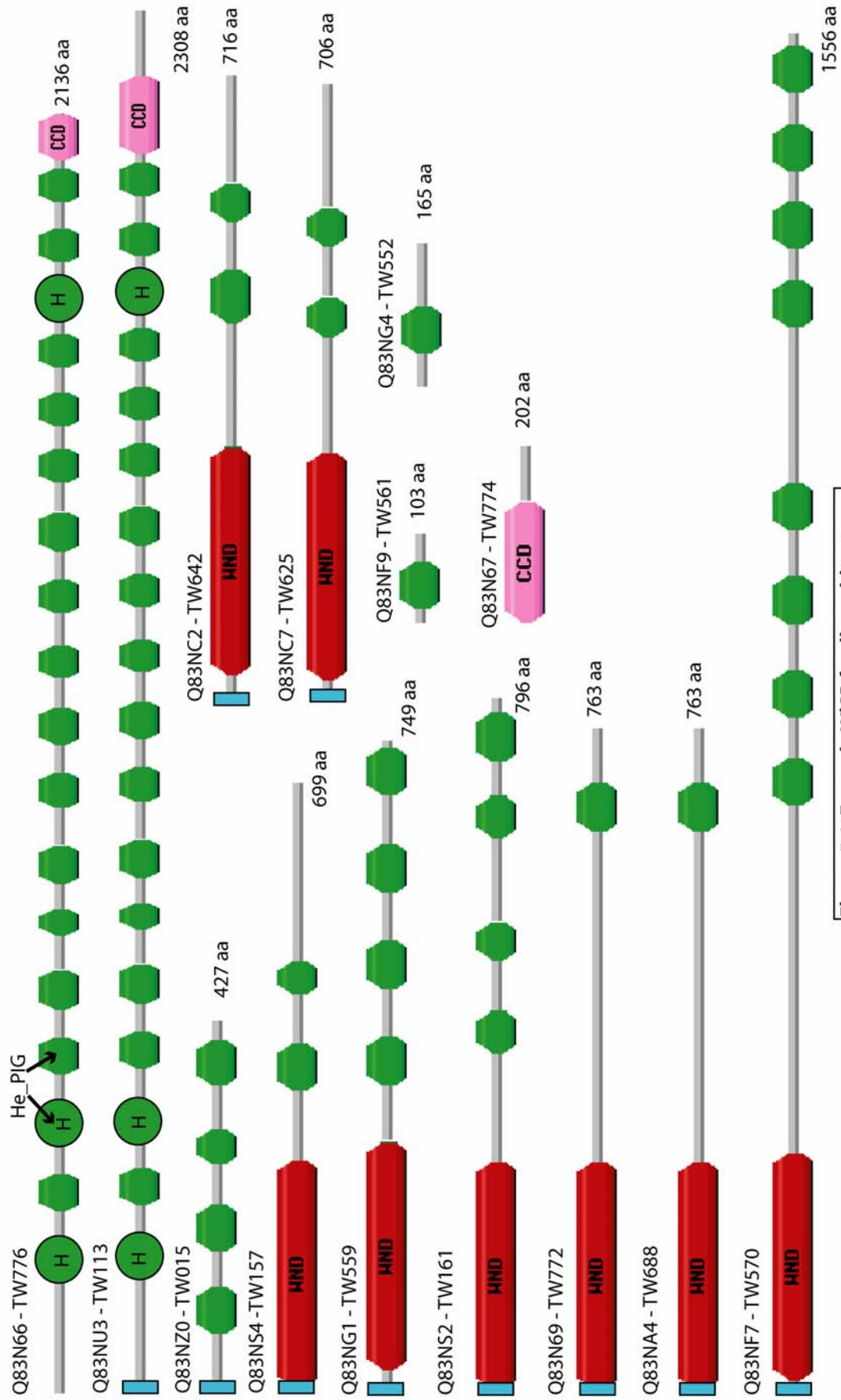


Figure 5.1: Example WISP family architectures

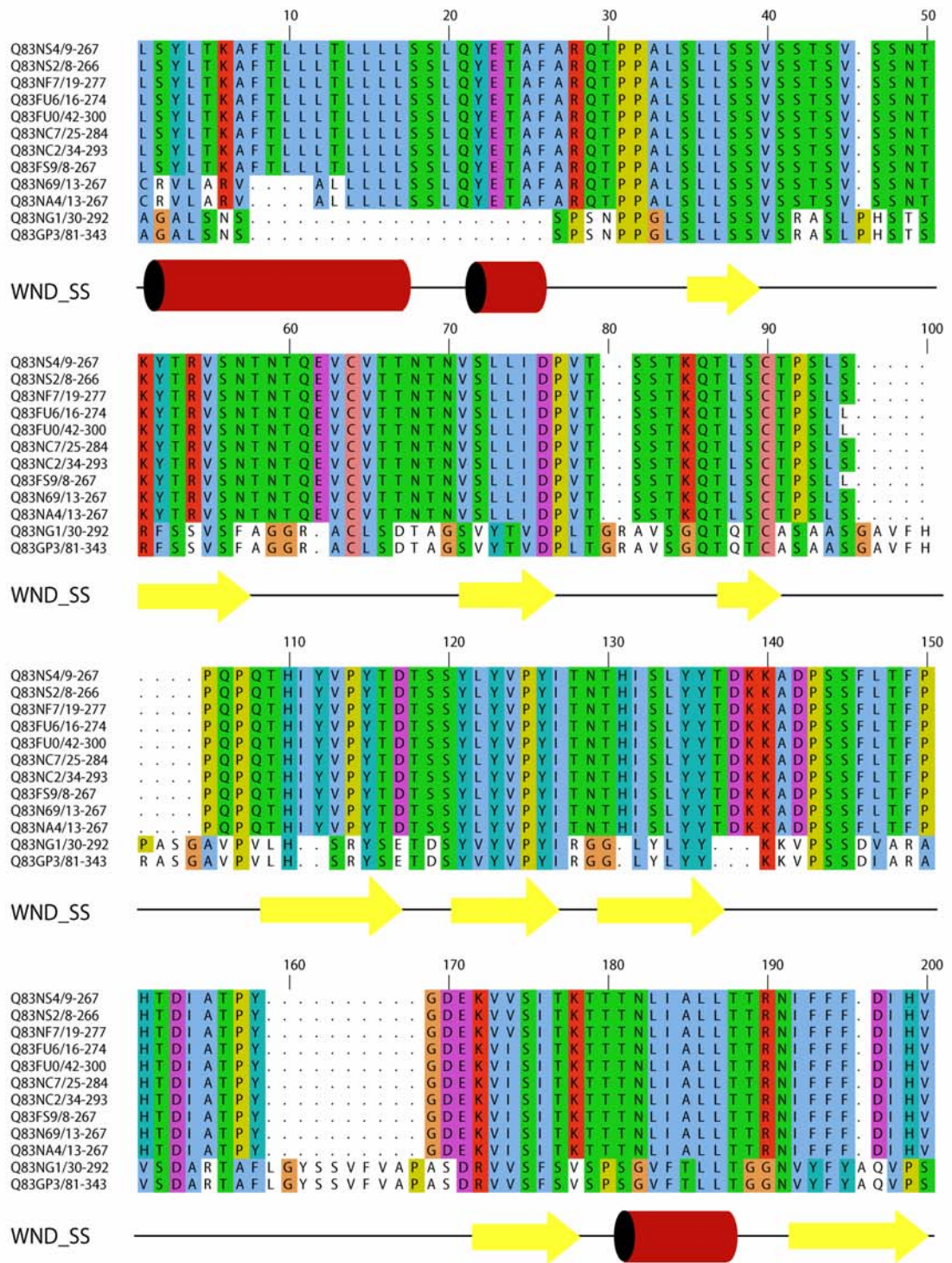


Figure 5.2: Example WND alignment (Page 1)

		210		220		230		240		250																																								
Q83FS9/8-267	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NC7/25-284	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83FU0/42-300	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NF7/19-277	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83FU6/16-274	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NG1/30-292	S	F	Q	S	S	V	T	V	N	V	K	Y	D	A	K	T	G	V	I	R	S	A	T	P	A	L	R	G	S	P	F	T	Y	S	L	S	T	P	V	A	G	V	R	L	D	A	N	T	G	A
Q83GP3/81-343	S	F	Q	S	S	V	T	V	N	V	K	Y	D	A	K	T	G	V	I	R	S	A	T	P	A	L	R	G	S	P	F	T	Y	S	L	S	T	P	V	A	G	V	R	L	D	A	N	T	G	A
Q83NS2/8-266	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83N69/13-267	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NS4/9-267	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NA4/13-267	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	
Q83NC2/34-293	T	E	K	P	K	I	T	V	P	I	H	K	Q	I	D	N	T	Y	L	S	D	I	P	S	L	R	N	S	R	Y	T	F	S	L	T	H	P	N	K	D	I	T	I	D	R	Y	T	G	Q	



		260		270		280		290																																
Q83FS9/8-267	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	L	A	D	E	P	R	V		
Q83NC7/25-284	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	L	A	D	E	P	R	V		
Q83FU0/42-300	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	H	N	P		
Q83NF7/19-277	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	H	N	P		
Q83FU6/16-274	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	H	N	P		
Q83NG1/30-292	L	S	G	S	V	K	D	A	T	V	G	A	H	G	L	T	A	T	A	V	H	Y	T	T	G	T	V	V	T	I	R	Y	L	F	.	D	S	P	V	P
Q83GP3/81-343	L	S	G	S	V	K	D	A	T	V	G	A	H	G	L	T	A	T	A	V	H	Y	T	T	G	T	V	V	T	I	R	Y	L	F	.	D	S	P	V	P
Q83NS2/8-266	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	H	N	P		
Q83N69/13-267	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	Q	Y	R		
Q83NS4/9-267	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	H	N	P		
Q83NA4/13-267	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	I	.	D	T	Q	Y	R		
Q83NC2/34-293	I	H	L	S	.	.	.	S	L	P	T	S	P	I	T	A	I	A	I	N	R	D	T	T	T	H	I	T	.	Y	A	L	A	D	E	P	R	V		



Figure 5.2: WND example alignment (Page 2)

entirely unknown; its secondary structure is predicted to be mostly unstructured with either one (PHD) or two (PROF) α -helices (see Figure 5.3).

He_PIG (Putative Ig fold Haemagglutinin; PF05345)

Repeat elements were identified in several members of the WiSP family using Dotter. Examples were aligned and used to iteratively search against the WISP family members in order to identify all the repeats. In total 67 copies were found in the WiSP family (see Figure 5.1 for architectures). This alignment was then used to search against UniProt. As of Pfam 15, there were 243 copies in UniProt; this is likely to be an underestimate since it was difficult to distinguish this domain family from two others of similar structure (see below). Many of the matches were to other long proteins with a similar repetitive nature, including the *Staphylococcus aureus* Biofilm Associated Protein (BAP; UniProt:AAK38834). The modular nature of the family suggests that they represent structural domains (see Figure 5.4 for some example architectures).

The domain has a median length of 107 residues, but only the central 35 residues seem to show strong conservation. Secondary structure predictions suggest that it consists mostly of β -strands (see Figure 5.5). This concurs with suggestive matches found to HYR (PF02494) and PKD (PF00801) domains, which are Ig-fold domains. As discussed briefly in chapter 1.4 Ig-like domains can bind nearly any compound, and so they are often found in cell surface proteins..

The identification of these domains in the WiSP proteins gives some clues as to their function. One possible role would be to mediate specific pathogen-host cell

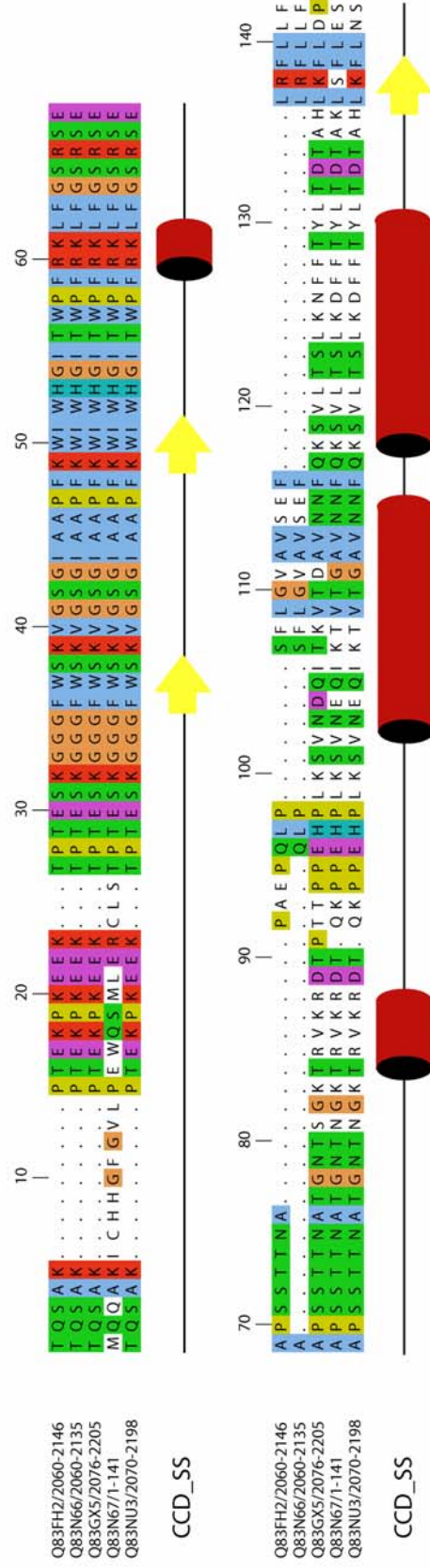


Figure 5.3: CCD example alignment

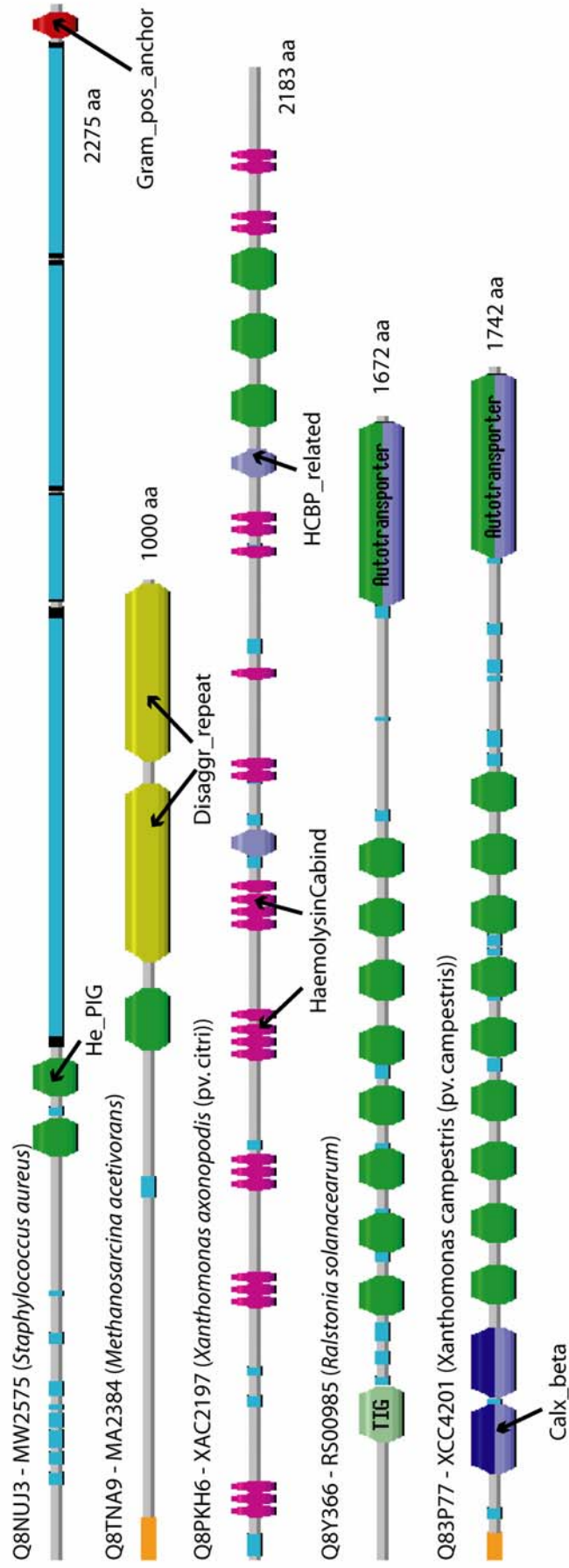


Figure 5.4: Example He_PIG architectures

interactions by direct contact - possibly by recognising specific structures on the target cells. A second role for substrate-binding domains on the cell surface is to mediate cell-cell interactions through an intermediate structural compound. For instance, in *Staphylococcus epidermis* the accumulation associated proteins, which consist of a chain of the N-acetylglucosamine-binding G5 domains (PF07501), mediate biofilm formation by binding a carbohydrate slime called polysaccharide intercellular adhesin (PIA). A third possibility is that the Ig-like domains bind other proteins, which then carry out an enzymatic or structural role. Which of these three roles, or possibly an unknown function, the WiSP domains play is not clear, but they clearly play an important role in the biology of *T. whipplei*, and, given its life-style, probably a role in pathogenesis.

To further elucidate the evolution of these proteins a Neighbour-Joining tree of each domain copy was built using Belvu (see Figure 5.6). The resulting tree does not show a single clear pattern of relationships between these copies as would be expected if these proteins had diverged from a single common ancestor, but rather a complex pattern of multiple internal duplications as well as whole gene duplication. Deletions are harder to find evidence for, but may also have happened.

All the repeats in TW776 and TW113 pair up in order down their whole length, showing that they are result of gene duplication. In contrast, the first four repeats of TW570 are more closely related to each other than any other repeats except one from TW157; the second four are not closely related to the first four, but show virtually no difference to each other or the He_PIG domains found in TW561 and TW562. There

are several potential mechanisms by which such a pattern could occur, but none of them are a straight forward process of divergence.

Another unusual pattern concerns TW015 in relation to TW776 and TW113. Each of the He_PIG domains found in TW015 is closely related to a pair from TW776 and TW113 (as noted above the He_PIG domains from these two are virtually identical); however, they are not in the same order. The TW015 domains are all roughly the same distance from their corresponding pair in TW776/TW113 (see Figure 5.6) suggesting that they separated from the ancestral sequence at the same time; however, they do not occur in the same order and they come from the middle of these proteins – despite also having a signal peptide. If we consider the central five He_PIG domains of TW776/TW113 to be named A-B-C-D-E then the four domains in TW015 occur in the order C-D-A-E.

There are two explanations for this pattern. One is that TW015 was formed by a duplication of the ancestor of TW776/TW113 and then went through some domain shuffling and loss event. The second explanation is that TW015 was constructed in a separate event to the TW776/TW113 ancestor, from a common source of He_PIG domains in which order is essentially arbitrary; this would support the novel gene conversion mechanism proposed by Bentley and co-workers. It is possible that both mechanisms are at work. Having many copies of the same domain in close location on the genome allows more scope for domain shuffling events, including semi-homologous recombination. This would also further maximise the rate of generating novel antigens.

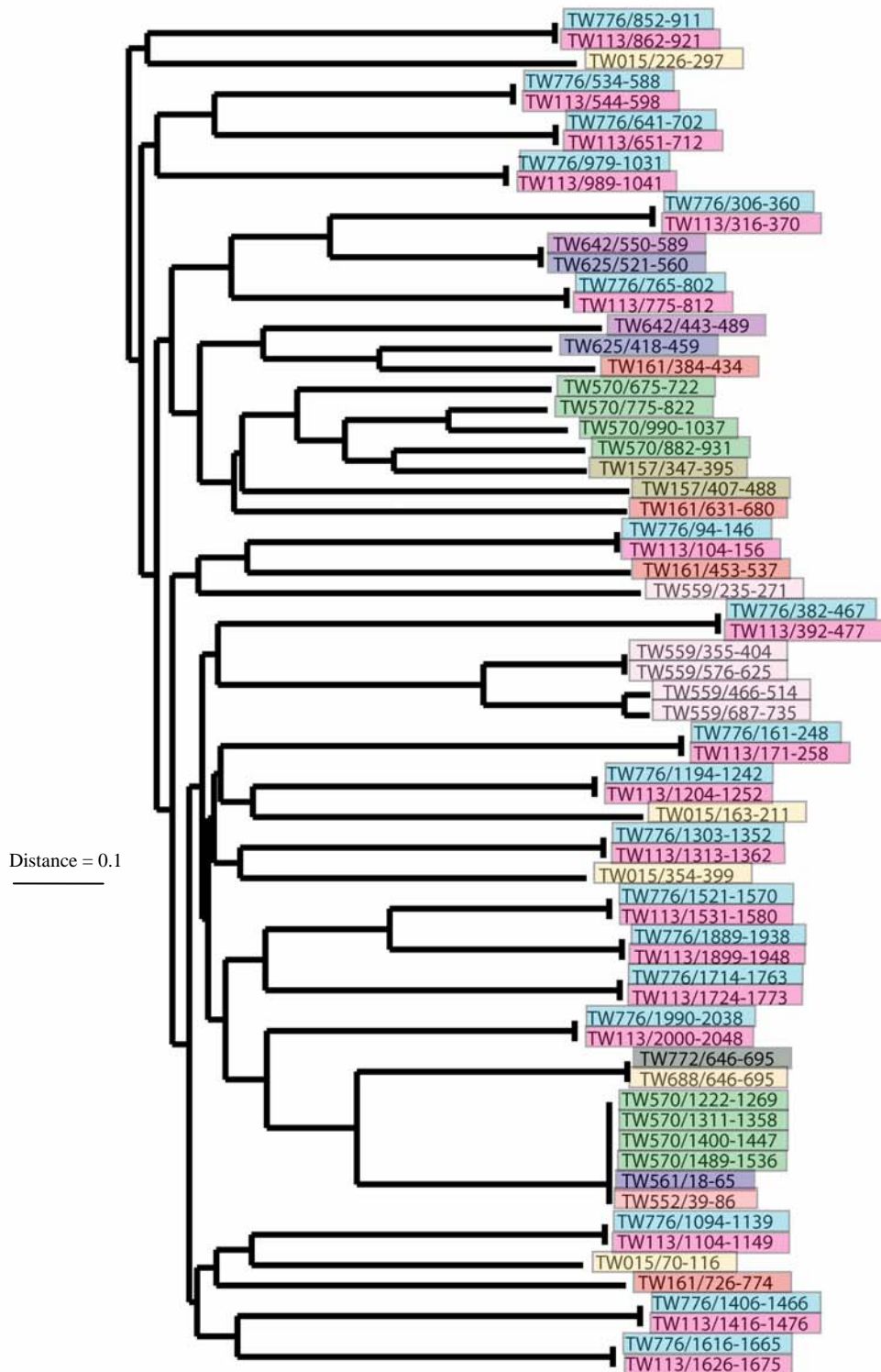


Figure 5.6: N-J Tree of all He_PIG domains from *Tropheryma whipplei*
 The tree was constructed in Belvu using uncorrected distances and the “center of tree” approach. The tree balance equals 0.0. Each leaf represents a He_PIG domain; the sequence name and coordinates are given. Each protein has been assigned its own colour for easier identification.

5.1.4 Implications for the Immune System

The WiSP family proteins do appear to be the major antigens of *Tropheryma whipplei* – there exists multiple copies that can be switched on and off, it is highly variable and also is apparently capable of rapid evolution. The novel gene conversion mechanism identified during sequencing of the genome indicates a method by which new forms can be introduced into these proteins. The complex pattern of similarities between the WiSP proteins also suggests that there may be frequent domains shuffling events, through which variation could be further distributed. Hence over several generations it would be possible for a clonal *T. whipplei* population to become a highly variant population with respect to their surface structures, making it possible for the organism to sustain a chronic infection.

5.2 Burkholderia pseudomallei (Holden *et al.*, 2004)

5.2.1 Background

Burkholderia pseudomallei is a Gram-negative soil-dwelling bacterium endemic to East Asia and Northern Australia (Chaowagul, White *et al.*, 1989). It is one of the primary causes of septicaemia in this region, and also can cause pneumonic disease when inhaled. The symptoms can vary greatly, leading the organism to be dubbed "the great mimicker", and an individual's response to the bacterium can vary greatly; this includes an instance of it lying dormant for 26 years before causing melioidosis (Koponen, Zlock *et al.*, 1991). However, overall mortality is around 40%, and the lack of a vaccine has led to the organism being classified as a category B agent on the US Centre for Disease Control's potential bioweapons list (<http://www.bt.cdc.gov/agent/agentlist.asp>).

The bacterium is rod shaped and has two chromosomes, one of 4.07 Mb and one of 3.17 Mb, encoding 3,460 and 2,395 genes respectively (Holden, Titball *et al.*, 2004). Core functions are primarily housed on the larger chromosome I and accessory or hypothetical genes are mostly found on Chromosome II. Since it is a saprophytic soil dwelling organism rather than an obligate pathogen, it encodes genes for the biosynthesis of many of the nutrient compounds it needs to survive and is adapted to commensal living in the roots of plants.

Given the size of the genome and the large selection of accessory genes it was thought that domain hunting may provide some valuable insights.

5.2.2 Novel Domains

SCPU (Spore Coat Protein U domain; PF05229)

This domain is around 60 residues in length, is predicted to have an all- β secondary structure (as predicted by PHD and PROF) and is found exclusively in the Proteobacteria. There are currently two recognised domain architectures, the difference being whether there is one or two SCPU domains (see Figure 5.7). In both architectures most of the proteins have signal peptides and/or transmembrane helices, suggesting a common function on the cell wall. In the literature two functions have been ascribed to SCPU. Firstly they are described as a component of the spore coat in *Myxococcus xanthus* (Gollop, Inouye *et al.*, 1991); secondly they are described as a component of a specialised type IV pili involved in biofilm formation on plastic and glass surfaces in the species *Acinetobacter baumannii* (Tomaras, Dorsey *et al.*, 2003) and *Pseudomonas aeruginosa* (Vallet, Diggle *et al.*, 2004). So it is likely that this

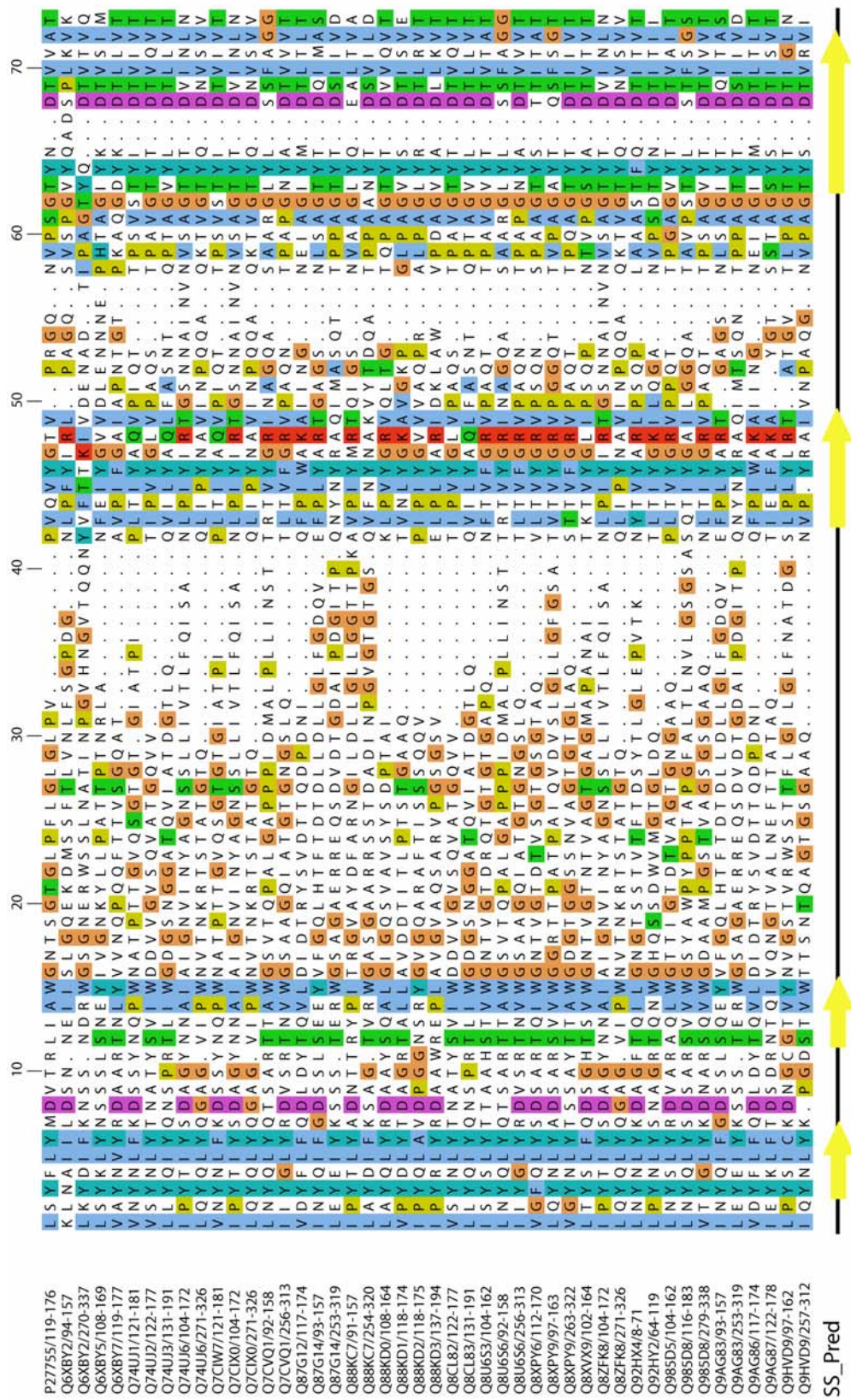
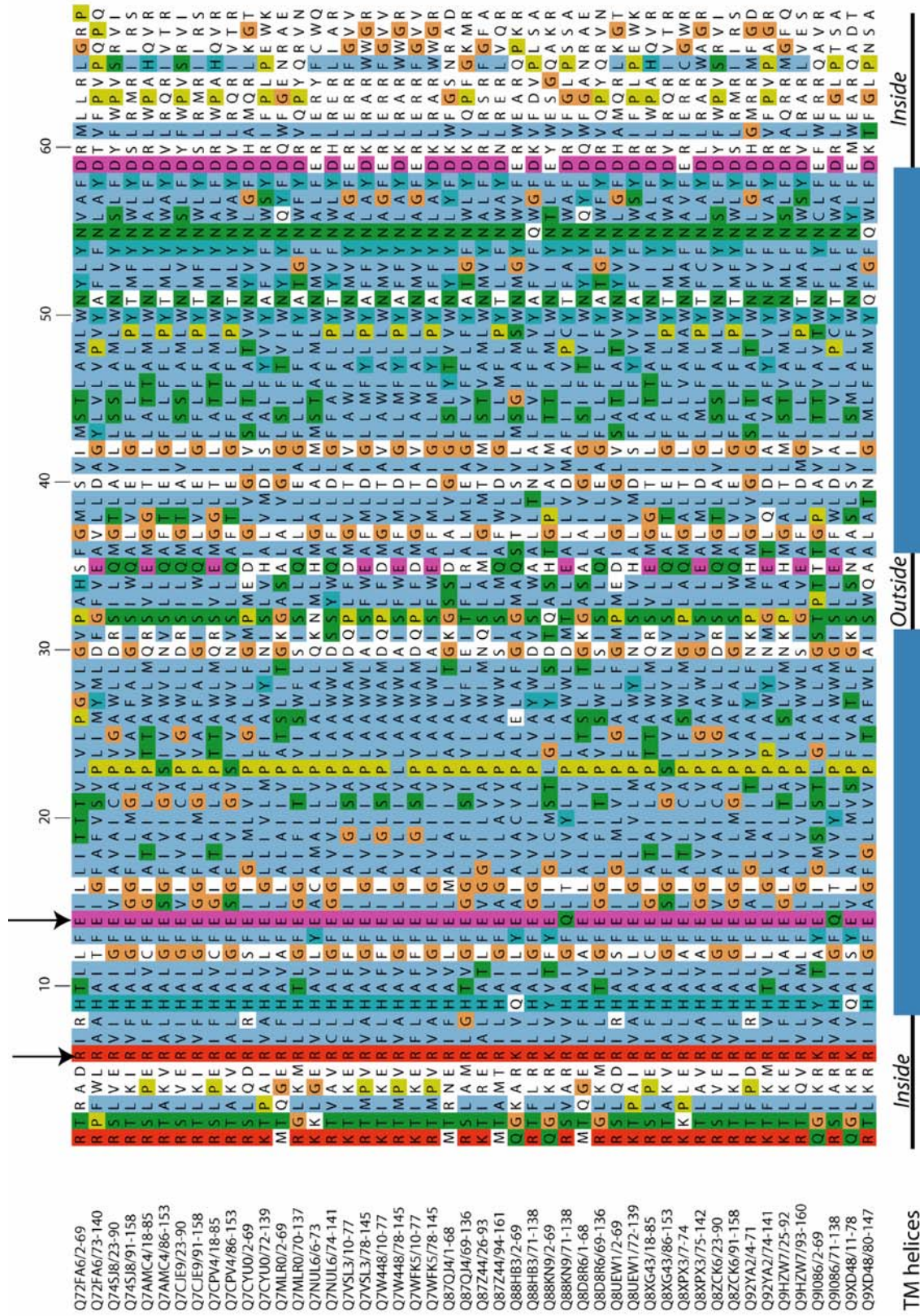


Figure 5.7: SCPU example alignment and architectures

domain is involved in attaching the bacterial cell to smooth surfaces in the case of *Myxococcus xanthus* as well. The domain itself may not directly attach to surfaces, but may bind a sticky intermediate. There are four mostly conserved tyrosines that may be functionally important.

BTP (Bacterial Transmembrane Pair family; PF05232)

This exclusively Proteobacterial family consists of a conserved pair of transmembrane helices with a short loop in between them (see Figure 5.8). All the BTP-containing proteins contain two copies of BTP and some also have a signal peptide, suggesting they are tightly associated with the outer membrane. Although none the family members have been experimentally annotated in any way the alignment shows some similarity (fs model E-value = 0.03) to a transmembrane region of a Ca⁺/Na⁺ antiporter (UniProt:Q9PW6, residues: 239-294), though this may be a spurious similarity caused by the medium compositional complexity of transmembrane helices. Whether there is a genuine evolutionary or functional link is not clear as several residues that show strong conservation in BTP are not conserved in the antiporter; though there are some fully conserved residues that are also found in the antiporter including an arginine residue, a glutamate and an aromatic residue. In all the proteins where BTP is found, the two BTP domains are very close together (i.e. one or zero residues separating them) and all the loops between the helices are the same length, suggesting fairly tight constraints on structural variance in this family. I propose that these proteins may form a pore in the cell wall, possibly a passive cation channel.



Q6LHE4 - PBPRB1419 (*Photobacterium profundum*) BTP BTP 140 aa

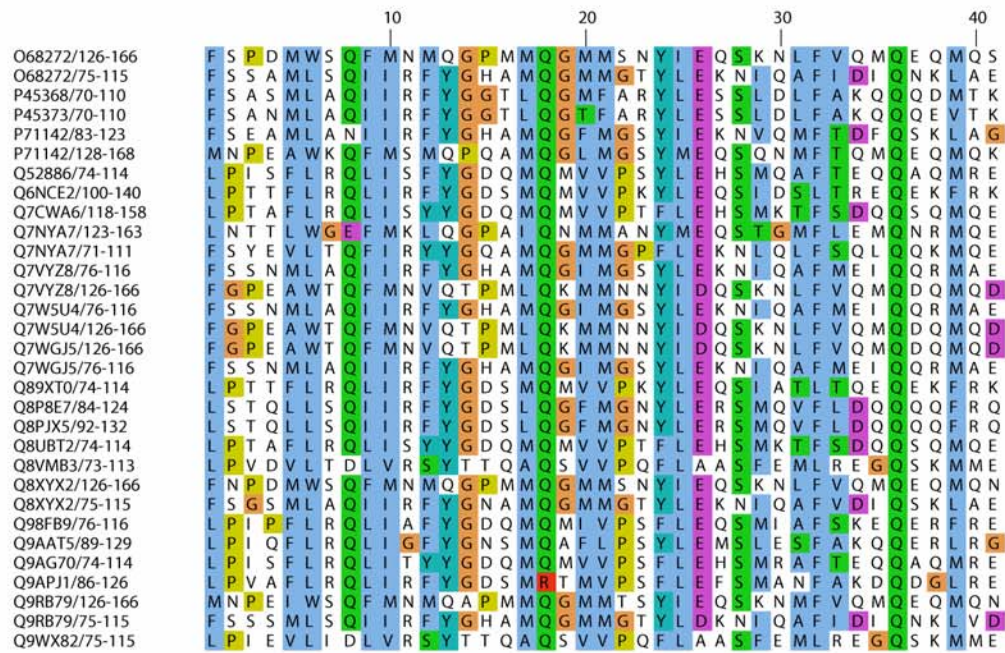
Figure 5.8: BTP example alignment and architectures
 The mostly invariant arginine and glutamine residues are marked by the black arrows

PHB_acc (PHB accumulation negative regulator; PF05233)

PHB_acc is a short region of around 35-40 residues in length and predicted to consist of two α -helices (see Figure 5.9). It occurs in a family of Proteobacterial regulators normally known as PhaF or PhbF, though the most characterised is PhaR of *Paracoccus denitrificans*. These regulators either have one or two copies of PHB_acc at their C-terminus and a more conserved region called PHB_acc_N at the N-terminus (see Figure 5.9). They are regulators of carbon flow in the Proteobacteria and are involved in controlling the generation of carbon stores in the form of poly-(beta-hydroxyalkanoate) copolymers (PHB) (e.g. Encarnacion, del Vargas *et al.*, 2002). Maehara, Taguchi *et al.* (2002) demonstrated that PhaR is able to bind PHB, which also causes it to disassociate from DNA. Since the N-termini of these proteins is conserved throughout the family, I would suggest that DNA-binding function resides there, while the PHB-binding function resides in the PHB_acc domains; as has been noted several times in this thesis, binding domains often vary in copy number so as to influence affinity.

The Repetitive β -helix Surface Structure Superfamily

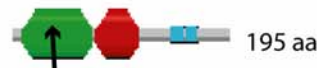
The cell surface proteins of bacteria are of great interest to biomedical research, and microbiology in general. The bacterial surface defines how the organism interacts with the environment, and in the case of pathogens the major surface proteins form both the sites that the host immune system recognises and the means by which they recognise target host cells. As has been seen in the case of the He_PIG proteins (see chapter 4.1) and the PPC domains (see chapter 2.1) cell surface proteins are often repetitive and modular in nature. This structure provides certain advantages to the



PHB_acc_SS



Q9WX82 - PHA responsive repressor (*Paracoccus denitrificans*)



PHB_acc_N

Q8XYX8 - RSc1634 (*Ralstonia solanacearum*)



PHB_acc

Figure 5.9: PHB_acc example alignment and architectures

bacterium. Extensive repetitive regions are more prone to various forms of replication error, leading to internal domain duplications and deletions; this in turn allows the cell surface to rapidly evolve new functions, refining their mechanisms for interacting with the environment, and evading immune system recognition.

Work carried out in *B. pseudomallei* led to the identification of two new families, Hep_Hag and Fil_haemagg, that were subsequently shown to be related to each other and to the Ice_nucleation family (see Figure 5.10 for examples of each family; see Figure 5.11 for example Fil_haemagg architectures). Work carried out in *Chlamydomophila abortus* led to the redefinition of the Chlam_PMP family and recognition that it is related to the other filamentous haemagglutinin families (see Chapter 4.3.2). The Ice_nucleation family is a small specific subset of the overall superfamily with little sequence variation; similarly, the Chlam_PMP represents a narrow family that has specifically expanded in the Chlamydia. The other two are extremely divergent and found in the Proteobacteria, Fusobacteria and the Firmicutes. All are predicted to form a β -helical structure. The structure of the filamentous haemagglutinin B (UniProt:P12255; PDB:1rwr) of *Bordetella pertussis* has been recently solved and confirms that they form a β -helix (Clantin, Hodak *et al.*, 2004).

Members of this very divergent superfamily have been implicated as being of critical importance in a wide-range of bacteria. For instance HecA of *Erwinia chrysanthemi* EC16 has been shown to be involved in attachment, aggregation and destruction of host (*Nicotiana clevelandii*) epidermal cells (Rojas, Ham *et al.*, 2002). *Bordetella pertussis* requires filamentous haemagglutinin A for invasion of respiratory tracts (Coutte, Alonso *et al.*, 2003). UspA1 of *Moraxella catarrhalis* is able to bind tissues

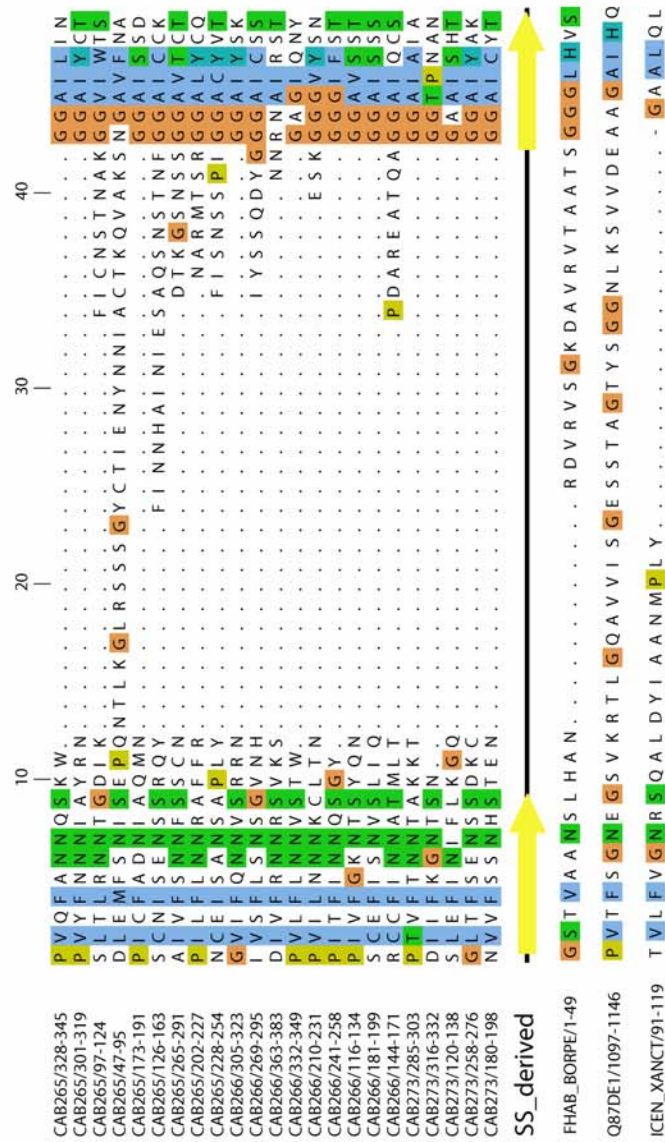


Figure 5.10: Example Chlam_PMP alignment with related sequences
 The main alignment shown here is of the Chlam_PMP family. Underneath are examples of the related Fil_hameagg, Hep_hag and Ice_nucleation families (respectively).

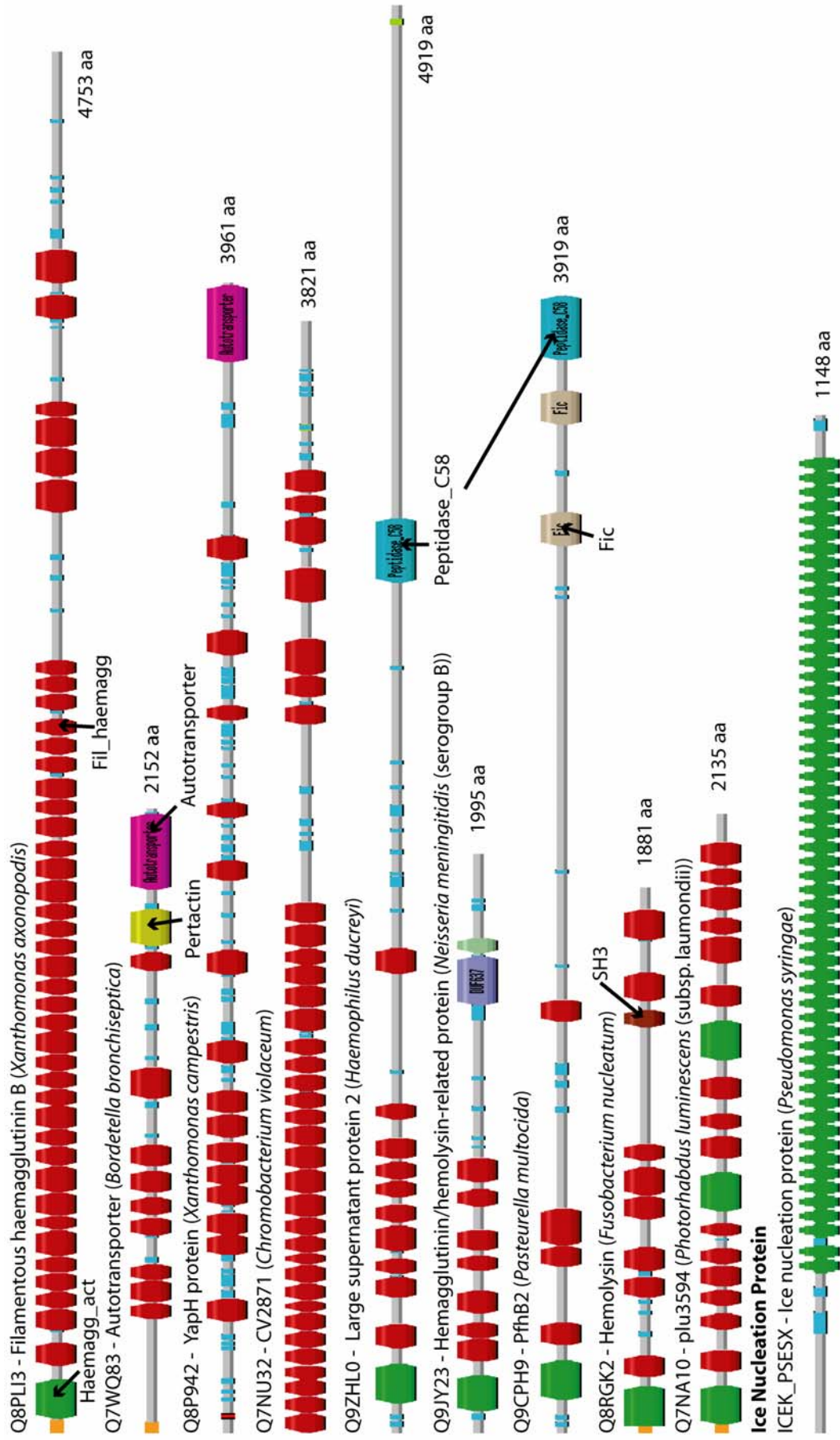


Figure 5.11: Fil_haemagg and Ice_nucleation example architectures

derived from human lung and middle ear (Holm, Vanlerberg *et al.*, 2003). YadA of *Yersinia enterocolitica* enables it to attach to and invade mammalian cells (Eitel and Dersch, 2002) through binding cell surface proteins.

Whilst this family shows a range of diversity for attaching to different cells and surfaces, the Ice_nucleation subfamily shows a surprising variance in function. This small family allows the bacterium to initiate the freezing of water at near zero degree centigrade temperatures; this enables the bacterium to cause frost damage to the host – presumably making cell invasion easier (Gurian and Lindow, 1992). Several studies support the presence of these proteins as a marker of virulence (i.e. Smirnova, Li *et al.*, 2001).

The reason for the high level of similarity (average of 65% calculated by alistat) of the Chlam_PMP subfamily is not clear, and they don't appear to have any functional quirks like the Ice_nucleation subfamily; it may be an artefact of them having descended from a single ancestral protein. This subfamily is discussed further in Chapter 5.3.

Whilst this family of proteins is fairly well studied, clarifying the nature of the repeats and improving the representative models will enable better characterisation of their variance and relationships, and allow better comparative modelling of their structures. These structural models will be of significant import as they are the major antigens of many of the species they occur in and will help direct discovery and refinement of pharmaceutical drugs and vaccines.

5.3 Chlamydomphila abortus (Manuscript under preparation)

5.3.1 Background

Chlamydomphila abortus is a pathogenic member of the Gram negative Chlamydiaceae and is endemic to ruminants. It is of particular economic concern as it resides in the placenta and triggers abortions in pregnant farm animals. It has also been seen that pregnant women in close contact with these animals can pick up the infection and miscarry (Longbottom and Coulter, 2003), so the bacterium represents a potential zoonose with possibly devastating consequences.

Like most of the Chlamydiaceae, it is an obligate intracellular pathogen with a biphasic life cycle. It enters cells as a small round infectious elementary body (EB), which then transforms into the larger replicative reticulate body (RB). While in the cell the bacteria live in small vacuoles called inclusion bodies; these vacuoles are able to avoid the endocytic pathway and instead join the exocytic pathway. The RB undergoes several rounds of binary fission, and the progeny transform into EB and are released into the body through cell lysis or exocytosis.

The genome is about 1.15 Mb and contains 961 coding sequences, of which 27 were pseudogenes (personal communication: N. Thomson). Of particular interest in the Chlamydiaceae are the polymorphic membrane proteins (PMPs) for several reasons: They are the major antigenic proteins (Cunningham and Ward, 2003). They are directly involved in pathogenesis (Wehrl, Brinkmann *et al.*, 2004); and they are the best current candidates for developing Chlamydia vaccines (Christiansen, Pedersen *et al.*, 2000).

5.3.2 The Chlamydial Polymorphic Membrane Protein Family

The major outer membrane proteins of the Chlamydia have been the primary target for vaccine development, and with the completion of several genomes it has become clear that these proteins belong to a single divergent family (Kalman, Mitchell *et al.*, 1999) – the Polymorphic Membrane Protein family (PMP) or Chlam_PMP in Pfam. Sequencing of *C. abortus* has shown that it also has 18 members of this family in its genome (see Figure 5.12). Evidence was found of phase variation within several PMP genes (Pedersen, Christiansen *et al.*, 2001), as well as a possible mechanism for recombining two different PMP operons, and also frequent duplication and deletion as evidenced by the variance found in gene number between different Chlamydiaceae (Gomes, Bruno *et al.*, 2004). To further characterise these proteins an investigation was carried out to refine their domain architectures.

The step was refining the Pfam model of the Chlam_PMP domain (Pfam 13) by redefining it as a set of tandem repeats rather than a single unit (see Figure 5.11). By correcting the boundaries, a relationship to the β -helix filamentous proteins was identified – as discussed in chapter 5.2.2 – allowing the determination that they form a β -helix. A novel domain (ChlamPMP_M) was also identified, which is discussed below.

ChlamPMP M (Chlamydia PMP Middle Region; PF07548)

Pfam families covering the N-terminal β -helix repeat region and the C-terminal Autotransporter domain had been previously created. However, a conserved, approximately 160 residue, region that occurs between these two regions had not been

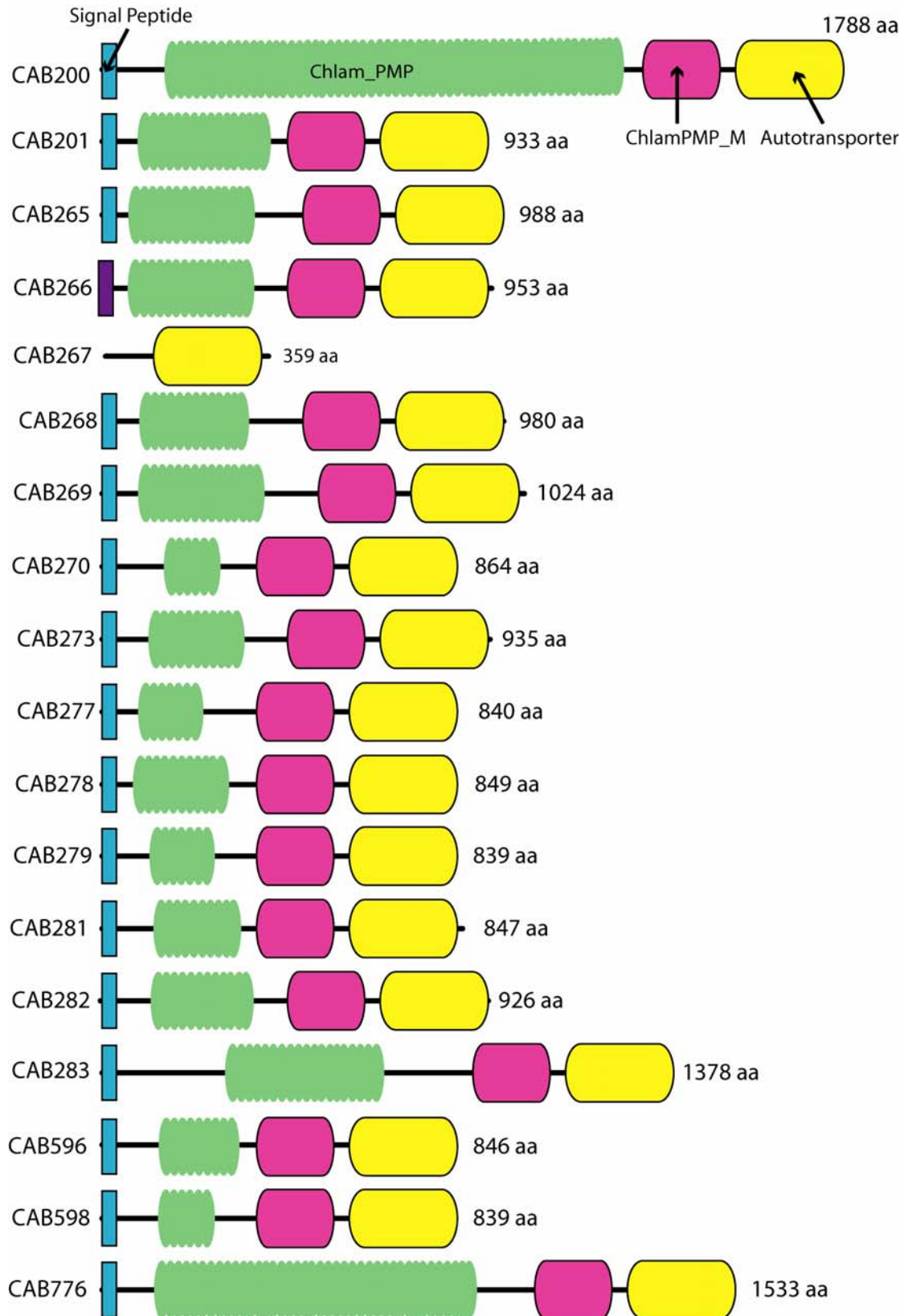


Figure 5.12: The *Chlamydomonas reinhardtii* polymorphic membrane protein family
 This family is not shown in the usual Pfam representation since the protein sequences are not yet available in UniProt. The signal peptide of CAB266 is shown in purple as there is some doubt over its validity. SignalP 3.0's two prediction methods gave conflicting results.

noted. Although the overall conservation of this region is quite low – approximately 27% average identity – several motifs and residues are nearly completely conserved (see Figure 5.13). The region is predicted to have an all- β structure. The function of this region is not known, but its discovery does fit with unexplained phenomena in the literature. Recently Wehrl, Brinkmann *et al.* (2004) observed two cleavage products from the *Chlamydomophila pneumoniae* PmpD protein subsequent to export from the cell. One part was the N-terminal region, which was closely associated with the membrane; the second was the middle region. This implies that this region is removed subsequent to secretion in order to form the final product. So the role of this region is likely to be as an aid to exporting the β -helical stem. Potential specific roles are either to occlude the haemagglutinin region and prevent it binding to proteins within the cell, or to aid localisation of the PMP_N region to the cell surface. Understanding how these proteins mature and are secreted may lead to insights on how to interfere with this organisms pathogenic abilities.

The ChlamPMP_M and Autotransporter regions can also be used to build a reliable evolutionary tree. The repeat regions show variation in length, which may bias any attempt to build a tree. This is because the similarity in length between some of these proteins may cause them to be scored as more similar to each other when in fact there are proteins of a different length that are more closely related. One way this could happen is if the repeat regions of two separate proteins duplicate themselves in independent events. Thus these two proteins would align better with each other than genuinely closer, but much shorter, relatives. Since these proteins always have a single ChlamPMP_M domain and a single Autotransporter domain in the same order and same position in the proteins, the C-terminal regions make a stable and

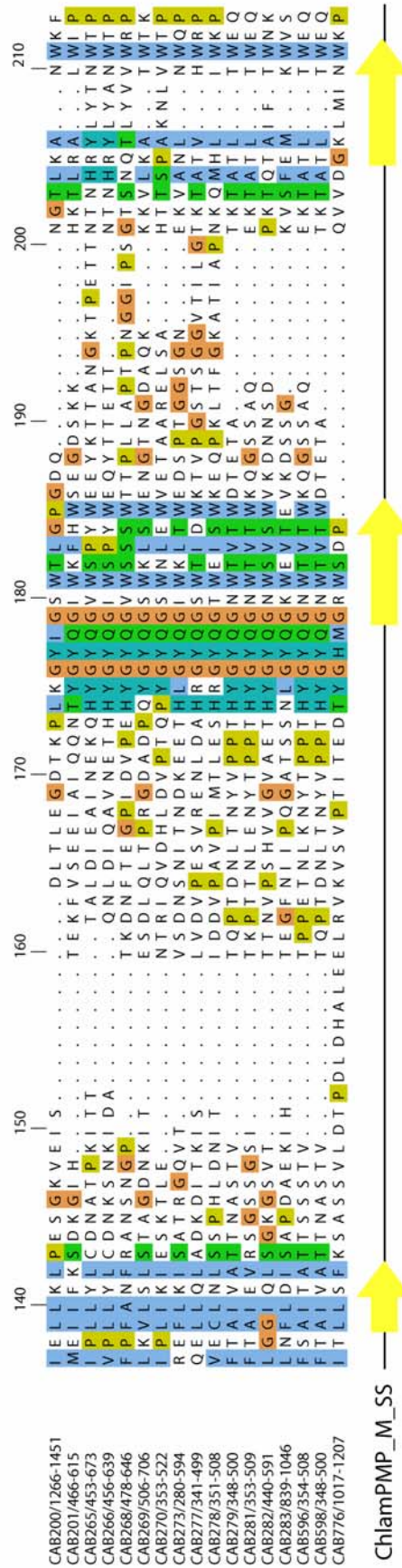


Figure 5.13: ChlamPMP_M example alignment (Page 2)

comparable set for building a phylogenetic tree. Currently an investigation is underway in conjunction with N. Thomson to examine the effect of building trees using different portions of these proteins, and whether this will help refine our understanding of their evolution.

5.4 Theileria annulata (Manuscript under preparation)

5.4.1 Background

The tick-borne eukaryote *Theileria annulata*, along with its relatives such as *Theileria parva*, is a major cause of cattle disease in tropical and sub-tropical regions. This makes it of major economic import to several developing countries, and it has the potential to cause a significant humanitarian disaster. It is a member of the Apicomplexa, and so is related to the malaria-causing Plasmodia; hence it is also hoped that it will provide some insight into malarial biology.

It also shows some highly unusual life cycle features (Dobbelaere and Kuenzi, 2004). *Theileria* species are the only known eukaryotic intracellular parasites that trigger cancer in order to maximise their replication, in a manner reminiscent of the bacterial plant pathogen *Agrobacterium tumefaciens*. *T. annulata* and its relative *T. parva* both show similar host species range and mechanisms of infection but show a different host cell specificity. *T. parva* infects T-cells, whereas *T. annulata* infects macrophages. The decision to sequence these organisms was made partly on the basis of their economic import, but also to try and determine the tumourogenic factors and what difference between the two closely related species causes the difference in host cell preference.

The genome of *T. annulata* consists of four haploid chromosomes of 4.5, 2.0, 1.9, and 1.8 Mb, encoding approximately 3800 protein coding genes (personal communication: A. Pain). The proteins were clustered by the genome annotation team using the Tribe-MCL algorithm to identify large or potentially important families. Potential clusters of interest were highlighted and the domain architectures investigated by me. I found that nearly all the major clusters were variants around a single theme - a family of proteins that consist of varying numbers of a single highly polymorphic domain. This domain is discussed below.

5.4.2 FAINT (Frequently Appears IN Theileria; PF04385)

The initial identification of this domain was made by W. Mifsud in representative proteins in UniProt. However, the initial boundaries were incorrectly assigned, and consequently the model had a low sensitivity; whilst they were approximately the correct periodicity, they were shifted along so that the C-terminus of the domain was recognised by the N-terminus of the model. Analysis of the repetitive nature of these proteins using Dotter enabled the assignment of better positioned boundaries, which enabled significant expansion of the family. The investigation was carried out using conservative judgements as the domain proved to be very variable (less than 20% average identity in *T. annulata* alone) and the searches were carried out solely against the *T. annulata* genome, so as to increase the significance of weak hits (see chapter 1.5 for a brief discussion on the effect of database size on E-values).

Eventually over 700 copies were identified in around 150 proteins - almost 5% of the species' proteins. An example alignment and architectures are shown in Figures 5.14

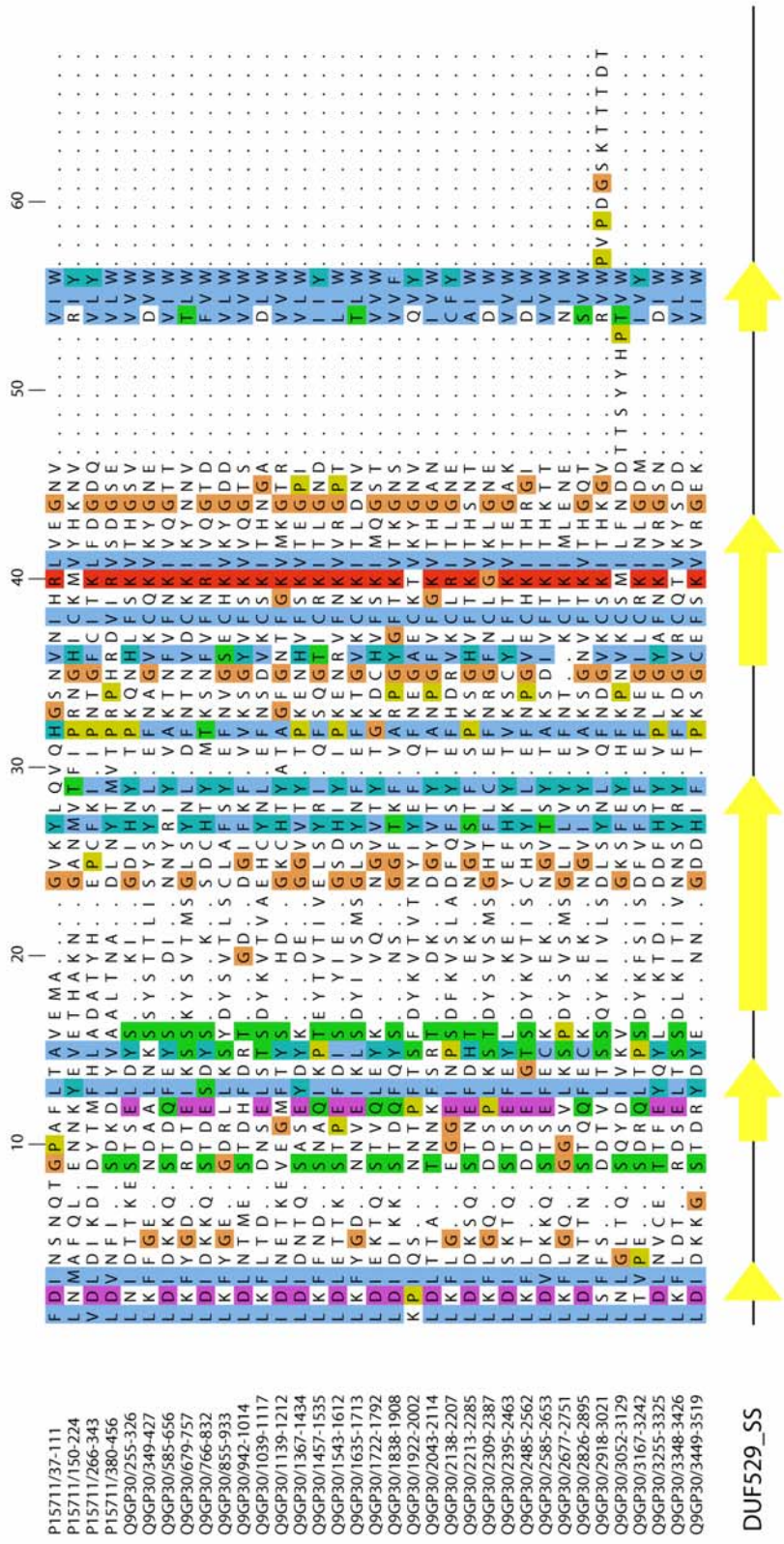


Figure 5.14: FAINT example alignment (Page 1)

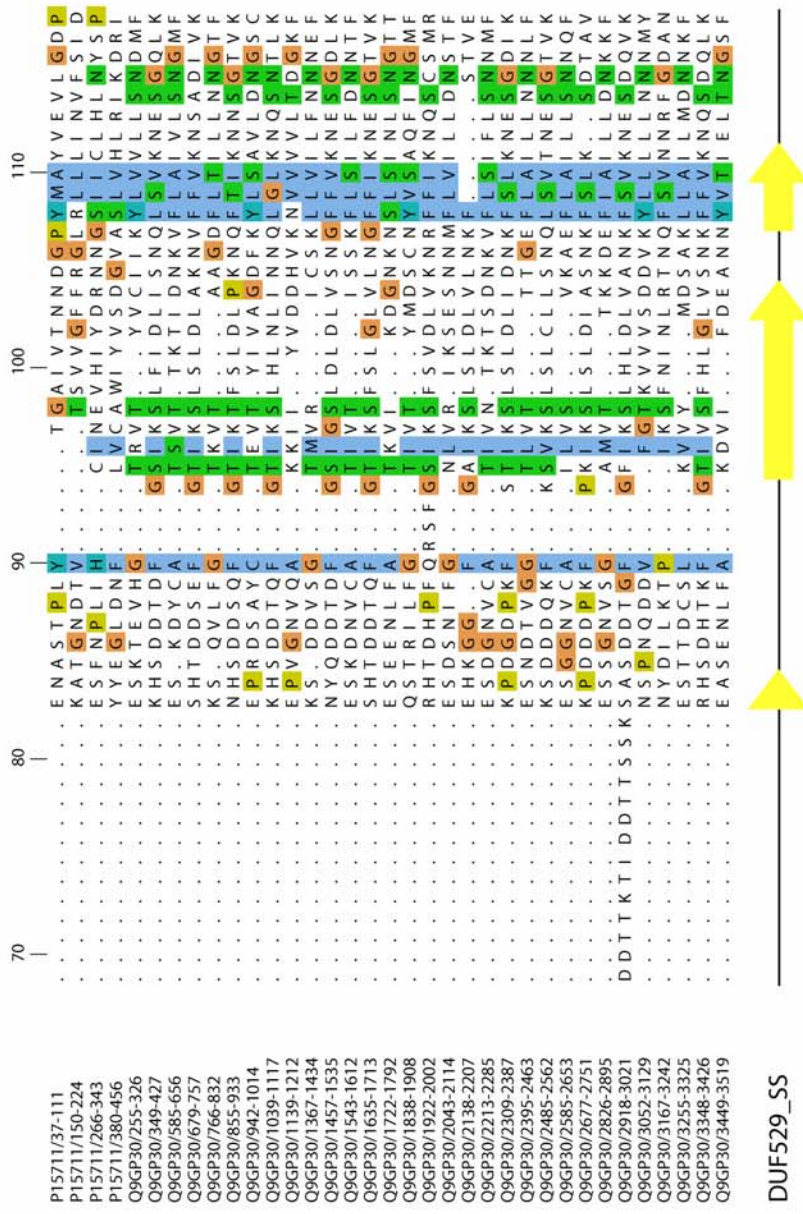


Figure 5.14: FAINT example alignment (Page 2)

and 5.15 respectively. FAINT-containing proteins ranged from containing a single copy (i.e. TA03165) up to 54 copies (i.e. TA16050), were secreted or cell wall-associated, and had no other domains. An exception is in the case of the TashAT proteins, which are discussed further below. The domain is around seventy residues in length and is predicted to have an all- β secondary structure. Searching against the *T. parva* genome demonstrated that it was present in a similar number of proteins; however, searching against UniProt identified only one homologue beyond the Theileria, in the closely related Piroplasmida, *Babesia equi*. Other completely sequenced Apicomplexa genomes, including *Plasmodium falciparum*, did not appear to encode this domain.

Previous studies had identified a small family of *T. annulata* proteins that were secreted during infection and localise to the host cell nucleus (Swan, Phillips *et al.*, 1999), and were called the TashAT proteins (and included SuaAT). These proteins contained AT-hook regions, which should allow them to interact with DNA, and also contained several motifs that may allow them to interact with the components of the cell cycle. These proteins also contained the FAINT domain at their N-terminus and the literature does not report any removal of this region during their export to the host nucleus.

Several of the antigens reported for *Theileria parva* turn out to be composed entirely of FAINT domains. For instance the polymorphic immunodominant molecule (PIM) protein contains at least 4 copies. The PIM gene locus has been reported as extremely polymorphic, with regular and rapid insertion and deletion events, driven by an unknown mechanism (Toye, Gobright *et al.*, 1995).

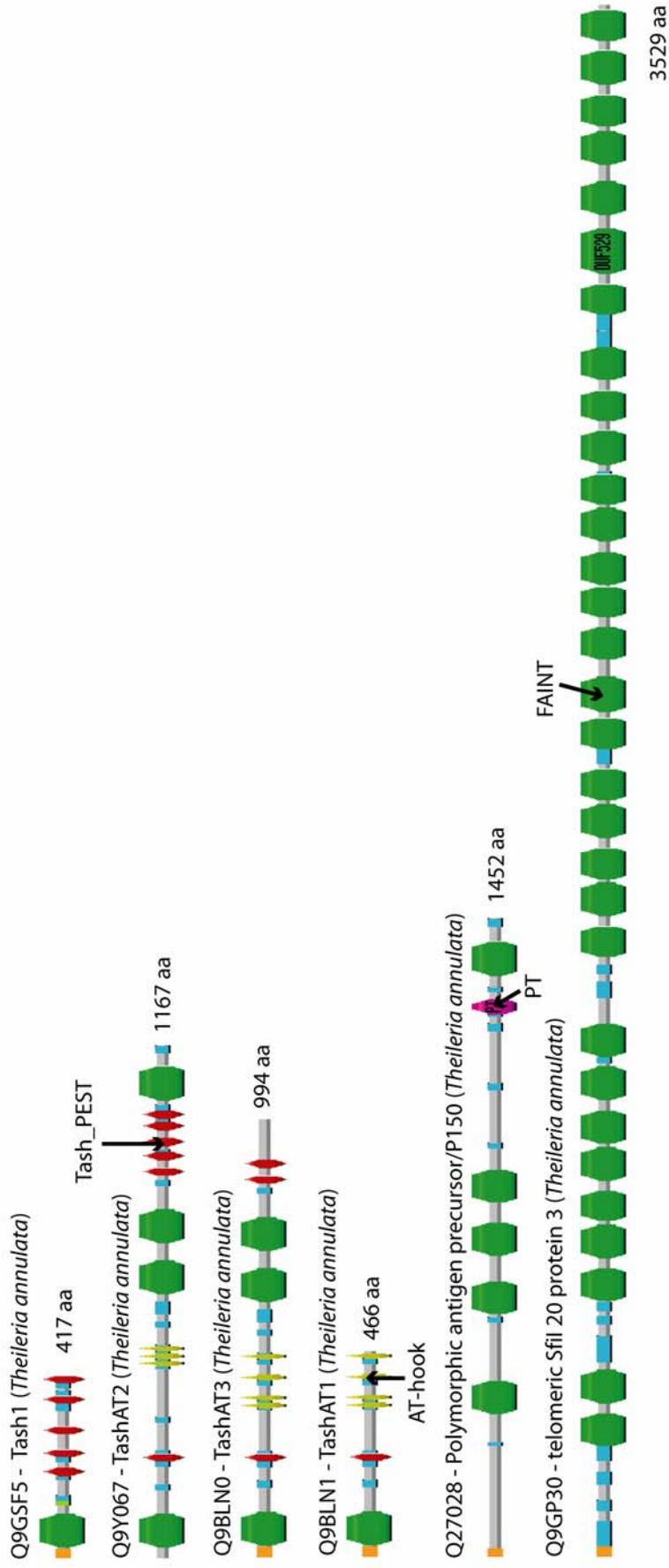


Figure 5.15: FAIN example architectures

The general lack of conserved residues in the alignment suggests that this domain performs a binding or structural role; there is a mostly conserved tryptophan, which is occasionally substituted by a tyrosine or a cysteine residue (see Figure 5.14). The FAINT domain's occupation of a significant proportion of the coding potential of the *Theileria* indicates that this domain is important, and the lack of obvious homologues in other species – excluding possibly the most related Apicomplexa – suggests that it is specifically important to the biology of *Theileria*. It would be interesting to find out if all these genes are expressed and if so, during which stage of the life cycle. Alternatively they may provide a means or source of domain copies to drive variation in the PIM protein.

5.4.1 The TASR Repeat Families

A second investigation into the protein family clusters focussed around the choline kinases of *T. parva* and *T. annulata*. These proteins were of particular interest to the genome annotation team as they were candidate tumorigenic factors, perhaps by forming part of the pathway by which the *Theileria* maintain the tumour state of the host cell (based on work in human cancer, such as that by Roberts, Stewart *et al.*, 2004). Aligning all the choline kinases found that both species contained a choline kinase with a large insert in the middle region of the domain. To confirm whether these two kinases were orthologues an NJ-tree was built using Belvu. The large inserts were masked so as to avoid the two proteins being grouped purely as a function of length, but instead to be grouped according to amino acid similarity. The resulting tree revealed that the choline kinases of the *Theileria* separate into two clear groups; one group containing the majority of the kinases and the other containing the two 'large insert' kinases (see Figure 5.16).

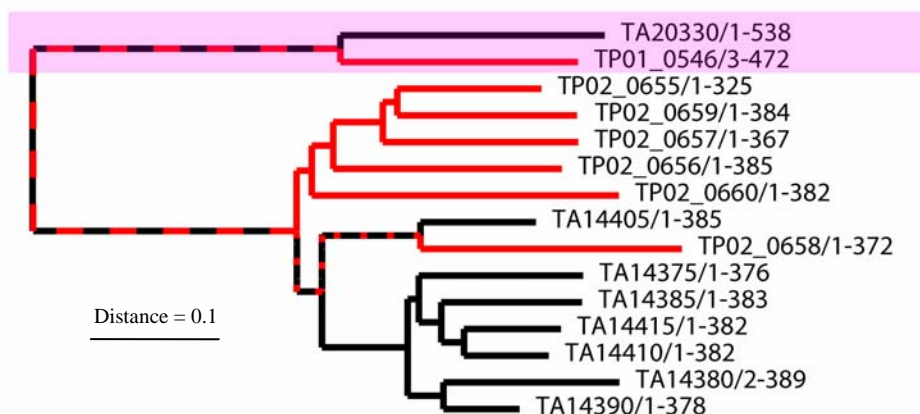


Figure 5.16: Neighbour Joining Tree of the choline kinases of *T. parva* and *T. annulata*
 The tree was constructed in Belvu using uncorrected distances and the “center of tree” approach. The tree balance equals 0.0. All the choline kinase sequences were aligned using MAFFT and then the large inserts in TA20330 and TP01_0546 were masked out, so that they did not influence the tree. These two proteins still clearly form an outgroup from the rest, supporting orthology (highlighted in purple). *T. annulata* proteins are marked by black lines, while *T. parva* proteins are marked by red.

Having established the evolutionary relationship between these two proteins the inserts were investigated for any distinctive characteristics that may shed light on their function. In both cases short repeats were visually identified, but at the amino acid level they showed no sequence similarity. For each set of repeats an alignment was built and searched against their respective genomes. The repeats were named TASR for Theileria Anomalous Short Repeat Families. The *T. annulata* repeat (TASR_1) is around 13 residues long and so contained enough information to build a reasonably discriminatory model; in contrast the *T. parva* repeat (TASR_2) was only three residues long, and so was extremely difficult to reliably identify. An alignment of six tandem repeats contained enough information to provide some specificity (see Figure 5.17 for alignments of the two repeats).

Iterative searching against the two genomes identified 103 proteins in *T. annulata* that contained TASR_1 and 67 in *T. parva* that contained TASR_2. The searches in *T.*

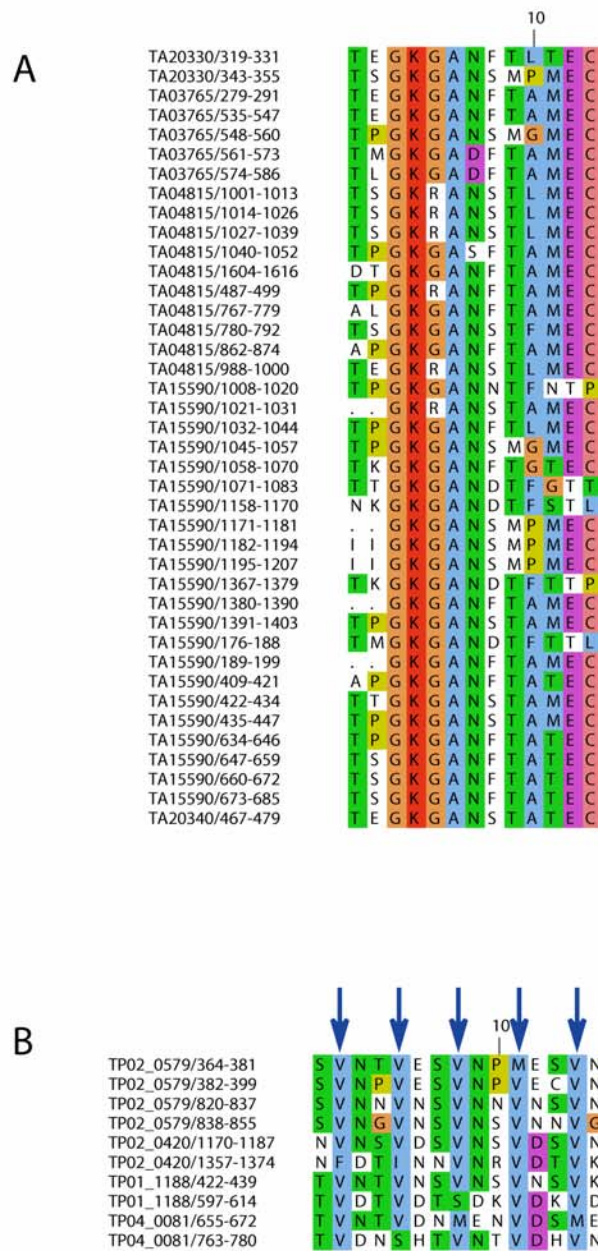
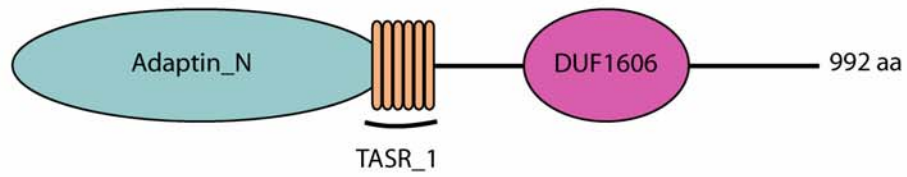
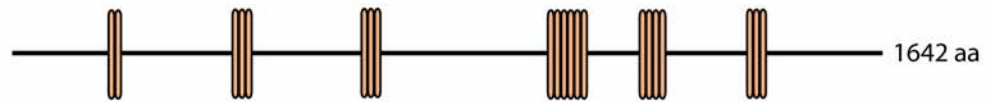


Figure 5.17: Example alignments for the TASR short repeats of *T. annulata* (A) and *T. parva* (B)
 The *T. annulata* TASR_1 repeat is 13 residues in length, and so can be reasonably confidently identified. Some of the repeats do appear to be 11 residues. The *T. parva* TASR_2 repeats are only three residues - the periodic valine residues are marked by the blue arrows. It is not clear whether these repeats are translated as part of the proteins or not.

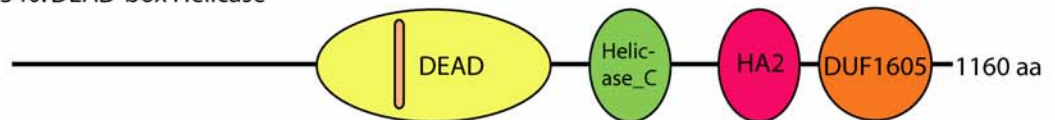
TA02765: Beta Coat Protein



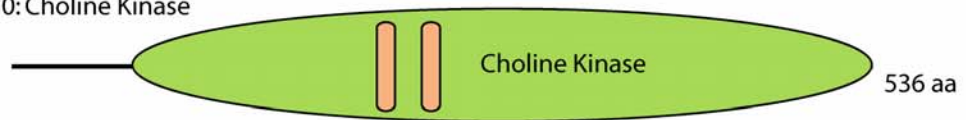
TA15590: Protein of Unknown Function



TA20340: DEAD-box Helicase



TA20330: Choline Kinase



TA07505: Ribosomal RNA adenine dimethylase



Figure 5.18: *Theileria annulata* TASR_1 example architectures
The normal Pfam view is not shown since these proteins were not yet available in UniProt. The TASR_1 repeats are depicted as orange bars. As can be seen they fall within, between and across the boundaries of domains.

parva used very low E-value thresholds (0.001) so as to attempt to ensure that no false positives were included. Searching with the *T. annulata* repeat against the *T. parva* genome demonstrated that it was not present; the reciprocal search was not sufficiently discriminatory to obtain a significant result. The orthologue of each protein that contained a TASR repeat was identified (personal communication: Arnab Pain) and then checked to see if it also contained a TASR repeat. Assuming that these repeats are not related, and hence that they were distributed around the genome in independent events, we can determine whether this overlap in orthologue sets is random by testing for significance against a binomial distribution. The test is performed each way, once for *T. parva* against *T. annulata* and once for *T. annulata* against *T. parva* (presented in Table 5.1).

<i>Annulata vs. Parva</i>				<i>Parva vs. Annulata</i>			
N ^o of orthologues with TASR_2 (A)*	P(success) = A/B	25	0.021	N ^o of orthologues with TASR_1 (A)*	P(success) = A/B	25	0.016
<i>T. parva</i> genome size (B)		4150		<i>T. annulata</i> genome size (B)		4000	
N ^o of TASR_1-containing proteins in <i>T. annulata</i>	N ^o of Trials	89		N ^o of TASR_1-containing proteins in <i>T. parva</i>	N ^o of Trials	65	
∴ assuming a binomial distribution the P(overlap by chance) = 0.00 (3 sig. figs).				∴ assuming a binomial distribution the P(overlap by chance) = 0.00 (3 sig. figs).			
* All proteins with a TASR, but no easily identifiable orthologue were discarded. This came to 15 proteins in all.							
Table 5.1: Statistics testing whether the overlap between the <i>T. parva</i> TASR_2-containing proteins and the <i>T. annulata</i> TASR_1-containing proteins is by chance.							

Doing the test for both the *T. annulata* set against *T. parva* and the *T. parva* set against *T. annulata* found very high levels of significance (P = 1) so it is safe to conclude that the overlap between these two sets is not by chance. Initial examination

of the DNA sequences failed to reveal any similarity, suggesting that their distributions have arisen after the separation of the two species and in two separate events. It also suggests that they would have been distributed by the same mechanism. The genome annotators noticed that the *Theileria* appeared to have several short DNA repeats around the genome (personal communication: Arnab Pain); these may also be TASR-like repeats, but require further characterisation. With regards to the type of proteins they were found in, visual examination of the domain structures suggests that RNA processing and vesicle formation was over-represented. However, this has not been definitively tested against the background genome.

Although we are currently unable to provide a definitive answer as to the function or nature of these repeats, or how they arise, there is clearly a biological process of interest occurring and possibly unlike anything previously described. There are several questions that are immediately apparent. For a start, it is not clear whether the TASRs are transcribed, or translated, or whether they perturb the protein structure. If they fulfil a role at the DNA level, what was it about this particular set of proteins that led to the insertion? They may even be the footprint of an invasive DNA sequence, like a transposon, and are deleterious to the organism. This phenomenon has become apparent with the sequencing of the genome and, it is hoped, that this investigation will serve as a starting point for its description.

6 Summary and Conclusions

At the start of this thesis I set out to find and describe novel protein domains that are of significant interest to biology in general. Through the identification of these sequence families I hoped to identify novel biological processes, elucidate previously unsuspected mechanisms in known processes, and to further characterise well studied proteins.

I chose to primarily study the proteins of Prokaryotes since there are many more sequenced prokaryotic genomes than eukaryotic genomes; this reduces any bias towards well studied families and allows increased focus on domains of general interest. However, the searches were not carried out exclusively in bacteria, and several domains that occur in Eukaryotes were identified - for instance the SCP domain and the Dabb domain.

The domain hunt methods I employed were generally aimed at identifying manageable (around 200) numbers of targets that may represent novel protein domains. One of the principal methods for identifying domains was based around identifying repeats within proteins; the second was based on clustering proteins of 100 residues or less in length.

Identification of sequence families allowed the construction of a multiple sequence alignment. Through these, several sources of information were related and interpreted, allowing the generation of new knowledge without recourse to laboratory experimentation. Data for each family was extracted from the published literature, from computational predictions - for example secondary structure predictions - and

from web sources. From the multiple sequence alignment itself, conserved regions and residues can be identified and correlated with previous observations. As an example the most conserved patch in the PepSY domain contains the loss-of-function mutation identified by Braun et al, 2000.

In total 41 domains, repeats, motifs and families are presented in this thesis and another 54 were identified but are not discussed. Most of these domains appear to be ligand-binding domains, or structural, with very few novel enzymes being uncovered. Many of them are also found on the bacterial cell surface. These are of particular interest, not only because of their involvement in pathogenicity, host interactions and antibiotic resistance, but also because of the difficulty in examining the structures of cell membrane associated proteins. By identifying the structural units it becomes simpler to excise them from the surrounding protein and solve their structure through crystallography or NMR; it also becomes possible to identify them in proteins that may be more amenable to laboratory-based investigation. Such an approach was successfully carried out by Wilson, Matsushita *et al.* (2003) to solve the structure of the PPC domain, and could work well with domains like He_PIG.

As has been found in the case of the PASTA domain, creating an alignment that links several species together can allow deeper interpretation of species specific information. In this case of PASTA I found that there appeared to be correlation between the number of PASTA-containing proteins in a genome and the different morphological types exhibited by that species. Specifically if a species has more than one PASTA-containing penicillin-binding protein (pPBP) or PASTA-containing serine/threonine protein kinase (pPSTK) then it will display more than one cell

morphology. *Streptomyces coelicolor* has three cell morphologies and three pPSTKs. This leads to the prediction that the *Bacillus cereus* group, which includes *Bacillus anthracis*, may have one more cell shape than other closely related Bacillales.

Other insights have been made into mechanisms of biofilm formation - the PepSY domain - and the immune system evasion methods of *Chlamydomonas abortus* and *Tropheryma whipplei*. Also, entirely novel processes have been uncovered, paving the way for new areas of research. As an example the short repeats found in the Theileria in chapter 5.4 have not been described before and initial characterisation suggests that they are the result of a process unlike anything previously described in the literature. Similarly the SCP domain has now been strongly implicated in the establishment of cell polarity and copper ion chelation, creating the framework for investigating the details of its function.

Whilst the merits of domain hunting have been long known, the work in this Thesis demonstrates that there is still much to be learned from this level of protein analysis. Certainly by no means have all the biologically important domains been identified, characterised and modelled, and it appears that many of the interesting ones remain. I have also shown that the multiple sequence alignment is an extremely powerful tool for relating the information from different proteins that were analysed in isolation and without consideration of the rest of the family.

I suggest that not only is the detailed analysis of homology a useful task to carry out subsequent to experimental work, but that experimental work should be guided by the multiple sequence alignment. By this I mean that the identification of a domain

family's global characteristics should be used in directing laboratory-based experimentation. From the alignment it should be possible to identify the biggest gaps in our description of a family and hence carry out the most useful experiment on the most effective example protein for further characterisation of the family. This approach will generate more information about a greater number of proteins than the current piecemeal approach that underlies many protein investigations. Whilst there is still clearly a place for the investigation of specific proteins in a specific context, the genomic age introduces a new paradigm. For instance further study into the bacterial secretin family should be guided by consideration of its, at least, 16 different domain architectures rather than the current organism-based approach.

Comparative analysis is already a powerful approach, and is increasing in power as more genomes are sequenced and individual families become better represented. However, there is also a need for improved automatic detection of novel domains. In Pfam there are over 7000 sequence families, and yet only around 50% of amino acid residues in UniProt are accounted for; furthermore many of these families may cover more than one domain, or even include partial domains. In order to speed up the rate of discovery, tools for splitting the known families into their subcomponents and for the *ab initio* prediction of domains are essential. Such tools would be useful for driving the type of semi-automatic investigations I have been carrying out as well as enabling effective analyses of large numbers of proteins; a task that will increase in importance as the rate of genome sequencing accelerates.

Bibliography

- Albers, S. V., Z. Szabo, *et al.* (2003). "Archaeal homolog of bacterial type IV prepilin signal peptidases with broad substrate specificity." Journal of Bacteriology **185**(13): 3918-3925.
- Aloy, P., A. Stark, *et al.* (2003). "Predictions without templates: New folds, secondary structure, and contacts in CASP5." Proteins-Structure Function and Genetics **53**(6): 436-456.
- Altschul, S. F., M. S. Boguski, *et al.* (1994). "Issues in searching molecular sequence databases." Nature Genetics **6**(2): 119-129.
- Altschul, S. F., T. L. Madden, *et al.* (1997). "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.
- Andrade, M. A., F. D. Ciccarelli, *et al.* (2002). "NEAT: A domain duplicated in genes near the components of a putative Fe³⁺ siderophore transporter from Gram-positive pathogenic bacteria." Genome Biology **3**(9): 47.
- Apweiler, R., A. Bairoch, *et al.* (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Research **32**: D115-D119.
- Aravind, L., V. M. Dixit, *et al.* (1999). "The domains of death: Evolution of the apoptosis machinery." Trends in Biochemical Sciences **24**(2): 47-53.
- Asmann, Y. W., F. Kosari, *et al.* (2002). "Identification of differentially expressed genes in normal and malignant prostate by electronic profiling of expressed sequence tags." Cancer Research **62**(11): 3308-+.
- Atrih, A. and S. J. Foster (2001). "Analysis of the role of bacterial endospore cortex structure in resistance properties and demonstration of its conservation amongst species." Journal of Applied Microbiology **91**(2): 364-372.

- Attwood, T. K., P. Bradley, *et al.* (2003). "PRINTS and its automatic supplement, prePRINTS." Nucleic Acids Research **31**(1): 400-402.
- Av-Gay, Y. and M. Everett (2000). "The eukaryotic-like Ser/Thr protein kinases of *Mycobacterium tuberculosis*." Trends in Microbiology **8**(5): 238-244.
- Bao, Q. Y., Y. Q. Tian, *et al.* (2002). "A complete sequence of the *T. tengcongensis* genome." Genome Research **12**(5): 689-700.
- Barquera, B., C. C. Hase, *et al.* (2001). "Expression and mutagenesis of the NqrC subunit of the NQR respiratory Na⁺ pump from *Vibrio cholerae* with covalently attached FMN." Febs Letters **492**(1-2): 45-+.
- Bateman, A., E. Birney, *et al.* (2000). "The Pfam protein families database." Nucleic Acids Research **28**(1): 263-266.
- Bateman, A. and M. Bycroft (2000). "The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD)." Journal of Molecular Biology **299**(4): 1113-1119.
- Bateman, A., L. Coin, *et al.* (2004). "The Pfam protein families database." Nucleic Acids Research **32**: D138-D141.
- Bendtsen, J. D., H. Nielsen, *et al.* (2004). "Improved prediction of signal peptides: SignalP 3.0." Journal of Molecular Biology **340**(4): 783-795.
- Bentley, S. D., K. F. Chater, *et al.* (2002). "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)." Nature **417**(6885): 141-147.
- Bentley, S. D., M. Maiwald, *et al.* (2003). "Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*." Lancet **361**(9358): 637-644.
- Berman, H. M., T. Battistuz, *et al.* (2002). "The Protein Data Bank." Acta Crystallographica Section D-Biological Crystallography **58**: 899-907.

- Bernstein, C., H. Bernstein, *et al.* (1999). "Bile salt activation of stress response promoters in *Escherichia coli*." Current Microbiology **39**(2): 68-72.
- Bernstein, F. C., T. F. Koetzle, *et al.* (1977). "The Protein Data Bank: A computer-based archival file for macromolecular structures." Journal of Molecular Biology **112**(3): 535-542.
- Bey, S. J., M. F. Tsou, *et al.* (2000). "The homologous terminal sequence of the *Streptomyces lividans* chromosome and SLP2 plasmid." Microbiology-Sgm **146**: 911-922.
- Bitter, W., M. Koster, *et al.* (1998). "Formation of oligomeric rings by XcpQ and PilQ, which are involved in protein transport across the outer membrane of *Pseudomonas aeruginosa*." Molecular Microbiology **27**(1): 209-219.
- Boekema, B. K., J. P. Van Putten, *et al.* (2004). "Host cell contact-induced transcription of the type IV fimbria gene cluster of *Actinobacillus pleuropneumoniae*." Infection and Immunity **72**(2): 691-700.
- Boitel, B., M. Ortiz-Lombardia, *et al.* (2003). "PknB kinase activity is regulated by phosphorylation in two Thr residues and dephosphorylation by PstP, the cognate phospho- Ser/Thr phosphatase, in *Mycobacterium tuberculosis*." Molecular Microbiology **49**(6): 1493-1508.
- Boland, F. M., A. Atrih, *et al.* (2000). "Complete spore-cortex hydrolysis during germination of *Bacillus subtilis* 168 requires SleB and YpeB." Microbiology-Uk **146**: 57-64.
- Bolotin, A., P. Wincker, *et al.* (2001). "The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp *lactis* IL1403." Genome Research **11**(5): 731-753.

- Bork, P., K. Hofmann, *et al.* (1997). "A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins." Faseb Journal **11**(1): 68-76.
- Boyington, J. C., B. J. Gaffney, *et al.* (1993). "The crystal structure of soybean lipoxygenase-1." Faseb Journal **7**(7): A1059-A1059.
- Braun, P., W. Bitter, *et al.* (2000). "Activation of *Pseudomonas aeruginosa* elastase in *Pseudomonas putida* by triggering dissociation of the propeptide-enzyme complex." Microbiology-(UK) **146**: 2565-2572.
- Brenot, A., D. Trott, *et al.* (2001). "Penicillin-binding proteins in *Leptospira interrogans*." Antimicrobial Agents and Chemotherapy **45**(3): 870-877.
- Brown, R. L., T. L. Haley, *et al.* (1999). "Pseudechetoxin: A peptide blocker of cyclic nucleotide-gated ion channels." Proceedings of the National Academy of Sciences of the United States of America **96**(2): 754-759.
- Brunskill, E. W., B. L. M. deJonge, *et al.* (1997). "The *Staphylococcus aureus* *scdA* gene: A novel locus that affects cell division and morphogenesis." Microbiology-Uk **143**: 2877-2882.
- Bucher, P., K. Karplus, *et al.* (1996). "A flexible motif search technique based on generalized profiles." Computers & Chemistry **20**(1): 3-23.
- Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." Journal of Molecular Biology **268**(1): 78-94.
- Chaowagul, W., N. J. White, *et al.* (1989). "Meloidosis - a major cause of community acquired septicemia in northeastern Thailand." Journal of Infectious Diseases **159**(5): 890-899.
- Chauvaux, S., F. Chevalier, *et al.* (2001). "Cloning of a genetically unstable cytochrome P-450 gene cluster involved in degradation of the pollutant ethyl

- tert-butyl ether by *Rhodococcus ruber*." Journal of Bacteriology **183**(22): 6551-6557.
- Chen, Z. R. (2003). "Assessing sequence comparison methods with the average precision criterion." Bioinformatics **19**(18): 2456-2460.
- Chopra P., B. Singh, *et al.* (2003) "Phosphoprotein phosphatase of *Mycobacterium tuberculosis* dephosphorylates serine-threonine kinases PknA and PknB." Biochemical and Biophysical Research Communications **311**(1):112-20.
- Chothia, C. (1992). "Proteins - 1000 families for the molecular biologist." Nature **357**(6379): 543-544.
- Christiansen, G., A. S. Pedersen, *et al.* (2000). "Potential relevance of *Chlamydia pneumoniae* surface proteins to an effective vaccine." Journal of Infectious Diseases **181**: S528-S537.
- Ciccarelli, F. D., R. R. Copley, *et al.* (2002). "CASH - a beta-helix domain widespread among carbohydrate- binding proteins." Trends in Biochemical Sciences **27**(2): 59-62.
- Clamp, M., J. Cuff, *et al.* (2004). "The Jalview Java alignment editor." Bioinformatics **20**(3): 426-427.
- Clantin, B., H. Hodak, *et al.* (2004). "The crystal structure of filamentous hemagglutinin secretion domain and its implications for the two-partner secretion pathway." Proceedings of the National Academy of Sciences of the United States of America **101**(16): 6194-6199.
- Clarke, S. E., L. C. Sieker, *et al.* (1979). "Mercury binding to hemerythrin. Coordination of mercury and its effects on subunit interactions." Biochemistry **18**(4): 684-689.
- Coutte, L., S. Alonso, *et al.* (2003). "Role of adhesin release for mucosal colonization by a bacterial pathogen." Journal of Experimental Medicine **197**(6): 735-742.

- Crosa, J. H. and C. T. Walsh (2002). "Genetics and assembly line enzymology of siderophore biosynthesis in bacteria." Microbiology and Molecular Biology Reviews **66**(2): 223-+.
- Cuff, J. A. and G. J. Barton (2000). "Application of multiple sequence alignment profiles to improve protein secondary structure prediction." Proteins-Structure Function and Genetics **40**(3): 502-511.
- Cunningham, A. F. and M. E. Ward (2003). "Characterization of human humoral responses to the major outer membrane protein and OMP2 of *Chlamydomphila pneumoniae*." Fems Microbiology Letters **227**(1): 73-79.
- Daniel, R. A., S. Drake, *et al.* (1994). "The *Bacillus subtilis* spovD gene encodes a mother-cell-specific penicillin-binding protein required for spore morphogenesis." Journal of Molecular Biology **235**(1): 209-220.
- Deloffre, L., B. Salzet, *et al.* (2003). "Antibacterial properties of hemerythrin of the sand worm *Nereis diversicolor*." Neuroendocrinology Letters **24**(1-2): 39-45.
- Dessen, A., N. Mouz, *et al.* (2001). "Crystal structure of PBP2x from a highly penicillin-resistant *Streptococcus pneumoniae* clinical isolate - a mosaic framework containing 83 mutations." Journal of Biological Chemistry **276**(48): 45106-45112.
- Dobbelaere, D. A. E. and P. Kuenzi (2004). "The strategies of the Theileria parasite: a new twist in host- pathogen interactions." Current Opinion in Immunology **16**(4): 524-530.
- Drake, S. L. and M. Koomey (1995). "The product of the pilQ gene is essential for the biogenesis of type IV pili in *Neisseria gonorrhoeae*." Molecular Microbiology **18**(5): 975-986.
- Drews, S. J., F. M. Hung, *et al.* (2001). "A protein kinase inhibitor as an antimycobacterial agent." Fems Microbiology Letters **205**(2): 369-374.

- Durbin, R., S. R. Eddy, *et al.* (1998). Biological Sequence Analysis. Cambridge, UK, Cambridge University Press.
- Echenique, J., A. Kadioglu, *et al.* (2004). "Protein serine/threonine kinase StkP positively controls virulence and competence in *Streptococcus pneumoniae*." Infection and Immunity **72**(4): 2434-2437.
- Edelman, G. M. and W. E. Gall (1969). "The antibody problem." Annual Reviews of Biochemistry **38**: 415-466.
- Edgar, R. C. (2004). "MUSCLE: Multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-1797.
- Eidne, K. A., J. Zabavnik, *et al.* (1994). "Calcium waves and dynamics visualized by confocal microscopy in *Xenopus* oocytes expressing cloned Trh receptors." Journal of Neuroendocrinology **6**(2): 173-178.
- Eitel, J. and P. Dersch (2002). "The YadA protein of *Yersinia pseudotuberculosis* mediates high- efficiency uptake into human cells under environmental conditions in which invasin is repressed." Infection and Immunity **70**(9): 4880-4891.
- Encarnacion, S., M. del Vargas, *et al.* (2002). "AniA regulates reserve polymer accumulation and global protein expression in *Rhizobium etli*." Journal of Bacteriology **184**(8): 2287-2295.
- Englander, J., E. Klein, *et al.* (2004). "DNA toroids: Framework for DNA repair in *Deinococcus radiodurans* and in germinating bacterial spores." Journal of Bacteriology **186**(18): 5973-5977.
- Enright, A. J., S. Van Dongen, *et al.* (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Research **30**(7): 1575-1584.

- Feng, D. F. and R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." Journal of Molecular Evolution **25**(4): 351-360.
- Fenollar, F. and D. Raoult (2001). "Whipple's disease." Clinical and Diagnostic Laboratory Immunology **8**(1): 1-8.
- Floriano, B. and M. Bibb (1996). "*afsR* is a pleiotropic but conditionally required regulatory gene for antibiotic production in *Streptomyces coelicolor* A3(2)." Molecular Microbiology **21**(2): 385-396.
- Fong, H. K. W., J. B. Hurley, *et al.* (1986). "Repetitive segmental structure of the transducin beta subunit - homology with the Cdc4 gene and identification of related messenger RNAs." Proceedings of the National Academy of Sciences of the United States of America **83**(7): 2162-2166.
- Galm, U., J. Schimana, *et al.* (2002). "Cloning and analysis of the simocyclinone biosynthetic gene cluster of *Streptomyces antibioticus* TO 6040." Archives of Microbiology **178**(2): 102-114.
- Galperin, M. Y., T. A. Gaidenko, *et al.* (2001). "MHYT, a new integral membrane sensor domain." Fems Microbiology Letters **205**(1): 17-23.
- Garciabustos, J. and A. Tomasz (1990). "A biological price of antibiotic resistance - major changes in the peptidoglycan structure of penicillin-resistant pneumococci." Proceedings of the National Academy of Sciences of the United States of America **87**(14): 5415-5419.
- George, R. A. and J. Heringa (2002). "Protein domain identification and improved sequence similarity searching using PSI-BLAST." Proteins-Structure Function and Genetics **48**(4): 672-681.

- Gerstein, M. (1998). "Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census." Proteins-Structure Function and Genetics **33**(4): 518-534.
- Gitai, Z., N. Dye, *et al.* (2004). "An actin-like gene can determine cell polarity in bacteria." Proceedings of the National Academy of Sciences of the United States of America **101**(23): 8643-8648.
- Goebel, M. and M. Yanagida (1991). "The TPR snap helix - a novel protein repeat motif from mitosis to transcription." Trends in Biochemical Sciences **16**(5): 173-177.
- Gollop, R., M. Inouye, *et al.* (1991). "Protein U, a late developmental spore coat protein of *Myxococcus xanthus*, is a secretory protein." Journal of Bacteriology **173**(11): 3597-3600.
- Golovin, A., T. J. Oldfield, *et al.* (2004). "E-MSD: an integrated data resource for bioinformatics." Nucleic Acids Research **32**: D211-D216.
- Gomes, J. P., W. J. Bruno, *et al.* (2004). "Recombination in the genome of *Chlamydia trachomatis* involving the polymorphic membrane protein C gene relative to *ompA* and evidence for horizontal gene transfer." Journal of Bacteriology **186**(13): 4295-4306.
- Gomis-Ruth, F.X., U. Gohlke, *et al.* (1996). "The helping hand of collagenase-3 (MMP-13): 2.7 Å crystal structure of its C-terminal haemopexin-like domain." Journal of Molecular Biology **264**:556-566.
- Gordon, E., N. Mouz, *et al.* (2000). "The crystal structure of the penicillin-binding protein 2X from *Streptococcus pneumoniae* and its acyl-enzyme form: Implication in drug resistance." Journal of Molecular Biology **299**(2): 477-485.

- Gouzy, J., F. Corpet, *et al.* (1999). "Whole genome protein domain analysis using a new method for domain clustering." Computers & Chemistry **23**(3-4): 333-340.
- Gracy, J. and P. Argos (1998). "Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities." Bioinformatics **14**(2): 174-187.
- Grant, A., D. Lee, *et al.* (2004). "Progress towards mapping the universe of protein folds." Genome Biology **5**(5): art. no.-107.
- Grebe, T. and R. Hakenbeck (1996). "Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of beta-lactam antibiotics." Antimicrobial Agents and Chemotherapy **40**(4): 829-834.
- Griffiths-Jones, S. and A. Bateman (2002). "The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs." Bioinformatics **18**(9): 1243-1249.
- Grishin, N. V. (2001). "Fold change in evolution of protein structures." Journal of Structural Biology **134**(2-3): 167-185.
- Gu, R. S., S. Fonseca, *et al.* (2004). "Transcript identification and profiling during salt stress and recovery of *Populus euphratica*." Tree Physiology **24**(3): 265-276.
- Gurian-Sherman, D. and S.E. Lindow (1993). "Bacterial ice nucleation: significance and molecular basis." FASEB Journal **7**(14): 1338-1343.
- Haft, D. H., J. D. Selengut, *et al.* (2003). "The TIGRFAMs database of protein families." Nucleic Acids Research **31**(1): 371-373.
- Hall, D. R., D. G. Gourley, *et al.* (1999). "The high-resolution crystal structure of the molybdate- dependent transcriptional regulator (ModE) from *Escherichia coli*: A novel combination of domain folds." Embo Journal **18**(6): 1435-1446.

- Haslam, R. J., H. B. Koide, *et al.* (1993). "Pleckstrin domain homology." Nature **363**(6427): 309-310.
- Hayashi, M., Y. Nakayama, *et al.* (2001). "FMN is covalently attached to a threonine residue in the NqrB and NqrC subunits of Na⁺-translocating NADH-quinone reductase from *Vibrio alginolyticus*." Febs Letters **488**(1-2): 5-8.
- Heck, L. W., K. Morihara, *et al.* (1986). "Degradation of soluble laminin and depletion of tissue-associated basement membrane laminin by *Pseudomonas aeruginosa* elastase and alkaline protease." Infection and Immunity **54**(1): 149-153.
- Henderson, J., S. Salzberg, *et al.* (1997). "Finding genes in DNA with a Hidden Markov Model." Journal of Computational Biology **4**(2): 127-141.
- Hinnebusch, J. and K. Tilly (1993). "Linear plasmids and chromosomes in bacteria." Molecular Microbiology **10**(5): 917-922.
- Hodgson, D. A. (2000). Primary metabolism and its control in streptomycetes: A most unusual group of bacteria. Advances in Microbial Physiology, Vol 42. **42**: 47-238.
- Hofmann, K. and P. Bucher (1995). "The FHA domain - a putative nuclear signaling domain found in protein kinases and transcription factors." Trends in Biochemical Sciences **20**(9): 347-349.
- Hogeweg, P. and B. Hesper (1984). "The alignment of sets of sequences and the construction of phyletic trees - an integrated method." Journal of Molecular Evolution **20**(2): 175-186.
- Holden, M., R. W. Titball, *et al.* (2004). " Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*." Proceedings of the National Academy of Sciences of the United States of America **101**(39): 14240-14245

- Holm, M. M., S. L. Vanlerberg, *et al.* (2003). "The hag protein of *Moraxella catarrhalis* strain O35E is associated with adherence to human lung and middle ear cells." Infection and Immunity **71**(9): 4977-4984.
- Hopwood, D. A. (1988). "The Leeuwenhoek lecture, 1987 - towards an understanding of gene switching in streptomyces, the basis of sporulation and antibiotic production." Proceedings of the Royal Society of London Series B-Biological Sciences **235**(1279): 121-&.
- Huang, C. H., Y. S. Lin, *et al.* (1998). "The telomeres of streptomyces chromosomes contain conserved palindromic sequences with potential to form complex secondary structures." Molecular Microbiology **28**(5): 905-916.
- Huang, X. Q. and W. Miller (1991). "A time-efficient, linear-space local similarity algorithm." Advances in Applied Mathematics **12**(3): 337-357.
- Hughey, R. and A. Krogh (1996). "Hidden Markov models for sequence analysis: Extension and analysis of the basic method." Computer Applications in the Biosciences **12**(2): 95-107.
- Hulo, N., C. J. A. Sigrist, *et al.* (2004). "Recent improvements to the PROSITE database." Nucleic Acids Research **32**: D134-D137.
- Huth, J. R., C. A. Bewley, *et al.* (1997). "The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif." Nature Structural Biology **4**(8): 657-665.
- Ito, H., I. Uchida, *et al.* (1993). "A cryptic DNA sequence, isolated from *Actinobacillus pleuropneumoniae*, confers a hemolytic activity upon *Escherichia coli* K12 strains." Journal of Veterinary Medical Science **55**(1): 173-175.

- Jacobson, M. R., K. E. Brigle, *et al.* (1989). "Physical and genetic map of the major Nif gene cluster from *Azotobacter vinelandii*." Journal of Bacteriology **171**(2): 1017-1027.
- Jarrell, K. F., J. D. Correia, *et al.* (1999). "Is the processing and translocation system used by flagellins also used by membrane-anchored secretory proteins in archaea?" Molecular Microbiology **34**(2): 395-398.
- Jiang, Y. X., A. Lee, *et al.* (2003). "X-ray structure of a voltage-dependent K⁺ channel." Nature **423**(6935): 33-41.
- Jouanneau, Y., H. S. Jeong, *et al.* (1998). "Overexpression in *Escherichia coli* of the rnf genes from *Rhodobacter capsulatus* - characterization of two membrane-bound iron-sulfur proteins." European Journal of Biochemistry **251**(1-2): 54-64.
- Kachlany, S. C., P. J. Planet, *et al.* (2000). "Nonspecific adherence by *Actinobacillus actinomycetemcomitans* requires genes wide-spread in Bacteria and Archaea." Journal of Bacteriology **182**(21): 6169-6176.
- Kalakoutskii, L. V. and S. N. Agre (1976). "Comparative aspects of development and differentiation in Actinomycetes." Bacteriological Reviews **40**(2): 469-524.
- Kalman, S., W. Mitchell, *et al.* (1999). "Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*." Nature Genetics **21**(4): 385-389.
- Katoh, K., K. Misawa, *et al.* (2002). "MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Research **30**(14): 3059-3066.
- Kessler, E., M. Safrin, *et al.* (1998). "Elastase and the LasA protease of *Pseudomonas aeruginosa* are secreted with their propeptides." Journal of Biological Chemistry **273**(46): 30225-30231.

- Kloda, A. and B. Martinac (2001). "Mechanosensitive channels in Archaea." Cell Biochemistry and Biophysics **34**(3): 349-381.
- Komatsuzawa, H., K. Ohta, *et al.* (2000). "Tn551-mediated insertional inactivation of the *fntB* gene encoding a cell wall-associated protein abolishes methicillin resistance in *Staphylococcus aureus*." Journal of Antimicrobial Chemotherapy **45**(4): 421-431.
- Kong, L. S. and S. Ranganathan (2004). "Delineation of modular proteins: Domain boundary prediction from sequence information." Briefings in Bioinformatics **5**(2): 179-192.
- Koonin, E. V., Y. I. Wolf, *et al.* (2000). Protein fold recognition using sequence profiles and its application in structural genomics. Advances in Protein Chemistry, Vol 54. **54**: 245-275.
- Koponen, M. A., D. Zlock, *et al.* (1991). "Melioidosis - forgotten, but not gone." Archives of Internal Medicine **151**(3): 605-608.
- Korf, I., M. Yandell, *et al.* (2003). BLAST, O'Reilly.
- Krogh, A., B. Larsson, *et al.* (2001). "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes." Journal of Molecular Biology **305**(3): 567-580.
- Kurtz, D. M. (1997). "Structural similarity and functional diversity in diiron-oxo proteins." Journal of Biological Inorganic Chemistry **2**(2): 159-167.
- Kurtz, D. M. (1999). Oxygen-carrying proteins: Three solutions to a common problem. Essays in Biochemistry, Vol 34, 1999. **34**: 85-100.
- La Scola, B., F. Fenollar, *et al.* (2001). "Description of *Tropheryma whipplei* gen. nov., sp nov., the Whipple's disease bacillus." International Journal of Systematic and Evolutionary Microbiology **51**: 1471-1479.

- Laible, G., B. G. Spratt, *et al.*, (1991) "Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*." Molecular Microbiology **5**(8): 1993-2002
- LaPointe, C. F. and R. K. Taylor (2000). "The type 4 prepilin peptidases comprise a novel family of aspartic acid proteases." Journal of Biological Chemistry **275**(2): 1502-1510.
- Laskowski, R. A. (2001). "PDBsum: Summaries and analyses of PDB structures." Nucleic Acids Research **29**(1): 221-222.
- Lauer, B., R. Russwurm, *et al.* (2001). "Molecular characterization of co-transcribed genes from *Streptomyces tendae* Tu901 involved in the biosynthesis of the peptidyl moiety and assembly of the peptidyl nucleoside antibiotic nikkomycin." Molecular and General Genetics **264**(5): 662-673.
- Lee, W., M. A. McDonough, *et al.* (2001). "A 1.2-angstrom snapshot of the final step of bacterial cell wall biosynthesis." Proceedings of the National Academy of Sciences of the United States of America **98**(4): 1427-1431.
- Letunic, I., R. R. Copley, *et al.* (2004). "SMART 4.0: Towards genomic data integration." Nucleic Acids Research **32**: D142-D144.
- Liechty, A., J. H. Chen, *et al.* (2000). "Origin of antibacterial stasis by polymyxin B in *Escherichia coli*." Biochimica Et Biophysica Acta-Biomembranes **1463**(1): 55-64.
- Lipman, D. J., A. Souvorov, *et al.* (2002). "The relationship of protein conservation and sequence length." BMC Evolutionary Biology **2**(1): 20.
- Liu, J. F. and B. Rost (2004). "CHOP: Parsing proteins into structural domains." Nucleic Acids Research **32**: W569-W571.
- Loewen, S. K., A. M. L. Ng, *et al.* (1999). "Identification of amino acid residues responsible for the pyrimidine and purine nucleoside specificities of human

- concentrative Na⁺ nucleoside cotransporters hCNT1 and hCNT2." Journal of Biological Chemistry **274**(35): 24475-24484.
- Loewen, S. K., S. Y. M. Yao, *et al.* (2004). "Transport of physiological nucleosides and anti-viral and anti- neoplastic nucleoside drugs by recombinant *Escherichia coli* nucleoside-H⁺ cotransporter (NupC) produced in *Xenopus laevis* oocytes." Molecular Membrane Biology **21**(1): 1-10.
- Longbottom, D. and L. J. Coulter (2003). "Animal chlamydioses and zoonotic implications." Journal of Comparative Pathology **128**(4): 217-244.
- Lupas, A., M. Vandyke, *et al.* (1991). "Predicting coiled coils from protein sequences." Science **252**(5009): 1162-1164.
- Lytle, B. L., F. C. Peterson, *et al.* (2004). "Letter to the Editor: Structure of the hypothetical protein At3g17210 from *Arabidopsis thaliana*." Journal of Biomolecular Nmr **28**(4): 397-400.
- Madec, E., A. Laszkiewicz, *et al.* (2002). "Characterization of a membrane-linked Ser/Thr protein kinase in *Bacillus subtilis*, implicated in developmental processes." Molecular Microbiology **46**(2): 571-586.
- Madera, M. and J. Gough (2002). "A comparison of profile hidden Markov model procedures for remote homology detection." Nucleic Acids Research **30**(19): 4321-4328.
- Madera, M., C. Vogel, *et al.* (2004). "The SUPERFAMILY database in 2004: additions and improvements." Nucleic Acids Research **32**: D235-D239.
- Maeda, T., J. Nishida, *et al.* (1999). "Expression pattern, subcellular localization and structure- function relationship of rat Tpx-1, a spermatogenic cell adhesion molecule responsible for association with Sertoli cells." Development Growth & Differentiation **41**(6): 715-722.

- Maehara, A., S. Taguchi, *et al.* (2002). "A repressor protein, PhaR, regulates polyhydroxyalkanoate (PHA) synthesis via its direct interaction with PHA." Journal of Bacteriology **184**(14): 3992-4002.
- Malho, R., N. D. Read, *et al.* (1995). "Calcium channel activity during pollen tube growth and reorientation." Plant Cell **7**(8): 1173-1184.
- Marie-Claire, C., B. P. Roques, *et al.* (1998). "Intramolecular processing of prothermolysin." Journal of Biological Chemistry **273**(10): 5697-5701.
- Massimi, I., E. Park, *et al.* (2002). "Identification of a novel maturation mechanism and restricted substrate specificity for the SspB cysteine protease of *Staphylococcus aureus*." Journal of Biological Chemistry **277**(44): 41770-41777.
- May, A. C. W. (2004). "Percent sequence identity: The need to be explicit." Structure **12**(5): 737-738.
- Meinhardt, L. W., H. B. Krishnan, *et al.* (1993). "Molecular cloning and characterization of a Sym plasmid locus that regulates cultivar-specific nodulation of soybean by *Rhizobium fredii* USDA257." Molecular Microbiology **9**(1): 17-29.
- Mesnage, S., T. Fontaine, *et al.* (2000). "Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation." Embo Journal **19**(17): 4473-4484.
- Meyer, I. M. and R. Durbin (2002). "Comparative *ab initio* prediction of gene structures using pair HMMs." Bioinformatics **18**(10): 1309-1318.
- Miller, J., A. D. McLachlan, *et al.* (1985). "Repetitive zinc-binding domains in the protein transcription factor Iiia from *Xenopus* oocytes." Embo Journal **4**(6): 1609-1614.

- Milne, T. J., G. Abbenante, *et al.* (2003). "Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily." Journal of Biological Chemistry **278**(33): 31105-31110.
- Miyoshi, S., H. Nakazawa, *et al.* (1998). "Characterization of the hemorrhagic reaction caused by *Vibrio vulnificus* metalloprotease, a member of the thermolysin family." Infection and Immunity **66**(10): 4851-4855.
- Mott, R. (2000). "Accurate formula for P-values of gapped local sequence and profile alignments." Journal of Molecular Biology **300**(3): 649-659.
- Moyle, M., D. L. Foster, *et al.* (1994). "A hookworm glycoprotein that inhibits neutrophil function is a ligand of the Integrin Cd11b/Cd18." Journal of Biological Chemistry **269**(13): 10008-10015.
- Mulder, N. J., R. Apweiler, *et al.* (2003). "The InterPro Database, 2003 brings increased coverage and new features." Nucleic Acids Research **31**(1): 315-318.
- Murzin, A. G. (1992). "Structural principles for the propeller assembly of beta-sheets - the preference for 7-fold symmetry." Proteins-Structure Function and Genetics **14**(2): 191-201.
- Murzin, A. G., S. E. Brenner, *et al.* (1995). "SCOP - a structural classification of proteins database for the investigation of sequences and structures." Journal of Molecular Biology **247**(4): 536-540.
- Mushegian, A. R. and E. V. Koonin (1996). "Sequence analysis of eukaryotic developmental proteins: ancient and novel domains." Genetics **144**(2): 817-828.
- Nagarajan, N. and G. Yona (2004). "Automatic prediction of protein domains from sequence information using a hybrid learning system." Bioinformatics **20**(9): 1335-1360.

- Neer, E. J., C. J. Schmidt, *et al.* (1994). "The ancient regulatory-protein family of WD-repeat proteins." Nature **371**(6500): 812-812.
- Neu, J. M., S. V. MacMillan, *et al.* (2002). "StoPK-1, a serine/threonine protein kinase from the glycopeptide antibiotic producer *Streptomyces toyocaensis* NRRL 15009, affects oxidative stress response." Molecular Microbiology **44**(2): 417-430.
- Neu, J. M. and G. D. Wright (2001). "Inhibition of sporulation, glycopeptide antibiotic production and resistance in *Streptomyces toyocaensis* NRRL 15009 by protein kinase inhibitors." Fems Microbiology Letters **199**(1): 15-20.
- Nielsen, H. and A. Krogh (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. Proceedings of the International Conference on Intelligent Systems for Molecular Biology.
- Nissen, M. S., T. A. Langan, *et al.* (1991). "Phosphorylation by Cdc2 kinase modulates DNA-binding activity of high mobility group-I nonhistone chromatin protein." Journal of Biological Chemistry **266**(30): 19945-19952.
- Notredame, C. (2002). "Recent progress in multiple sequence alignment: a survey." Pharmacogenomics **3**(1): 131-144.
- Notredame, C., L. Holm, *et al.* (1998). "COFFEE: An objective function for multiple sequence alignments." Bioinformatics **14**(5): 407-422.
- Nouwen, N., H. Stahlberg, *et al.* (2000). "Domain structure of secretin PulD revealed by limited proteolysis and electron microscopy." Embo Journal **19**(10): 2229-2236.
- Oh, J. T., Y. Cajal, *et al.* (2000). "Cationic peptide antimicrobials induce selective transcription of *micF* and *osmY* in *Escherichia coli*." Biochimica Et Biophysica Acta-Biomembranes **1463**(1): 43-54.

- Oke, M., R. Sarra, *et al.* (2004). "The plug domain of a neisserial TonB-dependent transporter retains structural integrity in the absence of its transmembrane beta-barrel." Febs Letters **564**(3): 294-300.
- Olson, J. H., X. Y. Xiang, *et al.* (2001). "Allurin, a 21-kDa sperm chemoattractant from *Xenopus* egg jelly, is related to mammalian sperm-binding proteins." Proceedings of the National Academy of Sciences of the United States of America **98**(20): 11205-11210.
- O'Toole, G., H. B. Kaplan, *et al.* (2000). "Biofilm formation as microbial development." Annual Review of Microbiology **54**: 49-79.
- Pahl, A., K. Brune, *et al.* (1997). "Fit for life? Evolution of chaperones and folding catalysts parallels the development of complex organisms." Cell Stress & Chaperones **2**(2): 78-86.
- Pearl, F. M. G., C. F. Bennett, *et al.* (2003). "The CATH database: an extended protein family resource for structural and functional genomics." Nucleic Acids Research **31**(1): 452-455.
- Pedersen, A. G., L. J. Jensen, *et al.* (2000). "A DNA structural atlas for *Escherichia coli*." Journal of Molecular Biology **299**(4): 907-930.
- Pedersen, A. S., G. Christiansen, *et al.* (2001). "Differential expression of Pmp10 in cell culture infected with *Chlamydia pneumoniae* CWL029." Fems Microbiology Letters **203**(2): 153-159.
- Perozo, E., A. Kloda, *et al.* (2002). "Physical principles underlying the transduction of bilayer deformation forces during mechanosensitive channel gating." Biophysical Journal **82**(1): 1298.
- Pohlmann, A., R. Cramm, *et al.* (2000). "A novel NO-responding regulator controls the reduction of nitric oxide in *Ralstonia eutropha*." Molecular Microbiology **38**(3): 626-638.

- Pons, T., R. Gomez, *et al.* (2003). "Beta-propellers: Associated functions and their role in human diseases." Current Medicinal Chemistry **10**(6): 505-524.
- Ponting, C.P., L. Aravind, *et al.* (1999) "Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer." Journal of Molecular Biology **289**(4):729-745.
- Ponting, C. P., R. Mott, *et al.* (2001). "Novel protein domains and repeats in *Drosophila melanogaster*: Insights into structure, function, and evolution." Genome Research **11**(12): 1996-2008.
- Qi, Y., J. M. Pei, *et al.* (2002). "C-terminal domain of gyrase A is predicted to have a beta- propeller structure." Proteins-Structure Function and Genetics **47**(3): 258-264.
- Qian, J., N. M. Luscombe, *et al.* (2001). "Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model." Journal of Molecular Biology **313**(4): 673-681.
- Rawlings, N. D., D. P. Tolle, *et al.* (2004). "MEROPS: The peptidase database." Nucleic Acids Research **32**: D160-D164.
- Read, T. D., R. C. Brunham, *et al.* (2000). "Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39." Nucleic Acids Research **28**(6): 1397-1406.
- Reichmann, P., A. Konig *et al.* (1996). "Penicillin-binding proteins as resistance determinants in clinical isolates of *Streptococcus pneumoniae*." Microbial Drug Resistance **2**(2): 177-181
- Roberts, C. J., A. J. Stewart, *et al.* (2004). "Human PHOSPHO1 displays high specific phosphoethanolamine and phosphocholine phosphatase activities: A means of generating inorganic phosphate in mineralising cells." Journal of Bone and Mineral Research **19**(6): 1041-1041.

- Roberts, K. P., K. M. Ensrud, *et al.* (2002). "A comparative analysis of expression and processing of the rat epididymal fluid and sperm-bound forms of proteins D and E." Biology of Reproduction **67**(2): 525-533.
- Rojas, C. M., J. H. Ham, *et al.* (2002). "HecA, a member of a class of adhesins produced by diverse pathogenic bacteria, contributes to the attachment,, aggregation, epidermal cell killing, and virulence phenotypes of *Erwinia chrysanthemi* EC16 on *Nicotiana clevelandii* seedlings." Proceedings of the National Academy of Sciences of the United States of America **99**(20): 13142-13147.
- Rossman, A. G. and A. Liljas (1974). "Recognition of structural domains in globular proteins." Journal of Molecular Biology **85**(1): 177-181.
- Rost, B. (1996). PHD: Predicting one-dimensional protein structure by profile- based neural networks. Computer Methods for Macromolecular Sequence Analysis. **266**: 525-539.
- Rost, B. and C. Sander (1993). "Prediction of protein secondary structure at better than 70 percent accuracy." Journal of Molecular Biology **232**(2): 584-599.
- Rutherford, K., J. Parkhill, *et al.* (2000). "Artemis: sequence visualization and annotation." Bioinformatics **16**(10): 944-945.
- Rzychon, M., A. Sabat, *et al.* (2003). "Staphostatins: An expanding new group of proteinase inhibitors with a unique specificity for the regulation of staphopains, *Staphylococcus* spp. cysteine proteinases." Molecular Microbiology **49**(4): 1051-1066.
- Scheffers, D. J., L. J. F. Jones, *et al.* (2004). "Several distinct localization patterns for penicillin-binding proteins in *Bacillus subtilis*." Molecular Microbiology **51**(3): 749-764.

- Servant, F., C. Bru, *et al.* (2002). "ProDom: Automated clustering of homologous domains." Briefings in Bioinformatics **3**(3): 246-251.
- Shapiro, L., H. H. McAdams, *et al.* (2002). "Generating and exploiting polarity in bacteria." Science **298**(5600): 1942-1946.
- Sheldon, P. J., S. B. Busarow, *et al.* (2002). "Mapping the DNA-binding domain and target sequences of the *Streptomyces peucetius* daunorubicin biosynthesis regulatory protein, DnrI." Molecular Microbiology **44**(2): 449-460.
- Sheu, Y. and M. Snyder (2001). Control of cell polarity and shape. Biology of the fungal cell. R. J. Howard and N. A. R. Gow. Berlin-Heidelberg-New York, Springer. **8**: 19-54.
- Silva, M. T., P. M. Macedo, *et al.* (1985). "Ultrastructure of bacilli and the bacillary origin of the macrophagic inclusions in Whipples disease." Journal of General Microbiology **131**(MAY): 1001-1013.
- Silverman-Gavrila, L. B. and R. R. Lew (2003). "Calcium gradient dependence of *Neurospora crassa* hyphal growth." Microbiology-Sgm **149**: 2475-2485.
- Skerker, J. M. and L. Shapiro (2000). "Identification and cell cycle control of a novel pilus system in *Caulobacter crescentus*." Embo Journal **19**(13): 3223-3234.
- Slakeski, N., S. M. Cleal, *et al.* (1999). "Characterization of a *Porphyromonas gingivalis* gene *prtK* that encodes a lysine-specific cysteine proteinase and three sequence-related adhesins." Oral Microbiology and Immunology **14**(2): 92-97.
- Smirnova, A., H. Q. Li, *et al.* (2001). "Thermoregulated expression of virulence factors in plant-associated bacteria." Archives of Microbiology **176**(6): 393-399.

- Sonnhammer, E. L. L. and R. Durbin (1995). "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein-sequence analysis." Gene-Combis **167**: 1-10.
- Sonnhammer, E. L. L. and D. Kahn (1994). "Modular arrangement of proteins as inferred from analysis of homology." Protein Science **3**(3): 482-492.
- Springer, T. A. (1997). "Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain." Proceedings of the National Academy of Sciences of the United States of America **94**(1): 65-72.
- Strong, M., T. G. Graeber, *et al.* (2003). "Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps." Nucleic Acids Research **31**(24): 7099-7109.
- Samatey F. A., H. Matsunami, *et al.* (2004) "Structure of the bacterial flagellar hook and implication for the molecular universal joint mechanism." Nature **431**(7012):1062-1068.
- Subramanian, G., E. V. Koonin, *et al.* (2000). "Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*." Infection and Immunity **68**(3): 1633-1648.
- Swan, D. G., K. Phillips, *et al.* (1999). "Evidence for localisation of a Theileria parasite AT hook DNA- binding protein to the nucleus of immortalised bovine host cells." Molecular and Biochemical Parasitology **101**(1-2): 117-129.
- Szyperski, T., C. Fernandez, *et al.* (1998). "Structure comparison of human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system." Proceedings of the National Academy of Sciences of the United States of America **95**(5): 2262-2266.

- Tang, B., S. Nirasawa, *et al.* (2003). "General function of N-terminal propeptide on assisting protein folding and inhibiting catalytic activity based on observations with a chimeric thermolysin-like protease." Biochemical and Biophysical Research Communications **301**(4): 1093-1098.
- Tatusov, R. L., N. D. Fedorova, *et al.* (2003). "The COG database: An updated version includes eukaryotes." Bmc Bioinformatics **4**: art. no.-41.
- Teichmann, S. A., J. Park, *et al.* (1998). "Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 14658-14663.
- Thompson, J. D., D. G. Higgins, *et al.* (1994). "Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Research **22**(22): 4673-4680.
- Thompson, J. D., F. Plewniak, *et al.* (1999). "BALiBASE: A benchmark alignment database for the evaluation of multiple alignment programs." Bioinformatics **15**(1): 87-88.
- Throup, J. P., F. Zappacosta, *et al.* (2001). "The *srhSR* gene pair from *Staphylococcus aureus*: Genomic and proteomic approaches to the identification and characterization of gene function." Biochemistry **40**(34): 10392-10401.
- Tomaras, A. P., C. W. Dorsey, *et al.* (2003). "Attachment to and biofilm formation on abiotic surfaces by *Acinetobacter baumannii*: Involvement of a novel chaperone-usher pili assembly system." Microbiology-Sgm **149**: 3473-3484.
- Toye, P., E. Gobright, *et al.* (1995). "Structure and sequence variation of the genes encoding the polymorphic, immunodominant molecule (PIM), an antigen of

- Theileria parva* recognized by inhibitory monoclonal antibodies." Molecular and Biochemical Parasitology **73**(1-2): 165-177.
- Umeyama, T., P. C. Lee, *et al.* (2002). "Protein serine/threonine kinases in signal transduction for secondary metabolism and morphogenesis in *Streptomyces*." Applied Microbiology and Biotechnology **59**(4-5): 419-425.
- Vallet, I., S. P. Diggle, *et al.* (2004). "Biofilm formation in *Pseudomonas aeruginosa*: Fimbrial cup gene clusters are controlled by the transcriptional regulator MvaT." Journal of Bacteriology **186**(9): 2880-2890.
- van der Woude, M. W. and A. J. Baumler (2004). "Phase and antigenic variation in bacteria." Clinical Microbiology Reviews **17**(3): 581-+.
- Vollack, K. U. and W. G. Zumft (2001). "Nitric oxide signaling and transcriptional control of denitrification genes in *Pseudomonas stutzeri*." Journal of Bacteriology **183**(8): 2516-2526.
- von Mering C., M. Huynen, *et al.* (2003) "STRING: a database of predicted functional associations between proteins." Nucleic Acids Research **31**(1):258-261.
- Voorhorst, W. G. B., R. I. L. Eggen, *et al.* (1996). "Isolation and characterization of the hyperthermostable serine protease, pyrolysin, and its gene from the hyperthermophilic archaeon *Pyrococcus furiosus*." Journal of Biological Chemistry **271**(34): 20426-20431.
- Wang, Z. X., S. M. Li, *et al.* (2000). "Identification of the coumermycin A(1) biosynthetic gene cluster of *Streptomyces rishiriensis* DSM 40489." Antimicrobial Agents and Chemotherapy **44**(11): 3040-3048.
- Wehrl, W., V. Brinkmann, *et al.* (2004). "From the inside out - processing of the Chlamydial autotransporter PmpD and its role in bacterial adhesion and activation of human host cells." Molecular Microbiology **51**(2): 319-334.

- Weichart, D., R. Lange, *et al.* (1993). "Identification and characterization of stationary phase- inducible genes in *Escherichia coli*." Molecular Microbiology **10**(2): 407-420.
- Wetlaufer, D. B. (1973). "Nucleation, rapid folding, and globular intrachain regions in proteins." Proceedings of the National Academy of Sciences of the United States of America **70**(3): 697-701.
- Wheeler, D. L., D. M. Church, *et al.* (2004). "Database resources of the National Center for Biotechnology Information update." Nucleic Acids Research **1**(32): D35-40.
- Whittle, G., N. B. Shoemaker, *et al.* (2002). "Characterization of genes involved in modulation of conjugal transfer of the Bacteroides conjugative transposon CTnDOT." Journal of Bacteriology **184**(14): 3839-3847.
- Wilson, J. J., O. Matsushita, *et al.* (2003). "A bacterial collagen-binding domain with novel calcium-binding motif controls domain orientation." Embo Journal **22**(8): 1743-1752.
- Wilson, K. H., R. Blichington, *et al.* (1991). "Phylogeny of the Whipples disease-associated bacterium." Lancet **338**(8765): 474-475.
- Wootton, J. C. and S. Federhen (1993). "Statistics of local complexity in amino acid sequences and sequence databases." Computers & Chemistry **17**(2): 149-163.
- Yamakawa, T., S. Miyata, *et al.* (1998). "cDNA cloning of a novel trypsin inhibitor with similarity to pathogenesis-related proteins, and its frequent expression in human brain cancer cells." Biochimica Et Biophysica Acta-Gene Structure and Expression **1395**(2): 202-208.
- Yeats, C. and A. Bateman (2003). "The BON domain: a putative membrane-binding domain." Trends in Biochemical Sciences **28**(7): 352-355.

- Yeats, C., S. Bentley, *et al.* (2003). "New knowledge from old: *In silico* discovery of novel protein domains in *Streptomyces coelicolor*." Bmc Microbiology **3**: art. no.-3.
- Yeats, C., R. D. Finn, *et al.* (2002). "The PASTA domain: A beta-lactam-binding domain." Trends in Biochemical Sciences **27**(9): 438-440.
- Yeats, C., N. D. Rawlings, *et al.* (2004). "The PepSY domain: A regulator of peptidase activity in the microbial environment?" Trends in Biochemical Sciences **29**(4): 169-172.
- Yim, H. H. and M. Villarejo (1992). "OsmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in *Escherichia coli*." Journal of Bacteriology **174**(11): 3637-3644.
- Yona, G., N. Linial, *et al.* (1999). "ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space." Proteins-Structure Function and Genetics **37**(3): 360-378.
- Zhang, J., P. C. Fitzjames, *et al.* (1993). "Cloning and characterization of a cluster of genes encoding polypeptides present in the insoluble fraction of the spore coat of *Bacillus subtilis*." Journal of Bacteriology **175**(12): 3757-3766.

Appendix A: List of All Domains in This Thesis

In this appendix a table containing all the domains found in this Thesis is presented.

The Pfam accession number and a short description line derived from the Pfam annotation are included. For more detailed information these domains can all be found in Pfam release 16.0.

7TMR-DISM_7TM	<i>PF07695</i>	7TM diverse intracellular signalling
ALF	<i>PF03752</i>	Short repeats of unknown function
AT_hook	<i>PF02178</i>	AT hook motif
Abhydrolase_1	<i>PF00561</i>	alpha/beta hydrolase fold
Adaptin_N	<i>PF01602</i>	Adaptin N terminal region
Arc_PepC	<i>PF06819</i>	Archaeal Peptidase A24 C-terminal Domain
Arc_PepC_II	<i>PF06847</i>	Archaeal Peptidase A24 C-terminus Type II
Autotransporter	<i>PF03797</i>	Autotransporter beta-domain
BNR	<i>PF02012</i>	BNR/Asp-box repeat
BON	<i>PF04972</i>	Putative phospholipid-binding domain
BTAD	<i>PF03704</i>	Bacterial transcriptional activator domain
BTP	<i>PF05232</i>	Bacterial Transmembrane Pair family
Big_2	<i>PF02368</i>	Bacterial Ig-like domain (group 2)
Big_3	<i>PF07523</i>	Bacterial Ig-like domain (group 3)
Big_4	<i>PF07532</i>	Bacterial Ig-like domain (group 4)
CBM_4_9	<i>PF02018</i>	Carbohydrate binding domain
CBM_5_12	<i>PF02839</i>	Carbohydrate binding domain
CBM_6	<i>PF03422</i>	Carbohydrate binding module (family 6)
CBS	<i>PF00571</i>	CBS domain
CCD	<i>PF07860</i>	WisP family C-Terminal Region
CW_binding_2	<i>PF04122</i>	Putative cell wall binding repeat 2
CXCXC	<i>PF03128</i>	CXCXC repeat
Calx-beta	<i>PF03160</i>	Calx-beta domain
CbiX	<i>PF01903</i>	CbiX
CheW	<i>PF01584</i>	CheW-like domain
ChlamPMP_M	<i>PF07548</i>	Chlamydia polymorphic membrane protein middle domain
Chlam_PMP	<i>PF02415</i>	Chlamydia polymorphic membrane protein (Chlamydia_PMP)
Choline_kinase	<i>PF01633</i>	Choline/ethanolamine kinase
Cleaved_Adhesin	<i>PF07675</i>	Cleaved Adhesin Domain
Coat_F	<i>PF07875</i>	Coat F domain
Coat_X	<i>PF07552</i>	Spore Coat Protein X and V domain
Cohesin	<i>PF00963</i>	Cohesin domain
Collagen	<i>PF01391</i>	Collagen triple helix repeat (20 copies)
CtnDOT_TraJ	<i>PF07863</i>	Homologues of TraJ from Bacteroides conjugative transposon
Cupin_2	<i>PF07883</i>	Cupin domain
DEAD	<i>PF00270</i>	DEAD/DEAH box helicase
DUF1034	<i>PF06280</i>	Domain of Unknown Function (DUF1034)
DUF1078	<i>PF06429</i>	Domain of unknown function (DUF1078)
DUF1533	<i>PF07550</i>	Protein of unknown function (DUF1533)
DUF1542	<i>PF07564</i>	Domain of Unknown Function (DUF1542)
DUF1605	<i>PF07717</i>	Domain of unknown function (DUF1605)
DUF1606	<i>PF07718</i>	Domain of unknown function (DUF1606)
DUF385	<i>PF04075</i>	Domain of unknown function (DUF385)
DUF637	<i>PF04830</i>	Possible hemagglutinin (DUF637)
Dabb	<i>PF07876</i>	Stress responsive A/B Barrel Domain
DiS_P_DiS	<i>PF06750</i>	Bacterial Peptidase A24 N-terminal domain
Disaggr_repeat	<i>PF06848</i>	Disaggregatase related repeat
Dockerin_1	<i>PF00404</i>	Dockerin type I repeat
E1-E2_ATPase	<i>PF00122</i>	E1-E2 ATPase

EAL	<i>PF00563</i>	EAL domain
EGF_CA	<i>PF07645</i>	Calcium binding EGF domain
F5_F8_type_C	<i>PF00754</i>	F5/8 type C domain
FAD_binding_2	<i>PF00890</i>	FAD binding domain
FAINT	<i>PF04385</i>	Domain of unknown function
FG-GAP	<i>PF01839</i>	FG-GAP repeat
FHA	<i>PF00498</i>	FHA domain
FMN_bind	<i>PF04205</i>	FMN-binding domain
FMN_red	<i>PF03358</i>	NADPH-dependent FMN reductase
FTP	<i>PF07504</i>	Fungalysin/Thermolysin Propeptide Motif
FecR	<i>PF04773</i>	FecR protein
FeoA	<i>PF04023</i>	FeoA domain
FeoB_C	<i>PF07664</i>	Ferrous iron transport protein B C terminus
FeoB_N	<i>PF02421</i>	Ferrous iron transport protein B
Fer4	<i>PF00037</i>	4Fe-4S binding domain
Fic	<i>PF02661</i>	Fic protein family
Fil_haemagg	<i>PF05594</i>	Haemagglutinin repeat
FlaE	<i>PF07559</i>	Flagellar basal body protein FlaE
Flagellin_IN	<i>PF07196</i>	Flagellin hook IN motif
FlgD	<i>PF03963</i>	Flagellar hook capping protein
Flg_bb_rod	<i>PF00460</i>	Flagella basal body rod protein
fn3	<i>PF00041</i>	Fibronectin type III domain
G5	<i>PF07501</i>	G5 domain
GA	<i>PF01468</i>	GA module
GGDEF	<i>PF00990</i>	GGDEF domain
Gate	<i>PF07670</i>	Nucleoside recognition
Glug	<i>PF07581</i>	The GLUG motif
Glyco_hydro_31	<i>PF01055</i>	Glycosyl hydrolases family 31
Glyco_hydro_43	<i>PF04616</i>	Glycosyl hydrolases family 43
Glyco_hydro_85	<i>PF03644</i>	Glycosyl hydrolase family 85
Gram_pos_anchor	<i>PF00746</i>	Gram positive anchor
HA	<i>PF03457</i>	Helicase associated domain
HA2	<i>PF04408</i>	Helicase associated domain (HA2)
HAMP	<i>PF00672</i>	HAMP domain
HATPase_c	<i>PF02518</i>	Histidine kinase-
HCBP_related	<i>PF06594</i>	Haemolysin-type calcium binding protein related domain
HTH_AraC	<i>PF00165</i>	Bacterial regulatory helix-turn-helix proteins
Haemagg_act	<i>PF05860</i>	haemagglutination activity domain
He_PIG	<i>PF05345</i>	Putative Ig domain
Helicase_C	<i>PF00271</i>	Helicase conserved C-terminal domain
Hemerythrin	<i>PF01814</i>	Hemerythrin HHE cation binding domain
HemolysinCabind	<i>PF00353</i>	Hemolysin-type calcium-binding repeat (2 copies)
HisKA	<i>PF00512</i>	His Kinase A (phosphoacceptor) domain
Hyaluronidase_2	<i>PF07555</i>	Hyaluronidase
Hydrolase	<i>PF00702</i>	haloacid dehalogenase-like hydrolase
Ice_nucleation	<i>PF00818</i>	Ice nucleation protein repeat
LCCL	<i>PF03815</i>	LCCL domain
LRR_1	<i>PF00560</i>	Leucine Rich Repeat
Lectin_C	<i>PF00059</i>	Lectin C-type domain
Lyase_8	<i>PF02278</i>	Polysaccharide lyase family 8

LysM	<i>PF01476</i>	LysM domain
MS_channel	<i>PF00924</i>	Mechanosensitive ion channel
MbtH	<i>PF03621</i>	MbtH-like protein
Myco_haema	<i>PF05692</i>	Mycoplasma haemagglutinin
NB-ARC	<i>PF00931</i>	NB-ARC domain
NEAT	<i>PF05031</i>	Iron Transport-associated domain
Nif11	<i>PF07862</i>	Nitrogen fixation protein of unknown function
Nucleos_tra2_C	<i>PF07662</i>	Na ⁺ dependent nucleoside transporter C-terminus
Nucleos_tra2_N	<i>PF01773</i>	Na ⁺ dependent nucleoside transporter N-terminus
Oxidored_FMN	<i>PF00724</i>	NADH:flavin oxidoreductase / NADH oxidase family
PA	<i>PF02225</i>	PA domain
PAC	<i>PF00785</i>	PAC motif PAC motif occurs C-terminal to a subset of all known PAS motifs.
PAS	<i>PF00989</i>	PAS domain
PASTA	<i>PF03793</i>	PASTA domain
PBP_dimer	<i>PF03717</i>	Penicillin-binding Protein dimerisation domain
PG_binding_1	<i>PF01471</i>	Putative peptidoglycan binding domain
PHB_acc	<i>PF05233</i>	PHB accumulation regulatory domain
PHB_acc_N	<i>PF07879</i>	PHB/PHA accumulation regulator DNA-binding domain
PKD	<i>PF00801</i>	PKD domain
PPC	<i>PF04151</i>	Bacterial pre-peptidase C-terminal domain
PT	<i>PF04886</i>	PT repeat
P_proprotein	<i>PF01483</i>	Proprotein convertase P-domain
PepSY	<i>PF03413</i>	Peptidase propeptide and YPEB domain
Peptidase_A24	<i>PF01478</i>	Type IV leader peptidase family
Peptidase_C25	<i>PF01364</i>	Peptidase family C25
Peptidase_C25_C	<i>PF03785</i>	Peptidase family C25; C terminal ig-like domain
Peptidase_C58	<i>PF03543</i>	Yersinia/Haemophilus virulence surface antigen
Peptidase_M26_C	<i>PF07580</i>	M26 IgA1-specific Metallo-endopeptidase C-terminal region
Peptidase_M26_N	<i>PF05342</i>	M26 IgA1-specific Metallo-endopeptidase N-terminal region
Peptidase_M28	<i>PF04389</i>	Peptidase family M28
Peptidase_M4	<i>PF01447</i>	Thermolysin metallopeptidase catalytic domain
Peptidase_M4_C	<i>PF02868</i>	Thermolysin metallopeptidase alpha-helical domain
Peptidase_M9	<i>PF01752</i>	Collagenase
Peptidase_S8	<i>PF00082</i>	Subtilase family
Pertactin	<i>PF03212</i>	Pertactin
Pkinase	<i>PF00069</i>	Protein kinase domain
Plug	<i>PF07715</i>	TonB-dependent Receptor Plug Domain
Polysacc_deac_1	<i>PF01522</i>	Polysaccharide deacetylase
Prenyltrans	<i>PF00432</i>	Prenyltransferase and squalene oxidase repeat
Pro_isomerase	<i>PF00160</i>	Cyclophilin type peptidyl-prolyl cis-trans isomerase
Reg_prop	<i>PF07494</i>	Two component regulator propeller

ResIII	<i>PF04851</i>	Type III restriction enzyme
Response_reg	<i>PF00072</i>	Response regulator receiver domain
RrnaAD	<i>PF00398</i>	Ribosomal RNA adenine dimethylase
SCP	<i>PF00188</i>	SCP-like extracellular protein
SCPU	<i>PF05229</i>	Spore Coat Protein U domain
SH3_1	<i>PF00018</i>	SH3 domain
SPDY	<i>PF03771</i>	Domain of unknown function (DUF317)
STN	<i>PF07660</i>	Secretin and TonB N terminus short domain
ScdA_N	<i>PF04405</i>	Domain of Unknown function (DUF542)
Secretin	<i>PF00263</i>	Bacterial type II and III secretion system protein
Secretin_N	<i>PF03958</i>	Bacterial type II/III secretion system short domain
Secretin_N_2	<i>PF07655</i>	Secretin N-terminal domain
ShTK	<i>PF01549</i>	ShTK domain
Sialidase	<i>PF02973</i>	Sialidase N-terminal domain
Subtilisin_N	<i>PF05922</i>	Subtilisin N-terminal Region
TIG	<i>PF01833</i>	IPT/TIG domain
TPR_1	<i>PF00515</i>	Tetratricopeptide repeat
TSP_1	<i>PF00090</i>	Thrombospondin type 1 domain
Tash_PEST	<i>PF07708</i>	Tash protein PEST motif
TonB_dep_Rec	<i>PF00593</i>	TonB dependent receptor
Trans_reg_C	<i>PF00486</i>	Transcriptional regulatory protein C terminal domain
Transgly	<i>PF00912</i>	Transglycosylase
Transpeptidase	<i>PF00905</i>	Penicillin binding protein transpeptidase domain
WND	<i>PF07861</i>	WisP family N-Terminal Region
YSIRK_signal	<i>PF04650</i>	YSIRK type signal peptide
Y_Y_Y	<i>PF07495</i>	Two component regulator three Y motif
zf-C3HC4	<i>PF00097</i>	Zinc finger C3HC4 type (RING finger)
zf-CHY	<i>PF05495</i>	CHY zinc finger