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.*

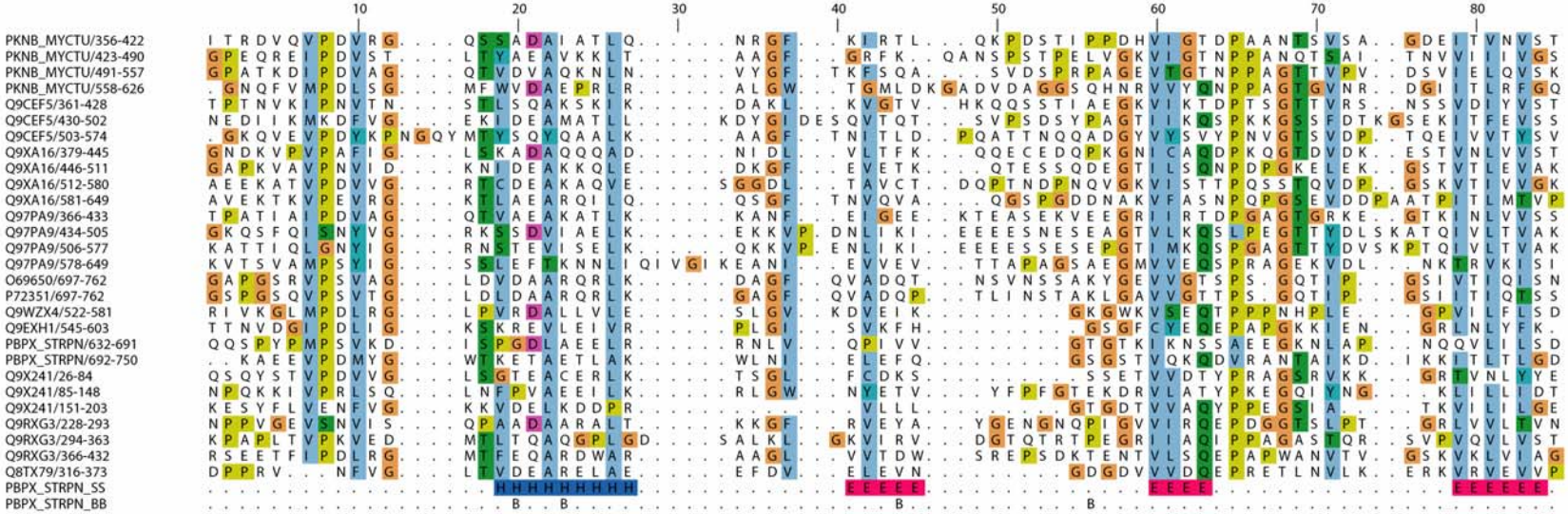

Figure 4.1: Example PASTA alignment.
Underneath the alignment the line PBP_STRPN_BB marks (B) the residues in closest proximity to cefuroxime in the crystal structure of PBP2X (PDB:1QMF).

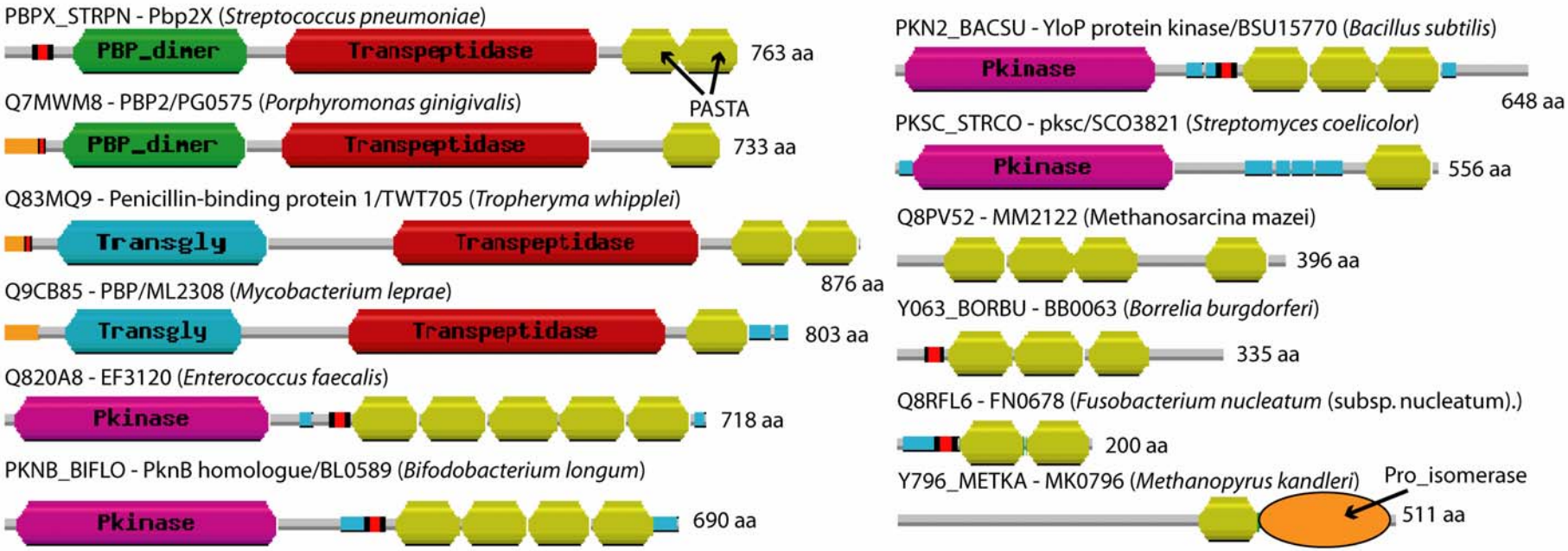

Figure 4.2: Example PASTA domain architectures

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).

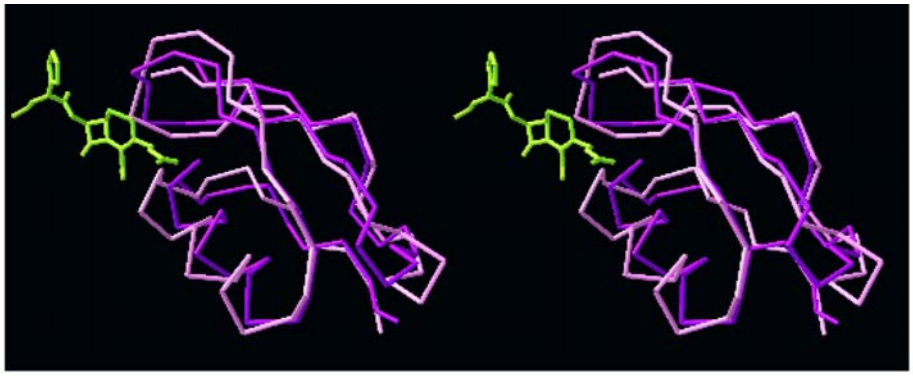

Figure 4.3: Stereo view of the two PASTA domains of *Streptococcus pneumoniae* PBP2X
The two PASTA domains of PBP2X are shown , one in purple (residues 636-691) and the second in pink (694-750), superposed on each other. As can be seen there is little variance in the structure between the two domains. 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. Eukaroytes 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 assignations 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

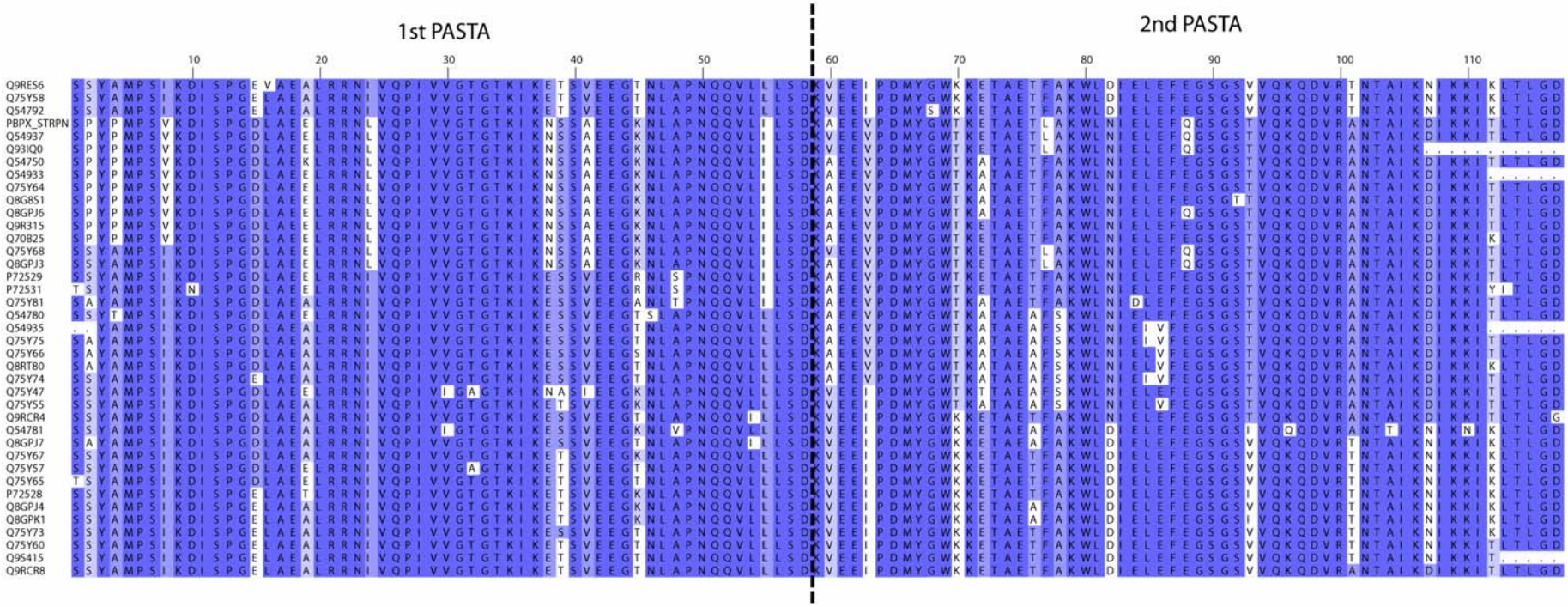

Figure 4.4: Distribution of resistance mutations in the PASTA domains of PBP2X
This alignment was created by taking the 86 copies of *Streptococcus pneumoniae* PBP2X C-terminus that included both PASTA domains from UniProt, aligning them with MAFFT, and making it non-redundant (no identical sequences). 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 $1.2x10^{-6}$. 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.

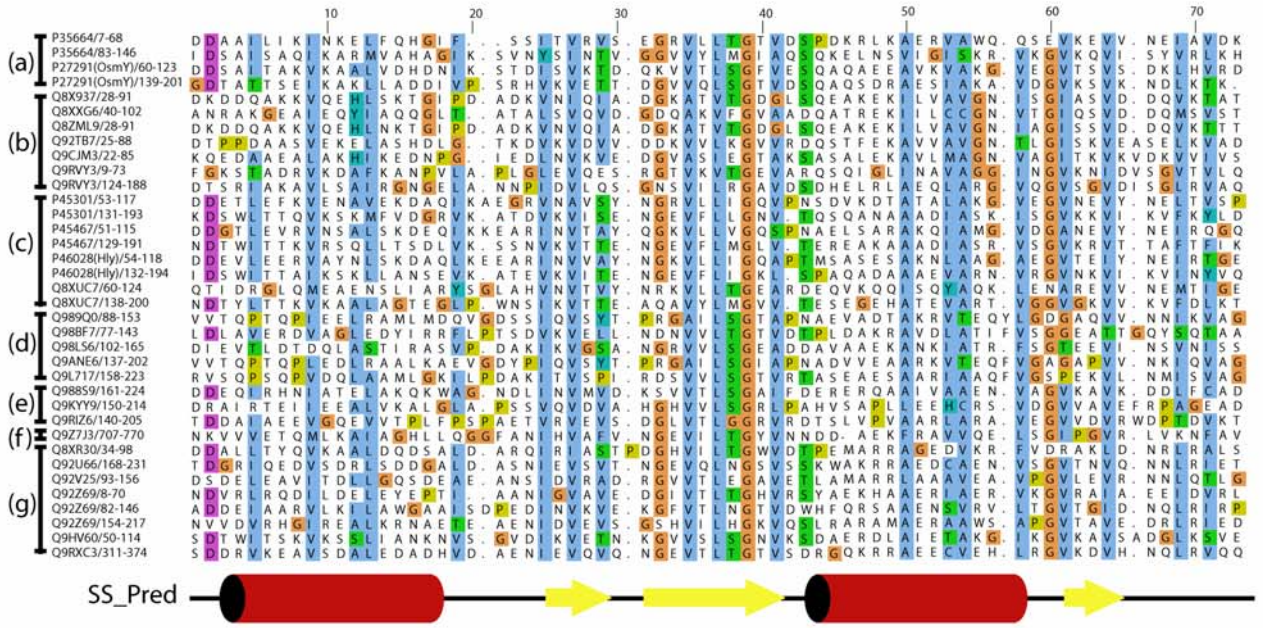

Figure 4.5: BON domain example alignment
The sequence alignment is grouped according to protein function. The groupings are as such: (a) OsmY protein-related (Swiss: P27291); (b) LysM domain-associated (Pfam: PF01476); (c) HLY protein-related (Swiss: P46028); (d) nodulation specificity; (e) CBS domain-associated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

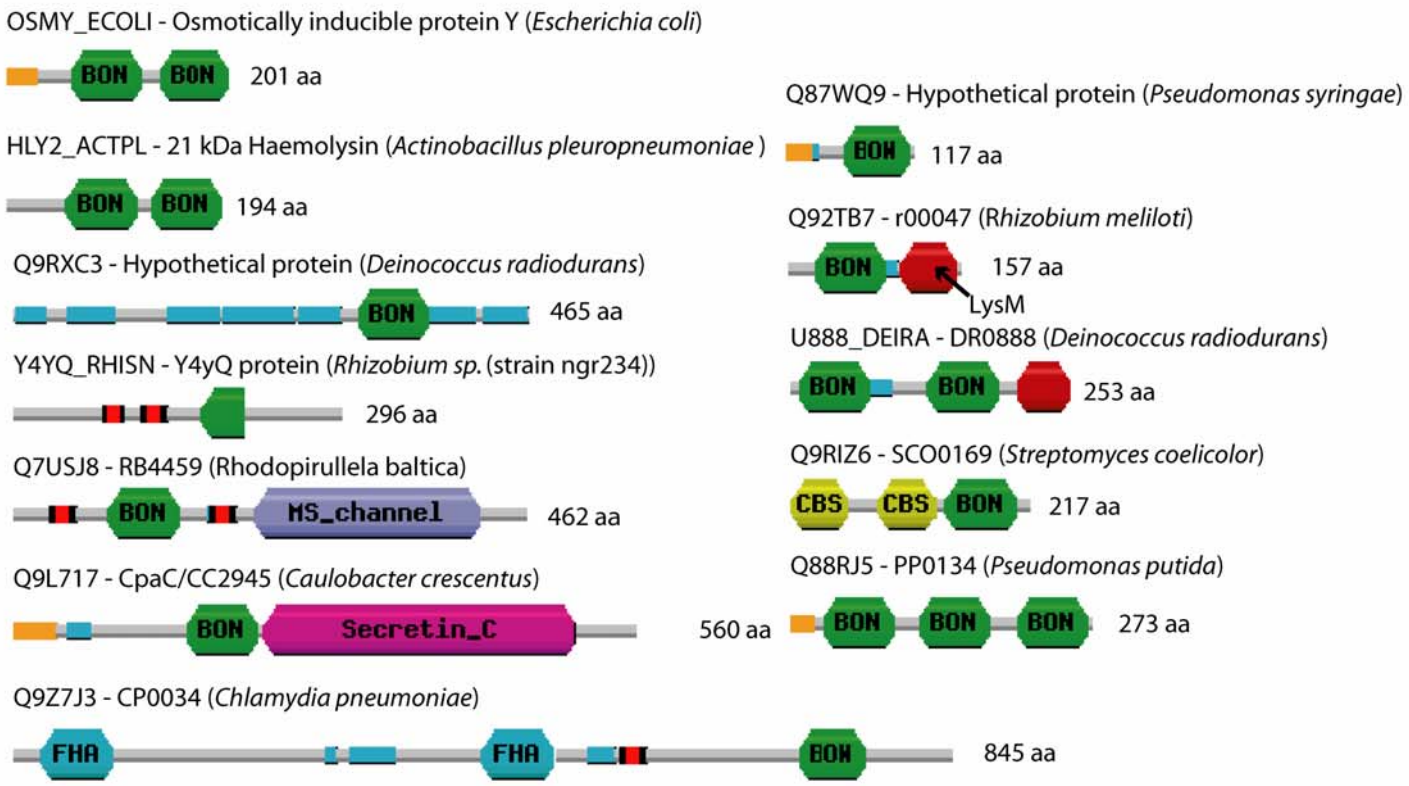

Figure 4.6: Example BON domain architectures

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

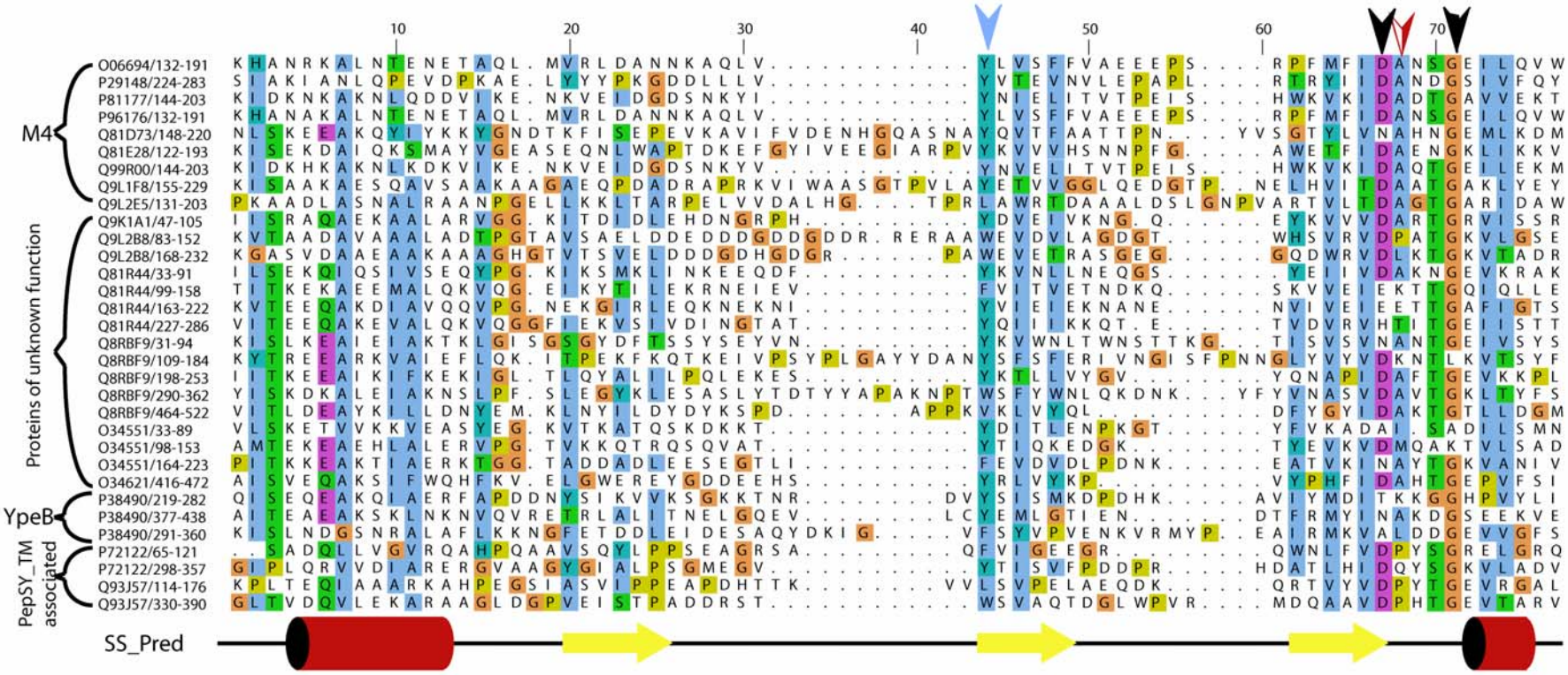

Figure 4.7: Example PepSY domain alignment
The particularly conserved region discussed in the main body of the text is marked by the two black arrows. The locus of the A183V mutation of pseudolysin (UniProt: P14756) described and characterised by Braun, Bitter, *et al.* (2000) is marked by the red arrow. The blue arrow marks the mostly conserved tyrosine or aromatic residues mentioned in the main body of the text.

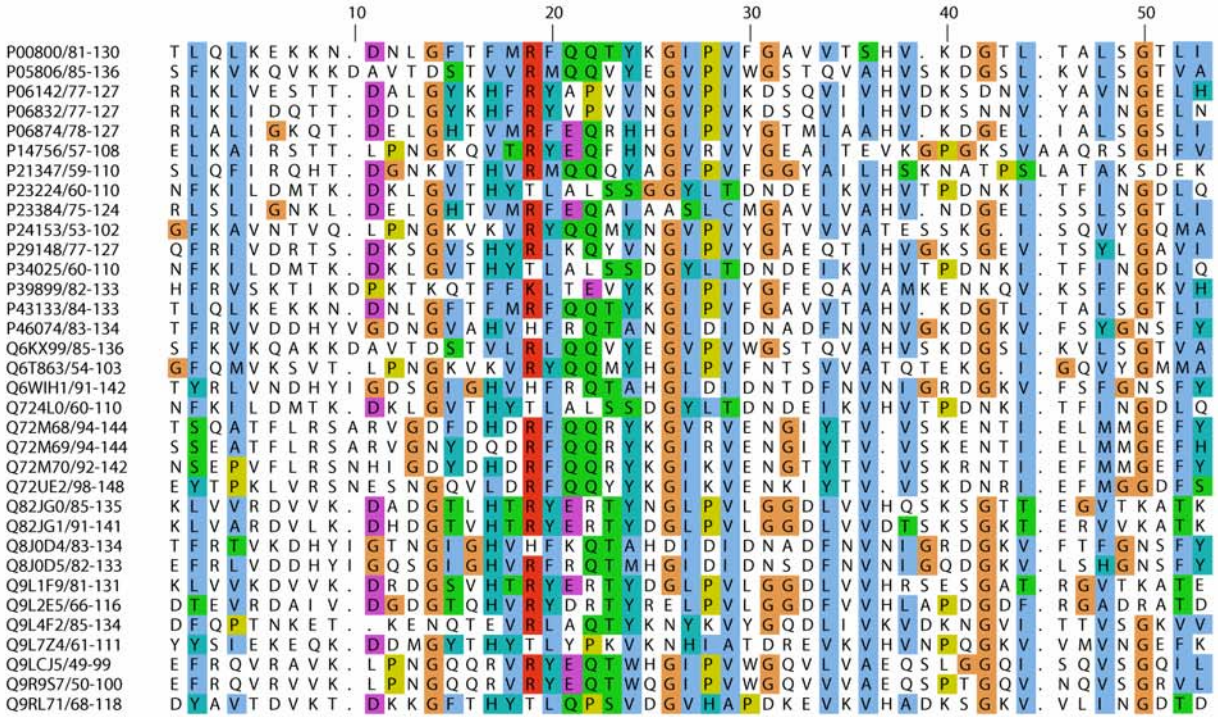

Figure 4.8: FTP motif example alignment

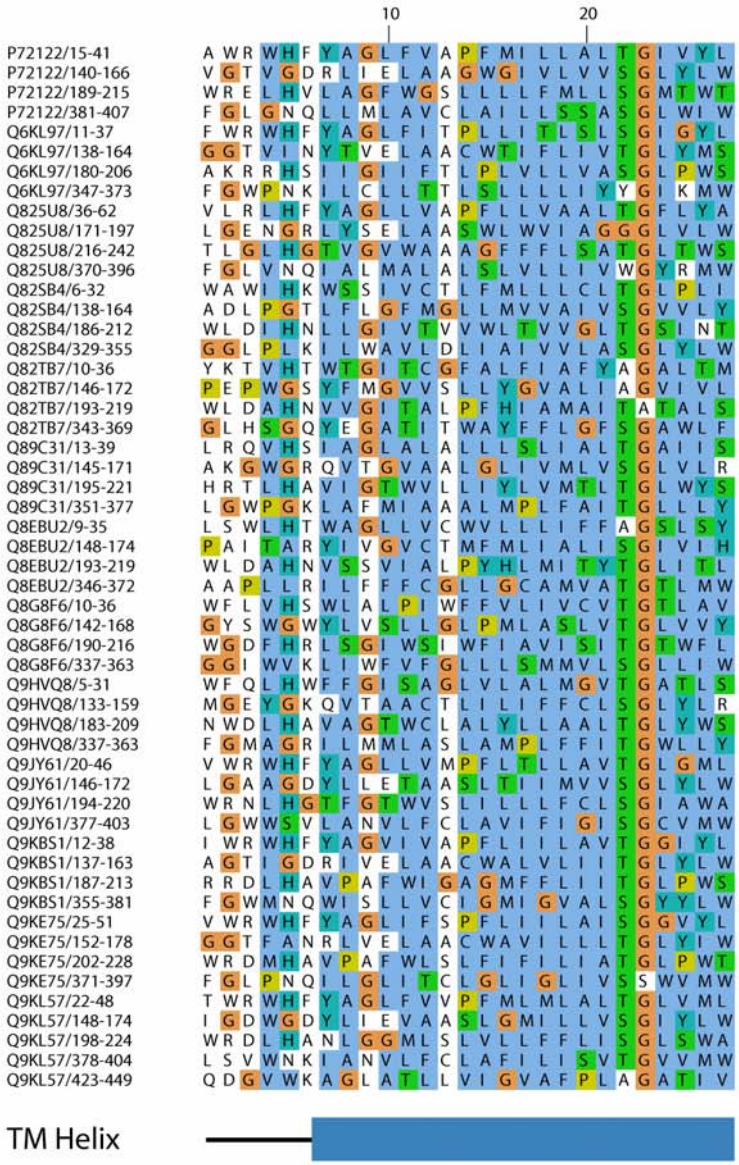

Figure 4.9: PepSY_TM example alignment
Below the alignment is a depiction of where the transmembrane helix occurs. The prediction was made using TMHMM.

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

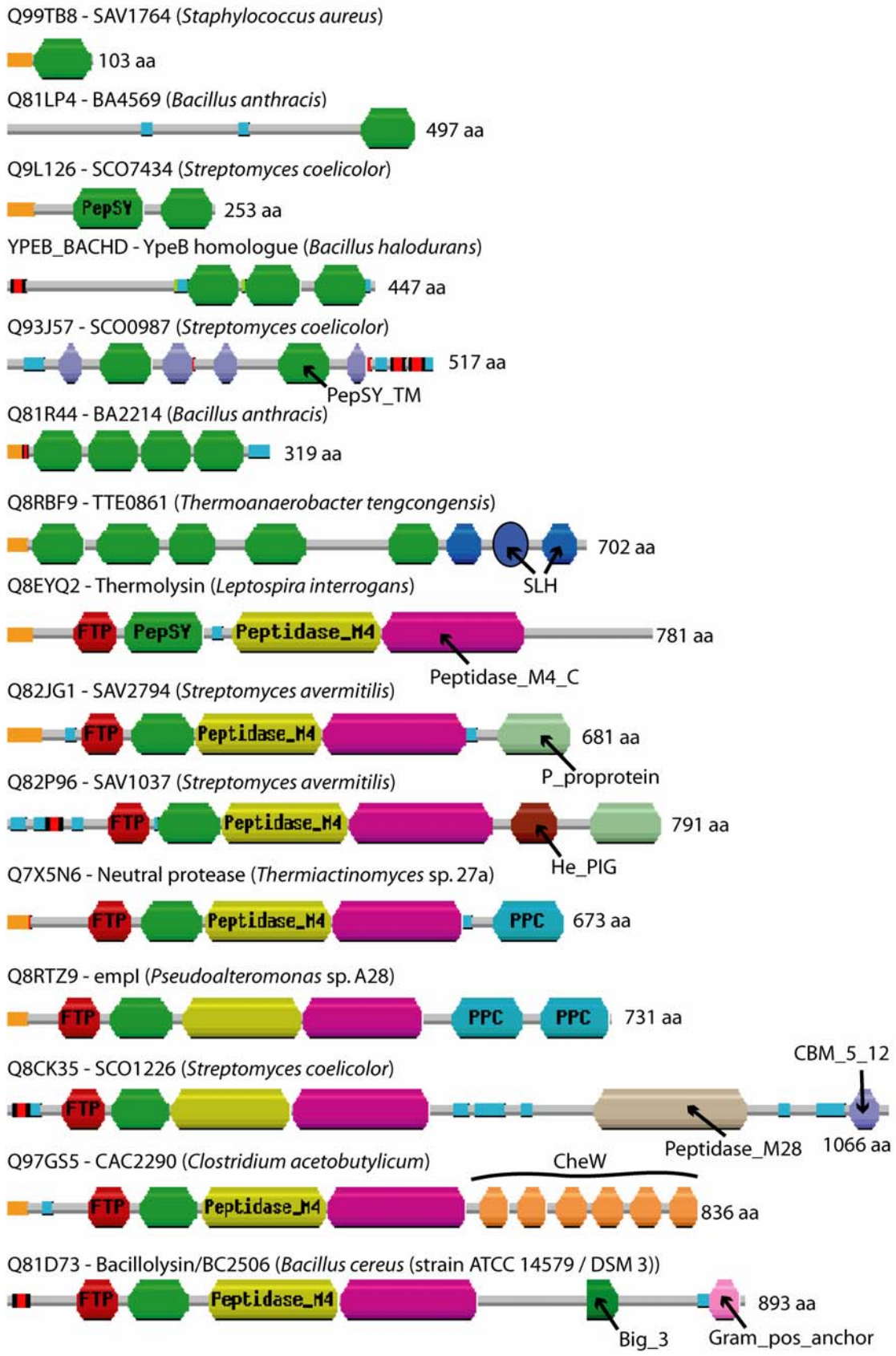

Figure 4.10: Example PepSY domain architectures

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

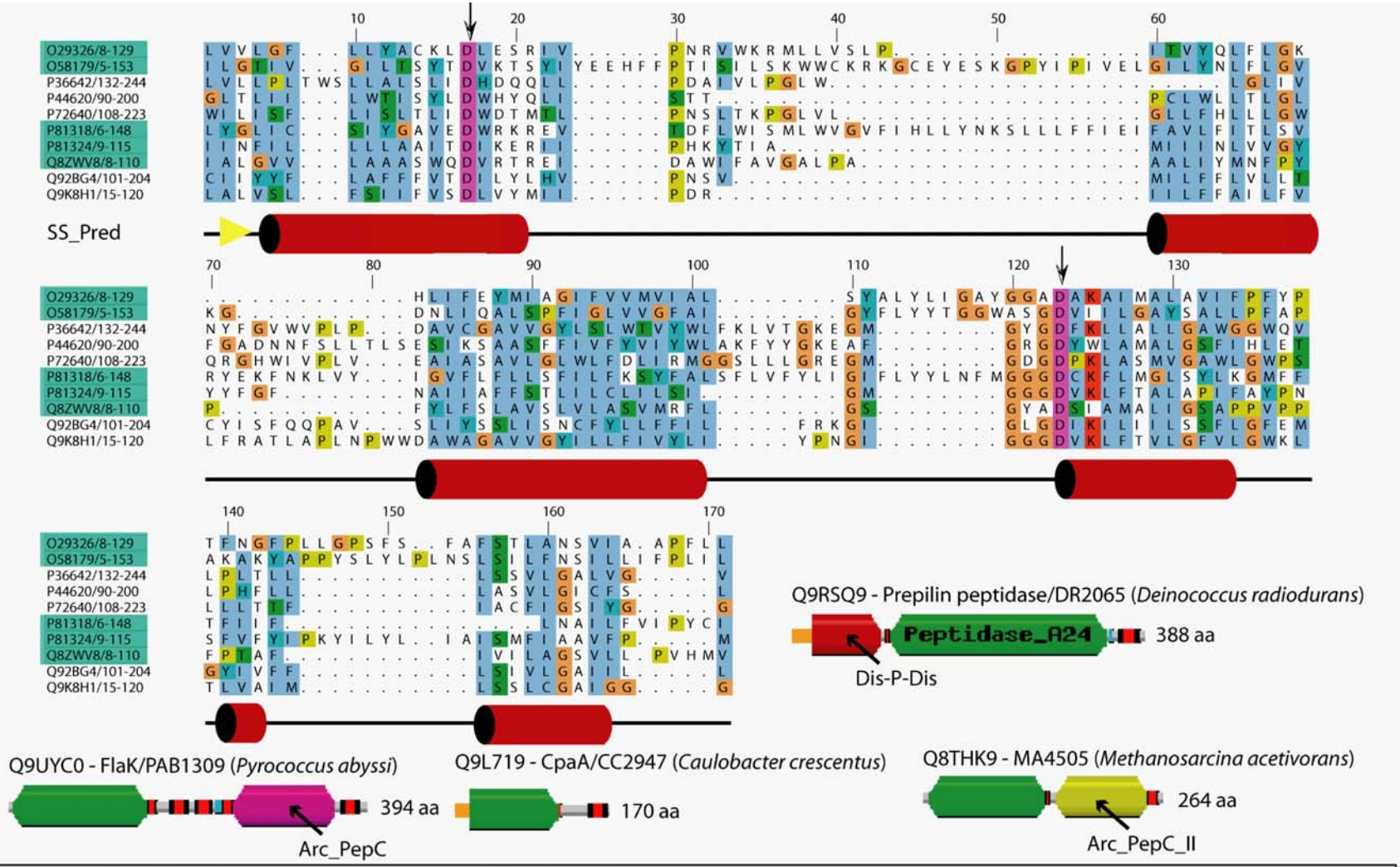

Figure 4.11: Peptidase_A24 example alignment and domain architectures
In the alignment the archaeal sequences are highlighted with green boxes. As can be seen they appear to be structurally similar to the eubacterial versions. The black arrows above the alignment indicate the two active site apartates

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. Foe 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 Vibrionacaea. 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.