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 affects 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 bar 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

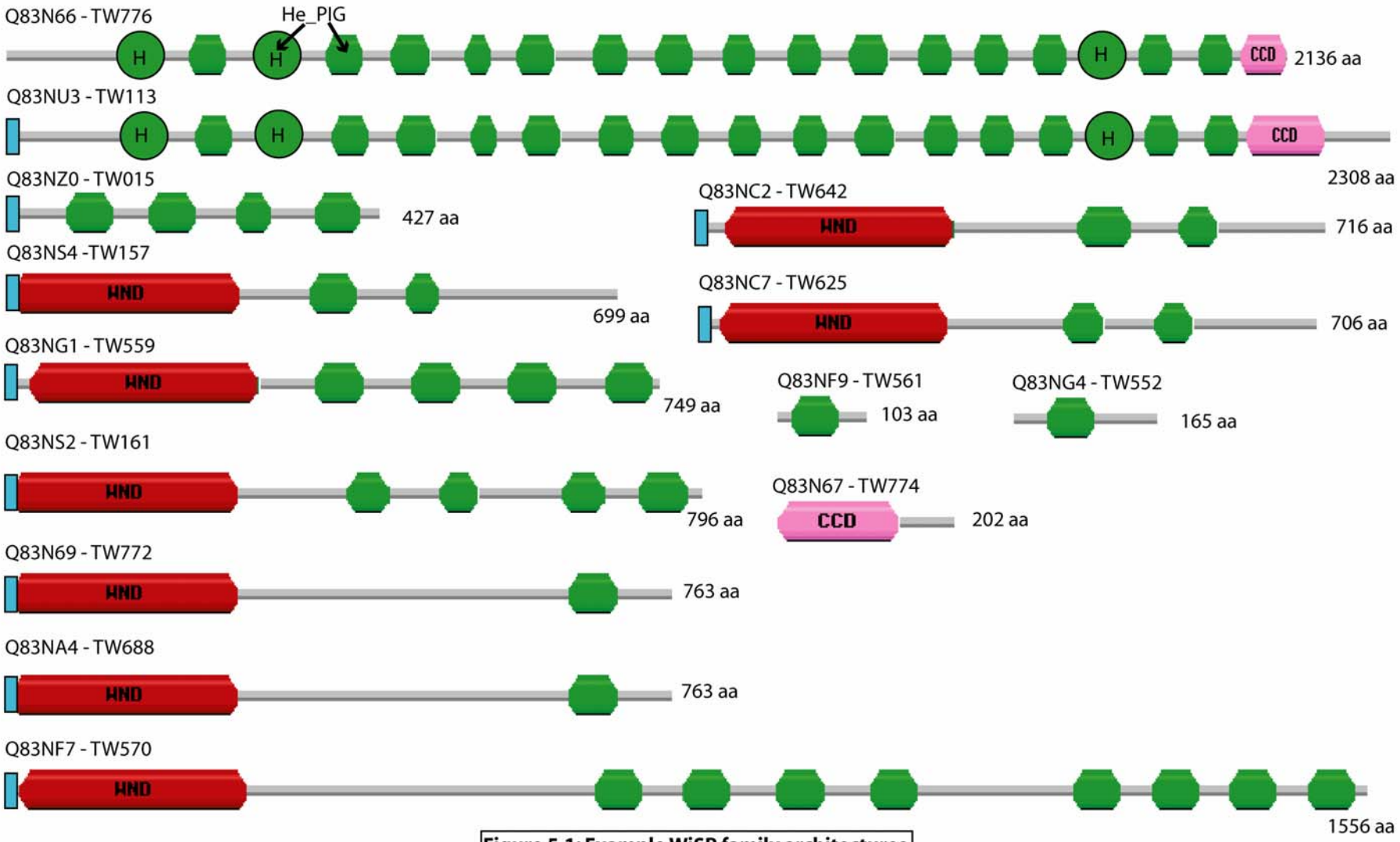


Figure 5.1: Example WiSP family architectures

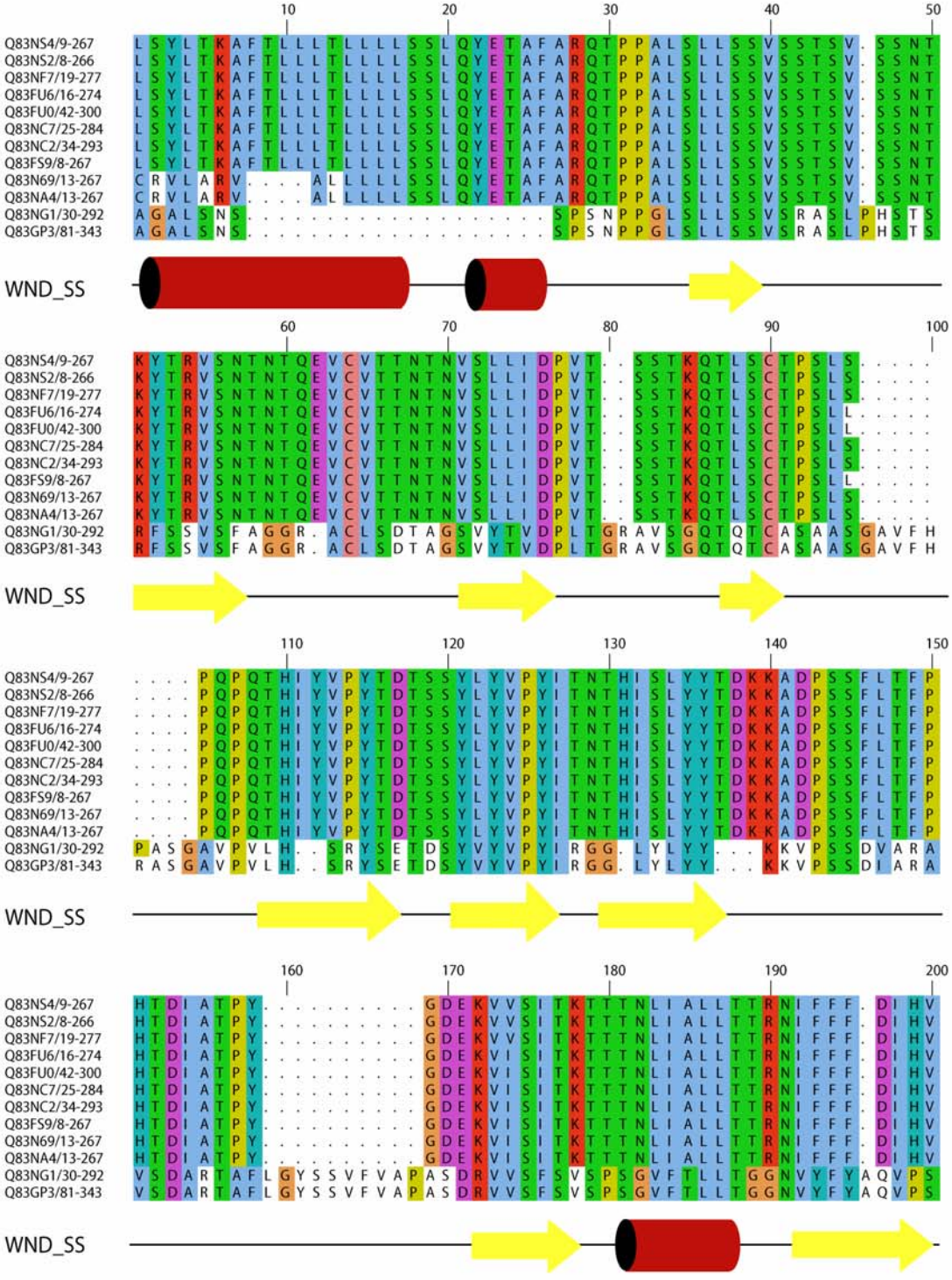


Figure 5.2: Example WND alignment (Page 1)

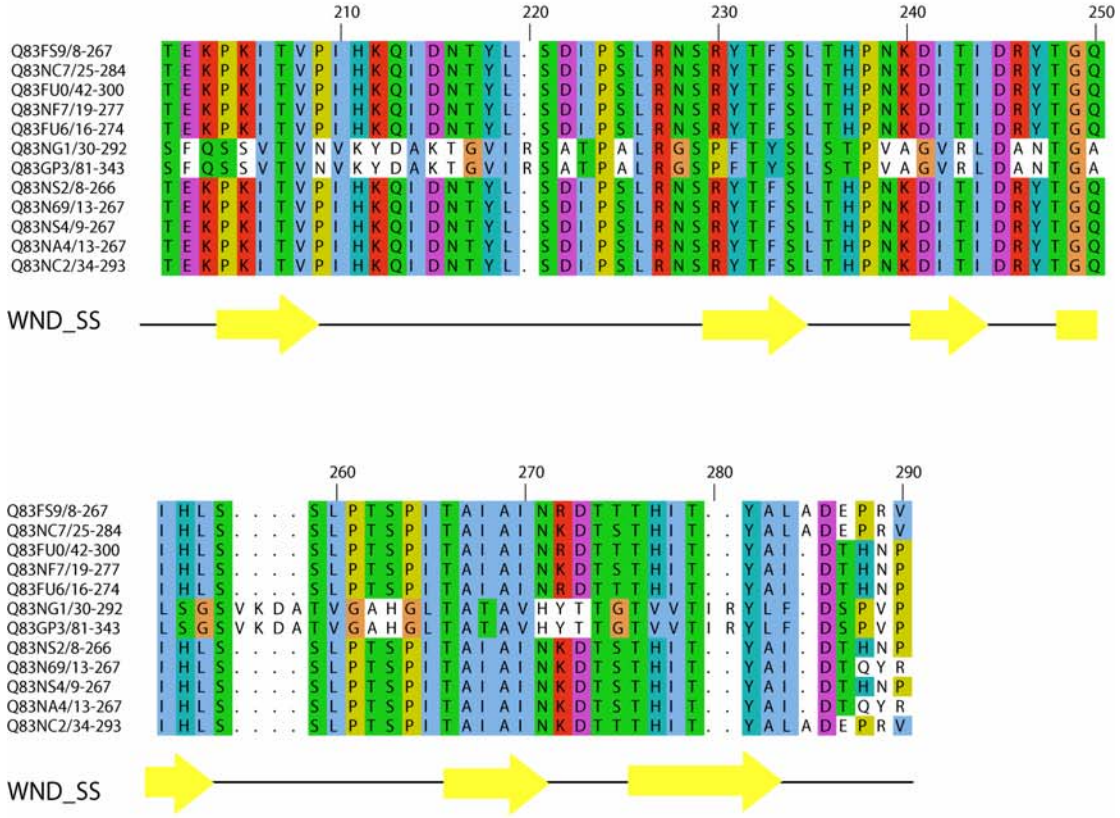


Figure 5.2: WND example alignment (Page 2)

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

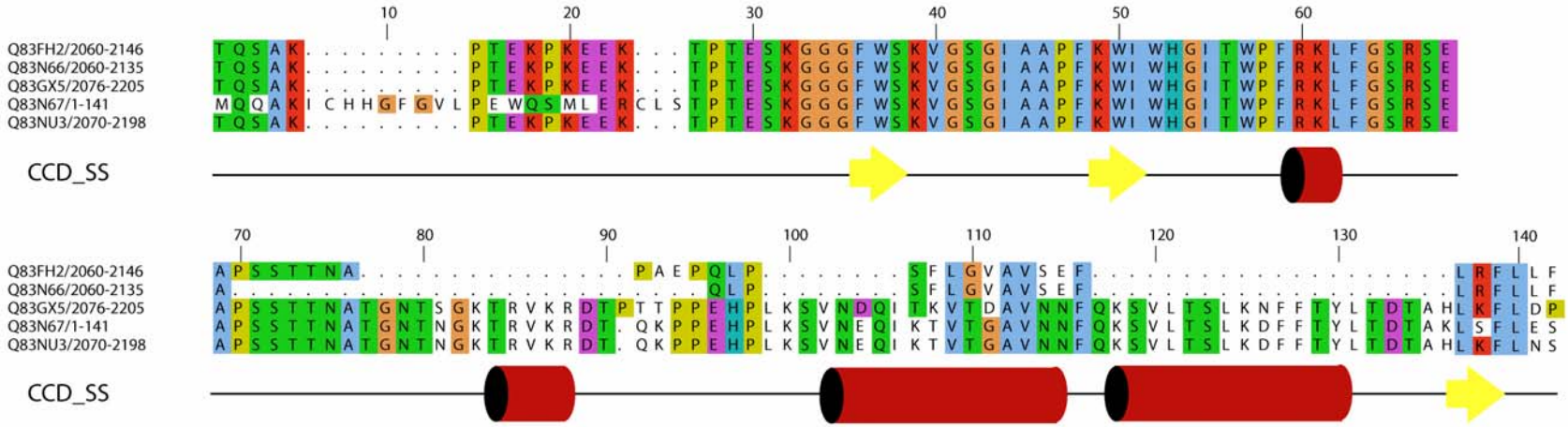


Figure 5.3: CCD example alignment

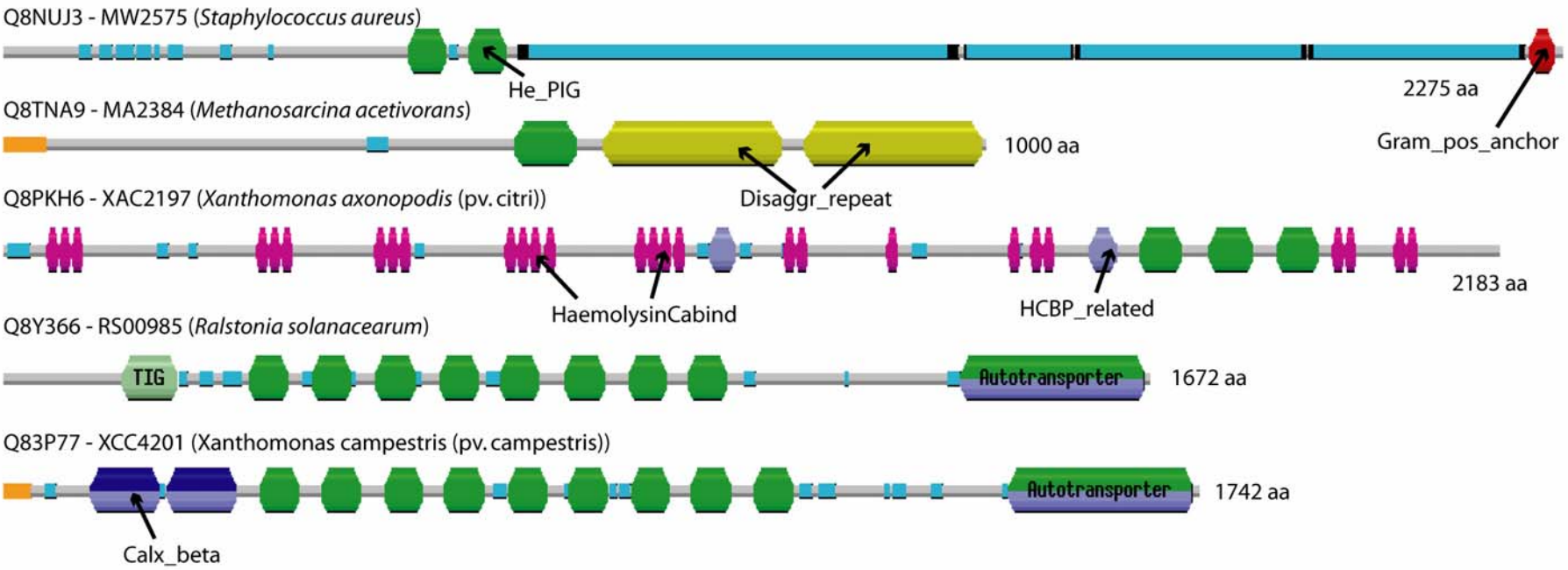


Figure 5.4: Example He_PIG architectures

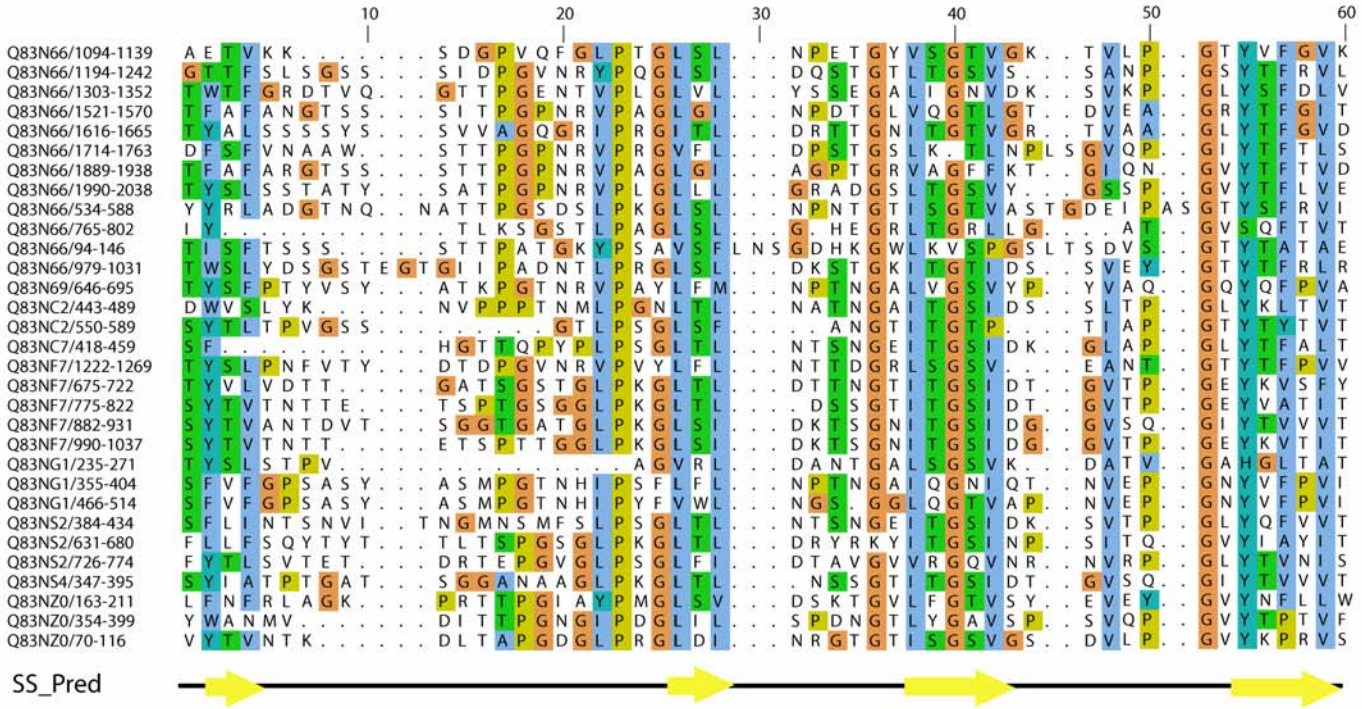


Figure 5.5: He_PIG example alignment

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.

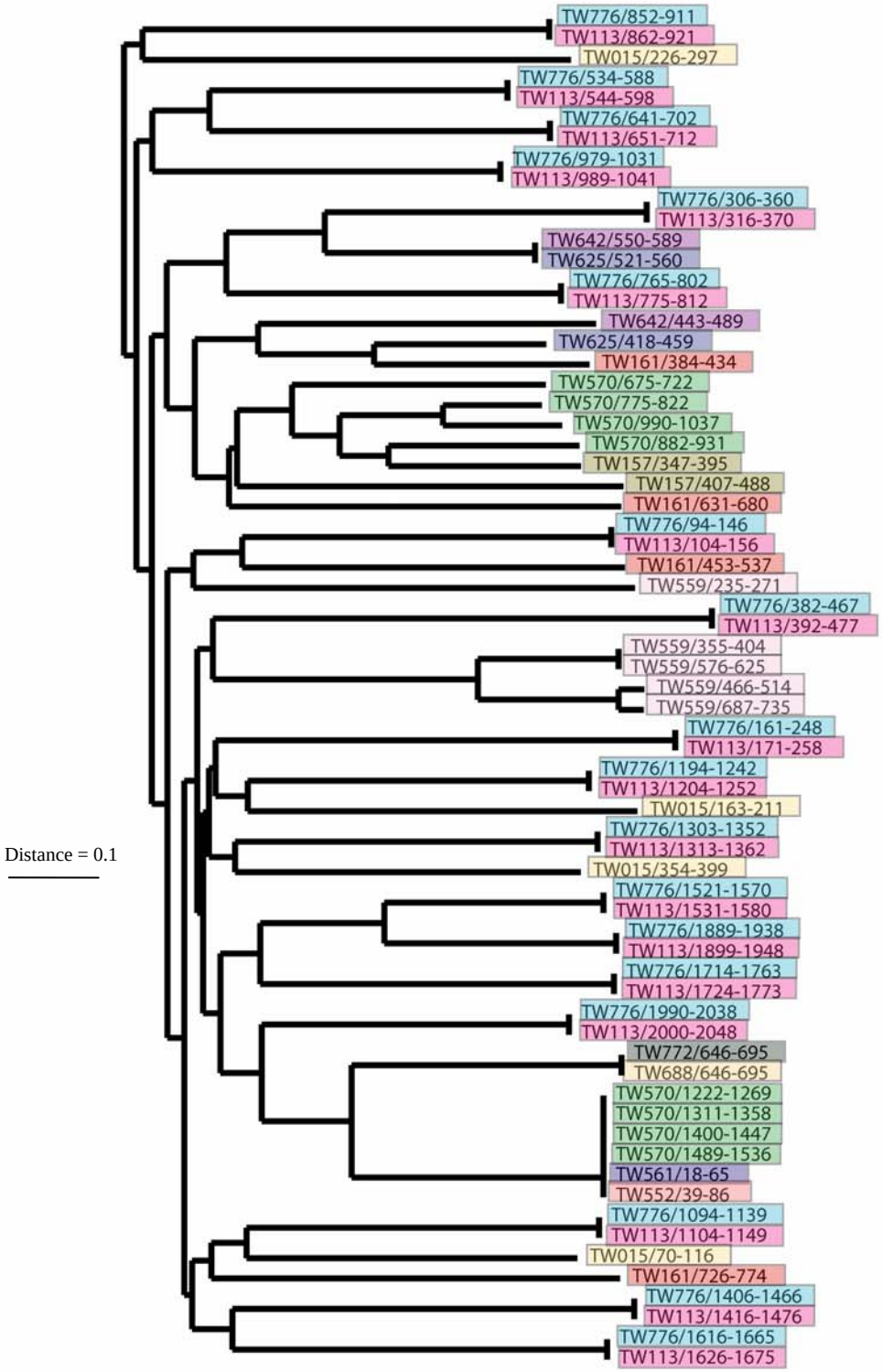


Figure 5.6: N-J Tree of all He_PIG domains from *Tropheryma whipplei*
The tree was constructed in Belvu using uncorrected distances and the "center of tree" approach. The tree balance equals 0.0. Each leaf represents a He_PIG domain; the sequence name and coordinates are given. Each protein has been 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 meliodosis (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

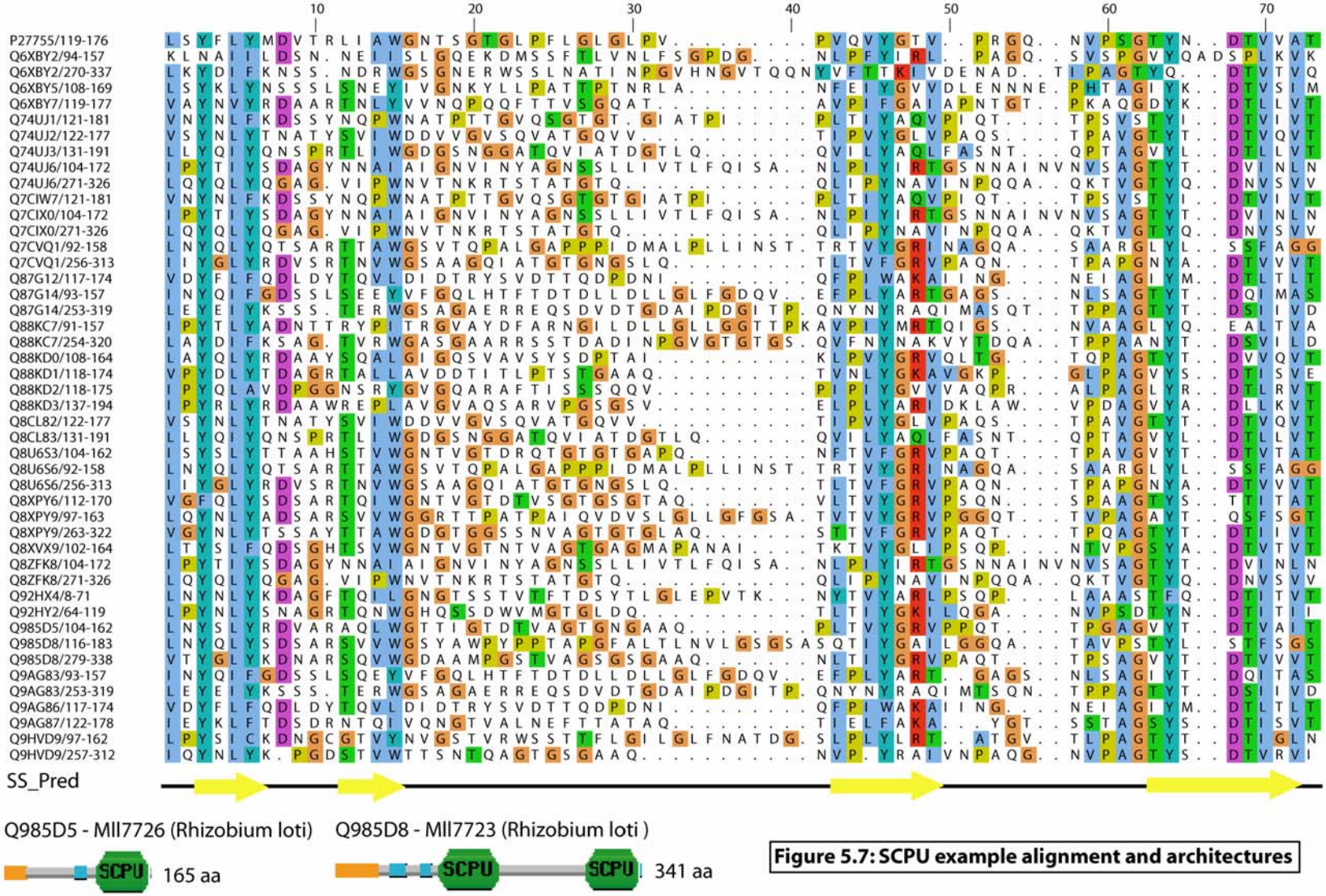


Figure 5.7: SCPU example alignment and architectures

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.

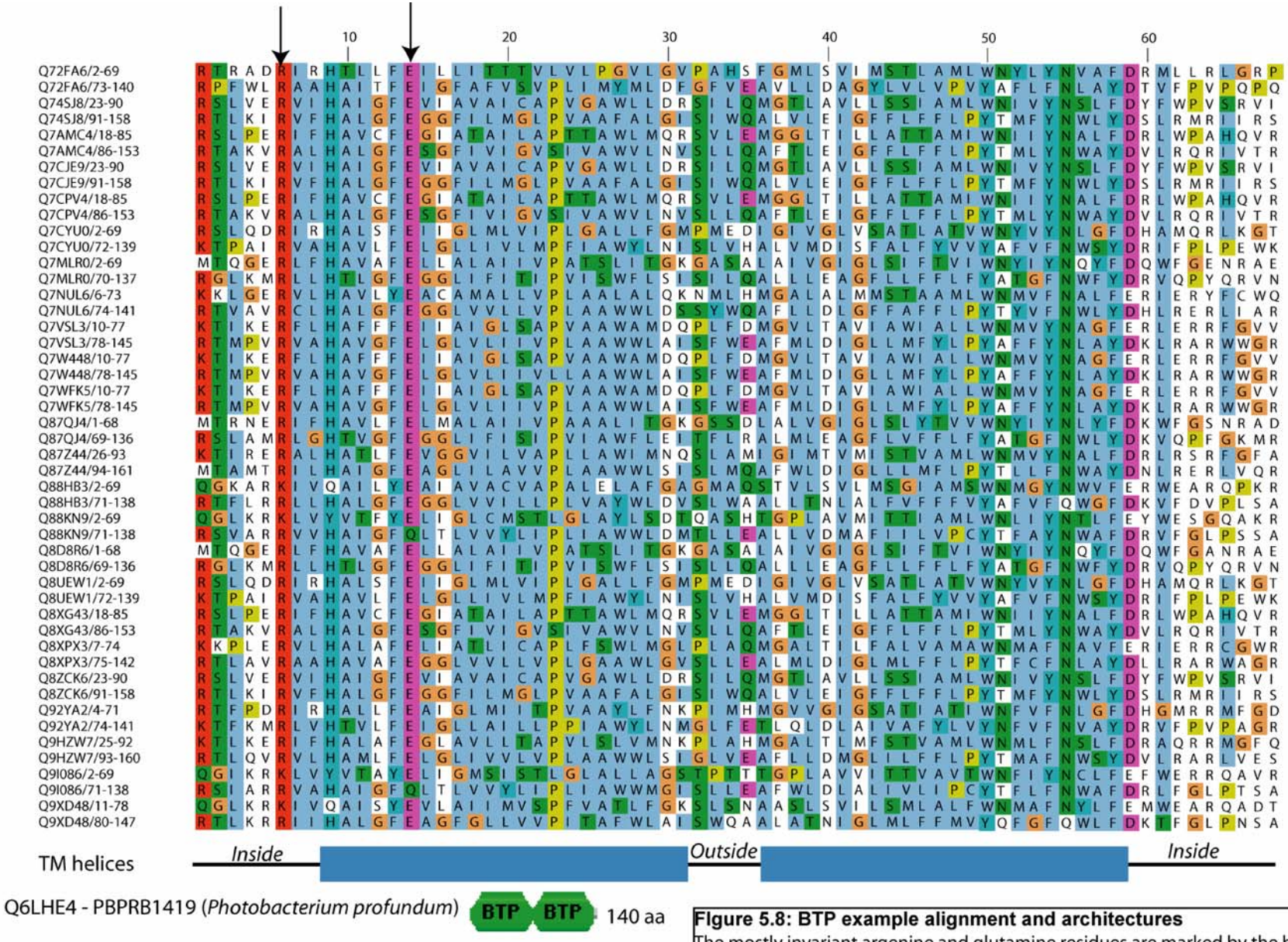


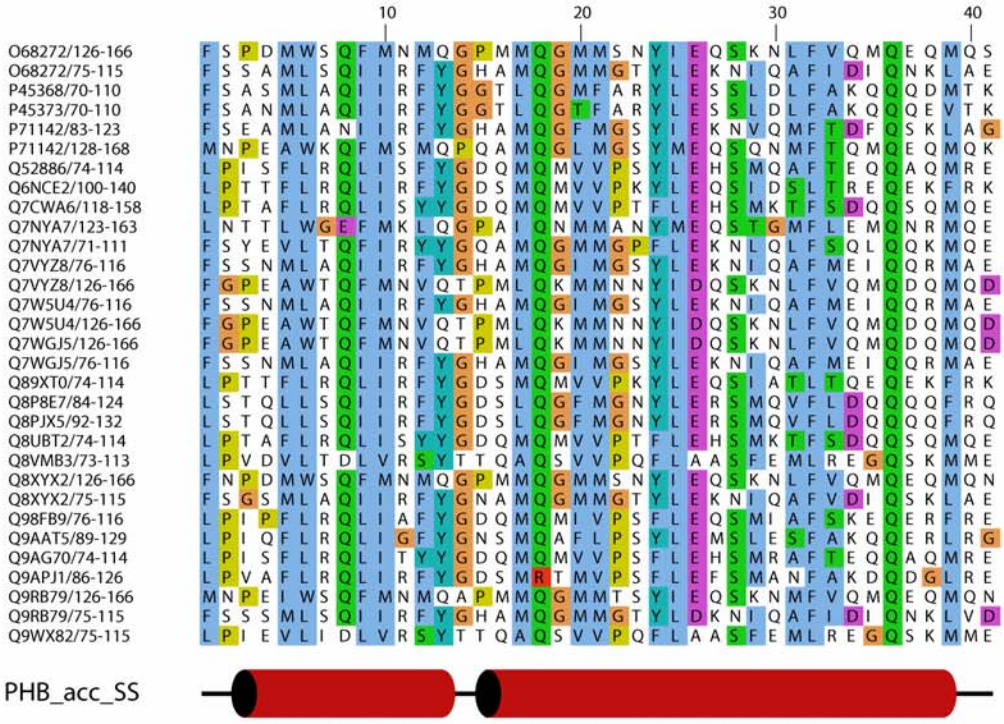
Figure 5.8: BTP example alignment and architectures
The mostly invariant argenine and glutamine residues are marked by the black arrows

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 dentrificans*. 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



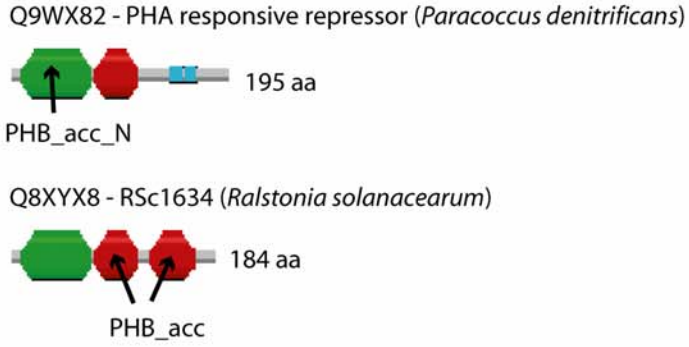


Figure 5.9: PHB_acc example alignment and architectures

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 *Chlamydophila 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

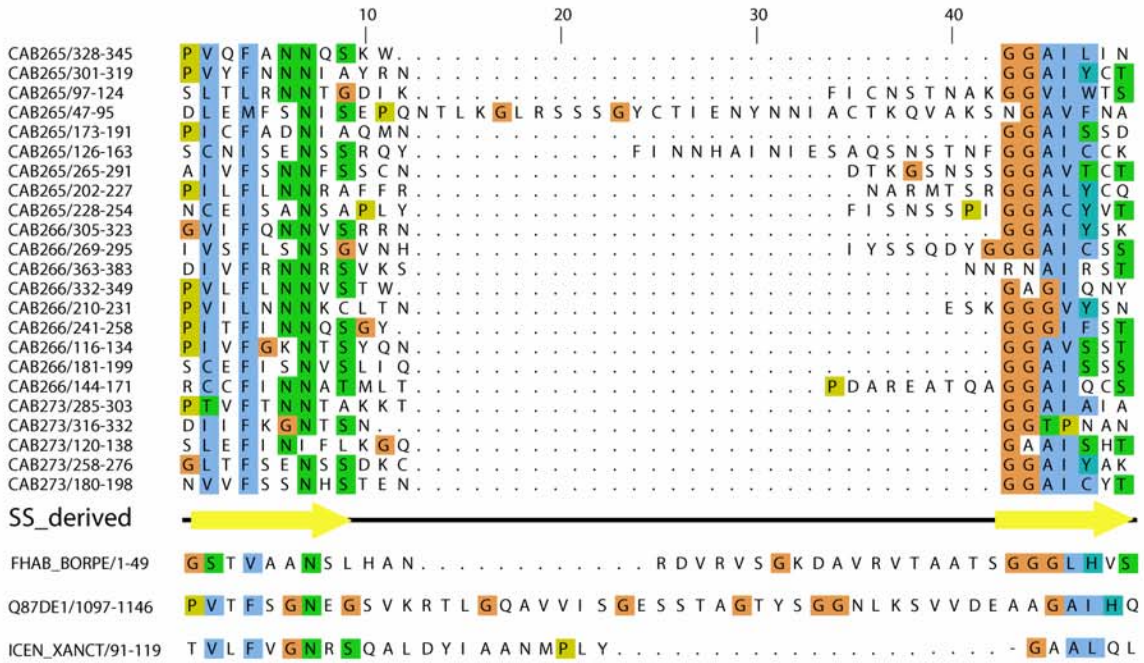


Figure 5.10: Example Chlam_PMP alignment with related sequences
The main alignment shown here is of the Chlam_PMP family. Underneath are examples of the related Fil_hameagg, Hep_hag and Ice_nucleation families (respectively).

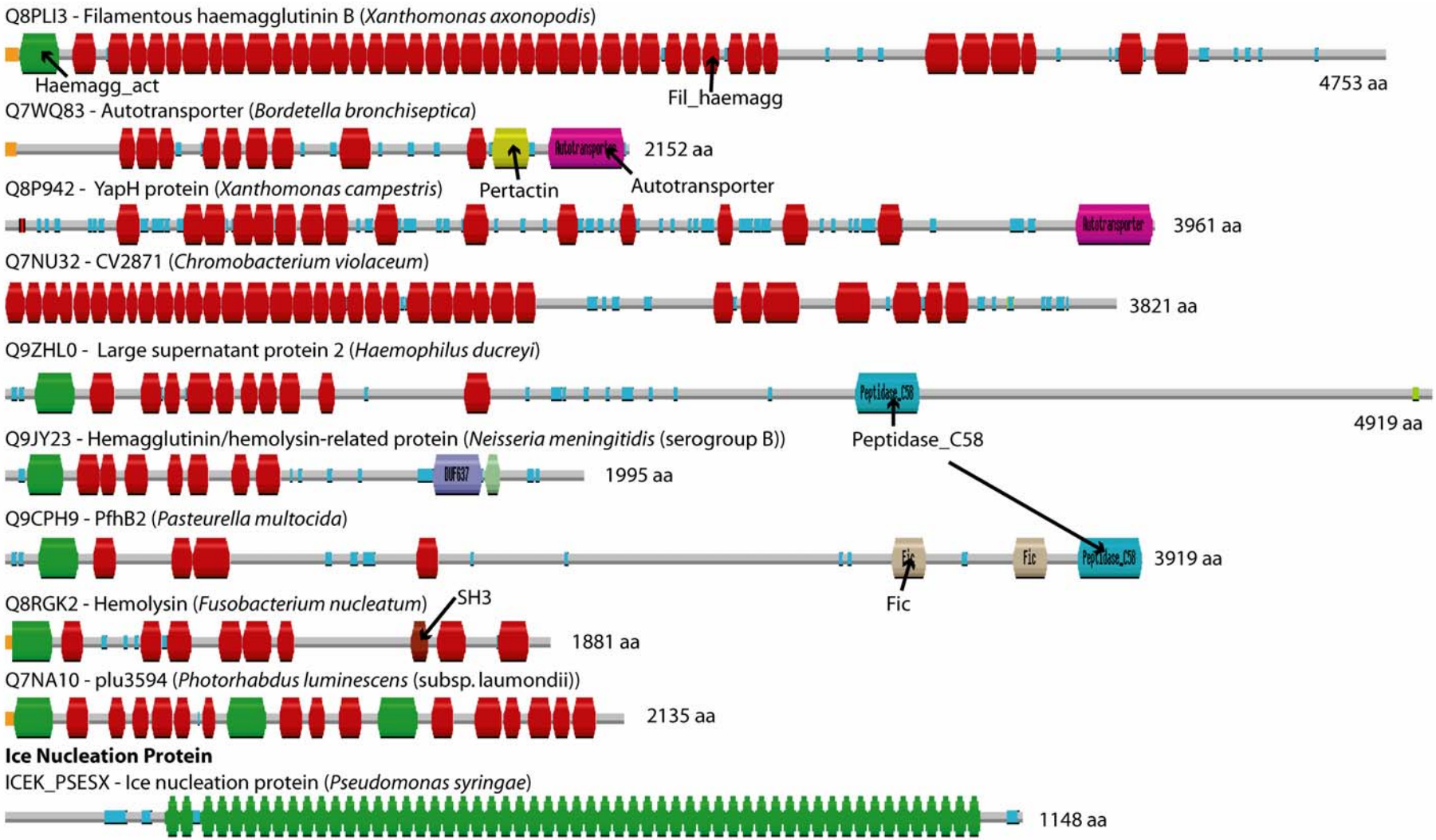


Figure 5.11: Fil_haemagg and Ice_nucleation example architectures

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 Chlamydophila abortus (Manuscript under preparation)

5.3.1 Background

Chlamydophila 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

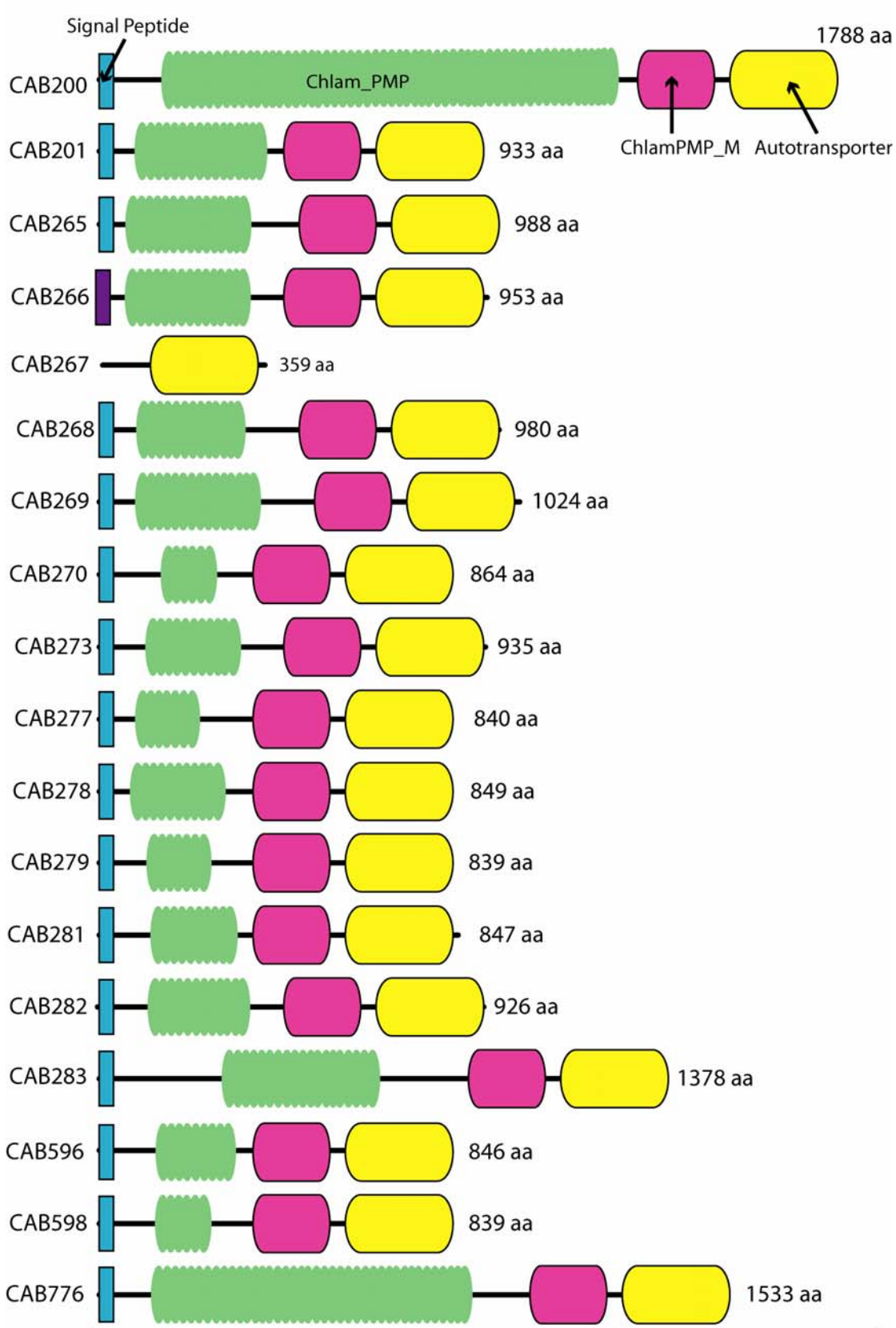


Figure 5.12: The *Chlamydophila abortus* polymorphic membrane protein family
This family is not shown in the usual Pfam representation since the protein sequences are not yet available in UniProt. The signal peptide of CAB266 is shown in purple as there is some doubt over its validity. SignalP 3.0's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

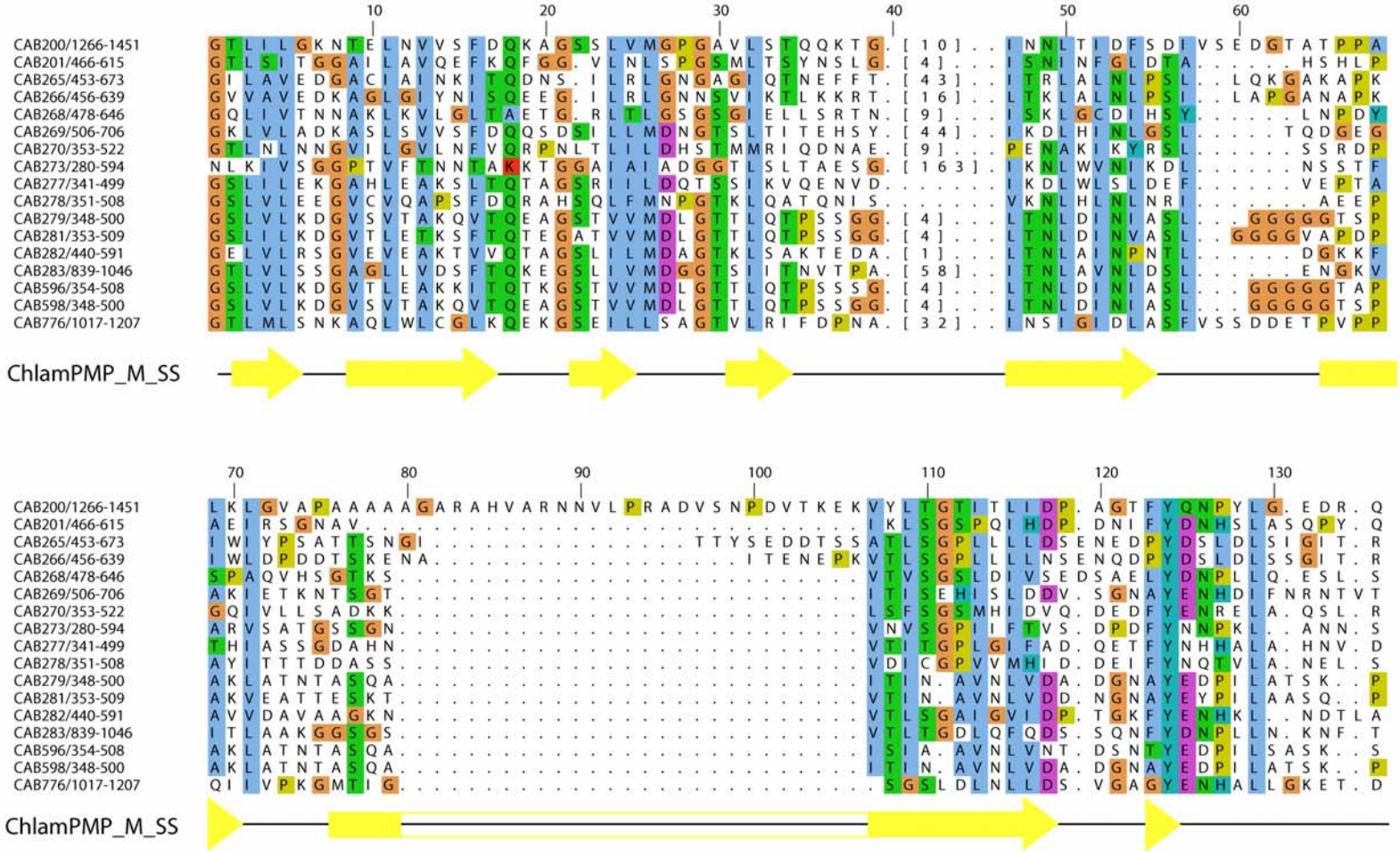


Figure 5.13: ChlamPMP_M example alignment (Page 1)

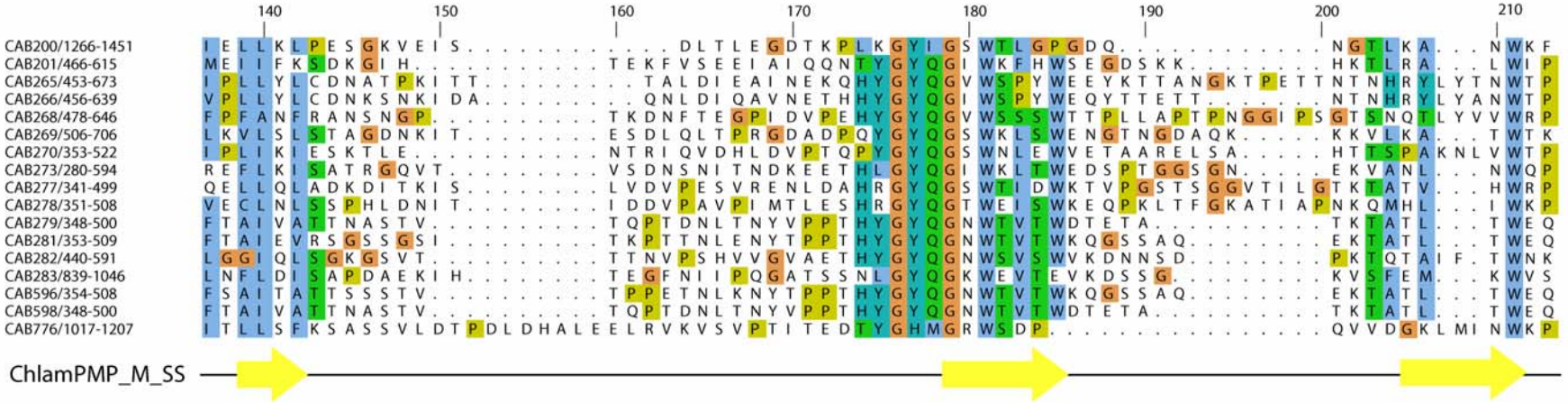


Figure 5.13: ChlamPMP_M example alignment (Page 2)

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

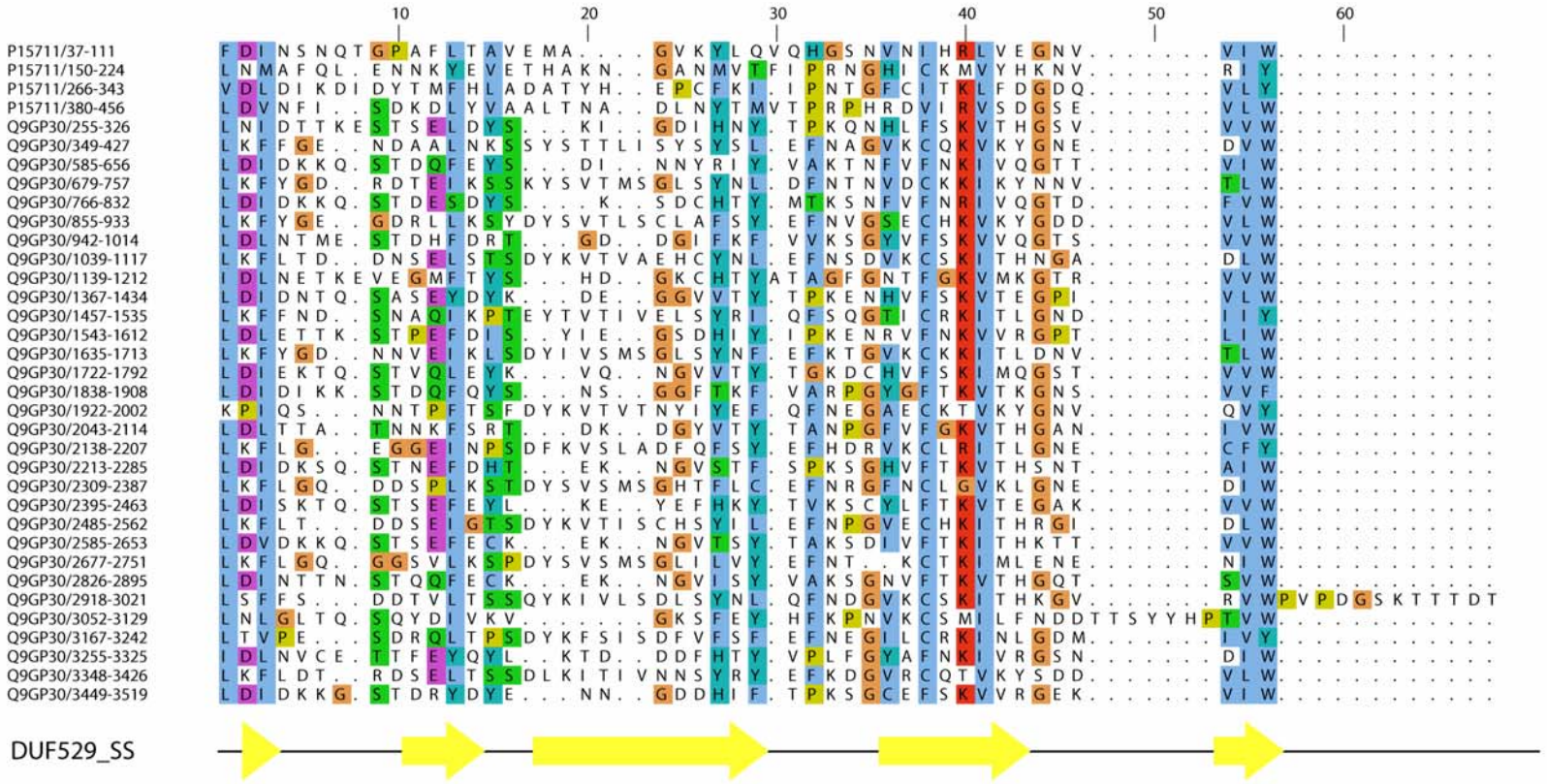


Figure 5.14: FAINT example alignment (Page 1)

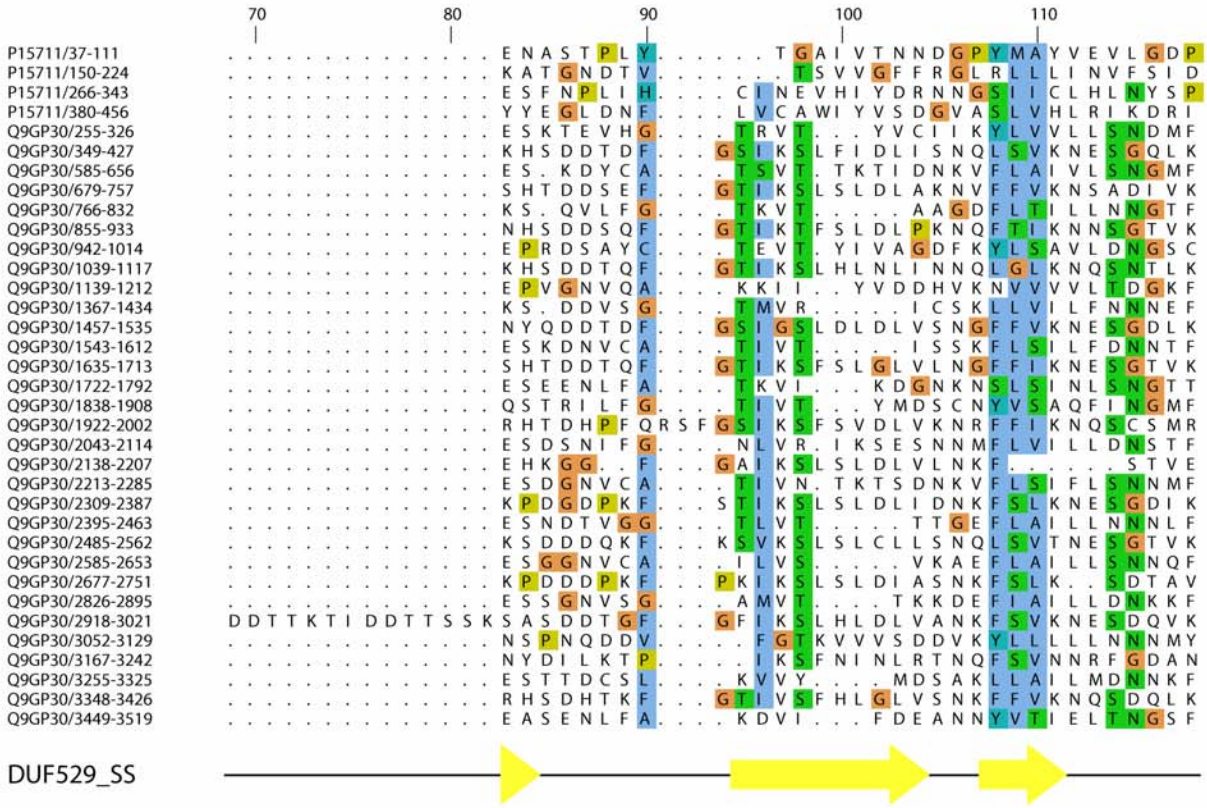


Figure 5.14: FAINT example alignment (Page 2)

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).

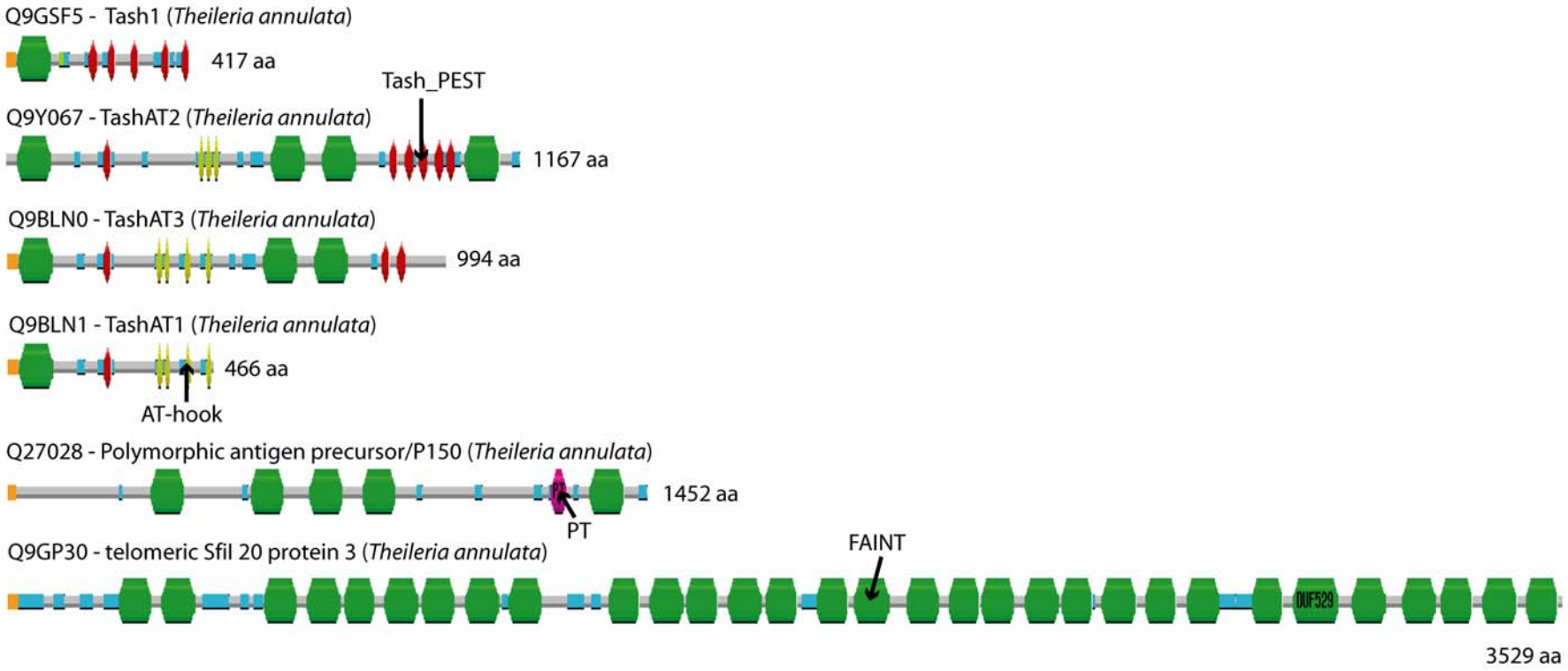


Figure 5.15: FAINT example architectures

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 tumourogenic 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).

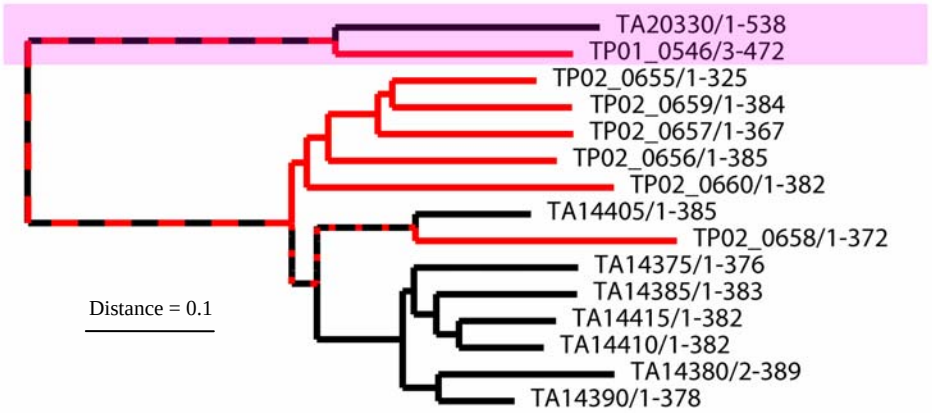


Figure 5.16: Neighbour Joining Tree of the choline kinases of *T. parva and T. annulata*
The tree was constructed in Belvu using uncorrected distances and the "center of tree" approach. The tree balance equals 0.0. All the choline kinase sequences were aligned using MAFFT and then the large inserts in 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.*

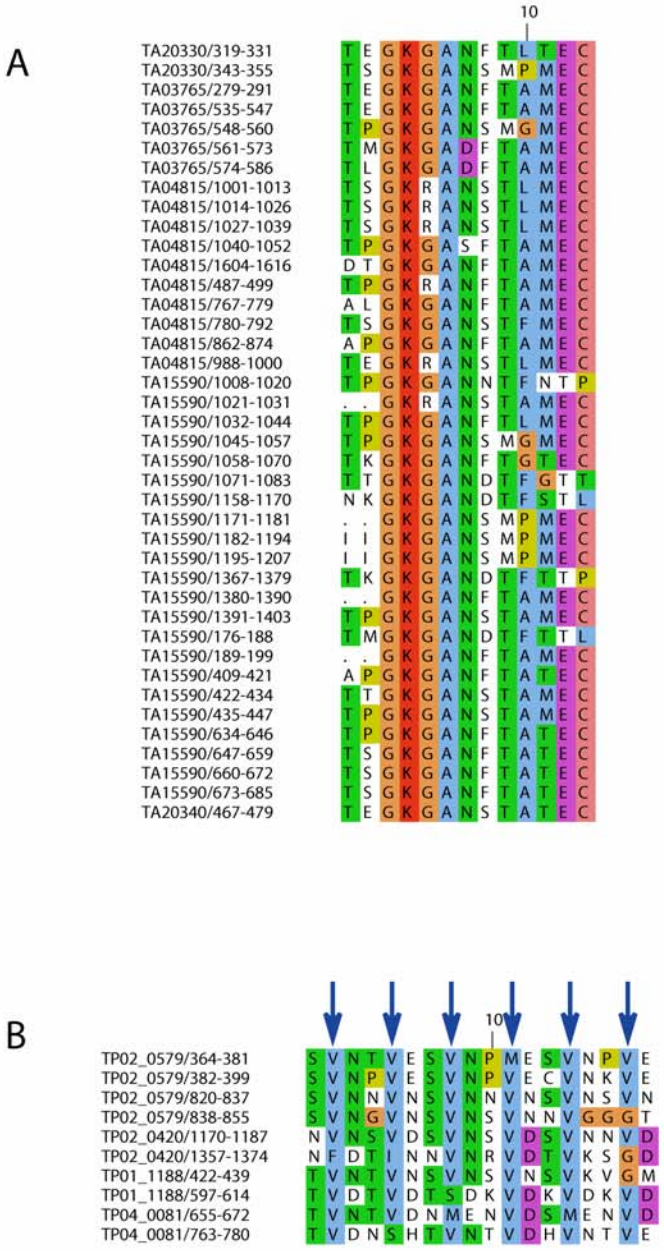


Figure 5.17: Example alignments for the TASR short repeats of *T. annulata* (A) and *T. parva* (B) The T. annulata TASR_1 repeat is 13 residues in length, and so can be reasonably confidently identified. Some of the repeats do appear to be 11 residues. The T. parva TASR_2 repeats are only three residues - 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.

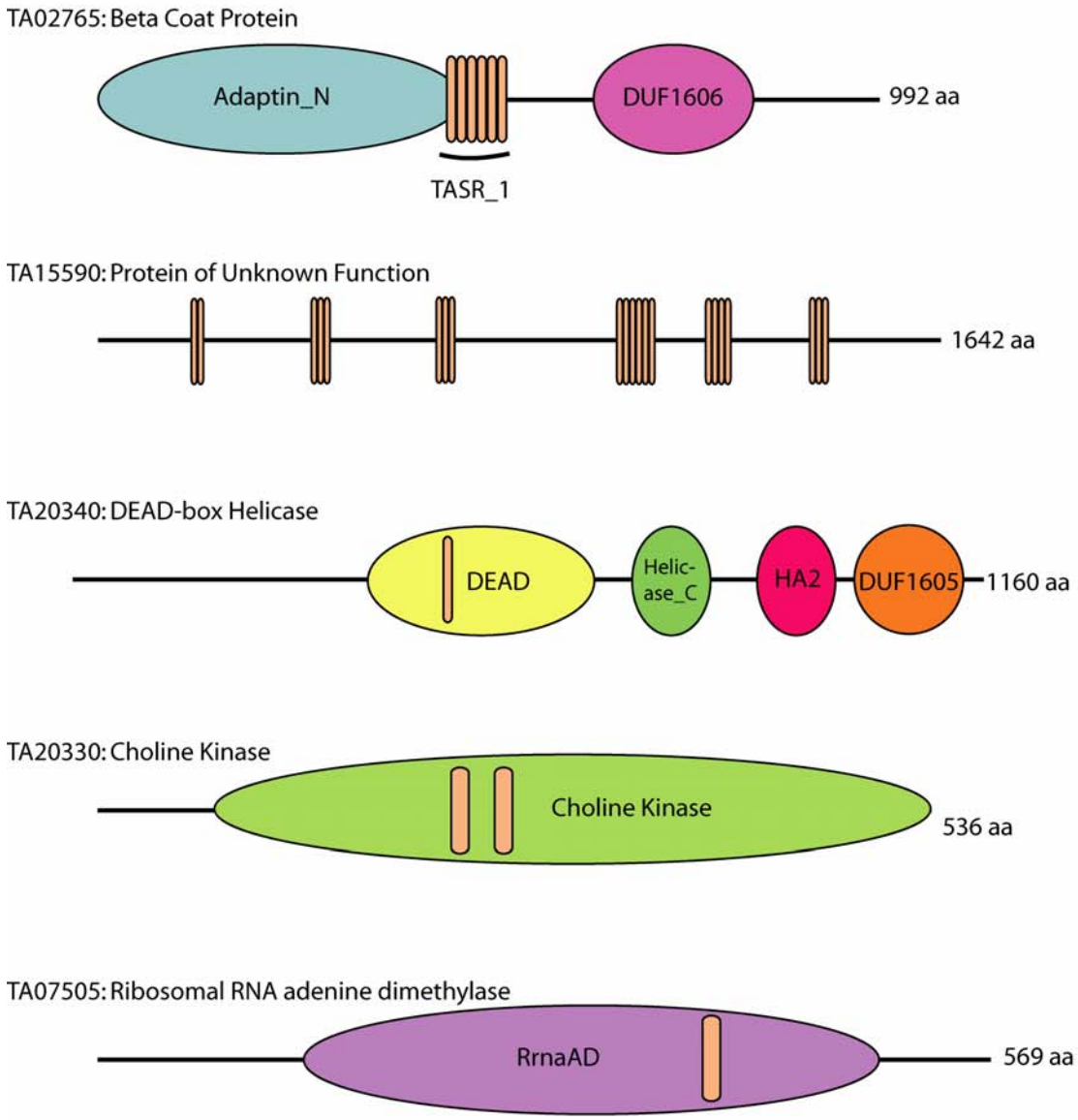


Figure 5.18: *Theileria annulata* TASR_1 example architectures
The normal Pfam view is not shown since these proteins were not yet available in UniProt. The TASR_1 repeats are depicted as orange bars. As can be seen they fall within, between and across the boundaries of domains.

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).

<table>
<tr><th colspan="4">Annulata vs. Parva</th><th colspan="4">Parva vs. Annulata</th></tr>
<tr><td rowspan="2">Nº of orthologues with TASR_2 (A)*</td><td rowspan="2">P(success) = A/B</td><td>25</td><td rowspan="2">0.021</td><td rowspan="2">Nº of orthologues with TASR_1 (A)*</td><td rowspan="2">P(success) = A/B</td><td>25</td><td rowspan="2">0.016</td></tr>
<tr><td>T. parva genome size (B)</td><td>4150</td><td>T. annulata genome size (B)</td><td>4000</td></tr>
<tr><td>Nº of TASR_1-containing proteins in T. annulata</td><td>Nº of Trials</td><td colspan="2">89</td><td>Nº of TASR_1-containing proteins in T. parva</td><td>Nº of Trials</td><td colspan="2">65</td></tr>
<tr><td colspan="4">:. assuming a binomial distribution the P(overlap by chance) = 0.00 (3 sig. figs).</td><td colspan="4">:. assuming a binomial distribution the P(overlap by chance) = 0.00 (3 sig. figs).</td></tr>
<tr><td colspan="8">* All proteins with a TASR, but no easily identifiable orthologue were discarded. This came to 15 proteins in all.</td></tr>
<tr><td colspan="8">Table 5.1: Statistics testing whether the overlap between the T. parva TASR_2-containing proteins and the T. annulata TASR_1-containing proteins is by chance.</td></tr>
</table>

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.