2 Identifying Novel Domains

2.1 Domain Hunt Methods

2.1.1 Introduction

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

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

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

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

2.1.2 Details of Methods

Small Protein Clustering (SPC)

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

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

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

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

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

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

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

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

Repeat Identification (RI)

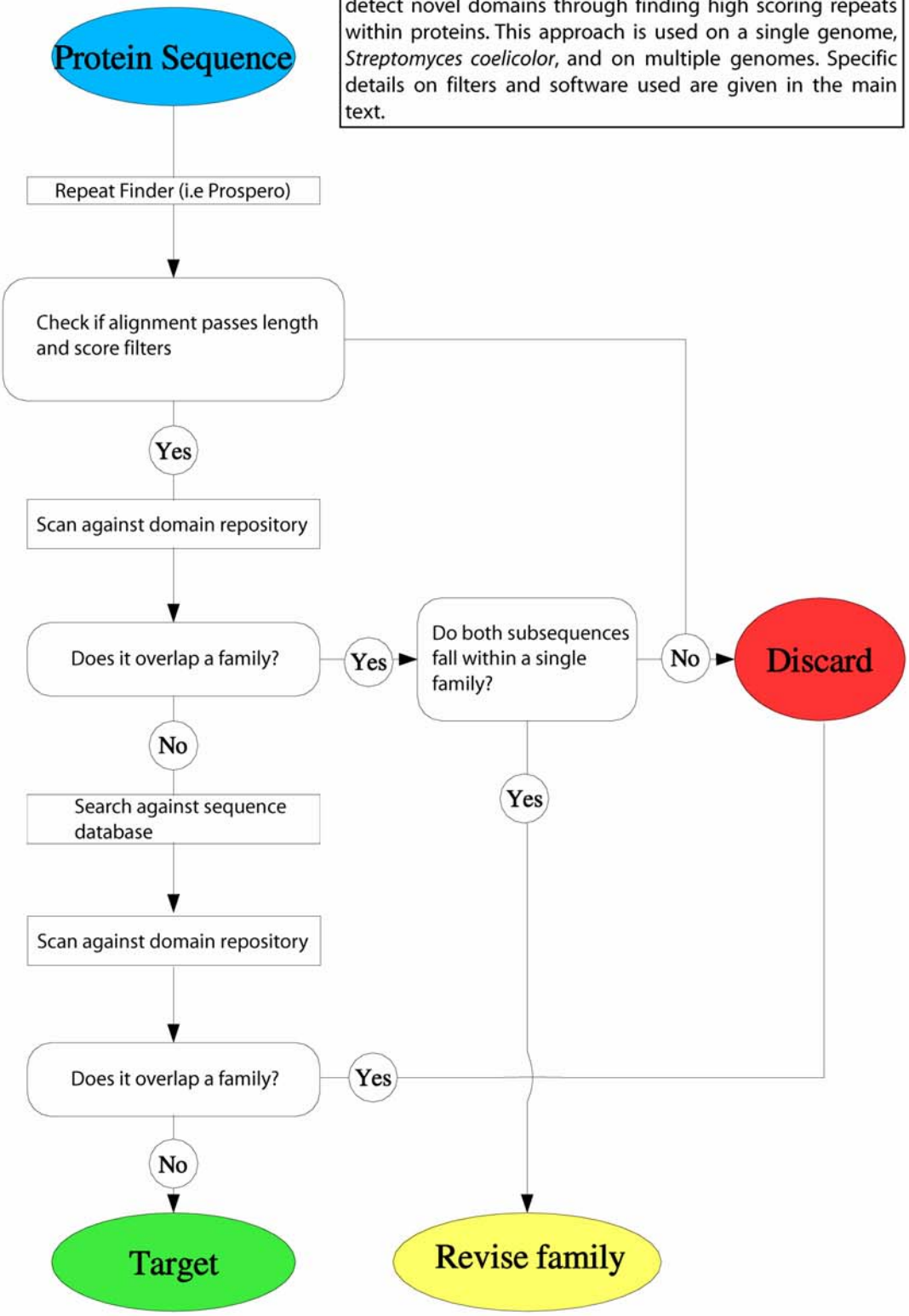
Repeat identification refers to the process of identifying domains through finding repeats within a protein, and is the method I've most used. In the past the discovery of internal repeats within a proteins sequence has led to the discovery of novel domains (e.g. Fong, Hurley *et al*., 1986 and Haslam,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

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

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

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

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



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

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

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

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

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

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

Sequence Fragmentation

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

"Blocky Alignments"

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

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

2.2 Domain Hunting in *Streptomyces coelicolor*

2.2.1 Introduction to *Streptomyces coelicolor* – a Complex Prokaryote

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

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

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

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

2.2.2 Methods

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

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

Repeat Identification

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

Short Protein Clustering

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

2.2.3 Summary of Results

Repeat Identification:

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

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

Small Protein Clustering

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

Pfam Accession No	*Family Name*	*Pfam Type*	*Basic Function*	*No of copies in S. coelicolor*	*Antibiotic biosynthesis*	*Cell Wall Biosynth*	*Cell Wall / Periplasm*	*Replication*	*Secreted*
A) Novel Families									
PF03457	HA	Domain	Putative RNA binding domain	21				X	
PF03621	MbtH	Domain	Possibly involved in antibiotic biosynthesis	2	X				
PF03625	DUF302	Domain	Unknown function	3			X		X
PF03640	Lipoprotein_15	Repeat	Unknown function	6			X		X
PF03703	DUF304	Domain	Unknown function	4			X		X
PF03704	BTAD	Family	Bacterial transcriptional activator domain	12	X				
PF03710	GlnE	Domain	Glutamate-ammonia ligase adenylyltransferase	2					
PF03713	DUF305	Domain	Unknown function	6			X		X
PF03714	PUD	Domain	Putative carbohydrate binding domain	2			X		X
PF03724	DUF306	Domain	Unknown function	2					X
PF03729	DUF308	Repeat	Unknown function	6			X		X
PF03733	DUF307	Domain	Unknown function	2			X		X
PF03752	ALF	Repeat	Putative signal transduction domains	16					X
PF03756	AfsA_repeat	Repeat	A-factor biosynthesis	2	X				
PF03771	SPDB	Domain	(Probably) mobile element replication	16					
PF03777	DUF320	Domain	Unknown function	11			X		X
PF03779	SPW	Repeat	Unknown function	2			X		
PF03793	PASTA	Domain	Cell wall peptidoglycan sensor domain	9		X	X	X	X
PF03794	HHE	Domain	Unknown function	7		X			
PF03795	YCII	Domain	Probably enzymatic domain	3					
PF03860	DUF326	Domain	Unknown function	6			X		
PF03984	DUF346	Repeat	Unknown function (β-propeller)	7			X		X
PF03988	DUF347	Repeat	Unknown function	4			X		
PF03990	DUF348	Domain	Unknown function	3			X		X
PF03992	ABM	Domain	Antibiotic biosynthesis monooxygenase	3	X				
PF03993	DUF349	Domain	Unknown function	3					
PF03994	DUF350	Domain	Unknown function	2			X		X
PF03995	DUF351	Domain	Unknown function	4					X
PF04151	PPC	Domain	PKD-like peptidase C-terminal domain	3					X
PF04205	FMN_bind	Domain	FMN-binding domain	2			X		X
PF05120	GvpG	Domain	Gas vesicle protein G	2					X
PF05121	GvpK	Domain	Gas vesicle protein K	2					
PF05122	SpdB	Domain	Mobile element transfer proteins	2					
B) Previously Described New Pfam Families									
PF03458	UPF0126	Domain	Unknown function	4			X		X
PF03459	TOBE	Domain	Transport-associated OB fold domain	9			X		
PF03707	MHYT	Repeat	Putative ligand receptor	6			X		X
PF03989	DNA_gyraseA_C	Repeat	DNA-binding β-propeller	8				X	
C) Significantly Extended Families									
PF00188	SCP	Domain	Unknown function	4			X		X
PF01839	FG-GAP	Repeat	Putative β-propeller	57			X		X

Table 2.1: Novel domains found in *Streptomyces coelicolor*

2.2.4 Notes on Table of All Novel Domains Identified

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

2.3 Descriptions of Novel Domain

HA (Helicase Associated domain; PF03457)

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

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

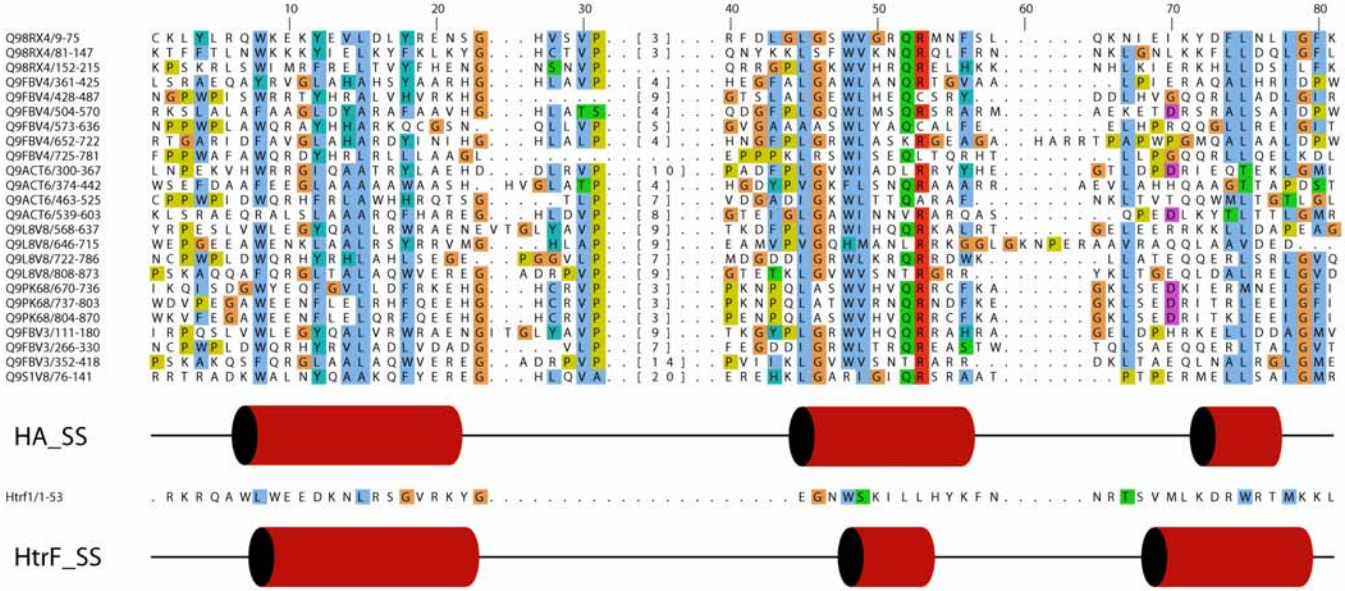


Figure 2.2: HA example alignment
As well as an example alignment of HA domains the DNA-binding domain of HtrF is shown below, along with its known secondary structure (PDB:1BA5). This sequence was aligned to the HA domains with T-Coffee. As can be seen there are some intriguing similarities between the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

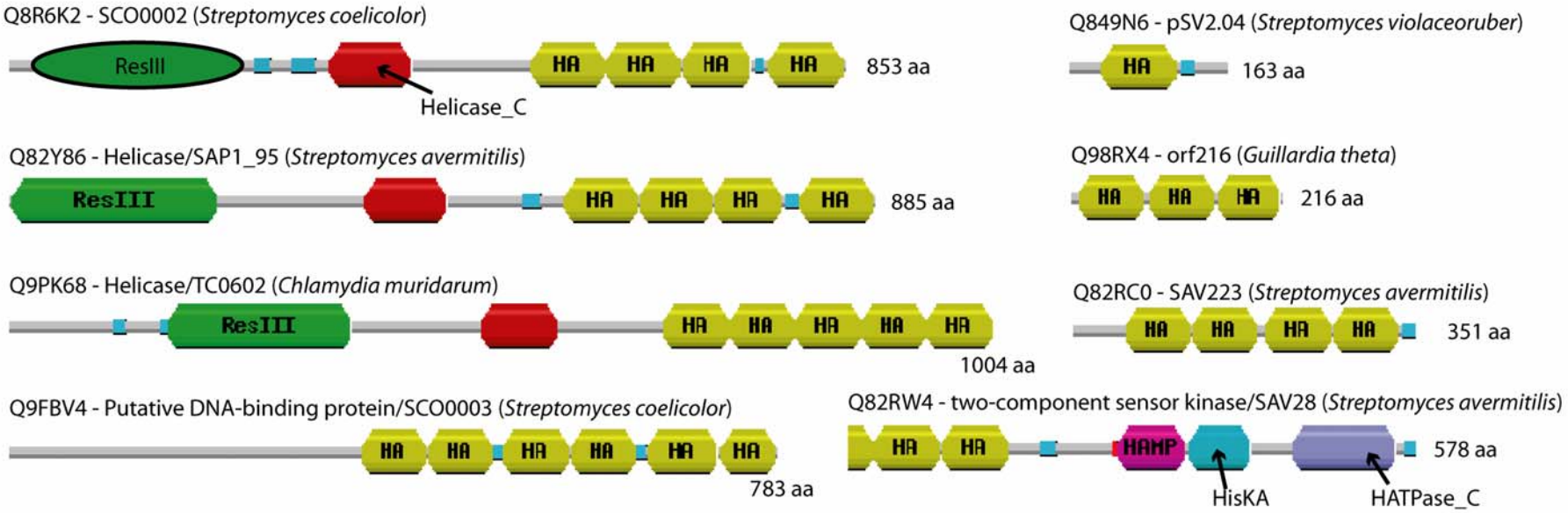


Figure 2.3: HA example domain architectures

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

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

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

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

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

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

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

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

BTAD (Bacterial transcriptional activator domain; PF03704)

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

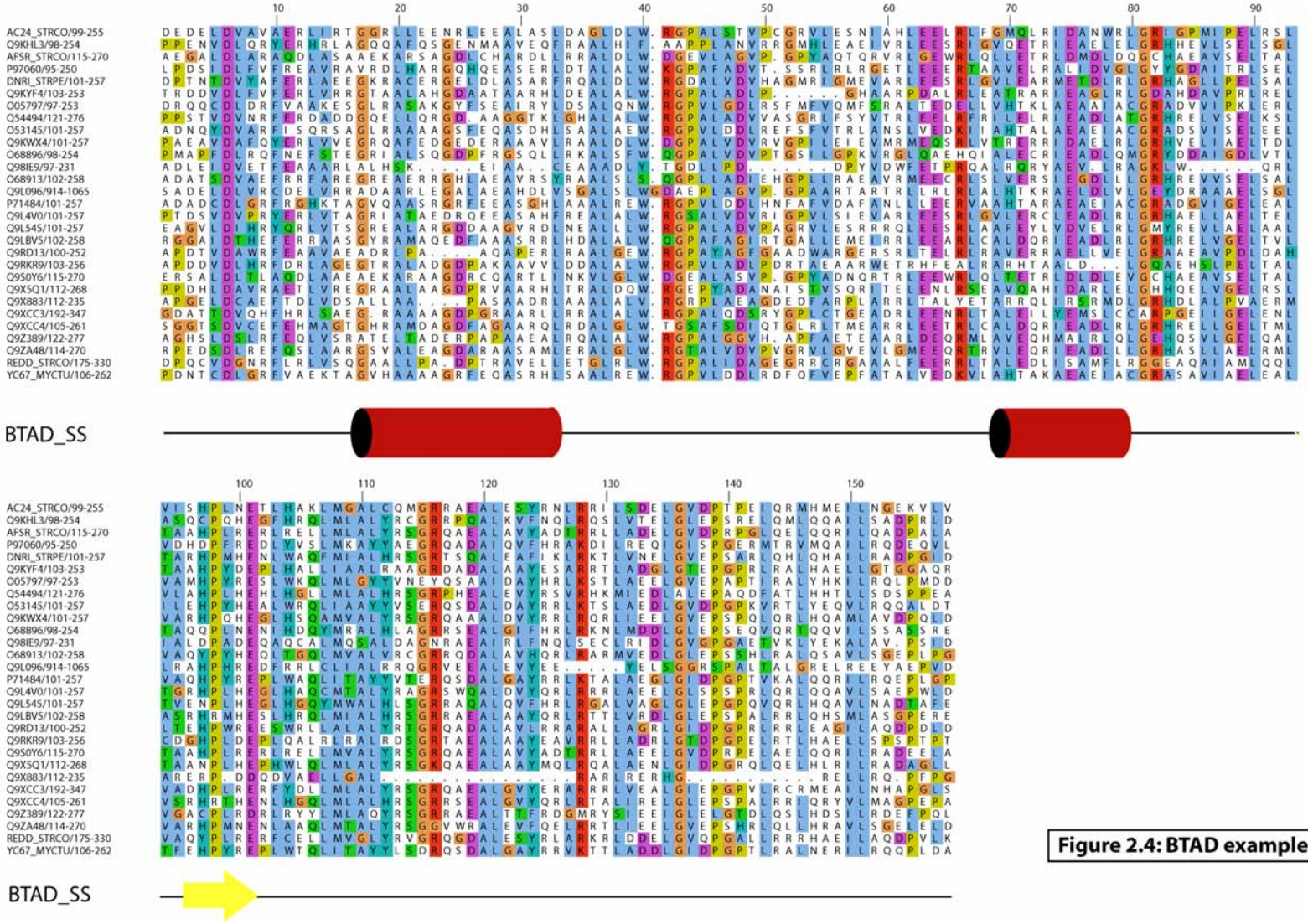


Figure 2.4: BTAD example alignment

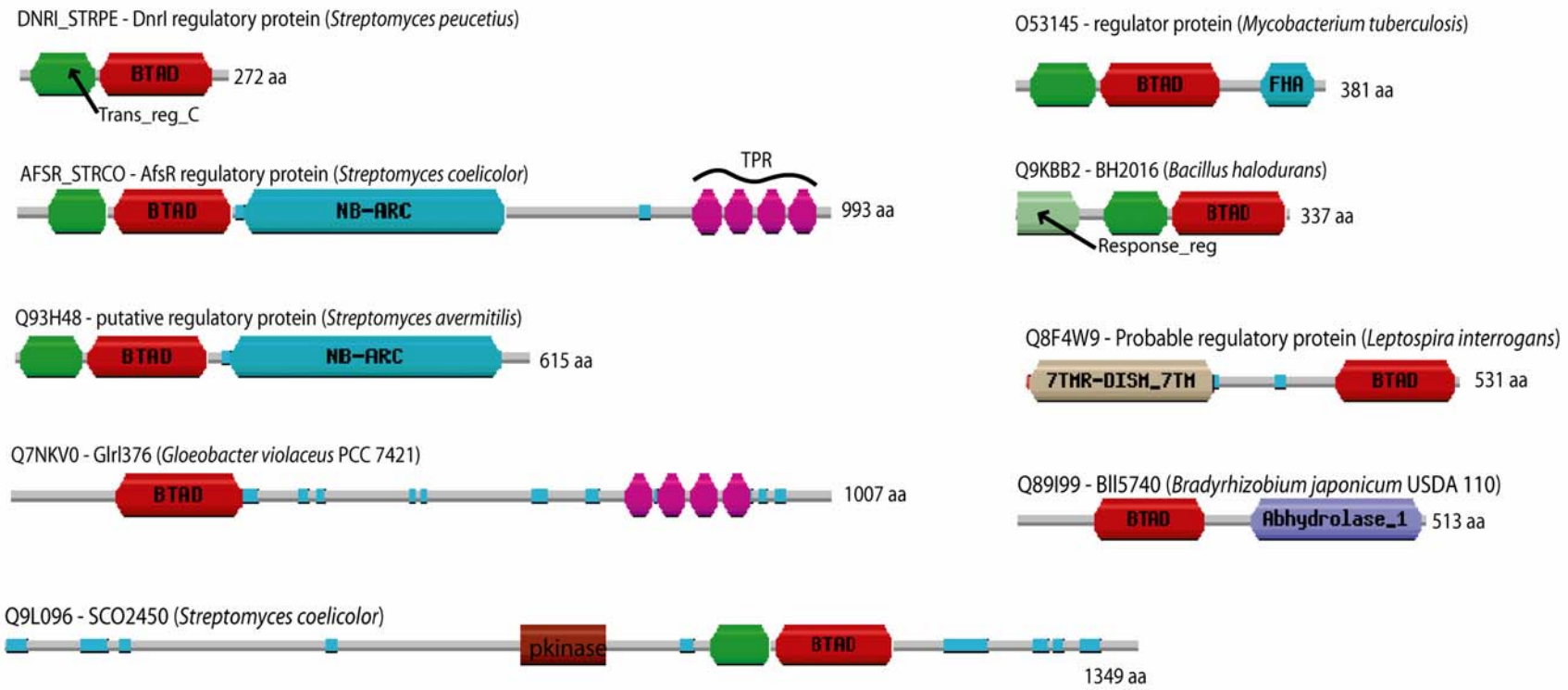


Figure 2.5: BTAD example architectures

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

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

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

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

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

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

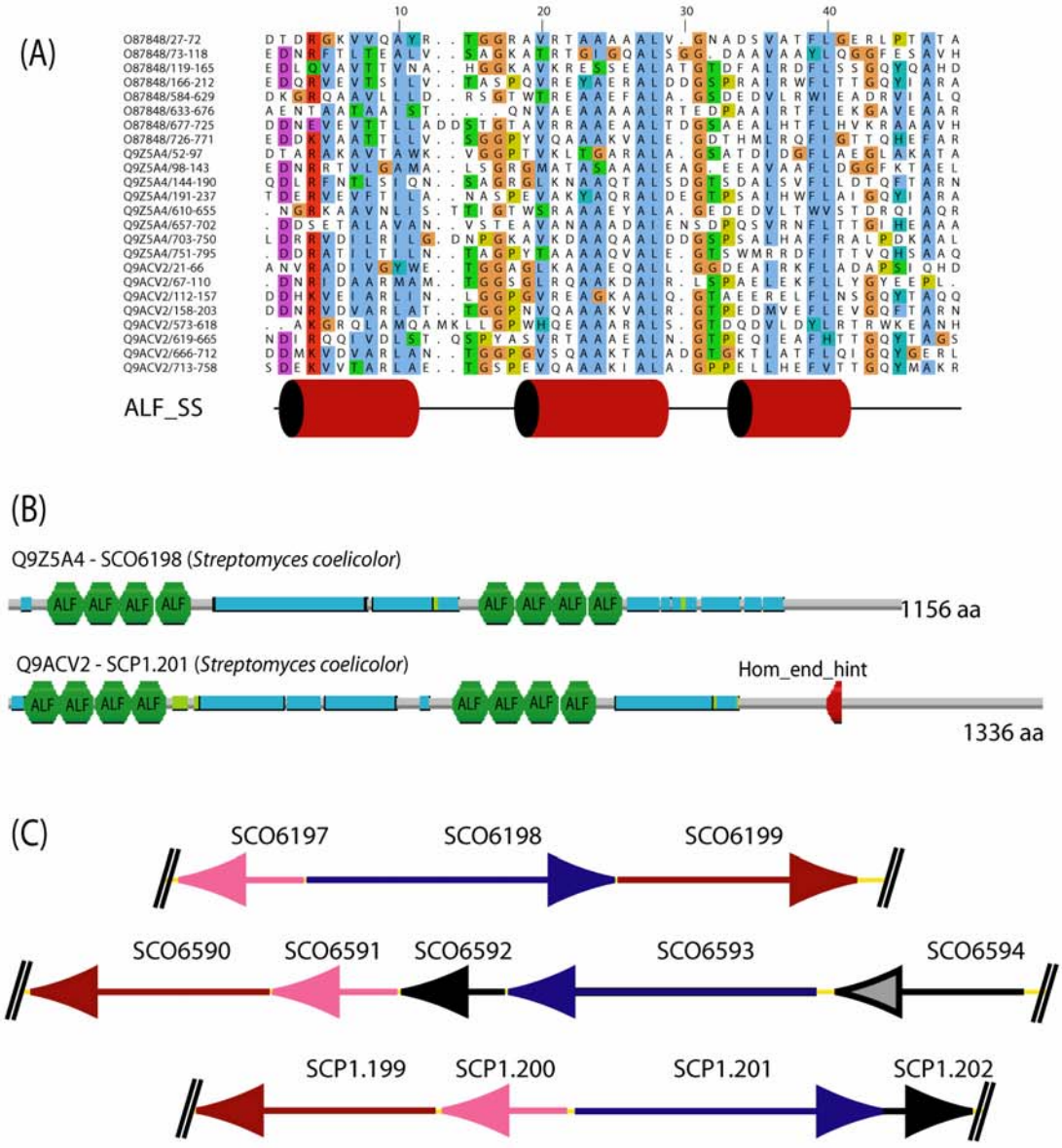


Figure 2.6: ALF alignment, architectures, and genome context
Parts A and B depict an example alignment and architectures for the ALF repeat as standard. Part C shows the ALF-containing proteins of Streptomyces coelicolor, along with their immediate gene neighbourhood.The arrows indicate the direction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

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

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

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

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

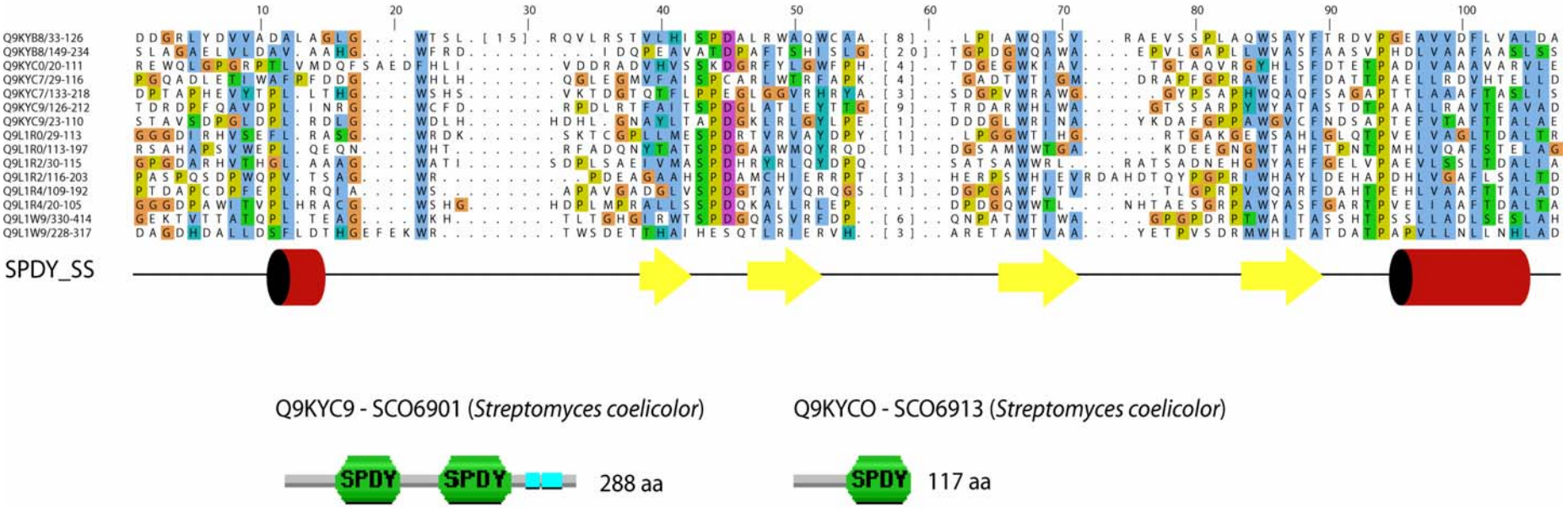


Figure 2.7: SPDY alignment and architectures

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

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

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

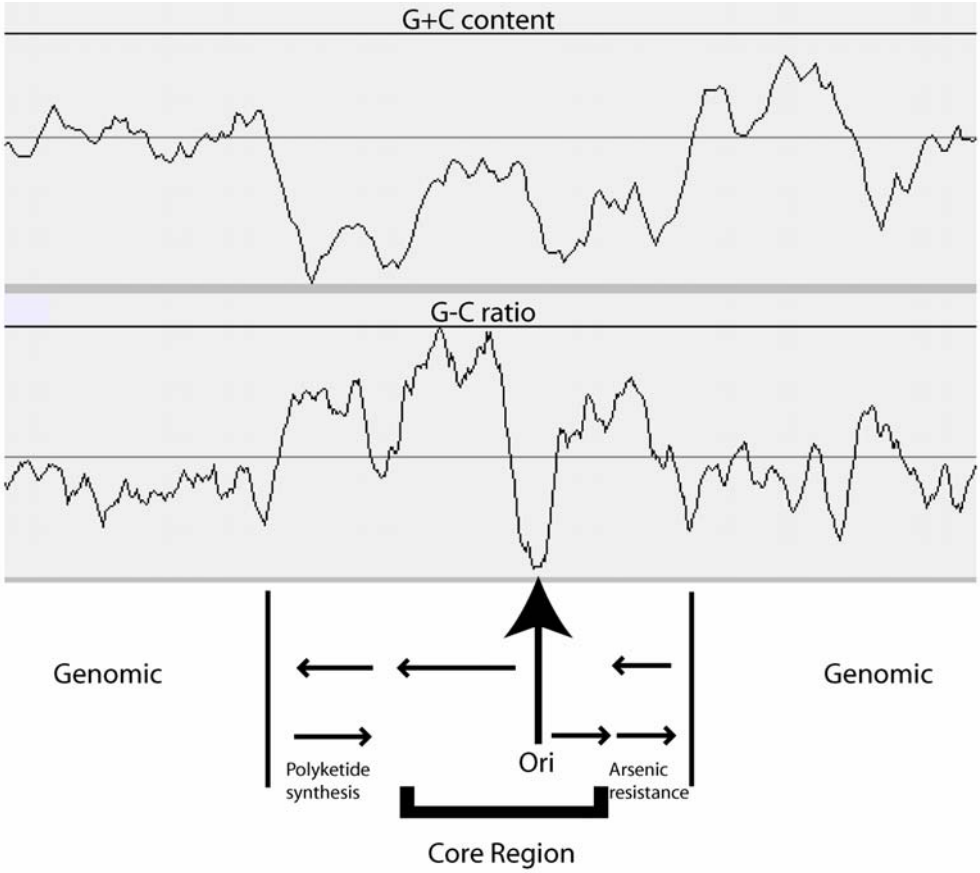


Figure 2.8: Evidence for the presence of a mobile DNA element
The SPDY domains are all found in a region of the genome believed to be a mobile element. As can be seen the G-C ratio and G+C content show a marked difference to the background genome. This element appears to carry genes for arsenic resistance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

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

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

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

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

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

HHE (Histidine-Histidine-Glutamate motif; PF03794)

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

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

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

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

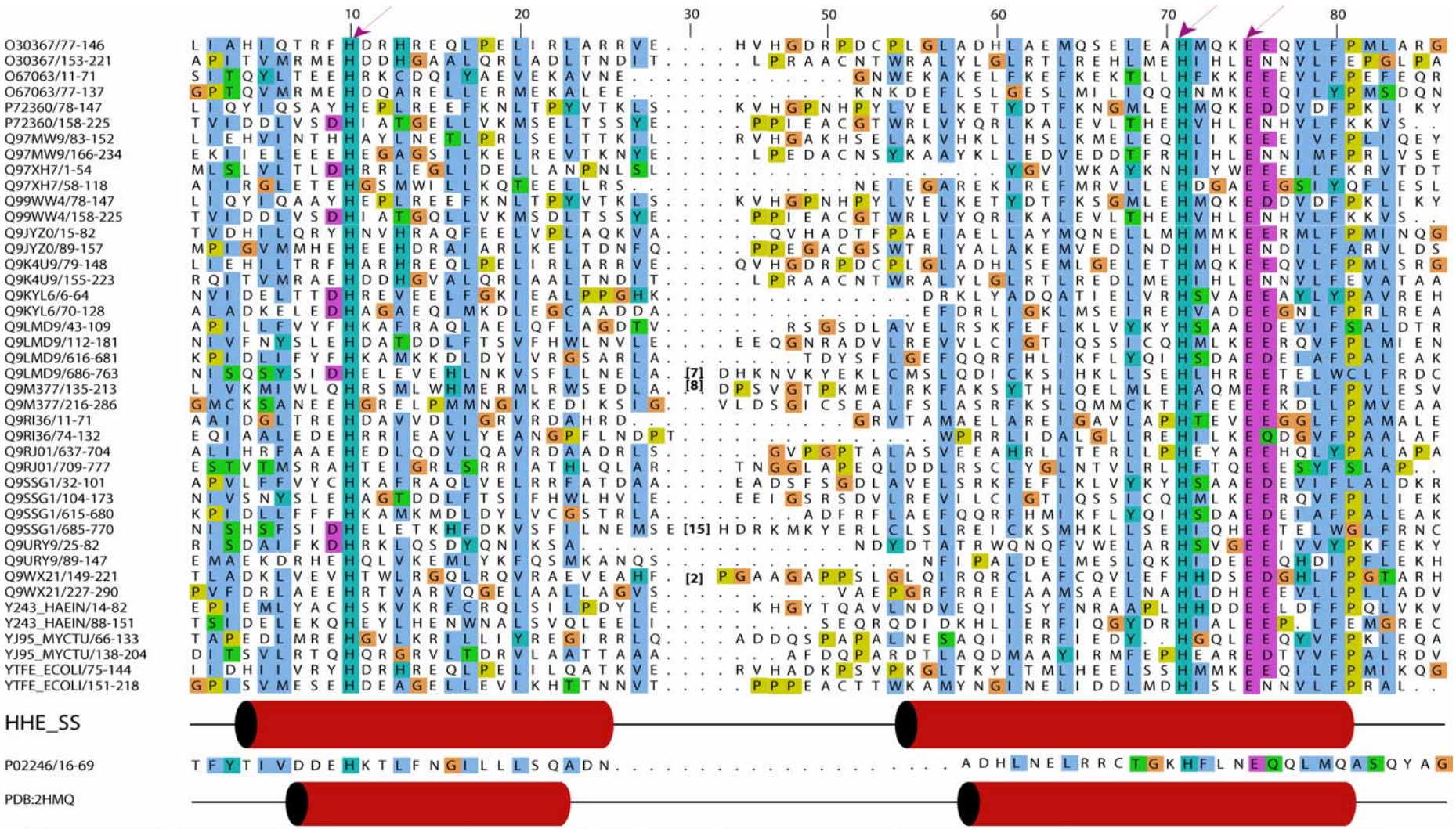


Figure 2.9: Alignment of original HHE domains and predicted secondary structure against a Hemerythrin domain and known structure
The ligand-binding residues are marked by the purple arrows. As can be seen these are very highly conserved

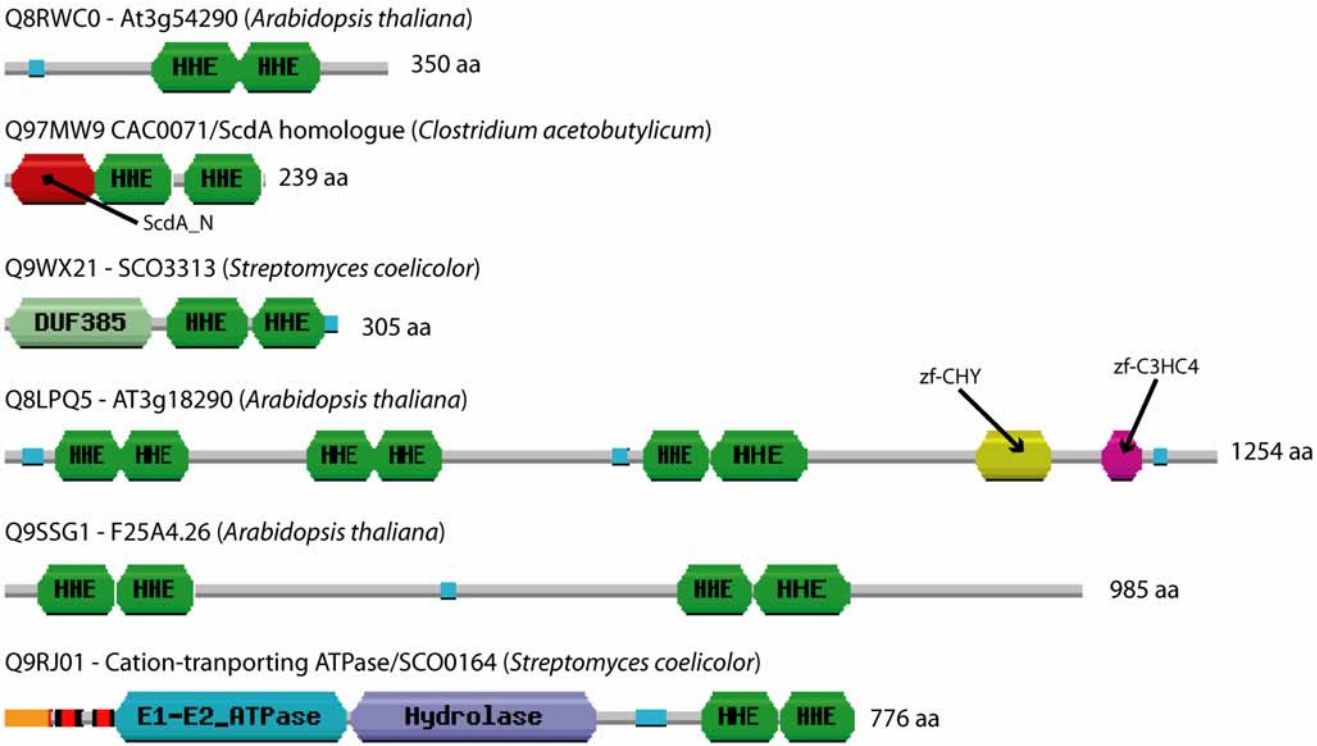


Figure 2.10: HHE architectures

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

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

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

to switch its metabolism from aerobic to anaerobic. While (Brunskill, deJonge *et al*., 1997) suggest that ScdA is a regulator of development, the evidence of its domain structure combined with its involvement in *S. aureus* survival when moving into anoxic environments suggests that its specific role may be either to scavenge O_2 from outside the cell or to provide an intracellular store. Alternatively it may function as a positive regulator and if there is no oxygen bound to the HHE domains it will up-regulate self-protective systems. The defects identified by Brunskill and colleagues, and noted above, may be ascribable to damage caused by oxidative stress. These proteins have another domain - ScdA_N - at the N-terminus, which does not have an identifiable function, but may transduce the signal from the sensor HHE domains to the next downstream element. Determination of the function of the ScdA_N domain should help to resolve how these proteins function.

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

PPC (Bacterial Prepeptidase C-terminal domain; PF04151)

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

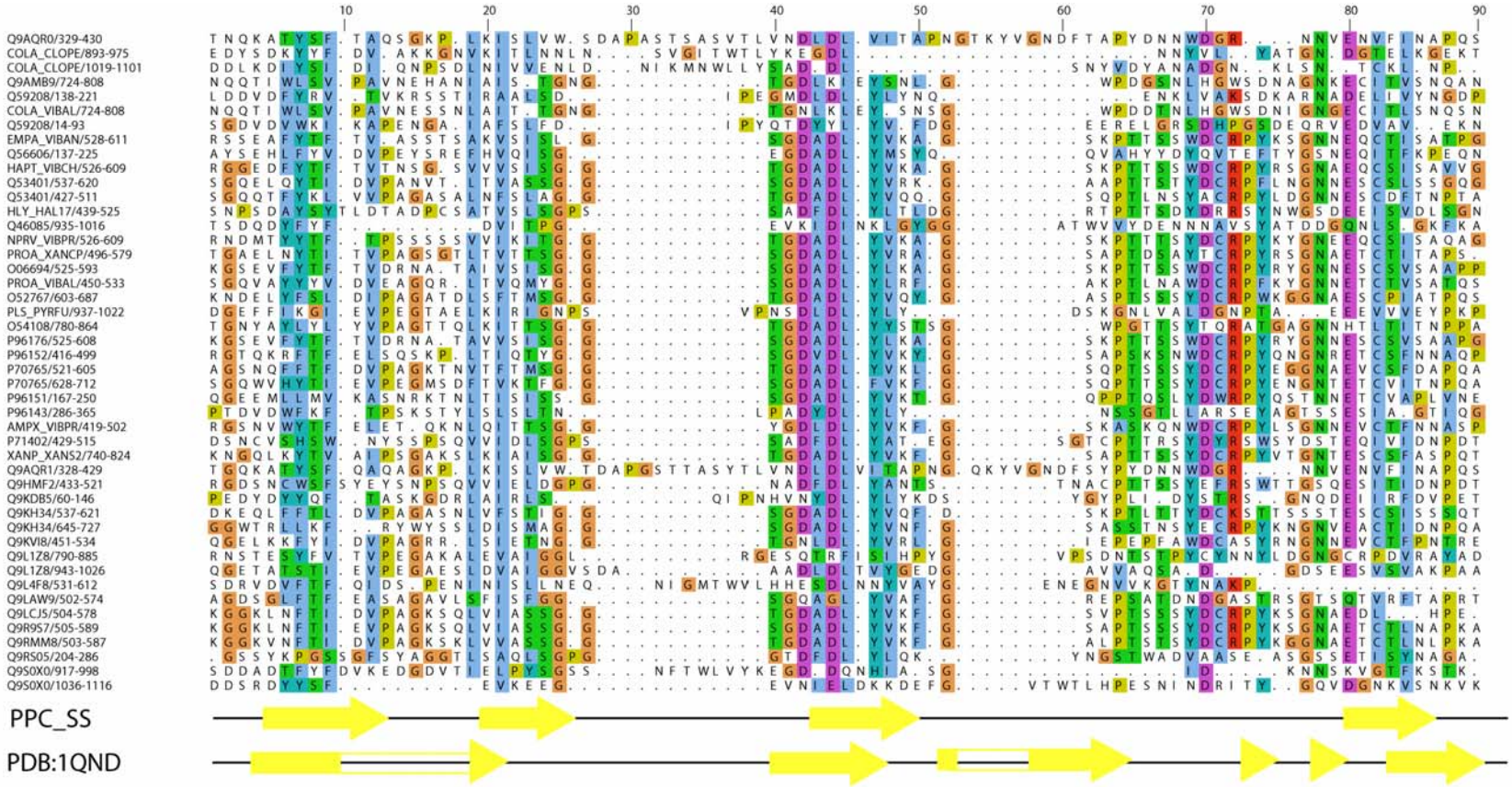


Figure 2.11: PPC alignment along with predicted and known secondary structure
Since the initial prediction of the structure of this domain the C-terminal PPC domain of Q9S0X0 (residues: 1035-1116) has been solved. The structure is largely in agreement with the predicted structure for the family. Where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

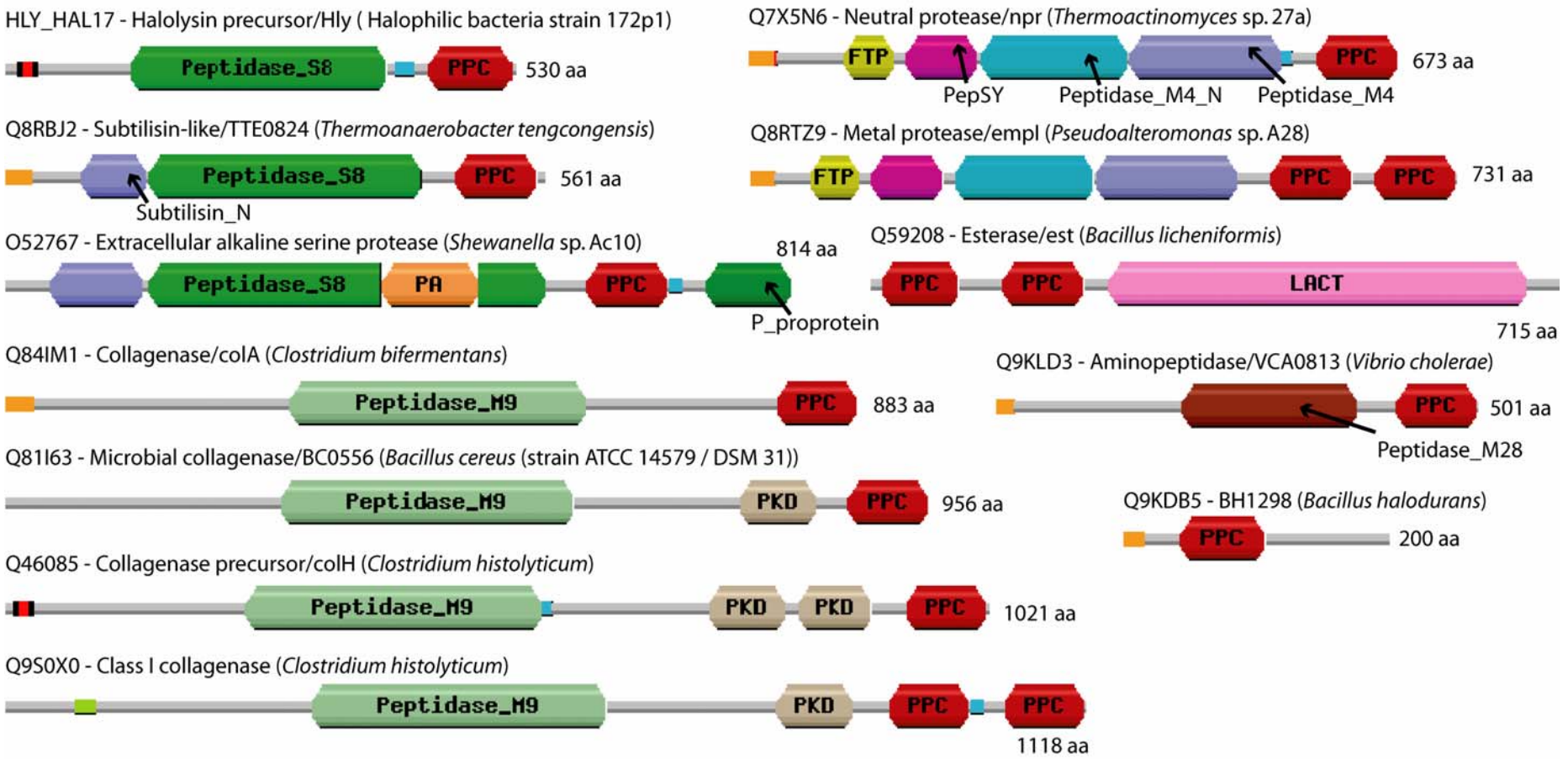


Figure 2.12: PPC domain architectures

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

The original publication of this domain (Yeats, Bentley *et al*., 2003) suggested that it was likely to belong to the Ig fold, based on the MSA (personal communication: A Bateman) and its apparent interchangeability with the PKD domain in various architectures (e.g. compare the architectures of Q81I63, Q899Y1, Q9S0X0 and Q46085). Furthermore it was predicted to either be involved in localisation of the enzyme or in acquisition of the substrate. Subsequent to these predictions, the crystal structure of a PPC domain was determined by Wilson, Matsushita *et al*. (2003), which showed that it actually to belong to the jelly roll fold. Their work has shown that this domain binds the triple-helix of collagen in a reaction mediated by calcium ions; however, the Ca^{2+}-binding site lay in a linker that does not fall within the PPC domain, so possibly the involvement of Ca^{2+} is restricted and not true of all PPC domains. A similar conclusion is reached by Wilson, Matsushita *et al*. (2003) on the basis of site-directed mutagenesis. It should also be noted that not all PPC domains necessarily bind collagen; further direct experimentation is needed to clarify their overarching function.

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

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

FMN_bind (Flavin MonoNucleotide-binding; PF04205)

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

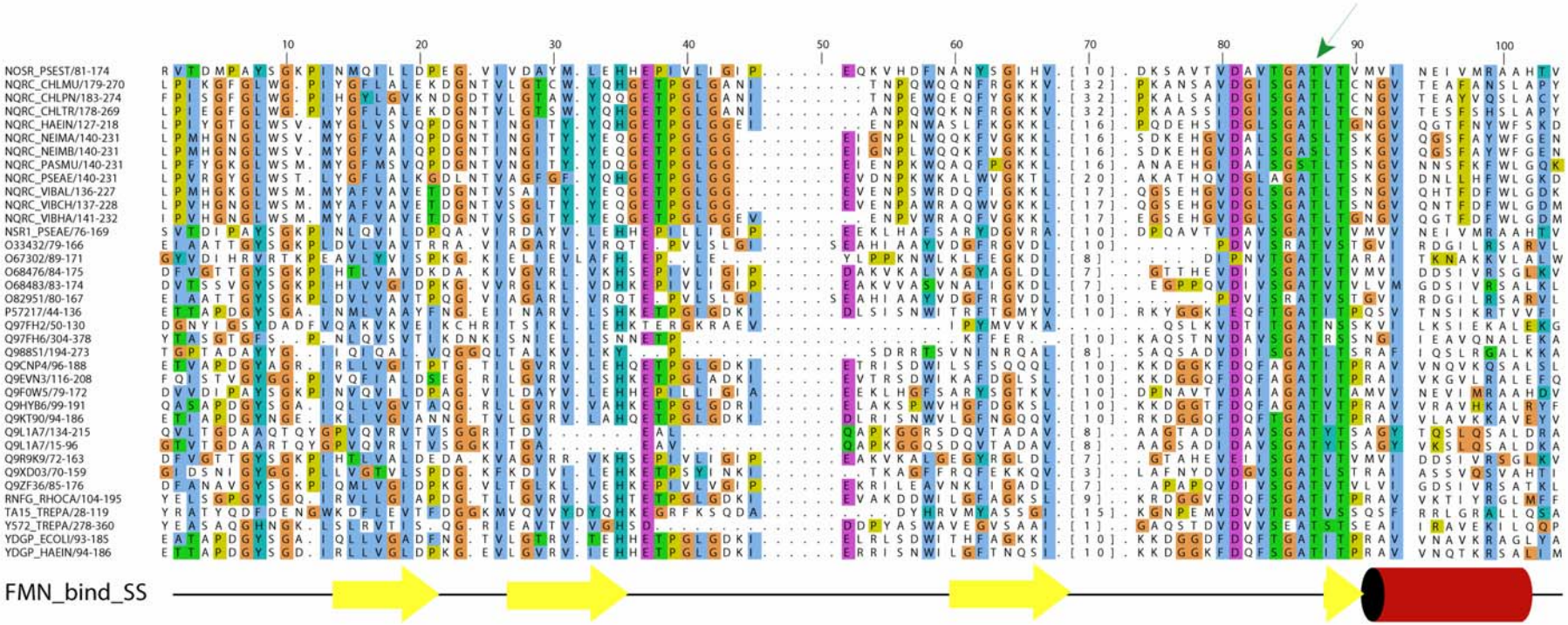


Figure 2.13: FMN_bind alignment
The green arrow marks the threonine residue found to bind FMN by Hayashi, Nakayama et al., 2001).

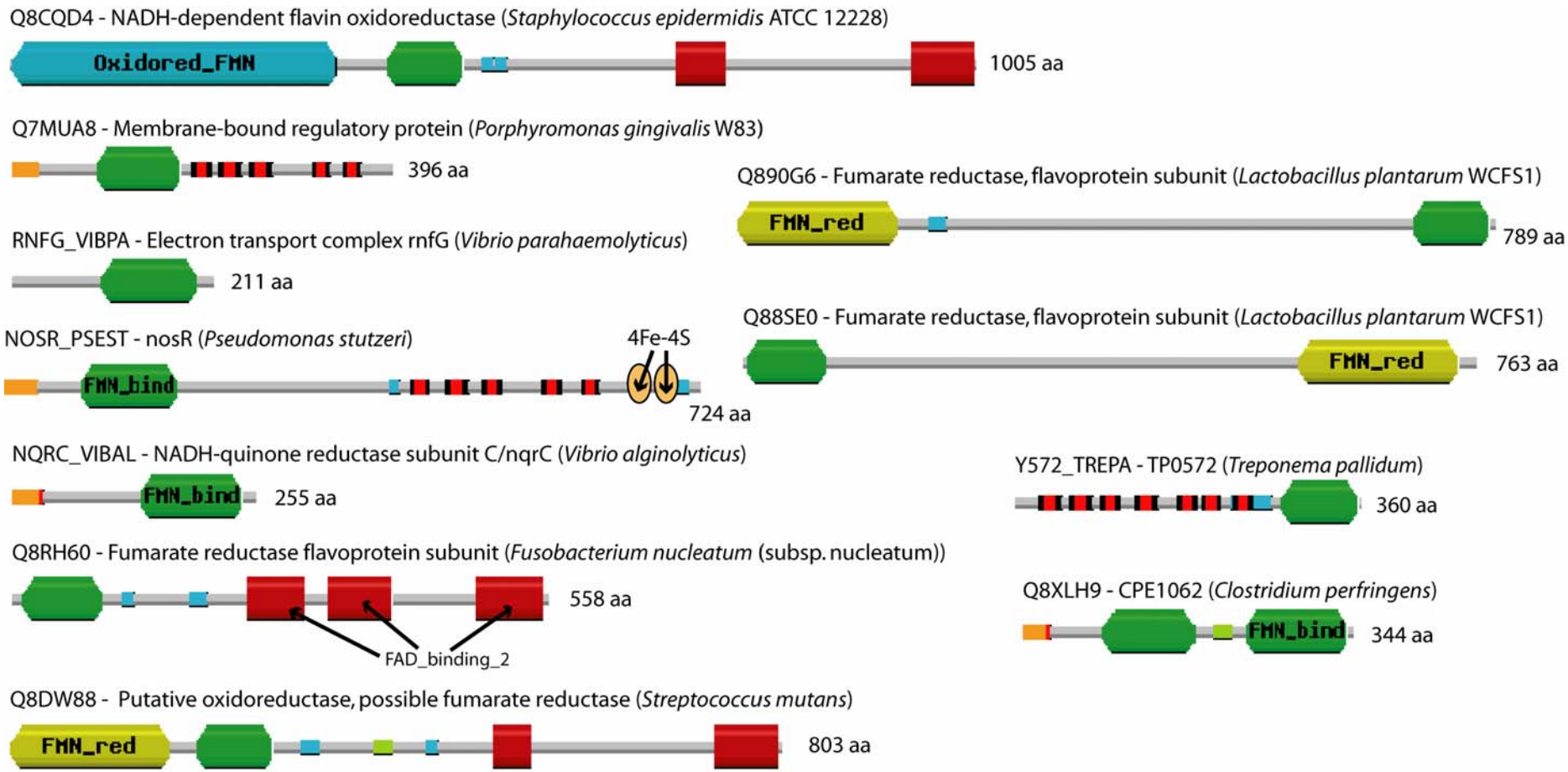


Figure 2.14: FMN_bind domain architectures

MbtH (MbtH-like proteins; PF03621)

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

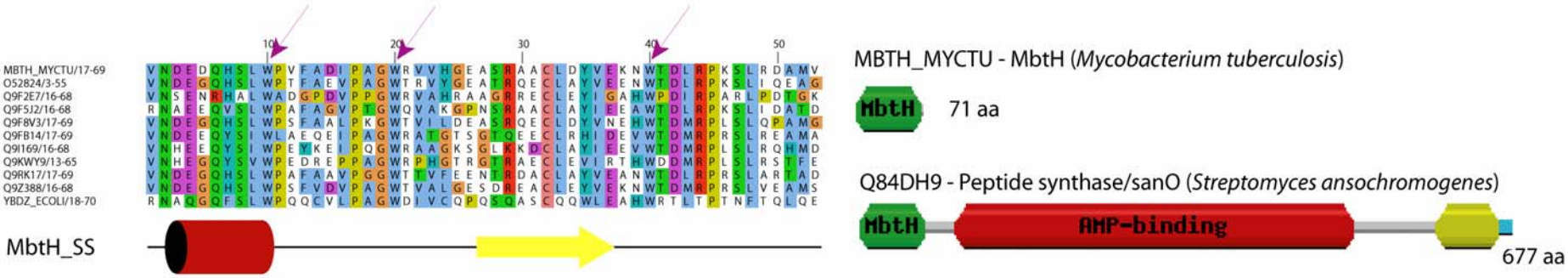


Figure 2.15: Alignment and architectures for the MbtH domain
The purple arrows above the alignment mark three invariant tryptophan residues that may have functional importance.

2.4 Significantly Extended Pfam Families

SCP (Secreted Cysteine-Rich Proteins; PF00188)

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

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

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

Review of the data available for this family had previously led to the conclusion that it was somehow involved in extracellular signalling; however, the protein is very large for a signalling molecule and, even though there is evidence for an active site, there is little evidence for the generation of a smaller signalling molecule. A recent paper by (Milne, Abbenante *et al*., 2003) suggested that Tex31, which contains a single SCP domain, is a Ca^{2+}-dependant protease. While the evidence for it being a protease is not definitive, mostly due to possible left-over impurities (personal communication: N Rawlings), the evidence for Ca^{2+}-binding is quite strong. This fits with the identification of the conserved histidines and glutamate (see HHE domain above), and also fits with the involvement of SCP-domain proteins in many diverse processes. For instance cell polarity has been well-established as being of fundamental importance in determining growth directions of pollen tubes and fungal hyphae.

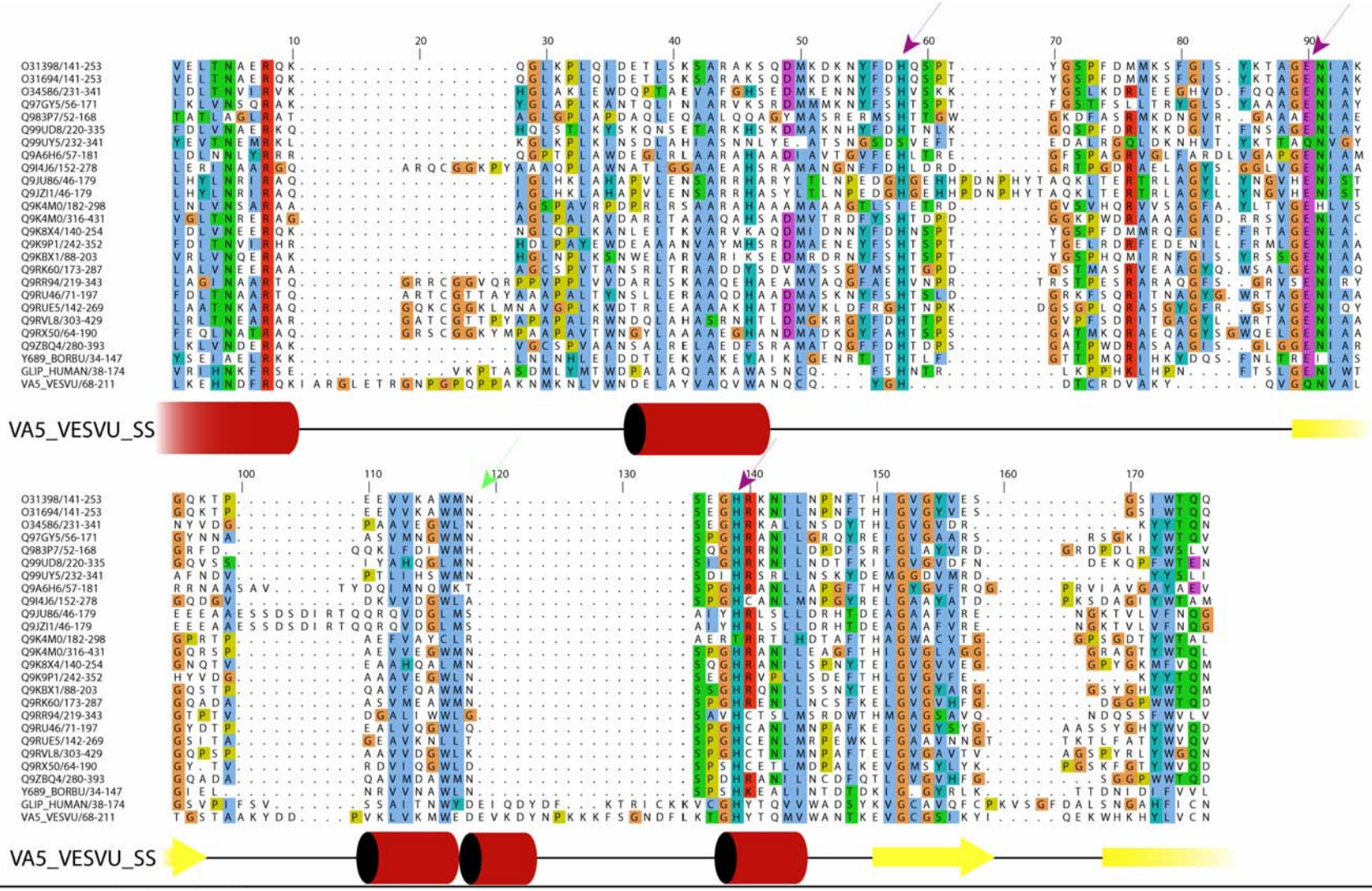


Figure 2.16: SCP domain alignment
The secondary structure provided in this figure is based on the three-dimensional structure of VA5_VESVU (PDB:1QNX). The arrows above the alignment mark the residues predicted by Szyperski et al (1998) to form an active site. One of these, marked by the green arrow, is not conserved between the bacterial and eukaryotic versions, ruling it out from being an important part of the active site.

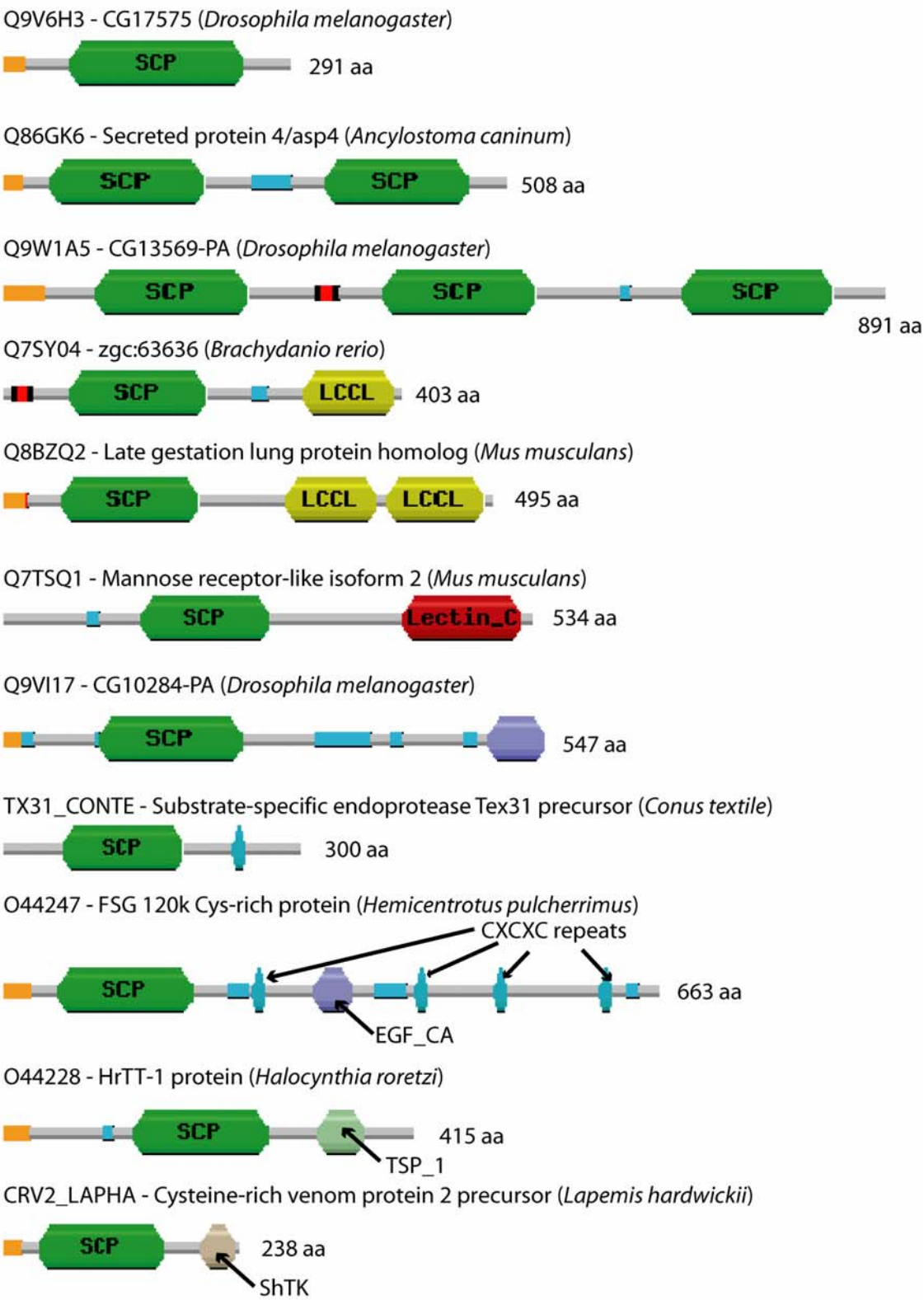


Figure 2.17: SCP domain architectures

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

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

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

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

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

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

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

FG-GAP (PF01839)

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

2.5 Concluding Comments

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

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

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

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