3 Multi-genome Domain Hunting

3.1 Rationale

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

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

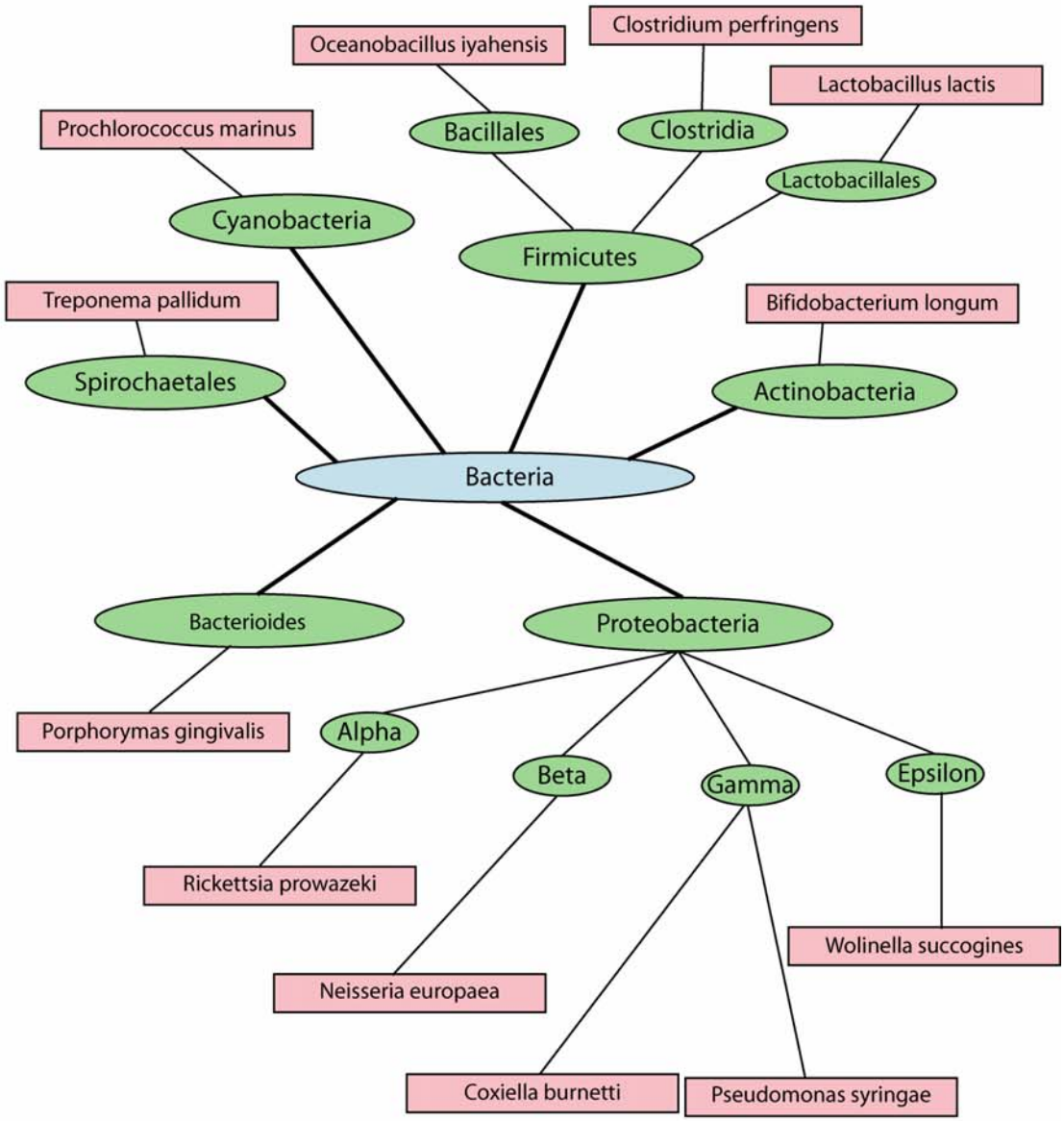

Figure 3.1: Simplified taxonomic tree of bacteria investigated in the multigenome hunt
The bacterial species investigated in the multigenome hunt are indictaed in the pink boxes on the leaves of the tree; the species groups are shown in the green ovals. When selecting these species an attempt was made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

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

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

3.2 Results

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

3.2.1 Summary of Results

Repeat Identification

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

Summary of Small Protein Clustering Results

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

3.2.2 Table of All Novel Domains and Families Identified

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

3.3 Descriptions of Novel Domains

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

Pfam Accession No	*Family Name*	*Pfam Type*	*Basic Function*	*No of copies in UniProt 44/27*	*Antibiotic biosynthesis*	*Spore Coat Formation*	*Cell Wall / Periplasm*	*Replication*	*Secreted*
A) Novel Families									
PF03413	PepSY	Domain	M4 Peptidase Inhibitor			X	X		X
PF03958	Secretin_N	Domain	Secretin N-terminal Domain				X		
PF07494	Reg_prop	Repeat	Regulatory β-propellor				X		
PF07495	Y_Y_Y	Motif	Unknown Regulatory Function				X		
PF07503	zf-HYPF	Domain	HypF-type Zinc Finger Domain						X
PF07550	DUF1533	Family	Unknown function						
PF07551	DUF1534	Family	Unknown function						
PF07552	Coat_X	Domain	Bacillales Coat X Domain			X	X		
PF07553	DUF1535	Domain	Unknown function				X		X
PF07554	FIVAR	Domain	NAG-binding Domain				X		X
PF07556	DUF1538	Family	Unknown function						
PF07559	FlaE	Domain	Flagellar Hook Protein Domain				X		X
PF07560	DUF1539	Family	Unknown function				X		
PF07561	DUF1540	Family	Unknown function						
PF07563	DUF1541	Family	Unknown function						
PF07577	DUF1547	Family	Unknown function						
PF07578	LAB_N	Family	Lipid A Biosynthesis N-terminal Domain				X		X
PF07581	Glug	Repeat	Short Cell Surface Repeat				X		X
PF07613	DUF1576	Family	Unknown Function				X		X
PF07615	Ykof	Family	Unknown Function						
PF07634	RtxA	Repeat	RTX toxin Repeat				X		
PF07655	Secretin_N_2	Domain	Secretin N-terminal Domain				X		
PF07660	STN	Domain	Secretin N-terminal Domain				X		
PF07670	Gate	Domain	Nucleoside Recognition				X		
PF07671	DUF1601	Family	Unknown function						
PF07675	Cleaved_adhesin	Family	RgpA-Kgp complex component				X		X
PF07853	CTnDOT_TraJ	Domain	Conjugative Transfer Protein J Domain				X		X
PF07862	Nif11	Domain	Nif11 Nitrogen Fixation Domain						
PF07865	DUF1652	Family	Unknown function						
PF07866	DUF1653	Family	Unknown function				X		
PF07867	DUF1654	Family	Unknown function						
PF07868	DUF1655	Family	Unknown function						
PF07869	DUF1656	Family	Unknown function						X
PF07870	DUF1657	Family	Unknown function				X		X
PF07871	DUF1658	Family	Unknown function						
PF07872	DUF1659	Family	Unknown function						
PF07873	YabP	Family	Unknown function						
PF07874	DUF1660	Family	Prophage Protein of Unknown Function						
PF07875	Coat_F	Domain	Bacillales Coat F Domain				X		
PF07876	Dabb	Domain	Stress-responsive Dimeric α/β Barrel				X		X
PF07877	DUF1661	Family	Unknown function						
PF07878	DUF1662	Family	Unknown function						

Table 3.1: Novel domains found in the multigenome hunt

3.3.1 Domains Identified Through Repeats

PepSY (PF03413 – M4 Peptidase inhibitory domain)

This domain is discussed in greater detail in chapter 4.3.

Gate (PF07670 – potential membrane channel specificity domain)

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

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

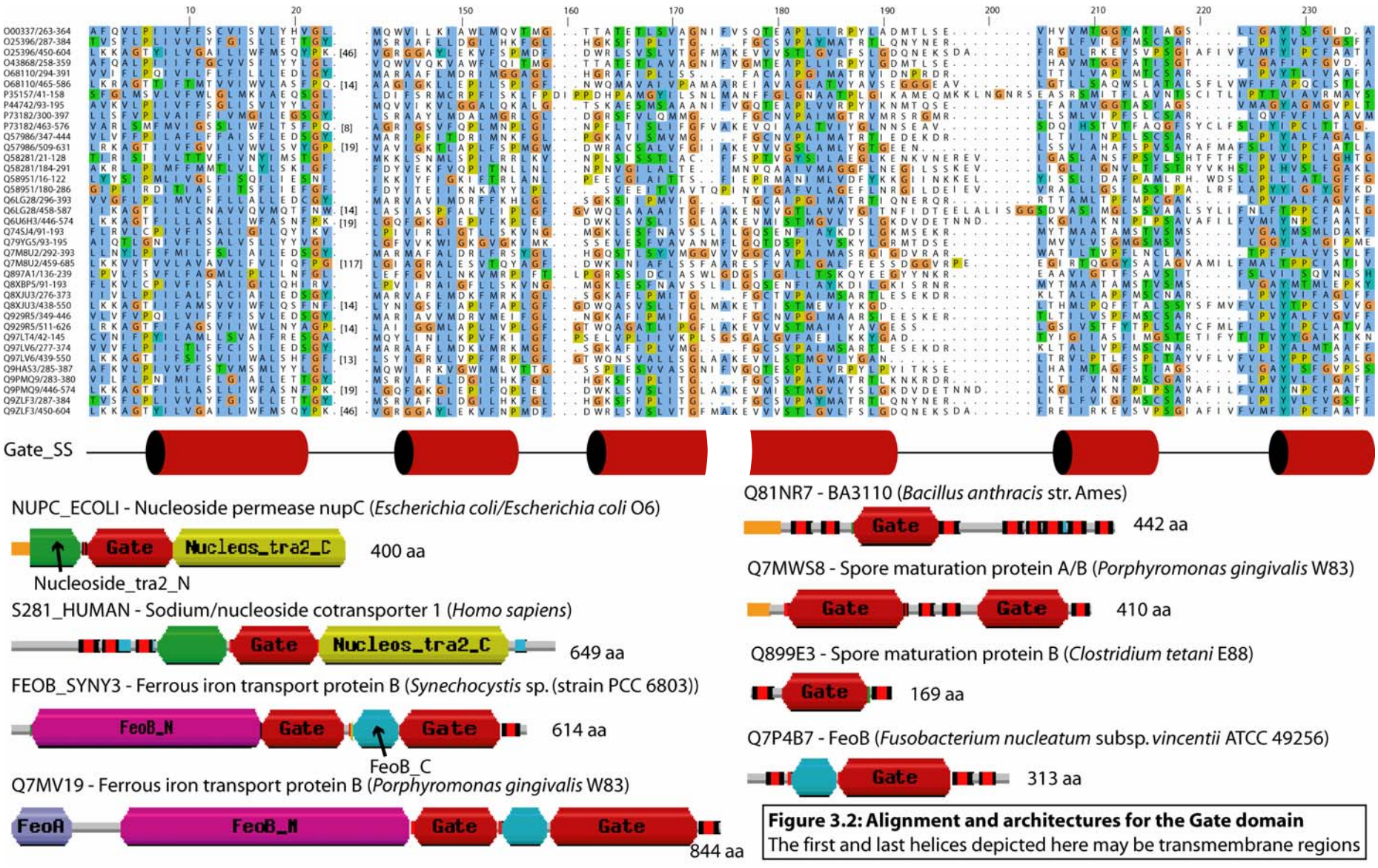

Figure 3.2: Alignment and architectures for the Gate domain
The first and last helices depicted here may be transmembrane regions

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

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

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

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

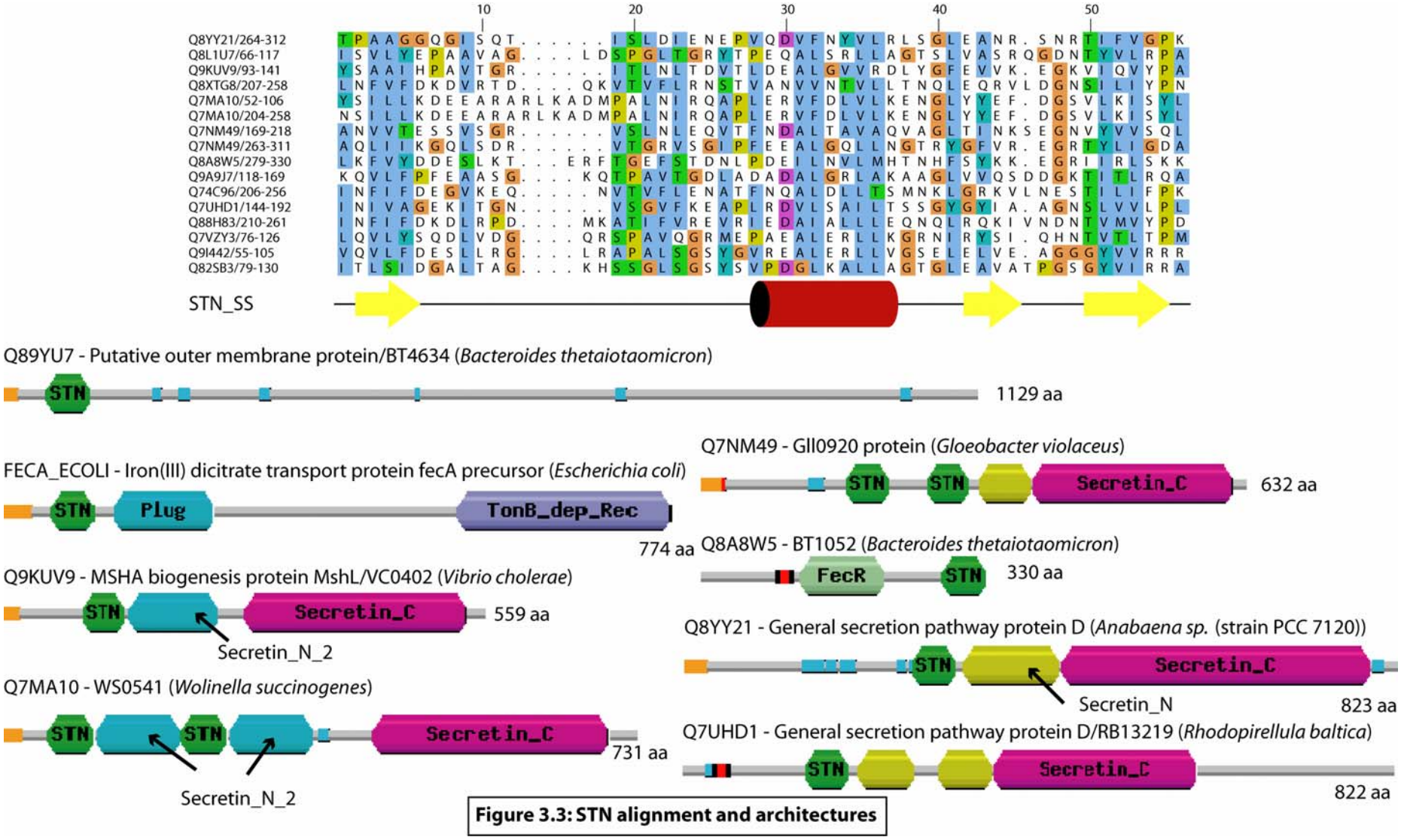

Figure 3.3: STN alignment and architectures

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

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

Secretin_N (Secretin N-terminal domain; PF03958)

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

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

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

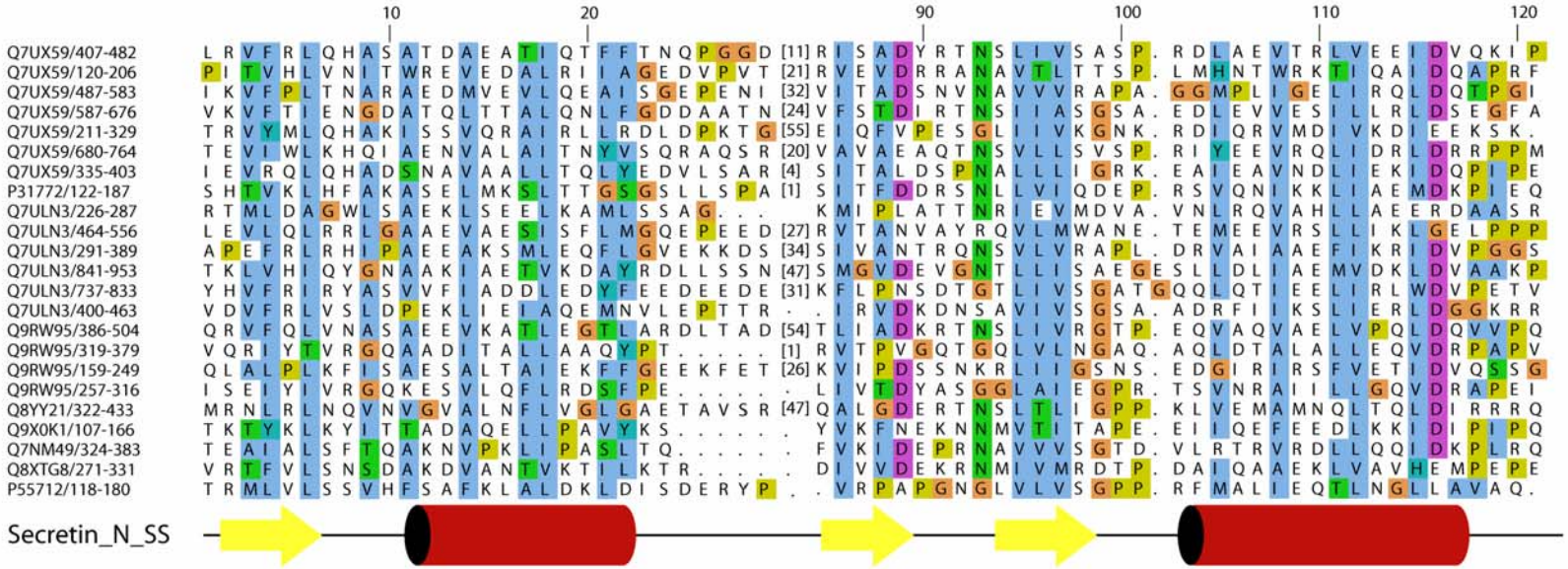

Figure 3.4: Secretin_N example alignment

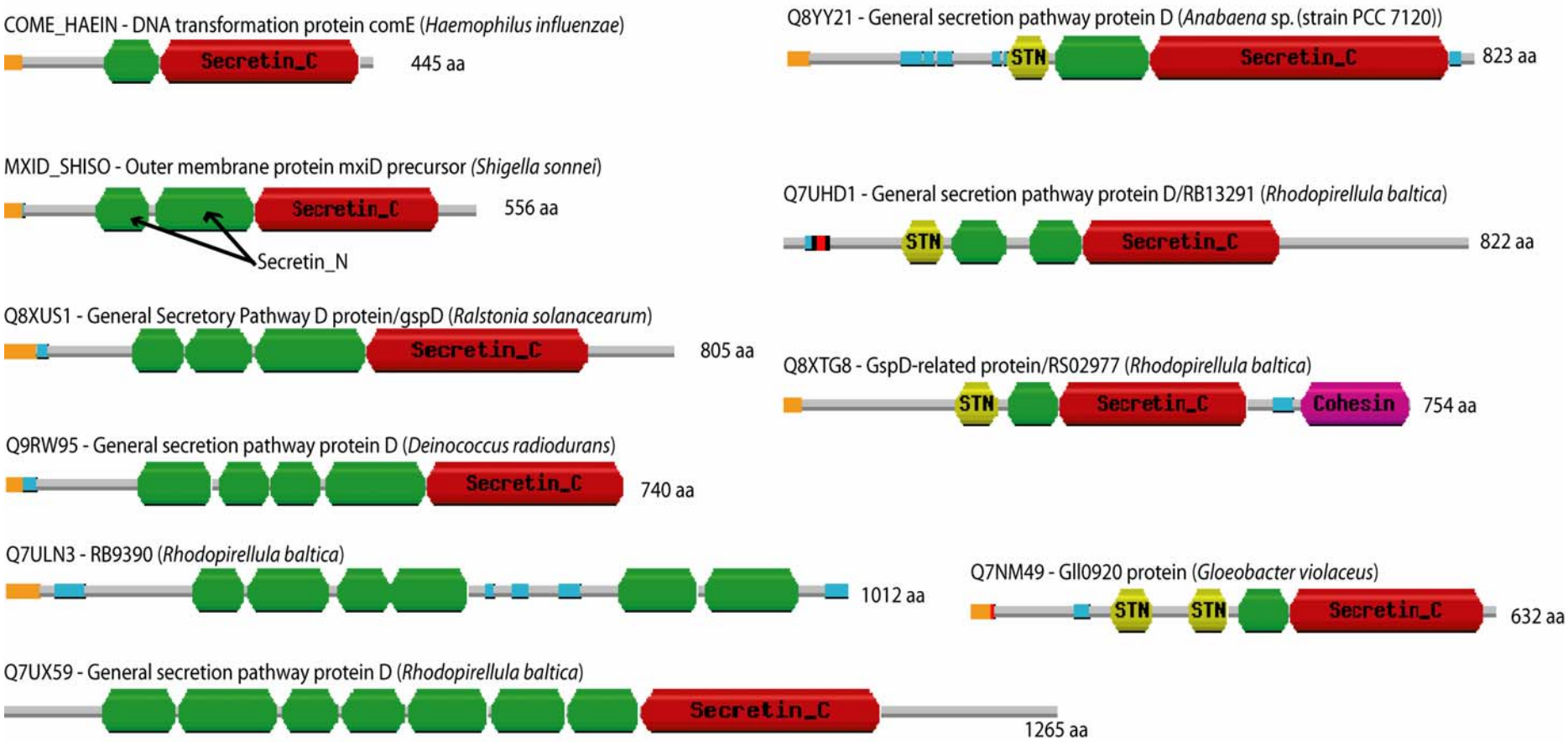

Figure 3.5: Secretin_N example architectures

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

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

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

Secretin_N_2 (Secretin N-terminal domain; PF07655)

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

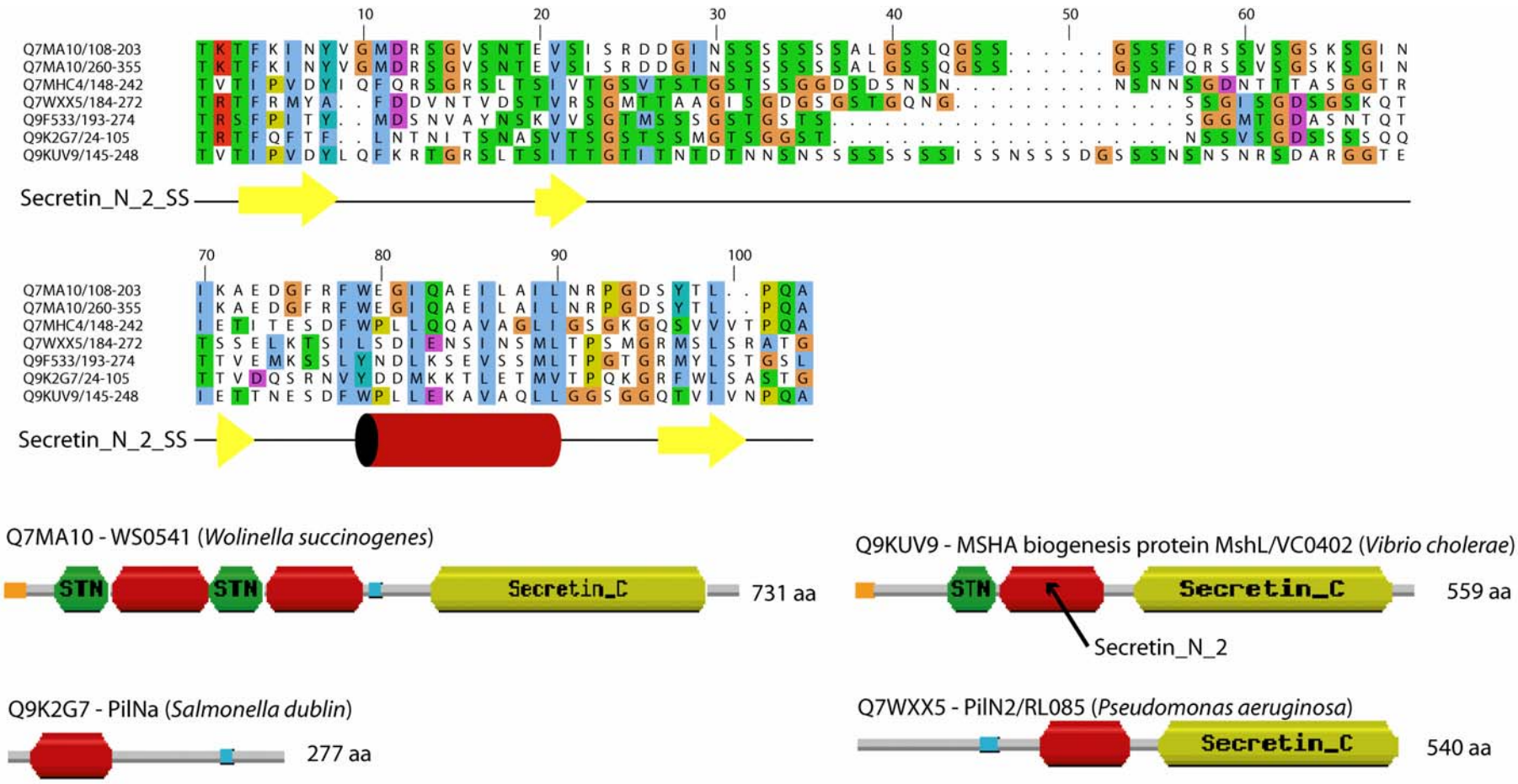

Figure 3.6: Secretin_N_2 alignment and architectures

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

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

Reg_prop (Regulatory Protein Propeller; PF07494)

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

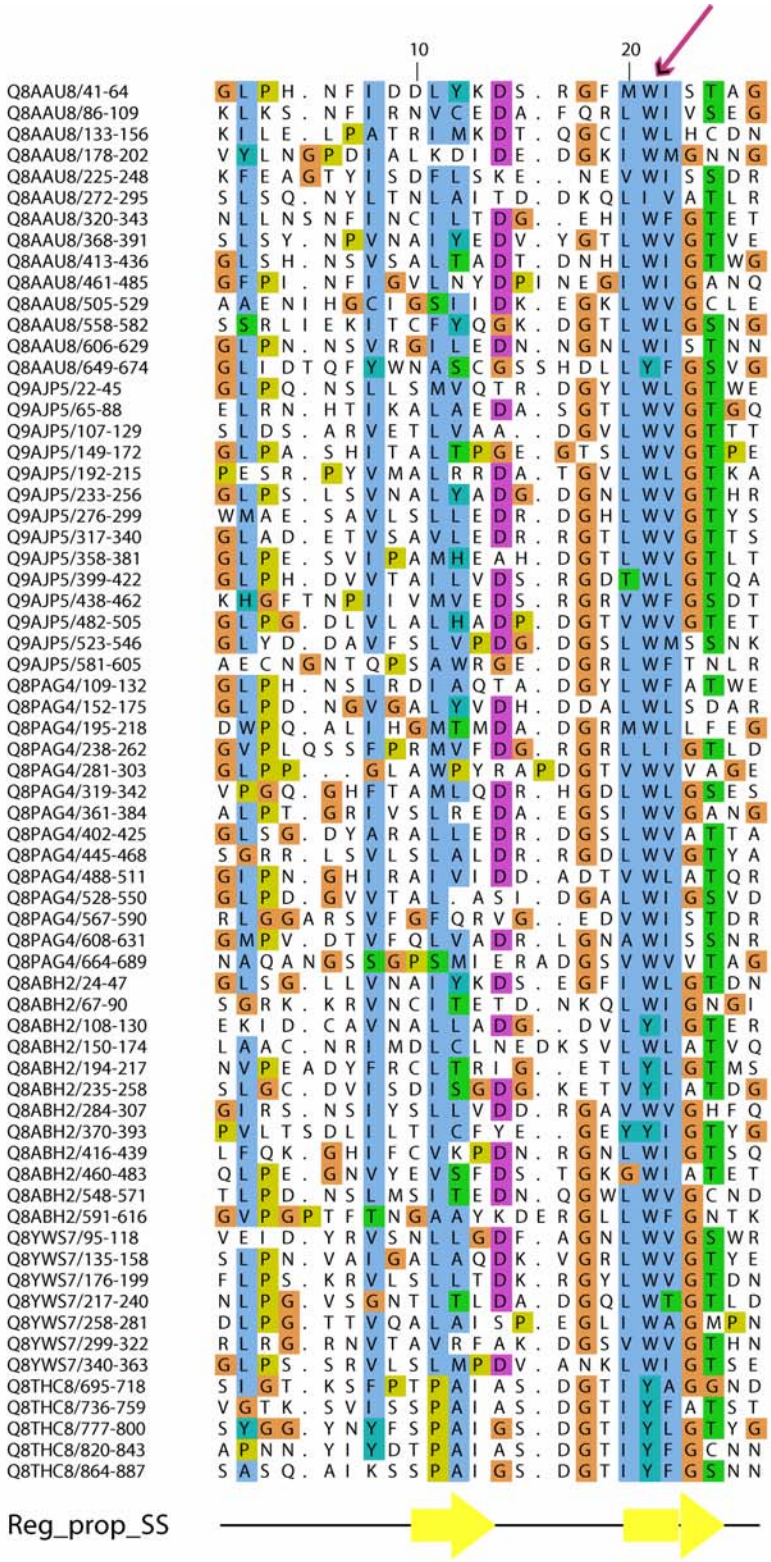

Figure 3.7: Example Reg_prop alignment
The purple arrow above the alignment marks a mostly invariant aromatic residue.

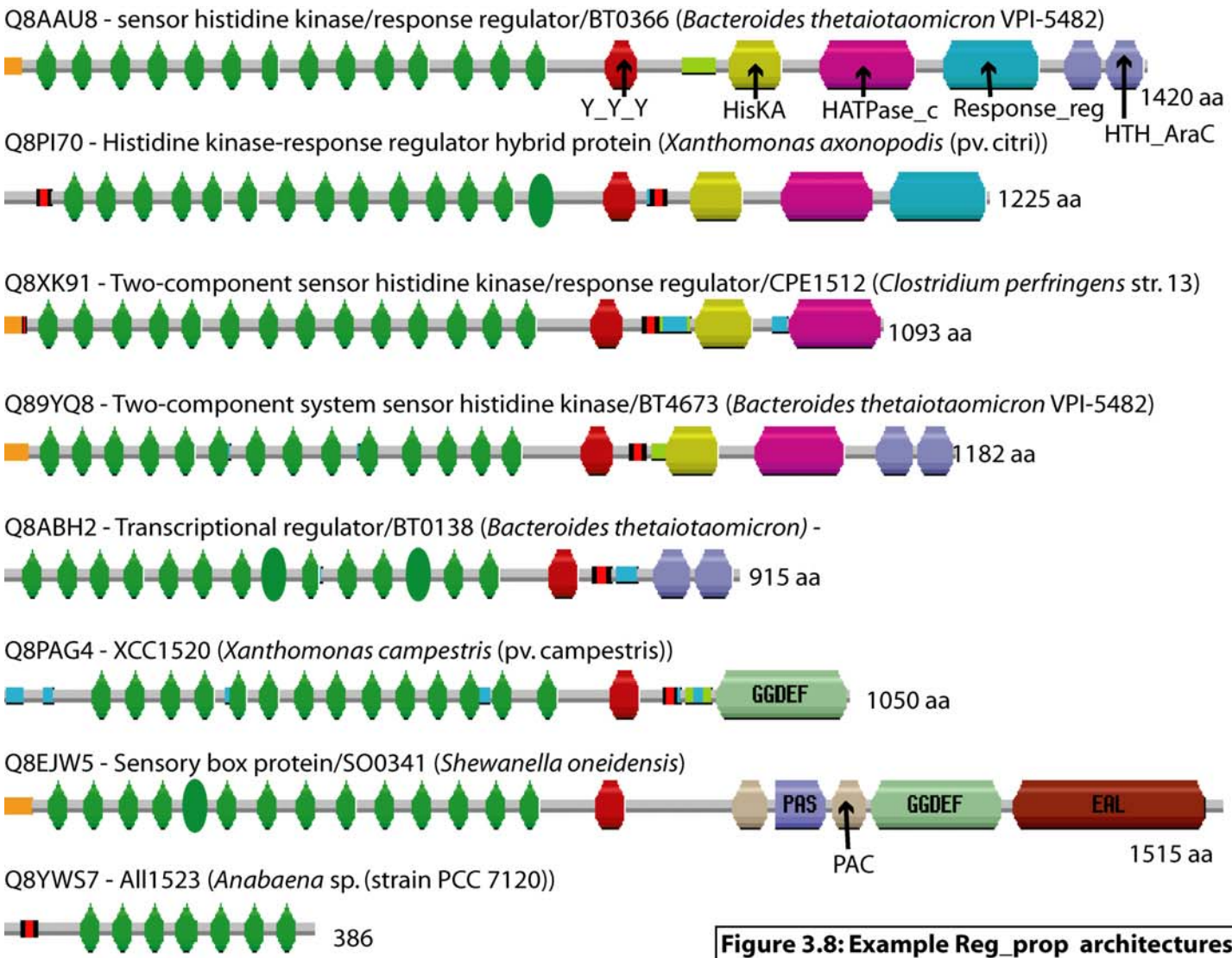

Figure 3.8: Example Reg_prop architectures

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

Y_Y_Y (Conserved Tyrosine Motif; PF07495)

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

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

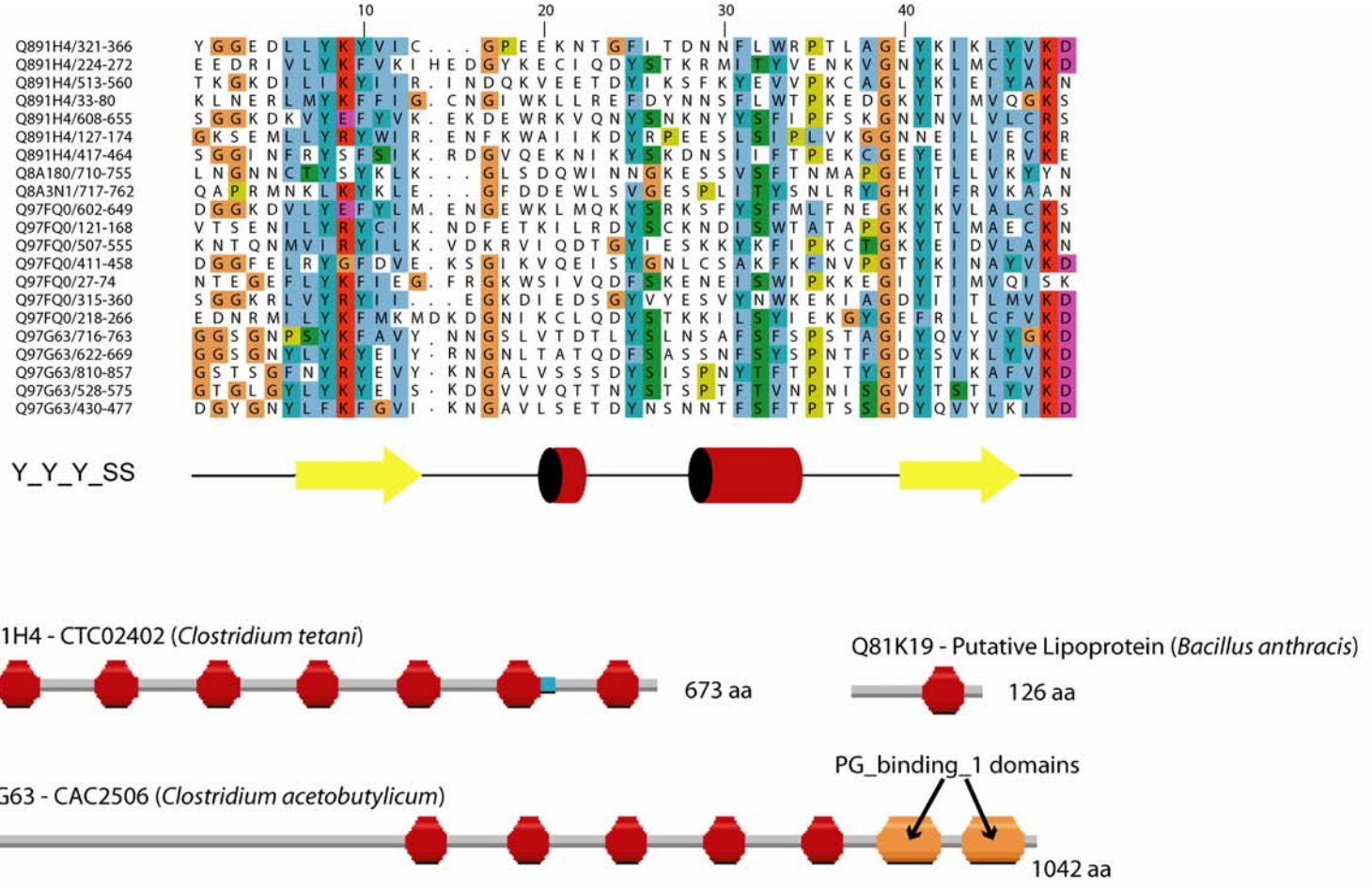

Figure 3.9: Y_Y_Y alignment and architectures

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

DUF1533 (Domain of Unknown Function 1533; PF07550)

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

Coat_X (Bacillus Coat Protein X domain; PF07552)

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

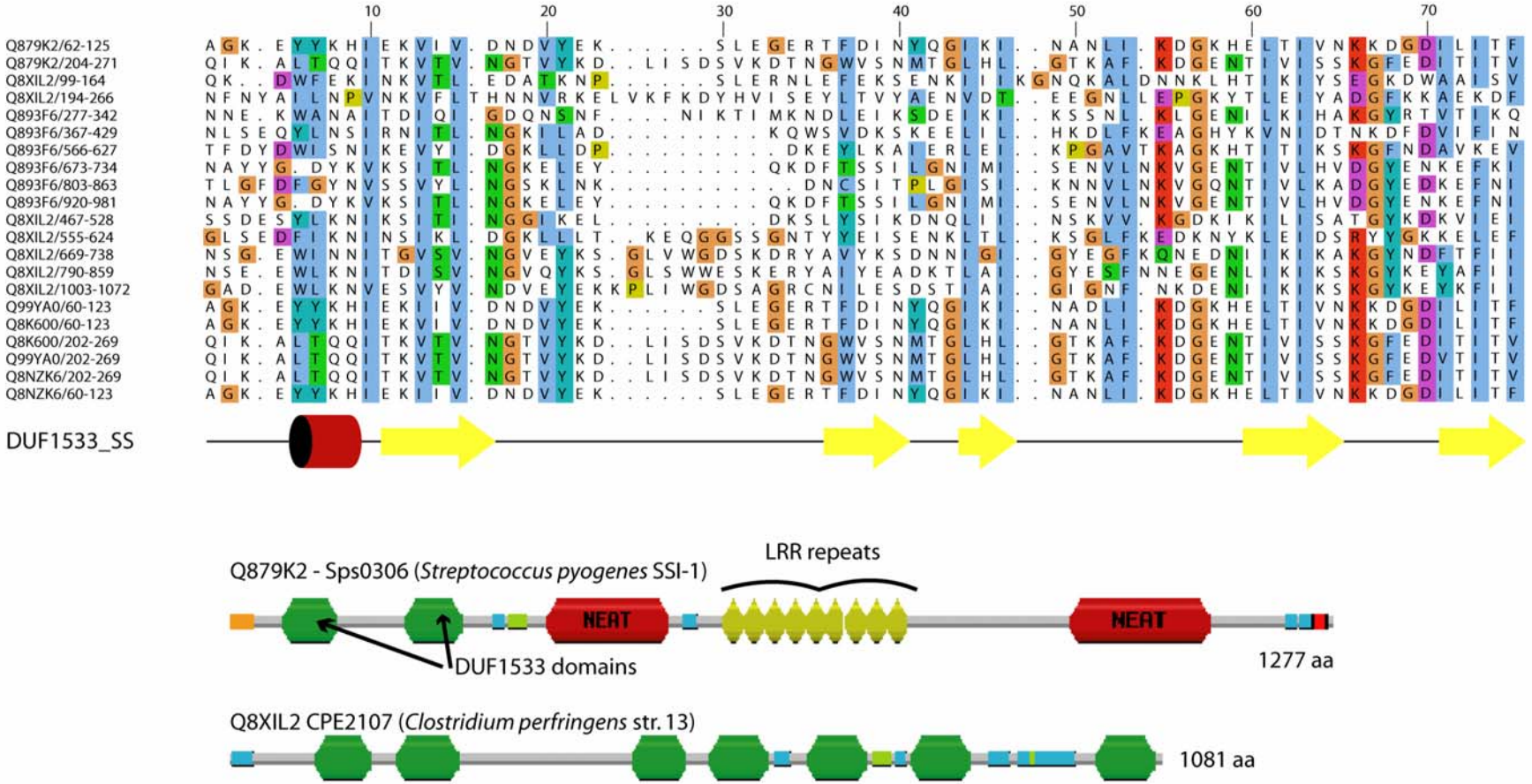

Figure 3.10: DUF1533 alignment and architectures

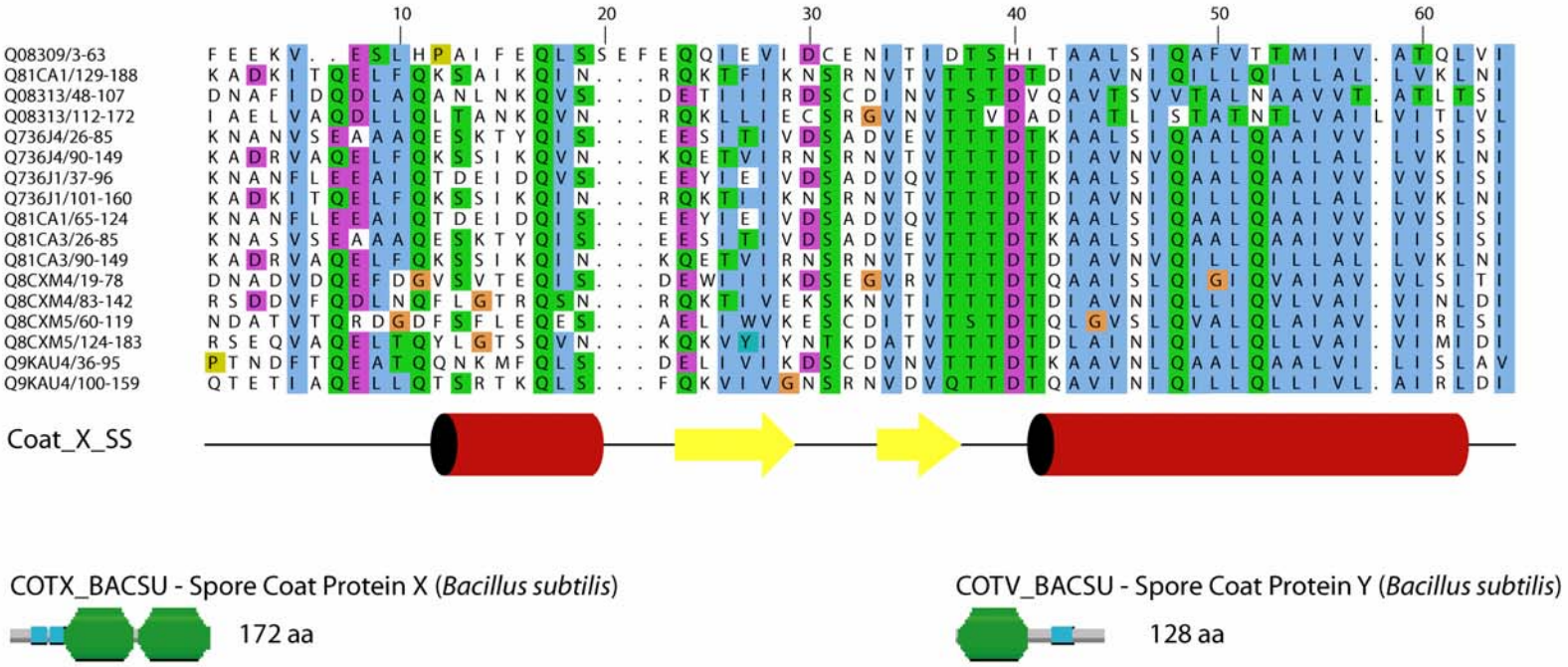

Figure 3.11: Coat_X alignment and architectures

Cleaved_Adhesin (PF07675)

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

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

FIVAR (likely NAG-binding motif; PF07554)

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

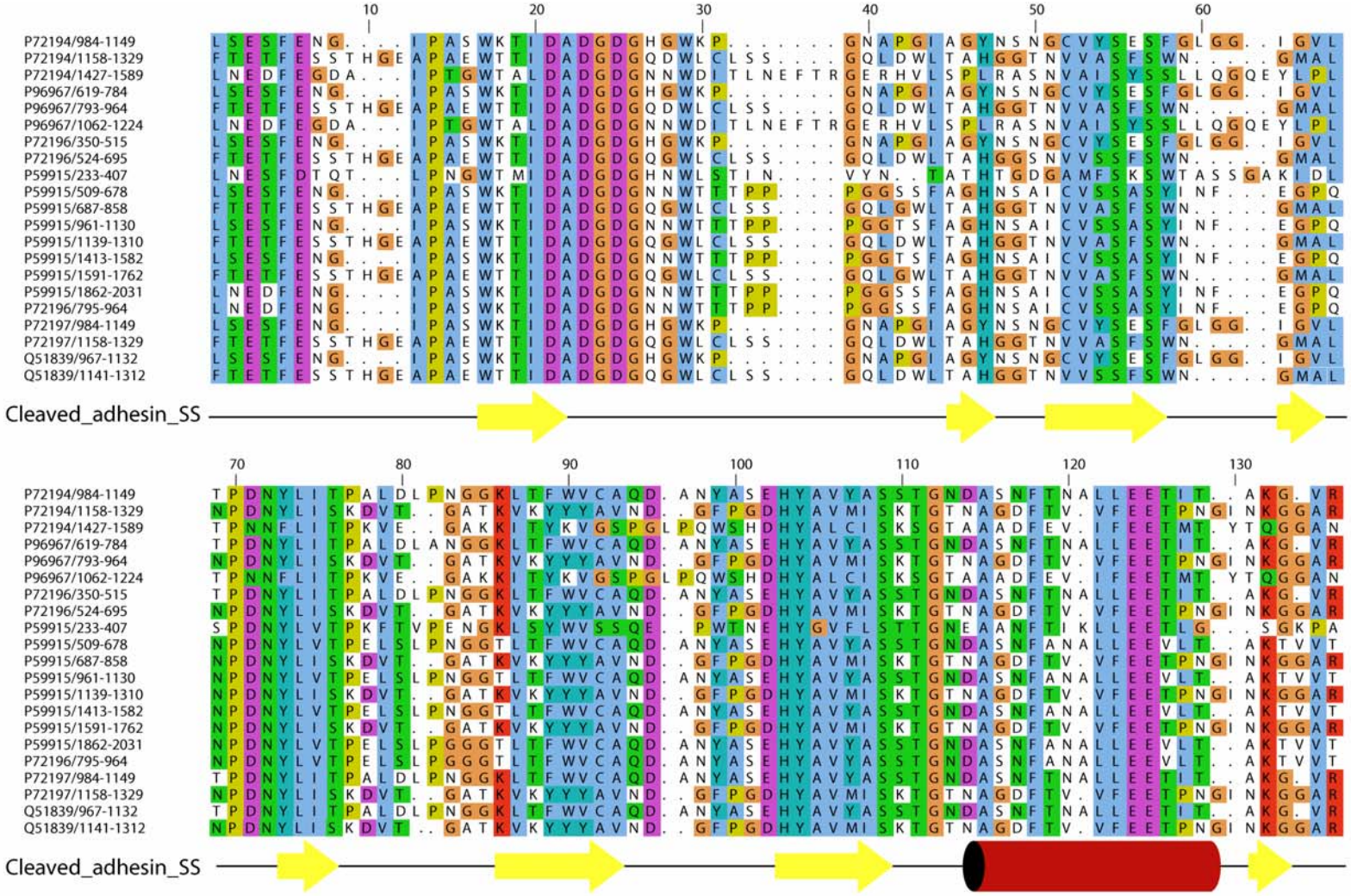

Figure 3.12: Cleaved_Adhesin alignment and architectures (Page 1)

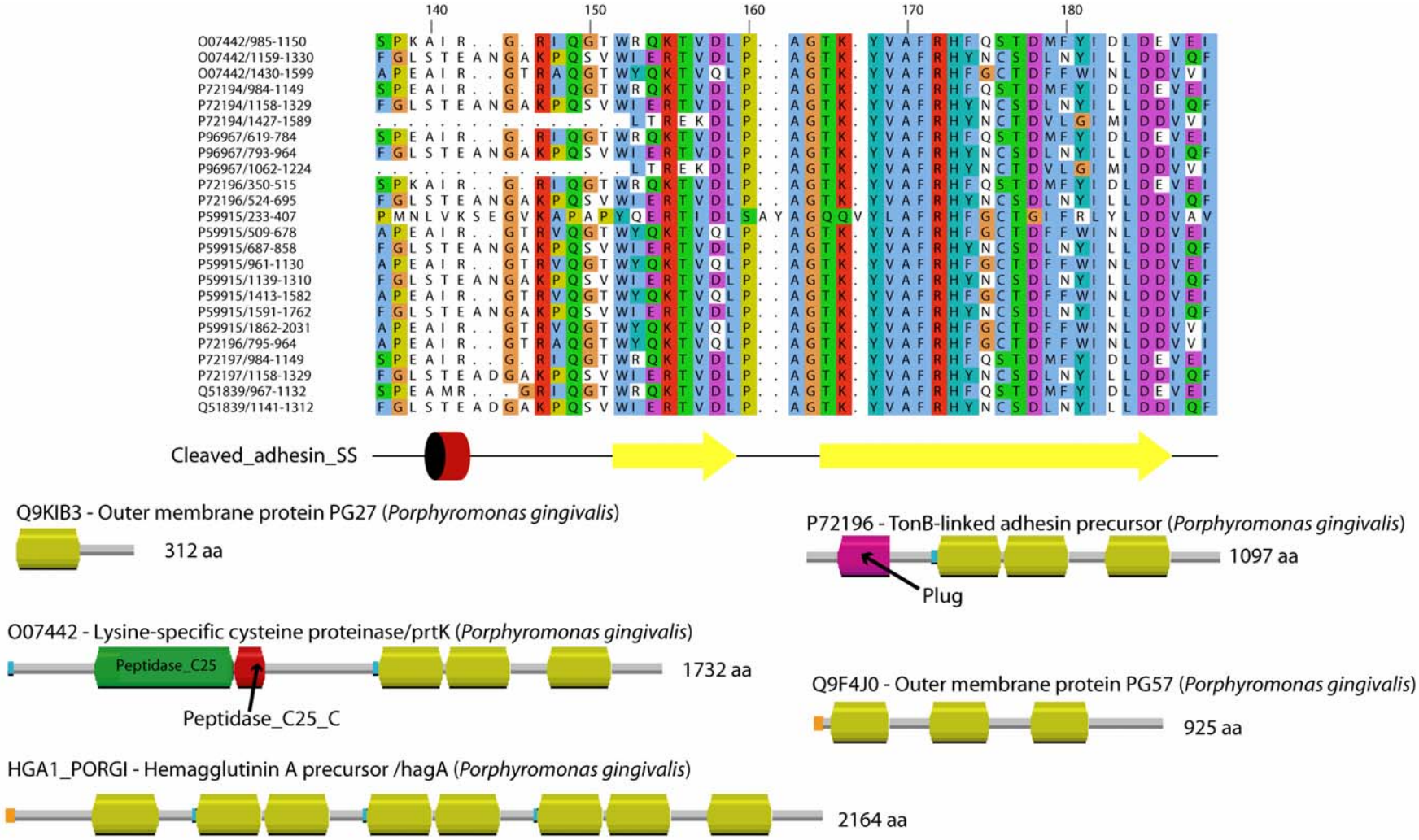

Figure 3.12: Cleaved_Adhesin example alignment and architectures (Page 2)

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

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

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

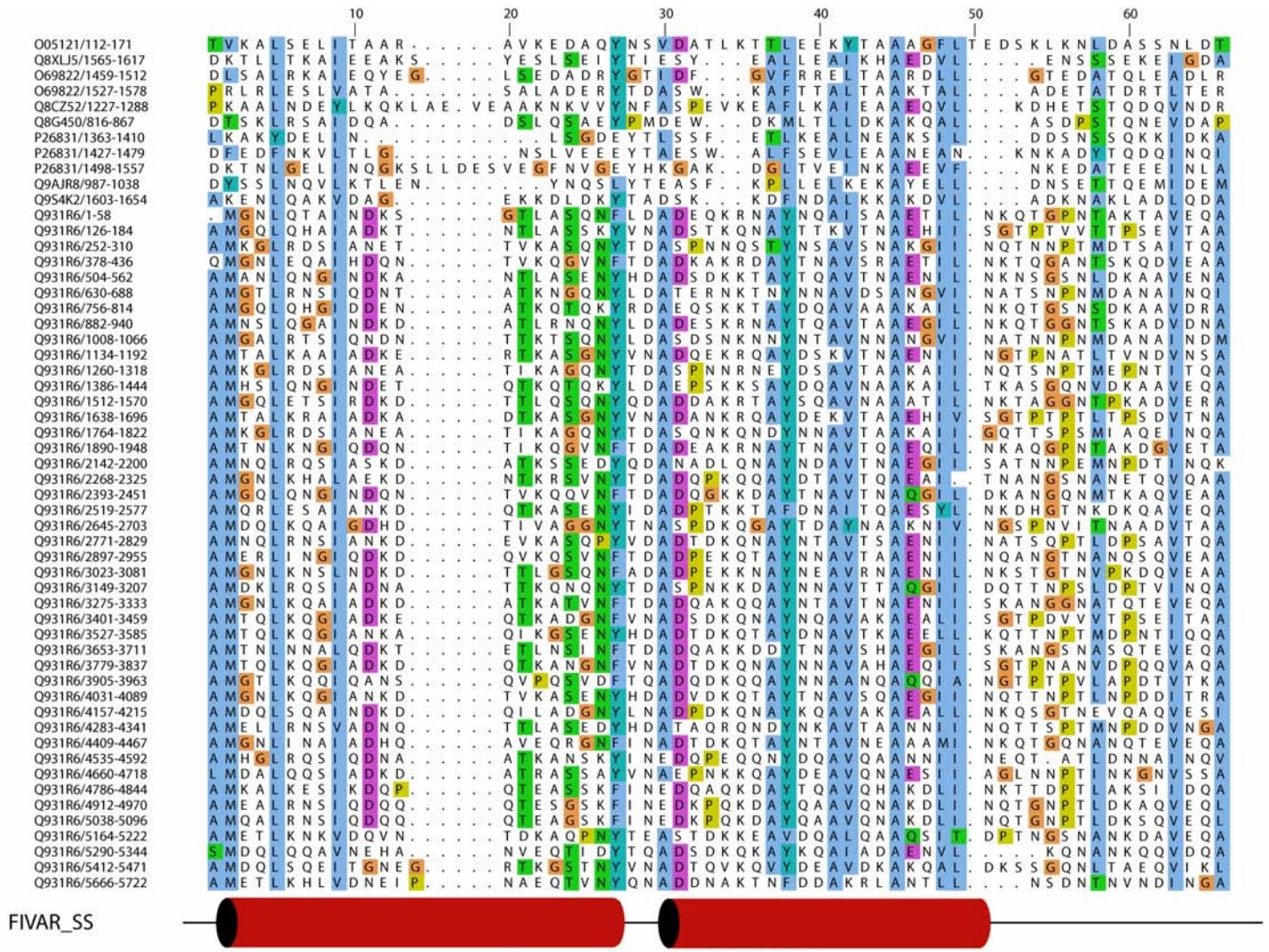

Figure 3.13: FIVAR alignment

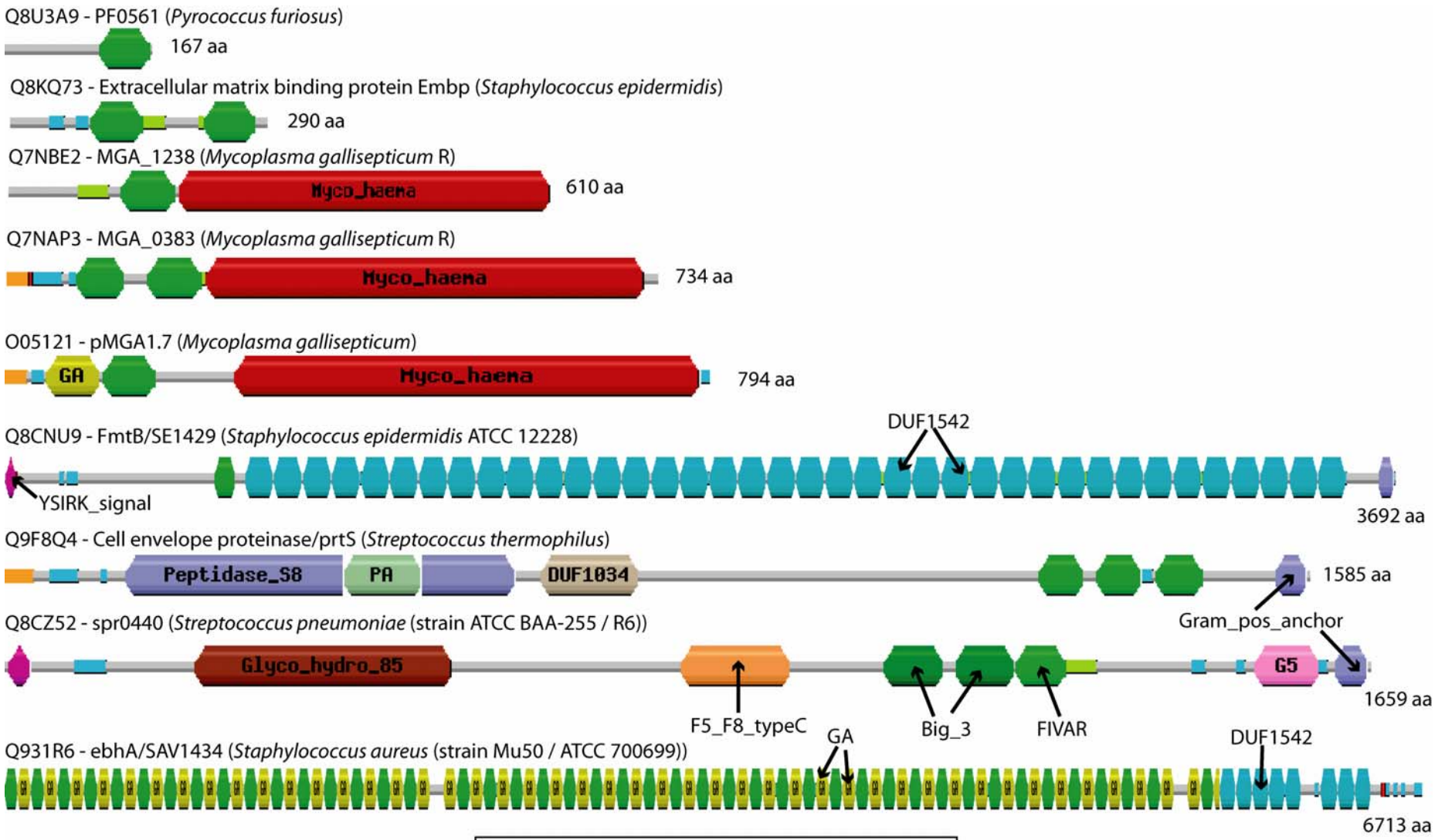

Figure 3.14: Example FIVAR Architectures (Page 1)

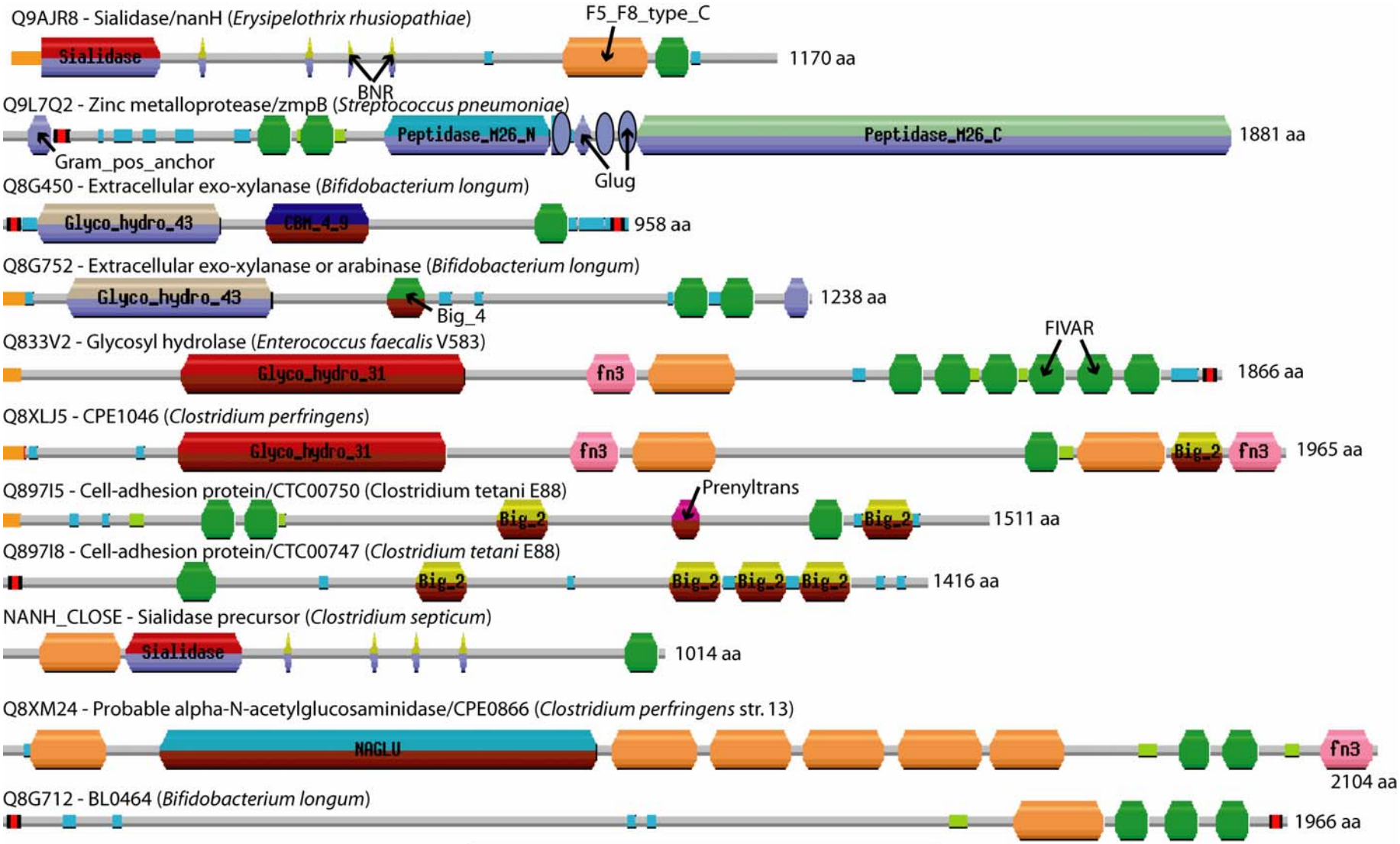

Figure 3.14: Example FIVAR Architectures (Page 2)

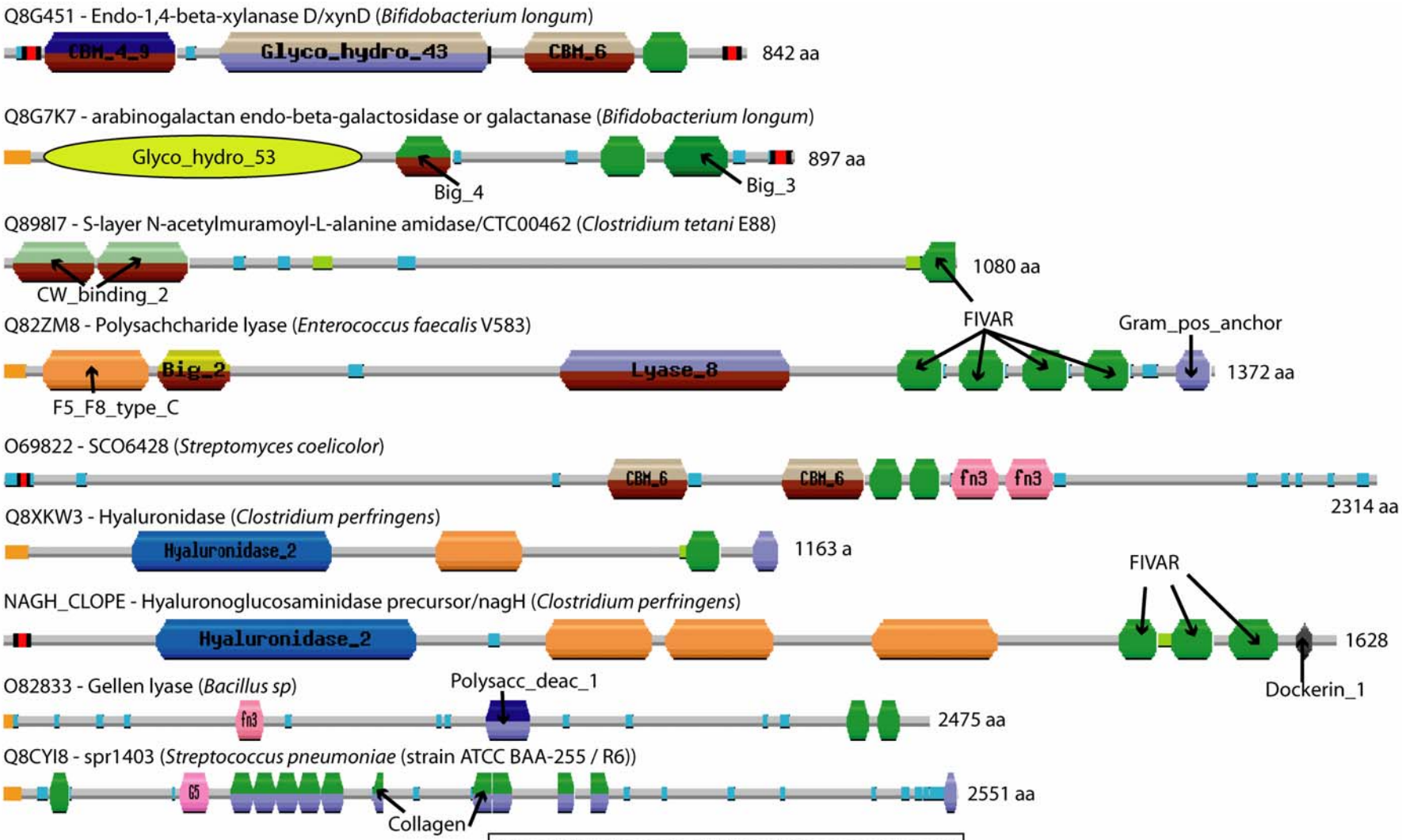

Figure 3.14: Example FIVAR architectures (Page 3)

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

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

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

FlaE (Flagella hook domain; PF07559)

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

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

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

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

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

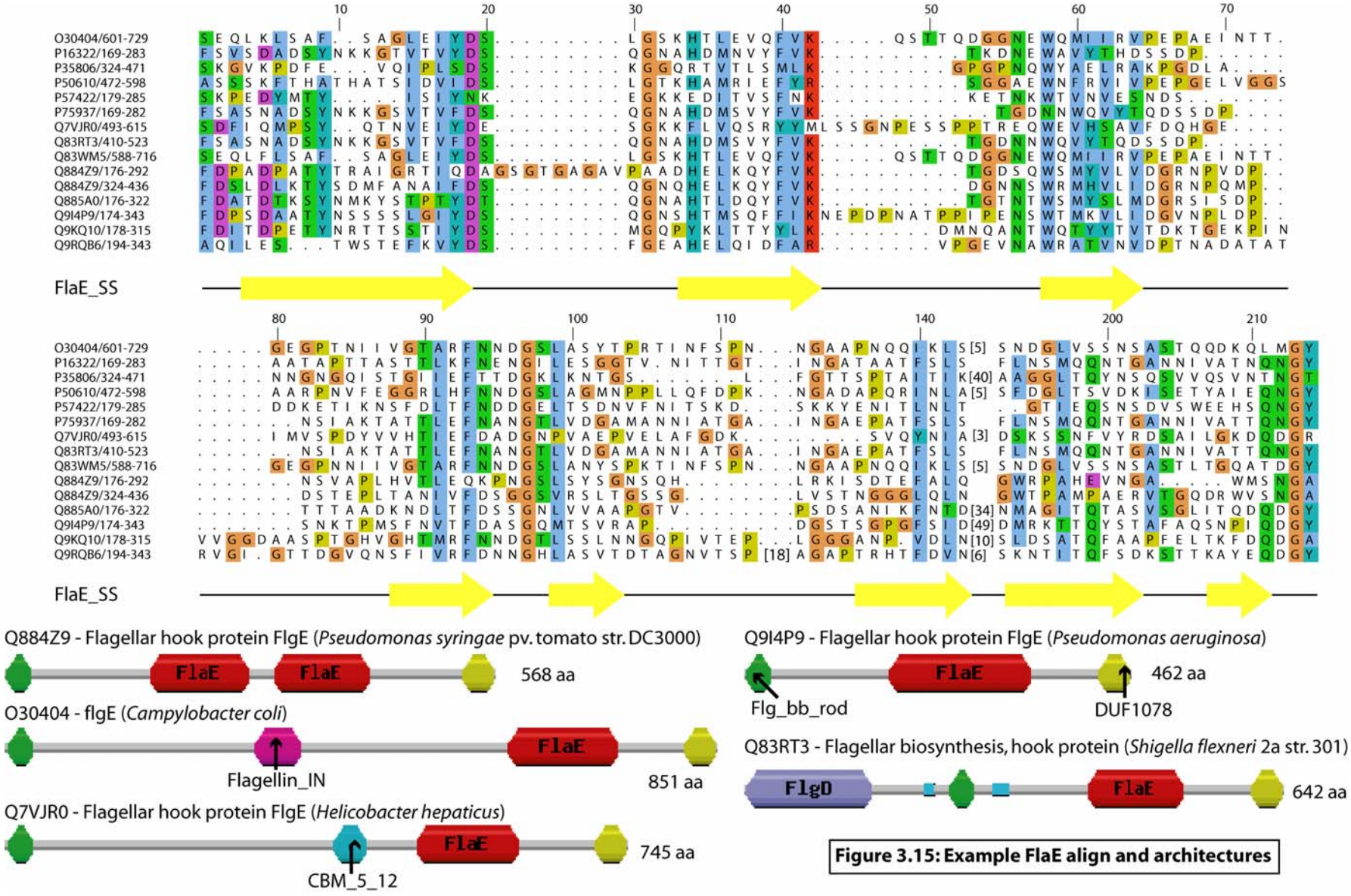

Figure 3.15: Example FlaE align and architectures

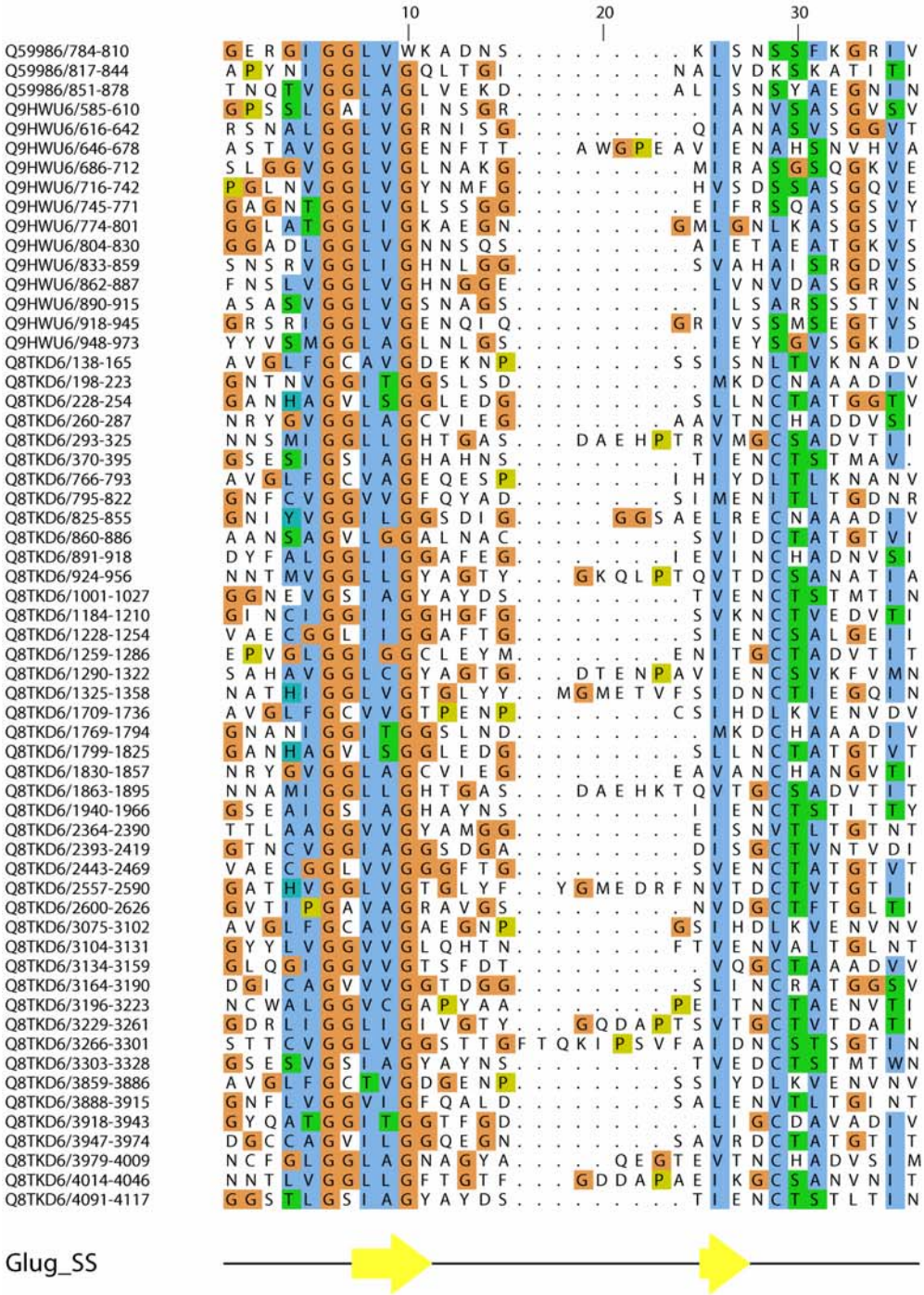

Figure 3.16: Example Glug repeat alignment

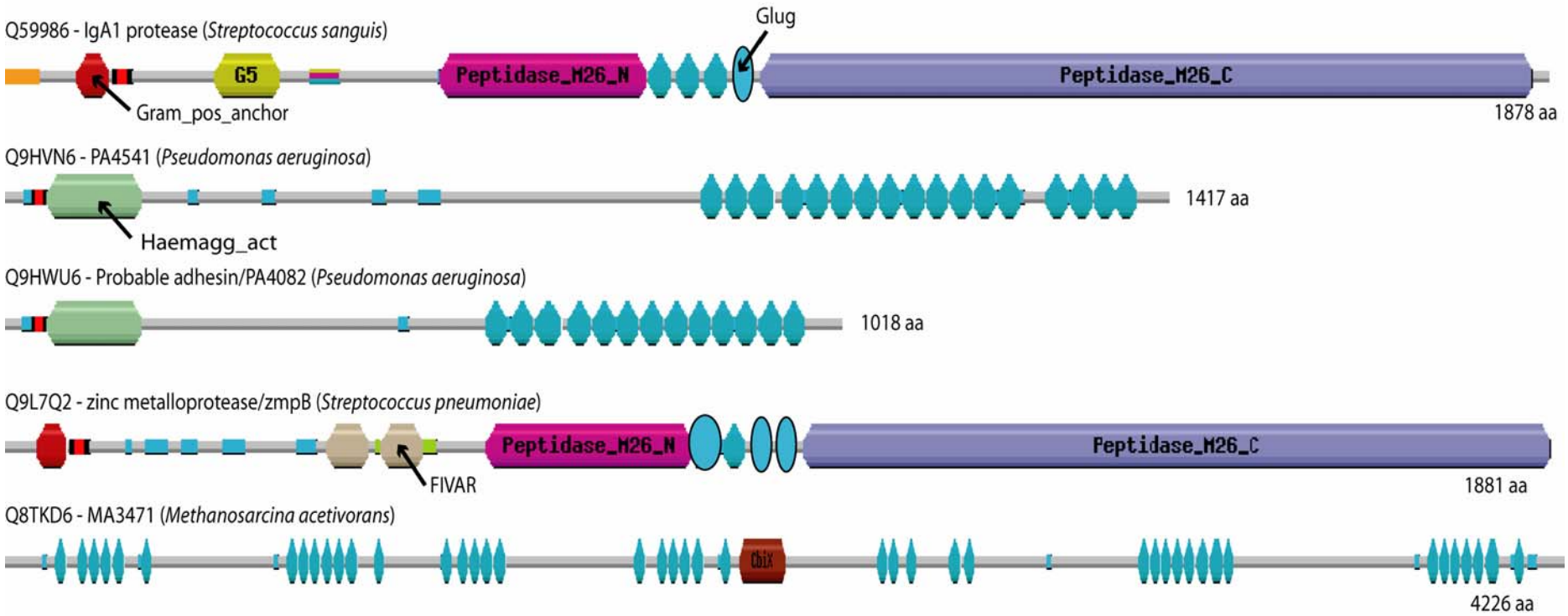

Figure 3.17: Example Glug repeat architectures

3.3.2 Domains Identified Through Protein Clustering

Coat_F (Coat protein F; PF07875)

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

CTnDOT_TraJ (Conjugative Transfer Protein J; PF07863)

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

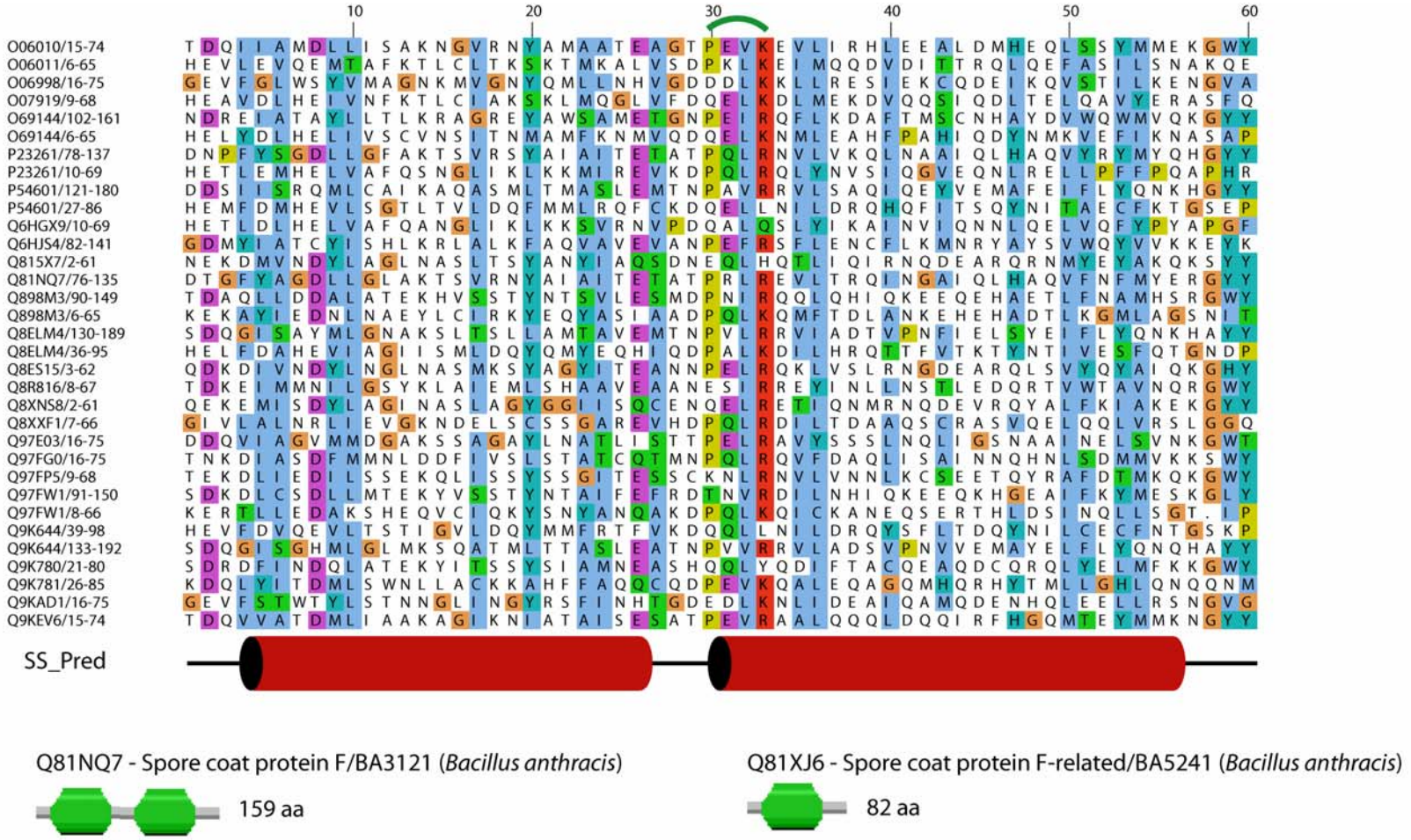

Figure 3.18: Example Coat_F alignment and architectures
The green bracket above the alignment marks the most conserved region of the family

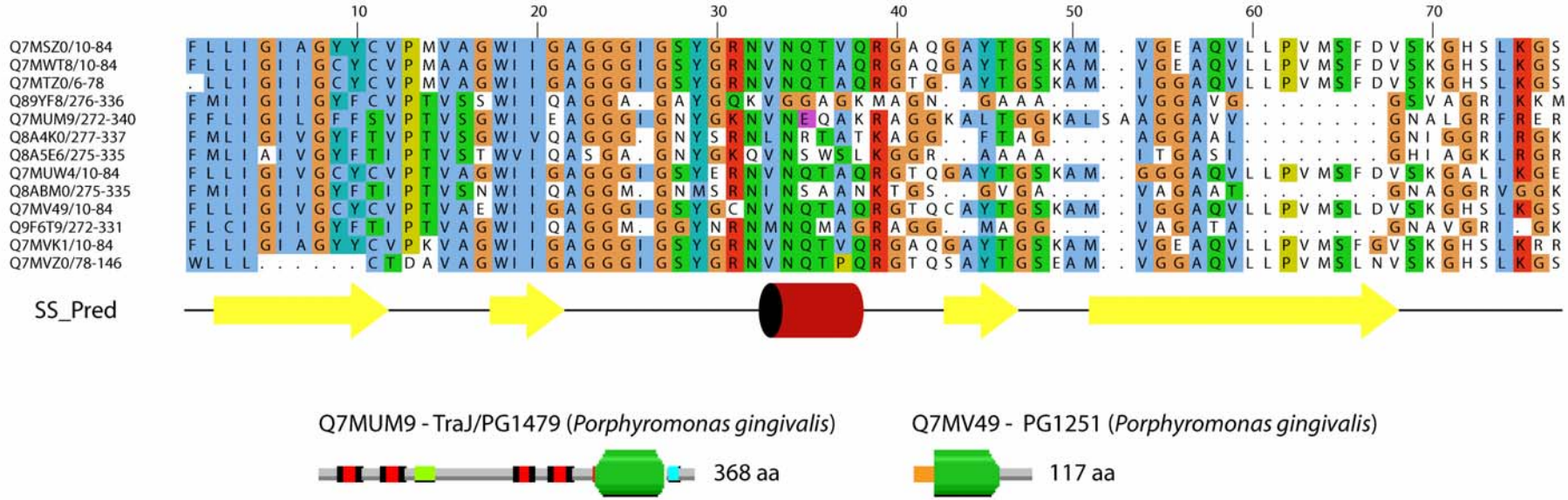

Figure 3.19: Example CTnDOT_TraJ alignment and architectures

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

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

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

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

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

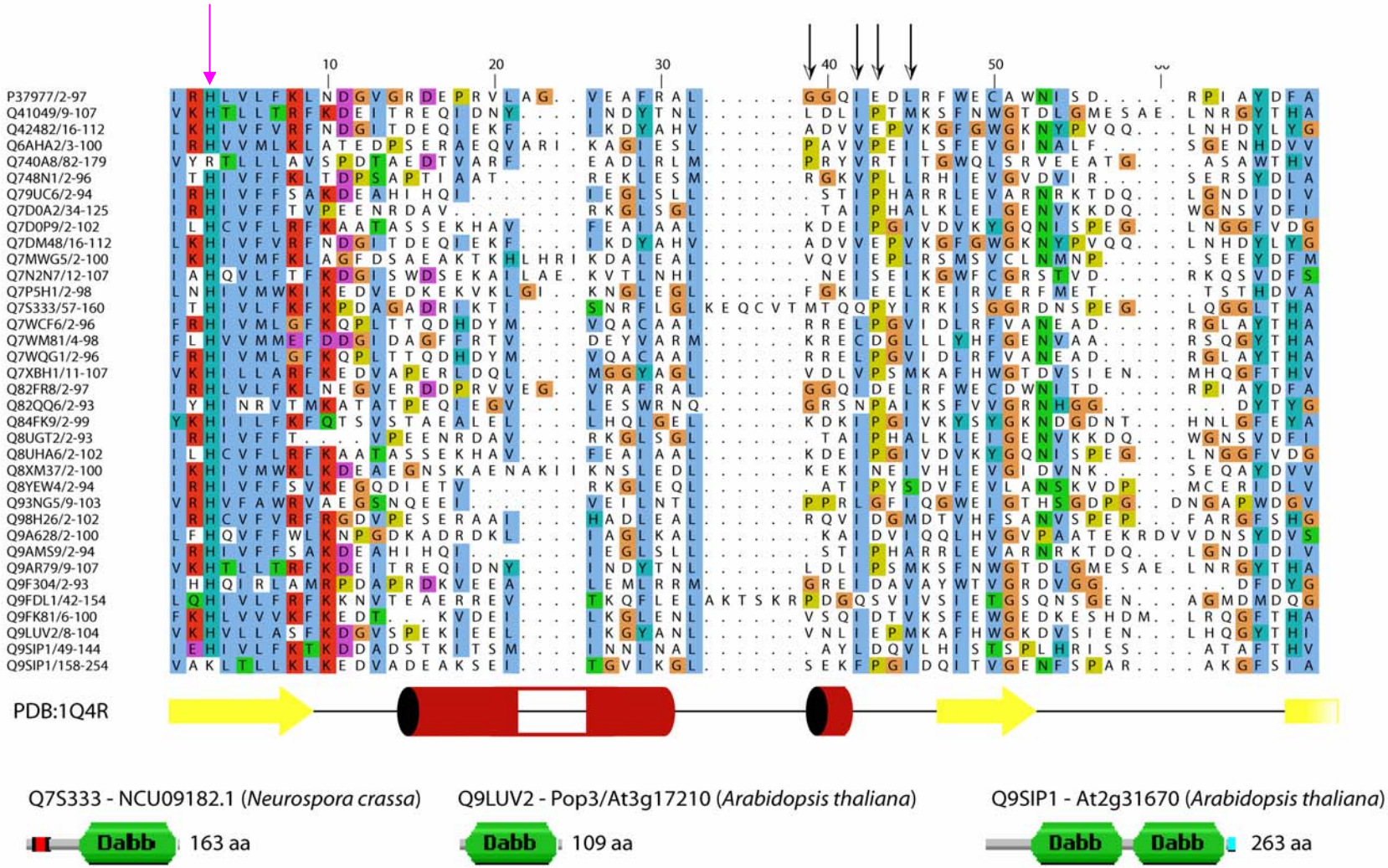

Figure 3.20: Example Dabb alignment and architectures (Page 1)
The residues that contact a magnesium ion in PDB1Q4R are marked by the black arrows above the alignment. PDB:1Q4R corresponds to UniProt:Q9LUV2

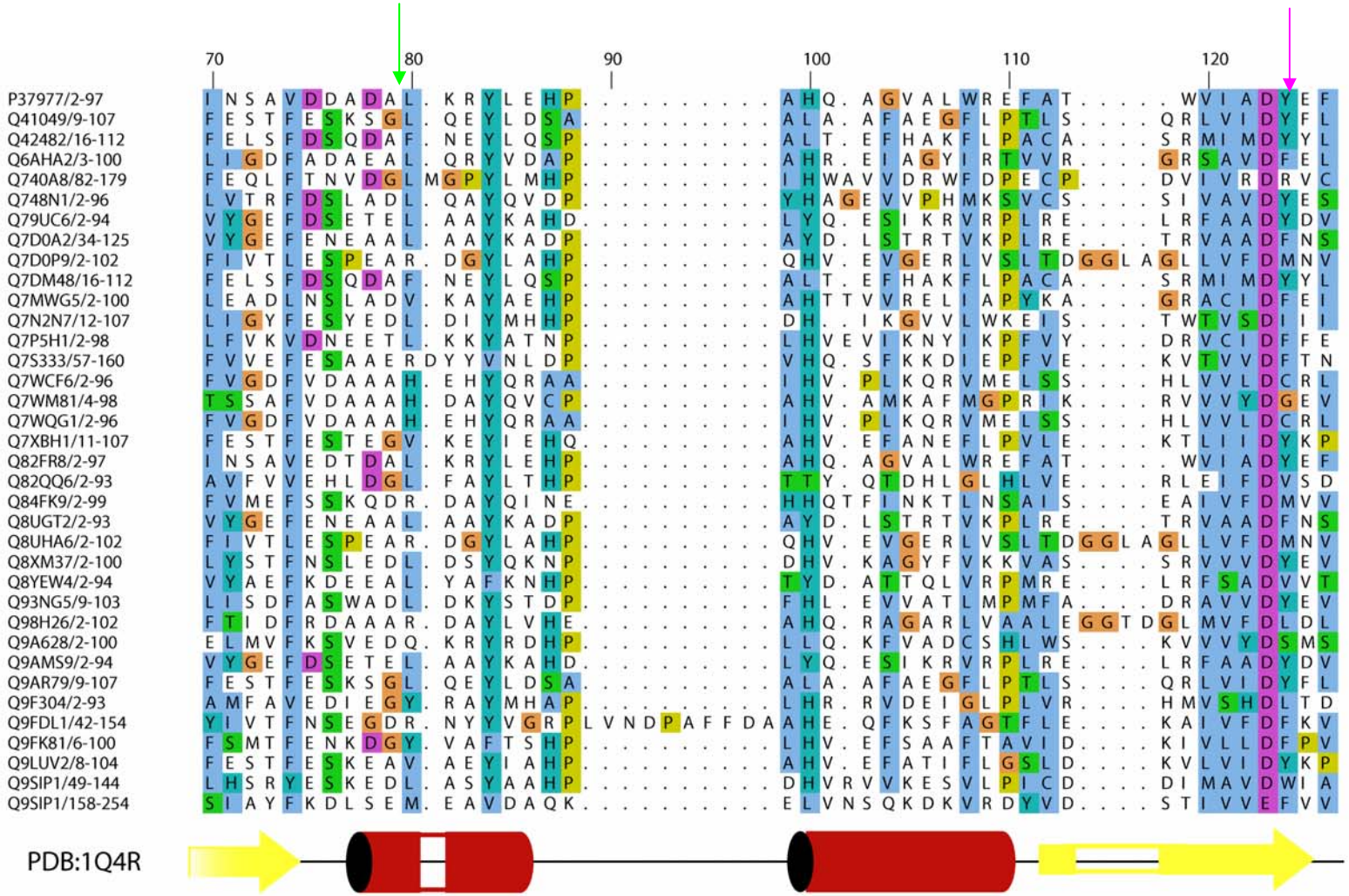

Figure 3.20: Example Dabb alignment and architectures (Page 2)

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

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

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

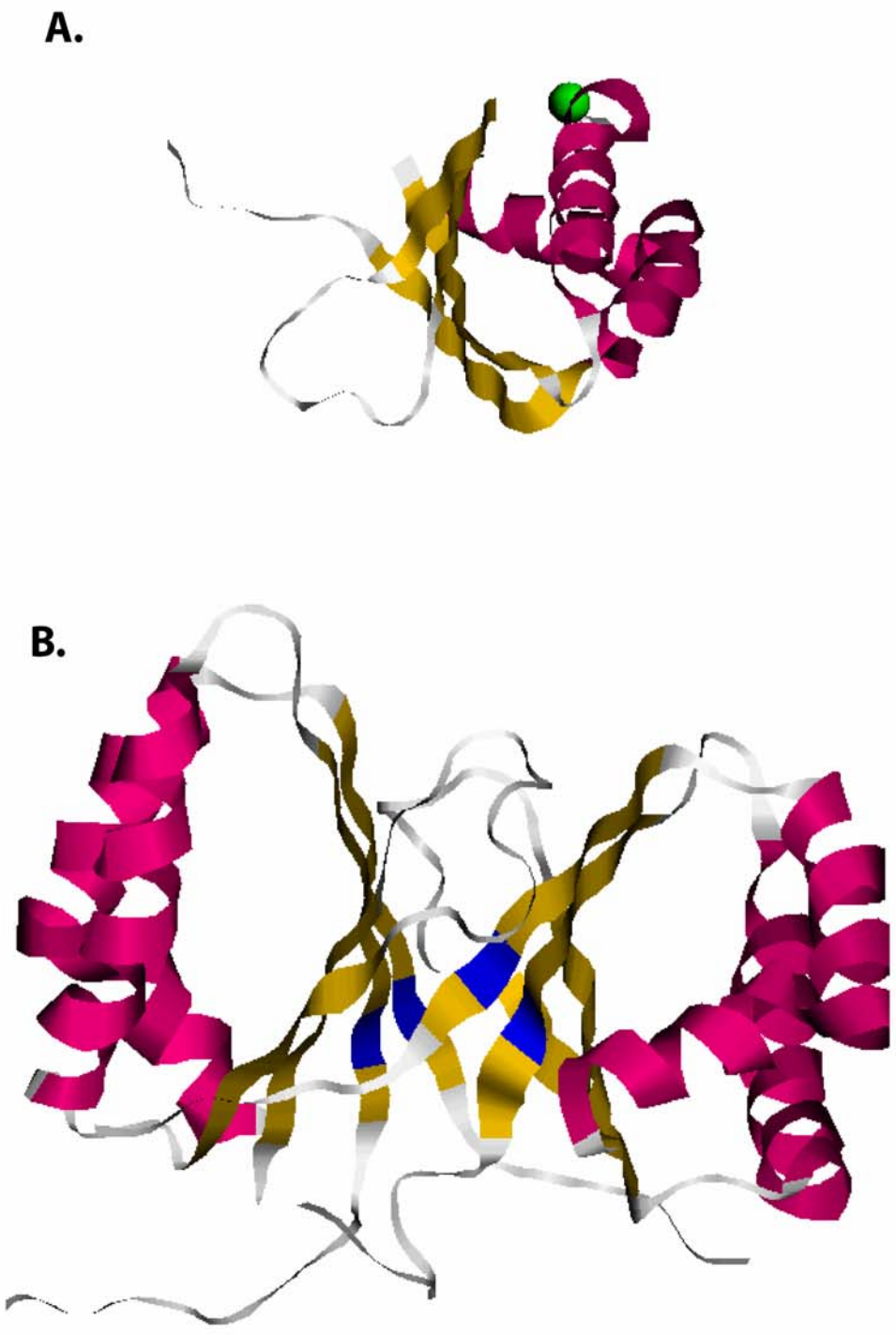

Figure 3.21: Structure of the Dabb dimeric barrel
Part (A) shows PDB:1Q4R. This image shows just a single Dabb protein, which only forms half the structure. The structure is coloureed according to the secondary structure; the magnesium ion is coloured green.

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

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

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

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

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

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

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

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

Nif11 (Nitrogen fixation 11; PF07862)

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

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

3.4 Other Potential Uses

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

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

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

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

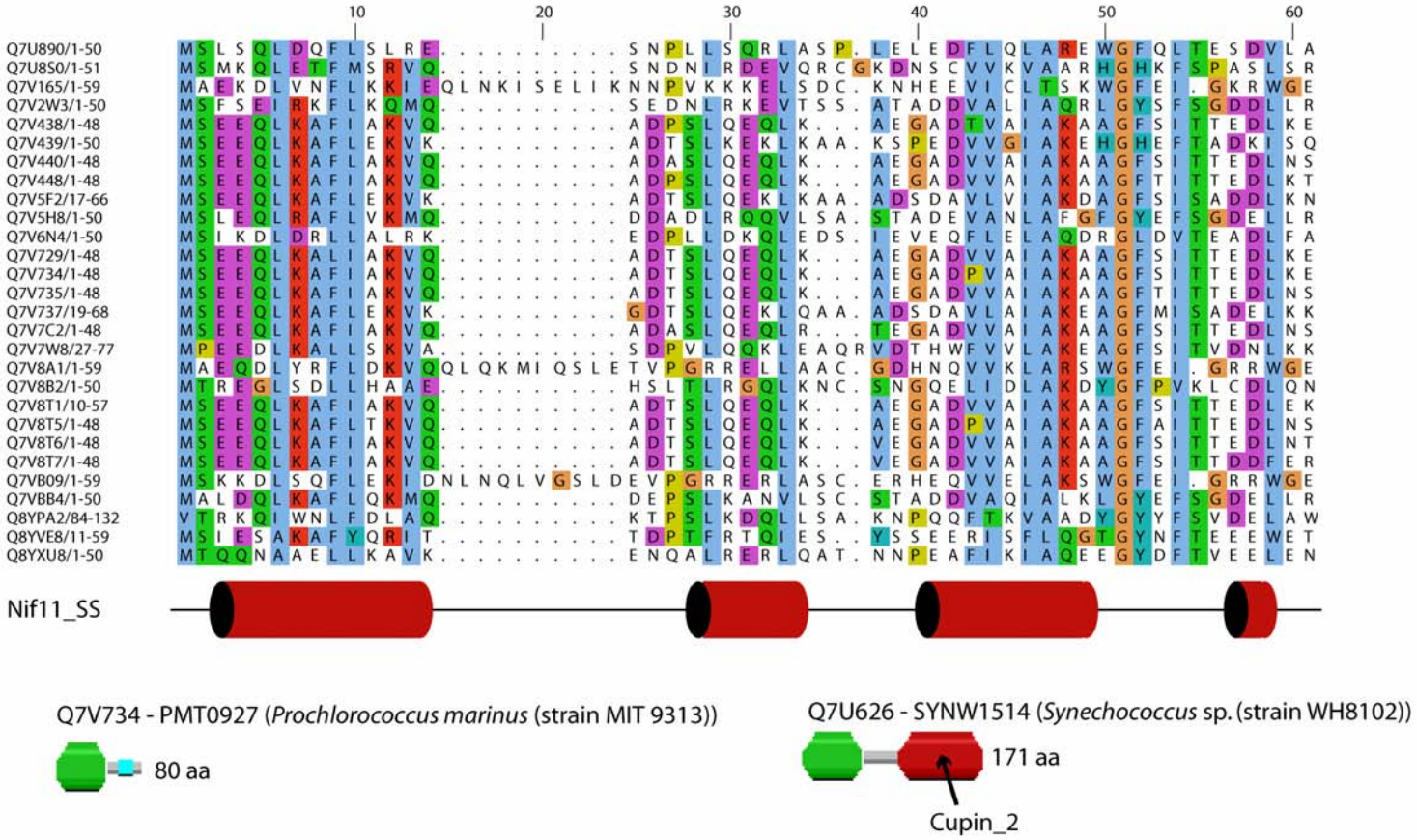

Figure 3.22: Example Nif11 architectures and alignment

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