BMC Microbiology

Research article

New Knowledge from Old: *In silico* discovery of novel protein domains in *Streptomyces coelicolor*

Corin Yeats*, Stephen Bentley and Alex Bateman

Address: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

Email: Corin Yeats* - cay@sanger.ac.uk; Stephen Bentley - sdb@sanger.ac.uk; Alex Bateman - agb@sanger.ac.uk

* Corresponding author

Published: 6 February 2003

BMC Microbiology 2003, **3**:3

This article is available from: http://www.biomedcentral.com/1471-2180/3/3

Received: 19 November 2002
Accepted: 6 February 2003

Abstract

Background: *Streptomyces coelicolor* has long been considered a remarkable bacterium with a complex life-cycle, ubiquitous environmental distribution, linear chromosomes and plasmids, and a huge range of pharmaceutically useful secondary metabolites. Completion of the genome sequence demonstrated that this diversity carried through to the genetic level, with over 7000 genes identified. We sought to expand our understanding of this organism at the molecular level through identification and annotation of novel protein domains. Protein domains are the evolutionary conserved units from which proteins are formed.

Results: Two automated methods were employed to rapidly generate an optimised set of targets, which were subsequently analysed manually. A final set of 37 domains or structural repeats, represented 204 times in the genome, was developed. Using these families enabled us to correlate items of information from many different resources. Several immediately enhance our understanding both of *S. coelicolor* and also general bacterial molecular mechanisms, including cell wall biosynthesis regulation and streptomycete telomere maintenance.

Discussion: Delineation of protein domain families enables detailed analysis of protein function, as well as identification of likely regions or residues of particular interest. Hence this kind of prior approach can increase the rate of discovery in the laboratory. Furthermore we demonstrate that using this type of *in silico* method it is possible to fairly rapidly generate new biological information from previously uncorrelated data.

Background

Streptomyces coelicolor – *a complex prokaryote*

Streptomyces coelicolor is a representative of a group of high G+C Gram positive bacteria whose successful adaptation to their niche is demonstrated by their almost ubiquitous presence in soil. 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.

Also unusual is the size and structure of streptomycete chromosomes. *Streptomyces coelicolor* has a linear chromosome which at 8,667,507 base pairs is the largest complete bacterial genome sequence currently available [1]. It is predicted to encode a remarkable 7825 proteins, around twice as many as most sequenced bacterial genomes and more than the eukaryote *Saccharomyces cerevisiae*. 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

Protein Domains

The direct functional and structural determination of all the proteins in an organism is prohibitively expensive and time consuming. The sequencing of a genome is a powerful aid to understanding the molecular biology of an organism even in the absence of direct experimental work on the organism. Given a complete genome sequence one can begin to ask global questions about the organism's metabolic potential as well as what molecular systems it contains. The transfer of information between related proteins is of fundamental importance into studies of the proteome. While comparison of whole protein sequences is a useful tool in finding close and direct relationships, it also misses the subtler relationships between proteins. A more sophisticated method of analysing proteins is through the determination of their domain content [2].

Protein domains are discrete stable amino acids structures, typically globular and formed from between 40 and 400 amino acids. Homologous domains exhibit highly similar tertiary structure, with the overall structure of the protein being a composite of its domains and connecting sections. To a varying extent biochemical and physiological functions can also be transferred between homologous domains. Some domain families exhibit a wide-range of activities, specificities or interactions, whereas others show far less variation. Of note, and analogous to domains, are structural repeats, such as the WD40 repeat. Typically such repeats are between 5 and 60 amino acid residues in length, and occur in a tandem array in a protein. These fold together to form stable, and often very regular, 3-dimensional structures. A common example is the β-propeller (covered in detail in [3]). It is important to realize that repeats are different from repeated domains. Repeated domains would be expected to be stable in isolation, contrasting with repeats which would not be.

Inference of Domain Function

In annotation of bacterial genomes a key step is to infer the function of a protein by similarity to other known proteins. This step usually takes each protein in the genome and searches a large non-redundant database using a sequence search method such as BLAST or FastA [4,5]. The list of matches is then examined to find if any similar protein has a function that can reliably be transferred. Care must be exercised in this process, as this approach can lead to missannotation. In cases of multidomain proteins the similarity to another protein may be due to a domain similarity. For example, in the original annotation of the *Methannococcus jannaschii* genome [6] several proteins were annotated as inosine-monophosphate dehydrogenase (IMPDH) enzymes. The similarity to IMPDH lay not in the enzymatic domain but to a regulatory domain [7]. Hence analysis of protein domain content is an important component of the annotation process.

In this paper we attempt to identify novel protein domains in *Streptomyces coelicolor*. To be useful in understanding the biology of *Streptomyces coelicolor* and other organisms we wish to infer the function of these novel domains. There are two complementary approaches to this problem. Firstly, similarity to other protein domains can be used. By examining the function of each protein containing the domain we try to infer what the common function might be between the proteins and hence the function of the domain. This process is often hampered by a lack of information about any of the proteins. Secondly and more recently methods using genomic context have been developed that allow increased confidence for functional prediction. These approaches include using gene order such as appearance of proteins in operons, the appearance of fusion proteins and phylogenetic profiles [8]. We can also use the knowledge of the biology of *Streptomyces coelicolor* to provide a species context. This allows interpretation of domains and proteins in the context of the whole organism's biology.

We use this principle to help elucidate putative biological mechanisms and deepen our understanding of described systems within the soil-dwelling prokaryote *Streptomyces coelicolor*. Firstly a set of novel domains is predicted using the recently completed genome sequence. Homologues in other organisms were searched for and descriptive information obtained through literature searching and other analytical tools. This information was then viewed within the context of the *Streptomyces coelicolor* organism. These results provide functions for many proteins leading to a number of testable hypotheses.

Methods

The Domain Hunt Methodology

The simplest way to accurately identify novel domains is through examination of high resolution protein structures, usually derived crystallographic studies; however only a small proportion of sequences have representative structures. To get maximum value from the large amounts of sequence data being produced, a variety of detailed sequence comparison methods are employed to predict domain families. Such predicted domains are actually representative of evolutionary conserved sequences rather than discrete protein structures; however experience shows that they mostly represent such structures. This finding has led to the consideration of domains as the building blocks of protein evolution (reviewed by [9]).

Predictions of novel domains are normally derived from one of two general methods. At one extreme a researcher will take a single protein sequence and search for partial matches against other sequences. They can then use these short matches as starting points for building new families. The success and ease of such manual building is often dependant on the experience of the researcher. At the other extreme are the fully automated methods that work on large protein sets. An example is the ProDom database [10], from which Pfam-B is derived. We used two methods to investigate the *S. coelicolor* genome, using a combination of rapid automatic identification of potential novel domains followed by detailed manual analyses. All derived families were deposited in the Pfam database [11].

Method One

A significant mechanism in the evolution of novel proteins is internal duplication. It has been suggested that some types of domain – especially ligand binding domains – often occur tandemly within a protein. Examples of this are PDZ (PF00595), ubiquitin (PF00240) and cadherin domains (PF00028). Self-self comparisons of proteins are a powerful way of taking advantage of this occurrence of internal duplications, providing greater sensitivity than all-against-all searching [12]. The reduction of the number of sequences being compared increases the likelihood that an apparent match is genuine and hence gives an increased sensitivity. An additional advantage is that duplications allow easier recognition of domain boundaries – often a difficult task. The approach described below for domain discovery has in essence been used previously, with noted success (for example see [12]). The following steps describe the procedure that we have implemented to identify novel domains by detecting internal protein duplications. These steps are also described in the flow diagram Figure 1.

Step 1

A set of 7846 potential and known coding sequences from *Streptomyces coelicolor* was used as the starting point. Low complexity regions were masked using 'seg' [13]. A comparison of each protein against itself was carried out using Prospero [14]. Prospero returns the highest scoring self-self matches with an E-value score measuring the significance of each alignment.

Step 2

Highest scoring matches were retained for each sequence and a series of filters were applied to remove matches that are unlikely to be novel domains. Firstly, all matches that had an E-value greater than 0.001 were discarded. Given the size of the Streptomyces coelicolor genome we would expect very few false alignments to be detected at this threshold. Secondly, alignments with a length of less than 30 residues were removed. Thirdly, alignments where the start points of each subsequence were separated by less than 45 residues ('shift') were discarded. Such short duplications are unlikely to be genuine domains. These are more likely to be structural repeats that are not stable in isolation. From this set any that overlapped a Pfam-A family were also discarded unless both subsequences occurred within the boundaries of single Pfam-A family. Such an occurrence indicates that the family contains more than one domain or repeat and needs refining. An overlap is defined as there being residues that occur in both the test alignment and the Pfam-A family alignment.

Step 3

The alignments generated by Prospero were used as an initial alignment to make profile-HMMs using the HMMER 2.2 software [15]. If the pair of sequences in the Prospero alignment overlapped then these overlap regions were removed from the alignment. Profile HMMs were built in local (fs) and global (ls) mode. The resulting profile-HMMs were scanned against the SWISS-PROT and TrEMBL databases [16]. An inclusion threshold of 0.01 was chosen and an alignment of all homologues detected was constructed using the hmmalign program from the HMMER package. This alignment was then compared again to the Pfam-A database to see if the profile-HMM searches had detected any similarities to known families. This step removed 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 analyze. The final step is a careful manual inspection of the family to extend its membership as well as improve the multiple sequence alignment and hopefully to determine the domains function. This analysis uses a wide variety of tools and methods (see below).

Method 2

A complementary method was also used to try to identify novel domains that may be of significance to the biology of *S. coelicolor*. The initial assumption of this process is that short proteins are likely to consist of single domain. Furthermore it seems likely that if a short protein family is represented multiple times in the genome, it should be of

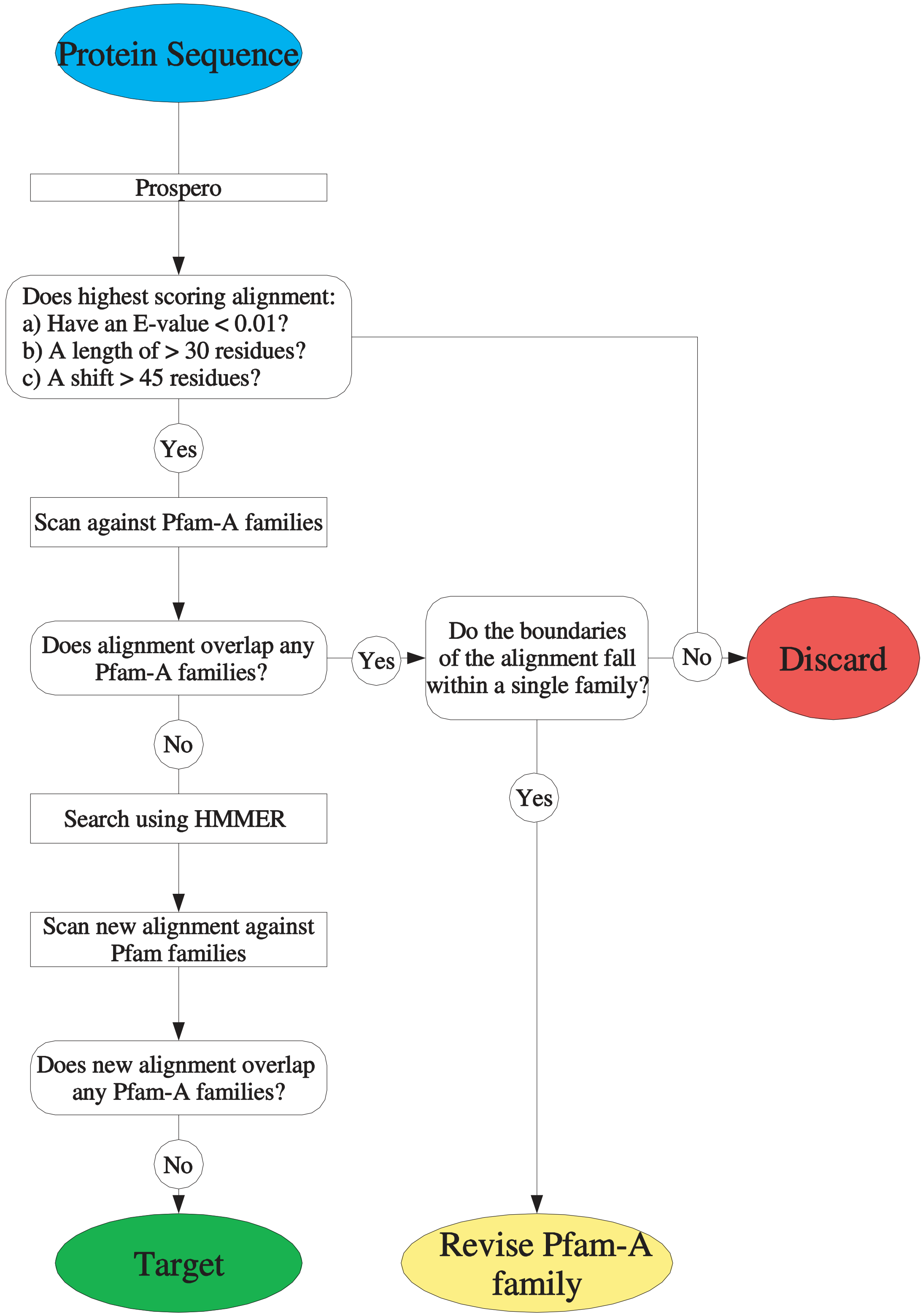

Protein Sequence
Prospero
Does highest scoring alignment:
a) Have an E-value < 0.01?
b) A length of > 30 residues?
c) A shift > 45 residues?
Yes
Scan against Pfam-A families
Does alignment overlap any Pfam-A families?
Yes
Do the boundaries of the alignment fall within a single family?
No
Discard
No
Yes
Search using HMMER
Scan new alignment against Pfam families
Does new alignment overlap any Pfam-A families?
No
Target
Revise Pfam-A family

Figure 1
Flowchart of the domain hunt process. Note: results that end up in the 'Revise Pfam-A' category are not discussed.

some importance. Using these principles we developed a second four-step process:

Step 1

A set of 597 short proteins (≤ 100 residues) was assembled. An all-against-all BLAST was carried out and the proteins clustered using single-linkage clustering with a cut-off threshold of 50 bits, which we determined was sufficiently high to prevent clustering of unrelated proteins.

Step 2

All clusters that corresponded to Pfam-A families and single proteins that did not cluster were then removed from the set. This step also provides a useful check on the stringency of the clustering cut-off score. The clustered sequences were then aligned using T-Coffee [17].

Step 3

The aligned clusters were then used as seeds for an iterative search process using HMMER 2.2, similar to above. The families were iterated until convergence. They were then realigned with T-Coffee and a single round of searching carried out. If any new family members were identified then the iterative search process was repeated.

Step 4

Manual analysis as carried out in Step 4 of Method 1 (also see below).

Software/Servers Used in Manual Analysis

All sequences provided in the alignments were obtained from SWISS-PROT/TrEMBL. Known domains were identified in these sequences using the SMART [18], ProSite [19], Pfam and InterPro [20].

To improve the accuracy of the sequence alignments, the automatic alignment software ClustalW [21] and T-Coffee were employed. These alignments were viewed using Belvu (Sonnhammer ELL) and manually edited with Jalview (Clamp M).

Although our primary interest is in detecting novel domains, other features are of interest. For each sequence in an alignment the following set of programs were run: SignalP [22], for secretory signal peptide prediction; TMHMM [23] to determine likely transmembrane regions; NCOILS [24] to predict coiled-coil regions.

The final domain alignments were submitted to the PredictProtein server and a secondary structure prediction made using PROF [25]. The results are shown in the sequence alignment figures for each domain provided.

In order to determine genomic context the position of the domains in the *S. coelicolor* genome was viewed using the Artemis [26] genome viewer.

Results and Discussion

Overview of the Novel Domains

Method 1 Results

From an initial set of 124 possible domain targets, 31 novel domains were identified, giving a 25% success rate. Sixteen targets were removed by the step 3 of the process. 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 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. SCP has not been previously reported in bacteria.

Method 2 Results

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 identified by both hunt processes. As a positive control the iterative search steps were carried out on the annotated clusters. All produced larger alignments that were simple to further develop to good approximations of the Pfam-A families. When run on the test set of clusters only one family significantly extended – the MbtH family (see below). Three small families (<10 sequences) – GvpG (PF05120), GvpK (PF05121) and spdb (PF05122) – were also produced.

Domains of Significant Interest

Novel Families

HA (Helicase Associated domain; PF03457)

See Figure 2 for example alignment. The domain is typically seventy residues in length and is predicted to have an α-helix-only 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 (Swiss:Q98RX4) in the lower eukaryote *Guillardia theta*. The gene in *C. muridarum* is likely to be a result of a lateral transfer event [27]. Examination of the position of the HA domain-containing proteins, using Artemis, on the *Streptomyces coelicolor* genome revealed a surprising result. From each end of the linear *S. coelicolor* chromosome the second and third ORFs contain HA domains. The second gene from each end is identical

Table 1: List of all domains identified by described methods, as well as their likely function and number in *S. coelicolor*.

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[1]	Domain	Unknown function	4			X		X

Table 1: List of all domains identified by described methods, as well as their likely function and number in *S. coelicolor*. *(Continued)*

PF03459	TOBE[2]	Domain	Transport-associated OB fold domain	9	X		
PF03707	MHYT[3]	Repeat	Putative ligand receptor	6	X		X
PF03989	DNA_gyraseA_C4	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

This table shows all new Pfam families added during this investigation. Part A shows entirely novel families. Part B shows families that are new to Pfam but have been previously described in the literature: (1) SWISS-PROT; (2) [54]; (3) [55]; (4) [56]. Part C shows families that have had significant extensions to them – for instance SCP was previously thought to be present only in eukaryotes. Domains highlighted in blue are discussed in further detail in section 3.2.

to the other (SCO0002 and SCO7845) as are the HA-containing genes third from each end (SCO003 and SCO7844). SCO0002 and SCO7845 have an N-terminal DEAH/D helicase 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://npsa-pbil.ibcp.fr/cgi-bin/primanal_hth.pl). One more gene encoding a single HA domain is found more centrally on the chromosome (SCO0034).

Specific complexes are required for maintaining the ends of the linear streptomycete chromosomes, and the appearance of the genes encoding these domains specifically at the ends suggests that the proteins may be involved in forming these complexes. This is further evidenced by the observation [28] 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 lay between the helicase domain and the HA domains, so it is possible that the helicases still retained functionality. The identification of an HA-containing helicase (SCP1.136) in the SCP1 plasmid, which is also linear and has the same type of telomere, further confirms this hypothesis.

There are no clear conserved catalytic residues in the alignment, 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 (human telomeric protein) against the Pfam SEED alignment with T-Coffee showed interesting similarities between them. Two of the three key tryptophan residues in Myb-like DNA binding domain align to tryptophan residues in HA; in the place of a third is a leucine, which is a structurally conservative replacement. The first helix appears to align well, however the second is longer in HA whereas 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.

BTAD (Bacterial transcriptional activator domain; PF03704)

The following family was an interesting case, and has been previously mentioned as an uncharacterized domain [29]. Although a repeat was detected with an E-value of 4.73 × 10^{-4} using Prospero on the masked sequence, the validity of the repeat could not be verified by other means. However the amino terminal region was related to a number of other bacterial proteins and was investigated further; see Figure 3 for alignment. The BTAD domain is found in small set of bacterial regulatory proteins that occur in the streptomycetes and the closely related Mycobacteria, though one is also found in *Rhizobium loti* (MLR2443/Q98IE9). One of the proteins it is found in – AfsR – is a global secondary metabolite regulator of *S. coelicolor* [30]. 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 (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 [31]; however there are no conserved serines or threonines in the BTAD domain so the phosphorylatyed residues are likely to occur in the DNA-binding domain. A mutation analysis by [32] on

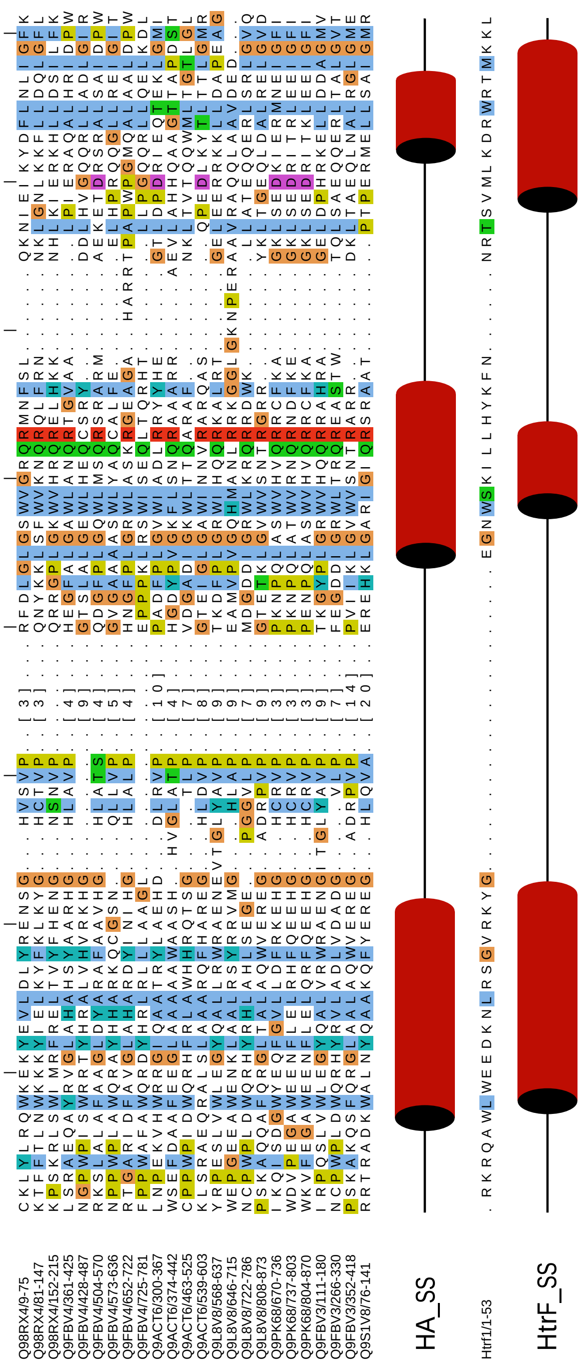


Figure 2

HA domain alignment. All proteins are from *S. coelicolor* except Q9PK68 (*Chlamydia muridarum*), Q9L8V8 (*S. lividans*), Q98RX4 (*Guillardia theta*). The line marked HA_SS is the predicted secondary structure of the HA domain. The line marked HtrF is the secondary structure of the HtrF protein (PDB: 1BA5).

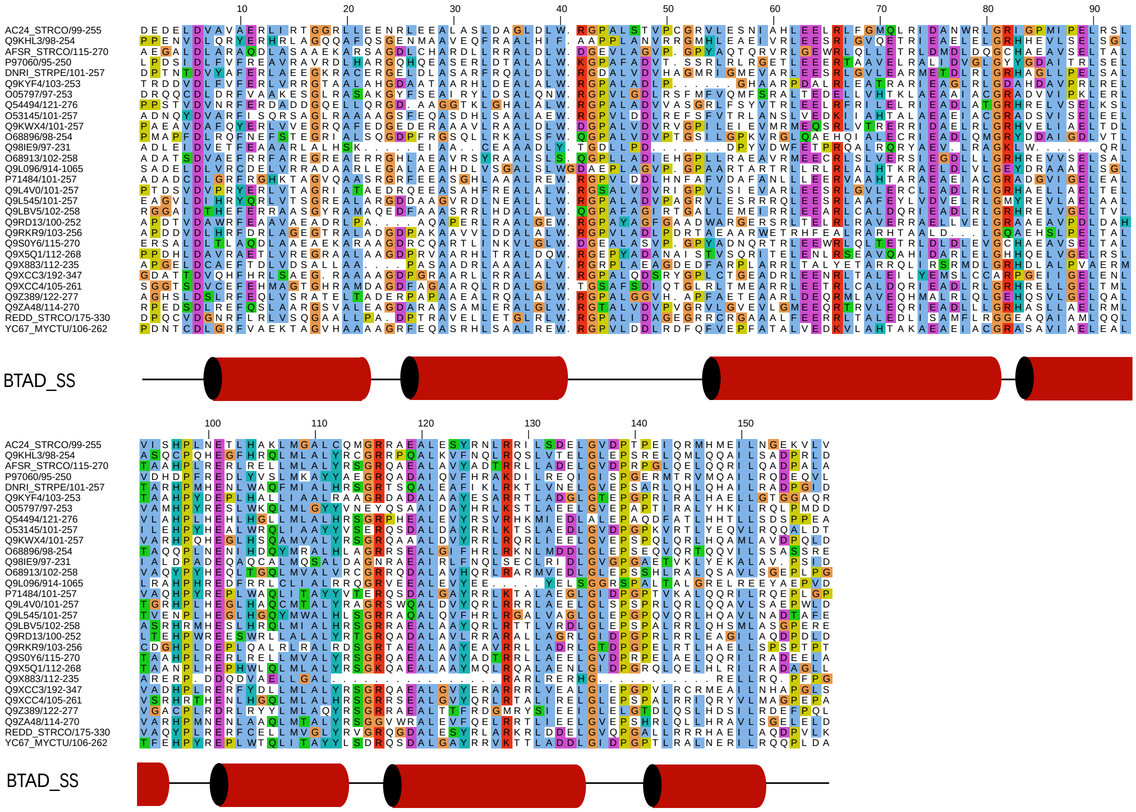


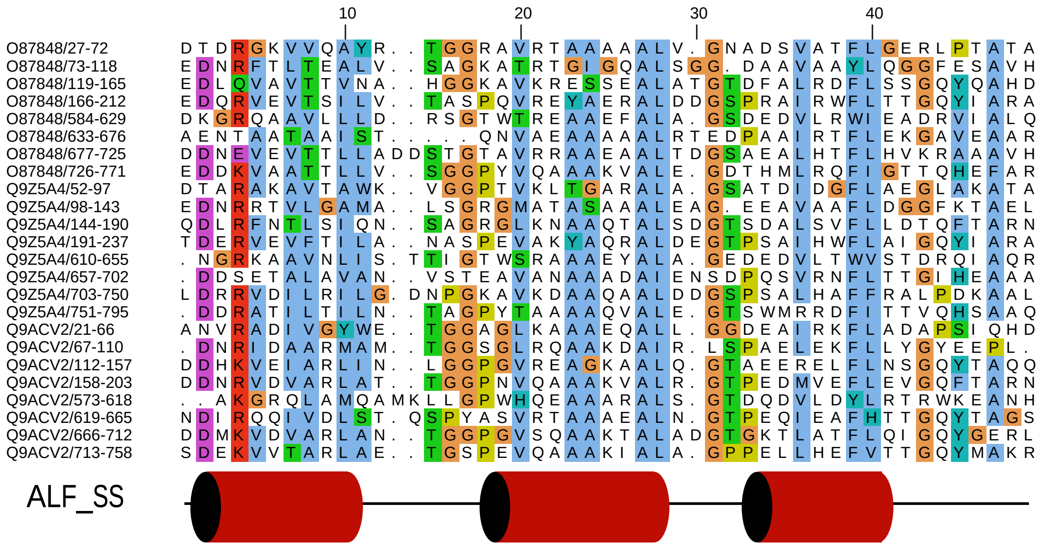
Figure 3

BTAD domain alignment. The predicted secondary structure is shown on the line BTAD_SS.

DnrI 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 mediates interactions between DnrI monomers that are binding tandem repeats in a promoter region. There are eleven pathway-specific regulatory proteins in *S. coelicolor* that contain this domain, including DnrI 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. The repeats have a predicted secondary structure of three α-helices (See Figures 4 &5). The unusual architecture of these proteins is of note. To the C-terminus of each set of repeats is a low-complexity or coiled-coil region. For all three proteins InterProScan http://www.ebi.ac.uk/interpro/scan.html finds a chemotaxis sensory transducer region (IPR:004089; PS50111) between the two ALF-repeat regions. However searching these regions with HMMER 2.2 against SWISS-PROT and TrEMBL found no significant homology to other chemotaxis proteins; similarly using PSI-BLAST at the NCBI found several false-positives (data not shown), but no chemotaxis signal transduction proteins. The sequence in this stretch is very alanine rich, and so could lead to 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. One of the proteins, SCP1.201 (Swiss: Q9ACV2), also contained an intein (N-terminus: SM00306, IPR003587; C-terminus: PS50818, IPR002203) at its C-terminus, which is the first identified in *S. coelicolor*.



```
                     10        20        30        40
                     |         |         |         |
O87848/27-72     DTDRGKVVQAYR..TGGRAVRTAAAAALV.GNADSVATFLGERLPTATA
O87848/73-118    EDNRFTLTEALV..SAGKATRTGIGQALSGG.DAAVAAYLQGGFESAVH
O87848/119-165   EDLQVAVTTVNA..HGGKAVKRESSEALATGTDFALRDFLSSGQYQAHD
O87848/166-212   EDQRVEVTSILV..TASPQVREYAERALDDGSPRAIRWFLTTGQYIARA
O87848/584-629   DKGRQAAVLLLD..RSGTWTREAAEFALA.GSDEDVLRWIEADRVIALQ
O87848/633-676   AENTAATAAIST....QNVAEAAAAALRTEDPAAIRTFLEKGAVEAAR
O87848/677-725   DDNEVEVTTLLADDSTGTAVRRAAEAALTDGSAEALHTFLHVKRAAAVH
O87848/726-771   EDDKVAATTLLV..SGGPYVQAAAKVALE.GDTHMLRQFIGTTQHEFAR
Q9Z5A4/52-97     DTARAKAVTAWK..VGGPTVKLTGARALA.GSATDIDGFLAEGLAKATA
Q9Z5A4/98-143    EDNRRTVLGAMA..LSGRGMATASAAALEAG.EEAVAAFLDGGFKTAEL
Q9Z5A4/144-190   QDLRFNTLSIQN..SAGRGLKNAAQTALSDGTSDALSVFLLDTQFTARN
Q9Z5A4/191-237   TDERVEVFTILA..NASPEVAKYAQRALDEGTPSAIHWFLAIGQYIARA
Q9Z5A4/610-655   .NGRKAAVNLIS.TTIGTWSRAAAEYALA.GEDEDVLTWVSTDRQIAQR
Q9Z5A4/657-702   .DDSETALAVAN..VSTEAVANAAADAIENSDPQSVRNFLTTGIHEAAA
Q9Z5A4/703-750   LDRRVDILRILG.DNPGKAVKDAAQAALDDGSPSALHAFFRALPDKAAL
Q9Z5A4/751-795   .DDRATILTILN..TAGPYTAAAAQVALE.GTSWMRRDFITTVQHSAAQ
Q9ACV2/21-66     ANVRADIVGYWE..TGGAGLKAAAEQALL.GGDEAIRKFLADAPSIQHD
Q9ACV2/67-110    .DNRIDAARMAM..TGGSGLRQAAKDAIR.LSPAELEKFLLYGYEEPL.
Q9ACV2/112-157   DDHKVEIARLIN..LGGPGVREAGKAALQ.GTAEERELFLNSGQYTAQQ
Q9ACV2/158-203   DDNRVDVARLAT..TGGPNVQAAAKVALR.GTPEDMVEFLEVGQFTARN
Q9ACV2/573-618   ..AKGRQLAMQAMKLLGPWHQEAAARALS.GTDQDVLDYLRTRWKEANH
Q9ACV2/619-665   NDIRQQIVDLST.QSPYASVRTAAAEALN.GTPEQIEAFHTTGQYTAGS
Q9ACV2/666-712   DDMKVDVARLAN..TGGPGVSQAAKTALADGTGKTLATFLQIGQYGERL
Q9ACV2/713-758   SDEKVVTARLAE..TGSPEVQAAAKIALA.GPPELLHEFVTTGQYMAKR
```

Figure 4
ALF repeat alignment. Predicted secondary structure is shown on the line ALF_SS.

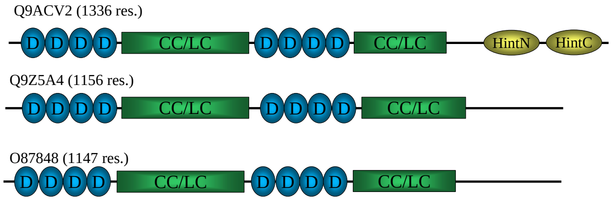


Figure 5
Domain architectures of the ALF-containing proteins. ALF repeats are represented by the blue ovals; the coiled-coil/low complexity regions are signified by the green boxes; Intein N and C-terminal domains are indicated by the yellow ovals.

Two of the proteins, SCO6198 (Swiss: Q9Z5A4) and SCO6593 (Swiss: O87848), are located on the chromosome 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 likely that they represent a conserved pathway and that the ALF regions act as a substrate- or product-recognition domain that passes a signal to or from the secreted esterases. The intein 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

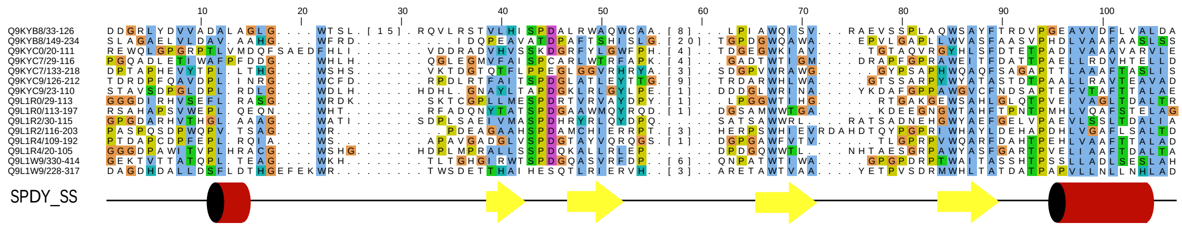

10
20
30
40
50
60
70
80
90
100
Q9KYB8/33-126 DDGRLYDVVADALAGLG....WTSL.[15].RQVLRSTVLHISPDALRWAQWCAA.[8]...LPIAWQISV...RAEVSSPLAQWSAYFTRDVPGEAVVDFLVALDA
Q9KYB8/149-234 SLAGAELVLDAV.AAHG....WFRD............IDQPEAVATDPAFTSHISLG.[20].TGPDGWQAWA....EPVLGAPLLWVASFAASVPHDLVAAFAASLSS
Q9KYC0/20-111 REWQLGPGRPTLVMDQFSAEDFHLI.......VDDRADVHVSSKDGRFYLGWFPH.[4]..TDGEGWKIAV.....TGTAQVRGYHLSFDTETPADIVAAAVARVLE
Q9KYC7/29-116 PGQADLETIWAFPFDDG....WHLH.......QGLEGMVFAISPCARLWTRFAPK.[4]...GADTWTIGM....DRAPFGPRAWEITFDATTPAELLRDVHTELLD
Q9KYC7/133-218 DPTAPHEVYTPL.LTHG....WSHS.......VKTDGTQTFLPPEGLGGVRHRYA.[3]..SDGPVWRAWG.......GYPSAPHWQAQFSAGAPTTLAAAFTASLIS
Q9KYC9/126-212 TDRDPFQAVDPL.INRG....WCFD.......RPDLRTFAITSPDGLATLEYTTG.[9]..TRDARWHLWA......GTSSARPYWYATASTDTPAALLRAVTEAVAD
Q9KYC9/23-110 STAVSDPGLDPL.RDLG....WDLH......HDHL.GNAYLTAPDGKLRLGYLPE.[1]..DDDGLWRINA....YKDAFGPPAWGVCFNDSAPTEFVTAFTTALAE
Q9L1R0/29-113 GGGDIRHVSEFL.RASG....WRDK.......SKTCGPLLMESPDRTVRVAYDPY.[1]...LPGGWTIHG.......RTGAKGEWSAHLGLQTPVEIVAGLTDALTR
Q9L1R0/113-197 RSAHAPSVWEPL.QEQN....WHT........RFADQNYTATSPDGAAWMQYRQD.[1]..DGSAMWWTGA.......KDEEGNGWTAHFTPNTPMHLVQAFSTELAG
Q9L1R2/30-115 GPGDARHVTHGL.AAAG....WATI.......SDPLSAEIVMASPDHRYRLQYDPQ.......SATSAWWRL....RATSADNEHGWYAEFGELVPAEVLSSLTDALIA
Q9L1R2/116-203 PASPQSDPWQPV.TSAG....WR..........PDEAGAAHSPDAMCHIERRPT.[3]..HERPSWHIEVRDAHDTQYPGPRIWHAYLDEHAPDHLVGAFLSALTD
Q9L1R4/109-192 PTDAPCDPFEPL.RQIA....WS........APAVGADGLVSPDGTAYVQRQGS.[1]..DGPGAWFVTV.......TLGPRPVWQARFDAHTPEHLVAAFTTALAD
Q9L1R4/20-105 GGGDPAWITVPLHRACG....WSHG......HDPLMPRALLSSPDQKALLRLEP.......DPDGQWWTL....NHTAESGRPAWYASFGARTPVELIAAFTDALTA
Q9L1W9/330-414 GEKTVTTATQPL.TEAG....WKH.......TLTGHGIRWTSPDGQASVRFDP..[6]..QNPATWTIWA......GPGPDRPTWAITASSHTPSSLLADLSESLAH
Q9L1W9/228-317 DAGDHDALLDSFLDTHGEFEKWR........TWSDETTHAIHESQTLRIERVH..[3]..ARETAWTVAA....YETPVSDRMWHLTATDATPAPVLLNLLNHLAD
SPDY_SS

Figure 6
SPDY domain alignment. Predicted secondary structure is shown on the line SPDY_SS.

in the *S. coelicolor* genome. This implies that the plasmid has passaged through another species that has mobile intein elements.

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 6). 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 [1]. The region appears to consist of two sections: a 'core' mobile element region with the essential replication genes and a flanking region containing a polyketide synthase and arsenic resistance genes. 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 [33]. It is a small (~70 residues) globular domain that binds cell wall peptidoglycan. With regards to *S. coelicolor*'s genome it shows an unusual distribution. Typically organisms that have PASTA domains have one PASTA-containing serine/threonine protein kinase (pPSTK), which is, putatively, the master regulator of cell wall peptidoglycan cross-linking and essential to growth and development, and one PASTA-containing penicillin-binding protein (pPBP), which is the primary cross-linking enzyme. For a type example see *Streptococcus pneumoniae*. 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. This also fits there being no pPBP as it would be specific to a single peptidoglycan structure; so we propose it uses an alternative localisation system, perhaps similar to that used by *Deinococcus radiodurans* or Gram-ve bacteria. Intriguingly *S. coelicolor* has three principle cell morphologies and it may be that each pPSTK regulates the development of each type.

HHE (Histidine-Histidine-Glutamate motif; PF03794)

This domain normally occurs as tandem repeats, is approximately 70 residues in length, and is predicted to be composed of 2 α-helices (See Figure 7). It is mostly found in prokaryotes, though four *Arabidopsis* proteins were identified with multiple HHE repeats and a *Schizosaccharomyces pombe* protein. Typically an HHE-containing protein consists of two HHE domains only, though there are exceptions like the *Arabidopsis* proteins (e.g. Q9LJQ1). There are two conserved histidines, both in the middle of predicted helices, and a conserved glutamate. It shows a slightly disparate phylogenetic distribution, but is found in eubacteria, archaea, fungi and plants. In several cases it appears to be involved in NO response – for instance DnrN from *Pseudomonas stutzeri* [34]. Deletion of *dnrN* leads to slower response to nitrite of the *nirSTB* operon, so it may be involved in regulation or signal recognition. However, in *Ralstonia eutropha* deletion of the HHE-containing genes *norA1* and *norA2*, despite being co-transcribed with the NO reductases-encoding *norB1* and *norB2*, does not appear to affect growth or ability to cope with NO stress [35]. It is also found in the ScdA protein of *Staphylococcus aureus*, which has been implicated in growth, development, and peptidoglycan cross-linking [36]. The two conserved histidines and the glutamate are suggestive of a cation-binding site, such as the binding of Zn^{2+} in Carboxypeptidase A. This hypothesis is supported

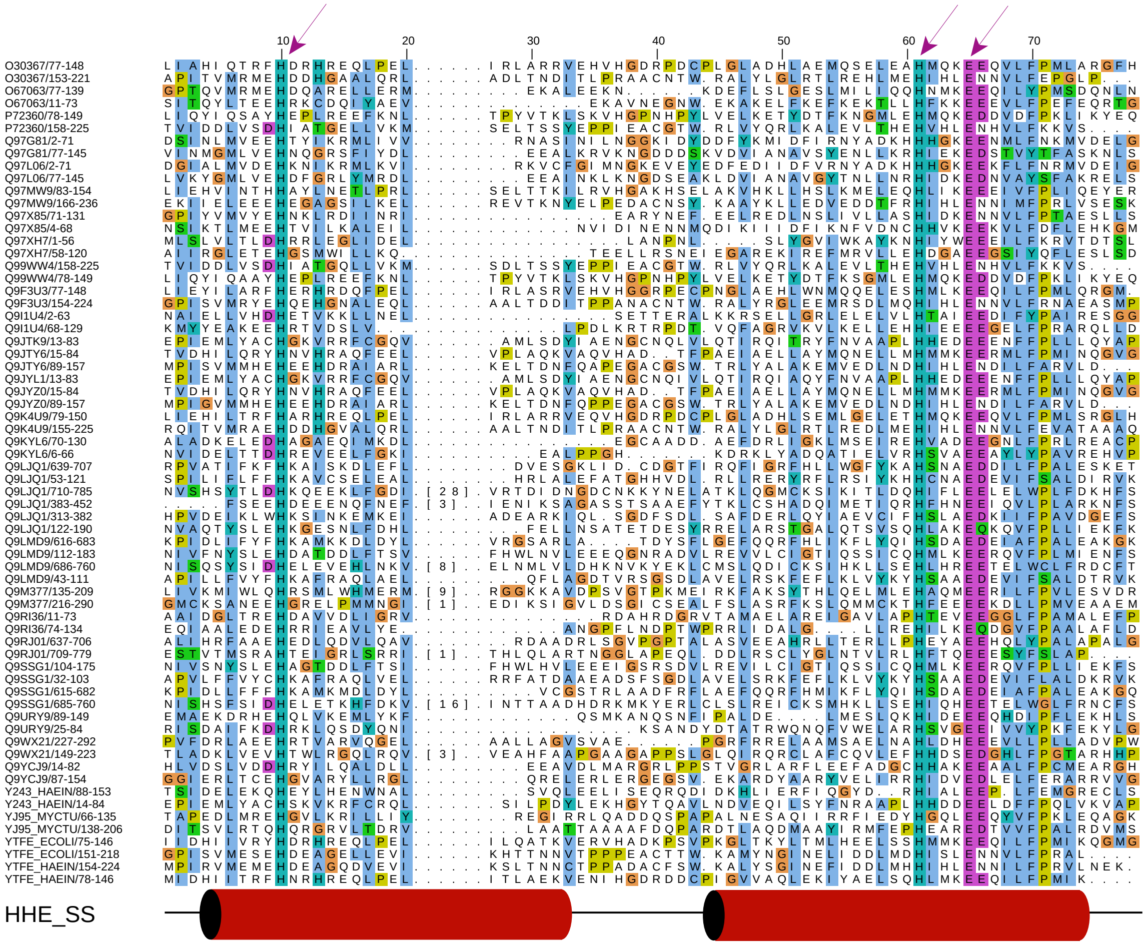


Figure 7
HHE domain alignment. The predicted secondary structure is shown in the line marked HHE_SS. The conserved histidines and glutamate are indicated with purple arrows.

by its occurrence in the putative cation-transporting ATPase SCO0164 (Swiss:Q9RJ01) where it might sequester cations for transport.

PPC (Bacterial Pre-peptidase C-terminal domain; PF04151)

These domains are typically ninety residues in length and found at the C-termini of secreted peptidases (See Figure 8). Surprisingly these domains are found in at least four different classes of peptidases. The PPC domain is found in some members of metallopeptidase families M4, M9 and M28 as well as the serine peptidase family S8 [37]. The PPC domains are cleaved off subsequent to secretion, but prior to activation of the peptidase. The actual function of them is not clear but they may aid secretion/localisation or inhibit the peptidase until needed. Visual inspection of the alignment, as well as predicted similarities in the secondary structure, suggests that it may be related to the PKD domain (PF00801), but no significant homology was detected using computational methods. They are often found in the same protein as the PKD domain and in very similar contexts, and it is tempting to suggest that they are functionally interchangeable (see Figure 9 for example domain architectures). PKD domains are thought to be involved in protein-protein interactions.

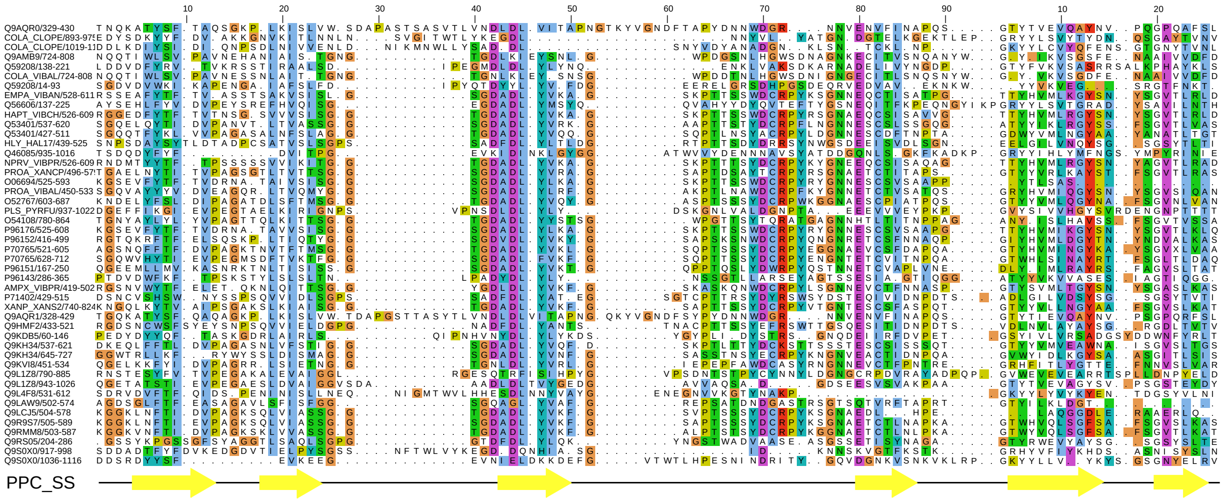


Figure 8
PPC domain alignment. The predicted secondary structure is shown in the line marked PPC_SS.

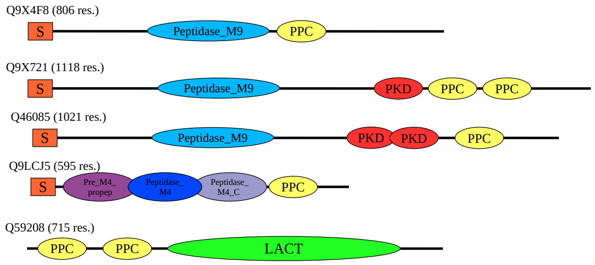


Figure 9
Example domain architectures of PPC-containing proteins. Example collagenase precursors Q9X4F8 (*Vibrio cholerae*), Q9X721 (*Clostridium histolyticum*), Q46085 (*Clostridium histolyticum*) and O54108 (*S. coelicolor*; SCO5912) demonstrate the apparent interchangeability of PPC and PKD domains. Q9LCJ5 (Protease precursor; *Aeromonas punctata*) represents a common protease architecture. Q59208 (esterase; *Bacillus licheniformis*) is an example of the PPC domain occurring at the N-terminus rather than the C-terminus. Domain names shown are Pfam identifiers.

Unlike the PKD domain the PPC domain is only found in bacteria and archaea, and not in eukaryotes.

FMN_bind (Flavin MonoNucleotide-binding; PF04205)

This domain represents a sixty residue region that includes an FMN-binding site (indicated in alignment, Figure 10), as determined in the NqrC proteins of *Vibrio cholerae* [38] and *Vibrio alginolyticus* [39]. Interestingly the NqrB proteins, which also bind FMN through a threonine residue and are part of the same complex, do not show any homology. The region is found in several electron transport chain proteins; for example the RnfG electron transport protein, part of a chain that supplies electrons to both nitrogen fixation and DNP reduction in *Rhodobacter capsulatus* [40]. Other examples include the NosR/NirI nitrous oxide reduction regulatory proteins. FMN_bind-containing proteins appear to split into two groups, which relate length to function. The shorter proteins, typically 200–350 residues, are components of electron transport chains whereas the longer proteins, typically 680–800 residues, have a regulatory function. The regulatory proteins typically have five transmembrane helices in the C-terminal half of the protein. Members of both groups often have 4Fe-4S domains present, suggesting that the regulatory mechanisms also involve charge movement.

MbtH (MbtH-like proteins; PF03621)

This domain is named after the MbtH protein from *Mycobacterium tuberculosis* (Swiss: O05821). The domain is

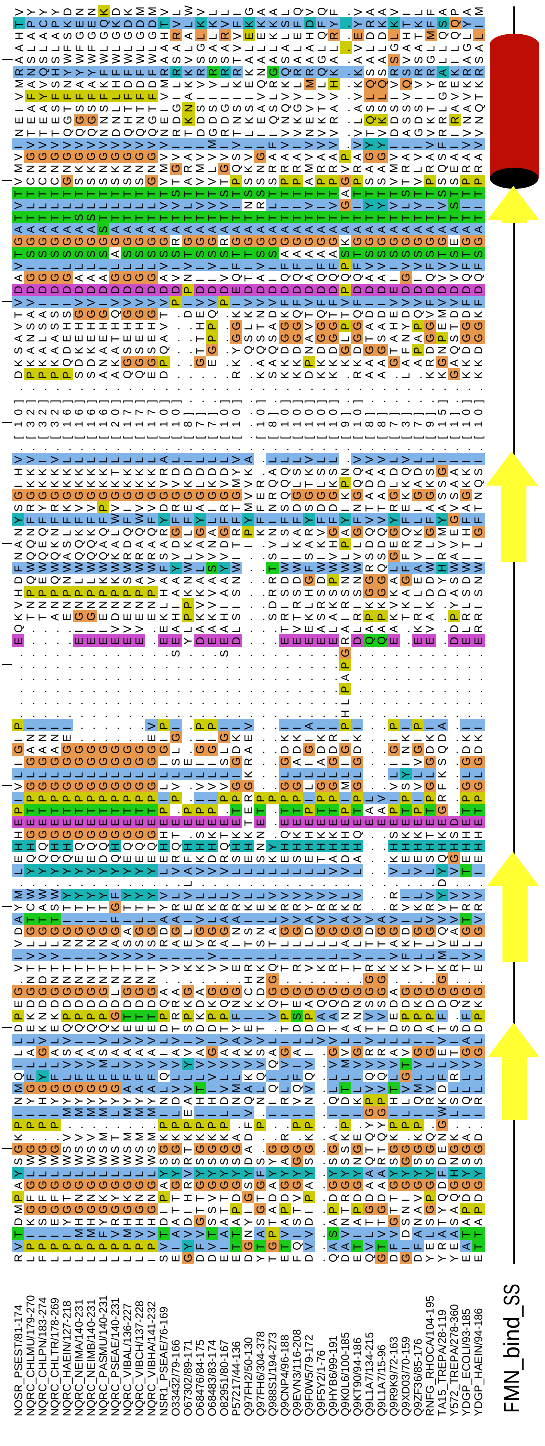


Figure 10
FMN_bind domain alignment. The FMN-binding residue is indicated by the green arrow. The predicted secondary structure is shown in the line marked FMN_bind_SS.

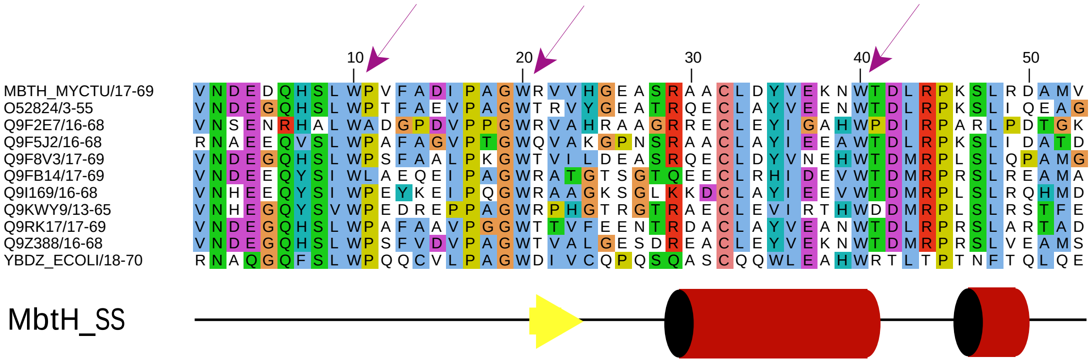

10
20
30
40
50
MBTH_MYCTU/17-69 VNDEDQHSLWPVFADIPAGWRVVHGEASRAACLDYVEKNWTDLRPKSLRDAMV
O52824/3-55 VNDEGQHSLWPTFAEVPAGWTRVYGEATRQECLAYVEENWTDLRPKSLIQEAG
Q9F2E7/16-68 VNSENRHALWADGPDVPPGWRVAHRAAGRRECLEYIGAHWPDIRPARLPDTGK
Q9F5J2/16-68 RNAEEQVSLWPAFAGVPTGWQVAKGPNSRAACLAYIEEAWTDLRPKSLIDATD
Q9F8V3/17-69 VNDEGQHSLWPSFAALPKGWTVILDEASRQECLDYVNEHWTDMRPLSLQPAMG
Q9FB14/17-69 VNDEEQYSIWLAEQEIPAGWRATGTSGTQEECLRHIDEVWTDMRPRSLREAMA
Q9I169/16-68 VNHEEQYSIWPEYKEIPQGWRAAGKSGLKKDCLAYIEEVWTDMRPLSLRQHMD
Q9KWY9/13-65 VNHEGQYSVWPEDREPPAGWRPHGTRGTRAECLEVIRTHWDDMRPLSLRSTFE
Q9RK17/17-69 VNDEGQHSLWPAFAAVPGGWTTVFEENTRDACLAYVEANWTDLRPRSLARTAD
Q9Z388/16-68 VNDEGQHSLWPSFVDVPAGWTVALGESDREACLEYVEKNWTDMRPRSLVEAMS
YBDZ_ECOLI/18-70 RNAQGQFSLWPQQCVLPAGWDIVCQPQSQASCQQWLEAHWRTLTPTNFTQLQE
MbtH_SS

Figure 11
MbtH domain alignment. Conserved tryptophans are marked with purple arrows. Predicted secondary structure is shown on the line MbtH_SS

typically 70 residues in length and covers the full length of the protein, though NikP1 from *Streptomyces tendae* (Swiss: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 *Rhizobium leguminosarum*. Several of these proteins have been implicated in antibiotic biosynthesis in several streptomycetes (for instance nikkomycins: [41]; simocyclinone: [42]; coumermycin A1: [43], and the formation of siderophores such as *E. coli*'s enterobactin or *M. tuberculosis*'s mycobactin (reviewed in [44]). 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 expression or transport of the siderophores out of the cell, with a similar role in antibiotic synthesis. There are several conserved residues, including three tryptophans that may have functional importance (See alignment in Figure 11).

Extended Families

SCP (PF00188)

This domain family has previously only been reported in eukaryotes, but in fact it contains a diverged sub-group that occurs in eubacteria as well. An alignment of the eukaryotic and prokaryotic versions show that the principle difference is the absence in bacteria of the conserved cysteine residues, which form disulphide bridges, whereas the proposed active site [45] (see Figure 12) is mostly conserved. In order to try and determine its function in bacteria a review of the information available for the eukaryotic domains was carried out.

So far all SCP-containing proteins appear to be secreted, and this is backed up by the consistent prediction of signal peptides at the N-terminus. There is very little direct evidence of their general function currently, however many examples have been found to be involved in signaling. For instance they are involved in several mammalian developmental processes, most notably sperm maturation [46] and sperm-egg fusion [47], and are up-regulated in several tumors ([48,49]). Clear evidence has been found, in *Xenopus*, of sperm following the concentration of 'Allurin' – an SCP-containing protein [50]. They are also commonly used by insects and reptiles as mammalian toxins.

However these proteins are very big for direct signaling molecules, typically being 200 or 400 residues (1 or 2 SCP domains). 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 [45], and three of the four residues predicted to make up the site are conserved between the eukaryotic and prokaryotic subfamilies (See

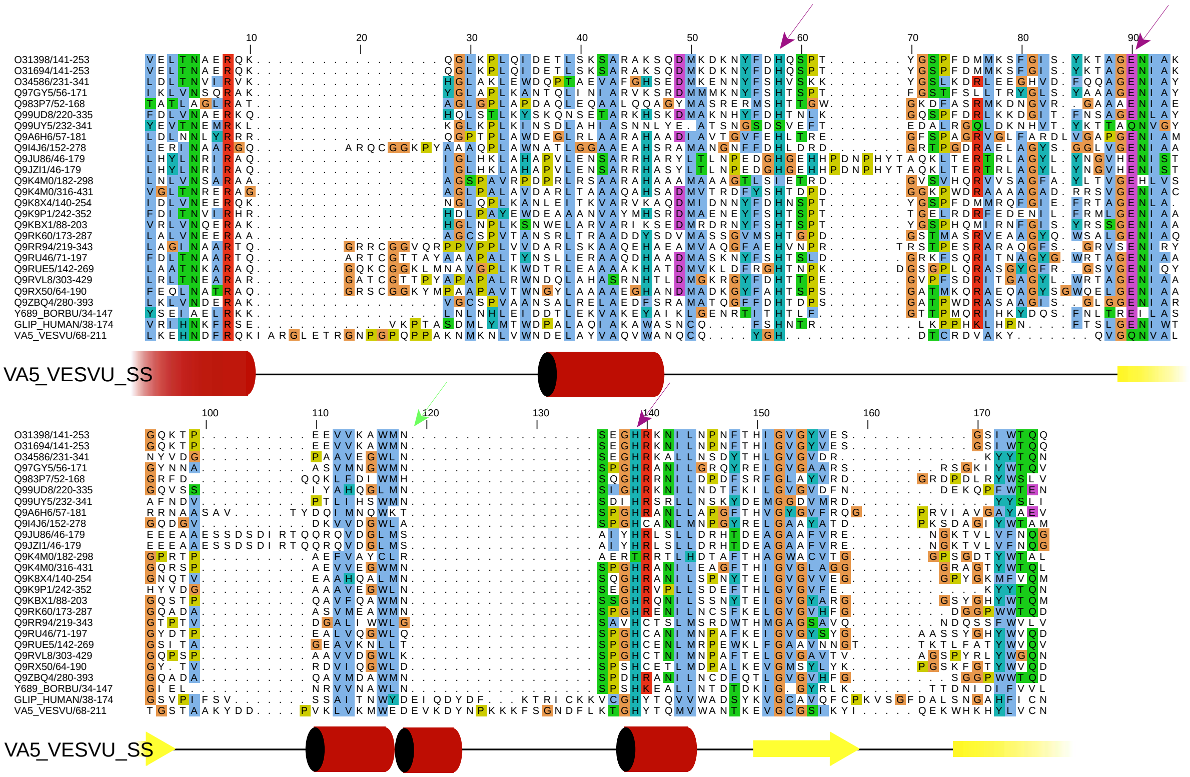


Figure 12

SCP domain alignment. All sequences shown are prokaryotic except GLIP_HUMAN and VA5_VESVU. The predicted active site residues, based on analysis of the eukaryotic domain [45], are marked by green or purple arrows. These residues are also almost fully conserved in the prokaryotic sequences except one which falls into an insert region not in the prokaryotic domains, which is marked by the green arrow. The secondary structure of the eukaryotic domain is shown on the line VA5_VESVU_SS.

Figure 12). This would imply that the domain generates a smaller signaling molecule. However no evidence has been found of such a molecule and several pieces of evidence conflict with this hypothesis. Firstly the nematode SCP-containing Neutrophil Inhibitory Factor (NIH) binds directly to integrins CD11b/CD18 on the neutrophil cell surface [51]. Secondly pseudochetoxin (from King brown snake) appears to bind the extracellular portion of cyclic-nucleotide gated ion channels (CNG channels) blocking their function [52]. In the second case there does appear to be time-lag between association or disassociation and blocking or release of the gate. This does seem to suggest that its mode of action is not simply as a steric block.

SCP-containing proteins are involved in a tremendously wide range of processes, and found to be essential in plants (PR1-like proteins), mammals, lizards, insects (venom allergens) and nematodes. It appears likely that they are similarly important and similarly multi-function in bacteria, and hence are an important target for further analysis.

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; the updated family in Pfam 7.5 is found in thirty nine bacterial proteins – including fourteen in S. coelicolor (see Figure 13 for distribution of FG-GAP repeats in bacteria). An extra thirty-four eukaryotic family members are also identified, as well as an archaeal protein (Swiss:O28333). The FG-GAP repeats have been predicted

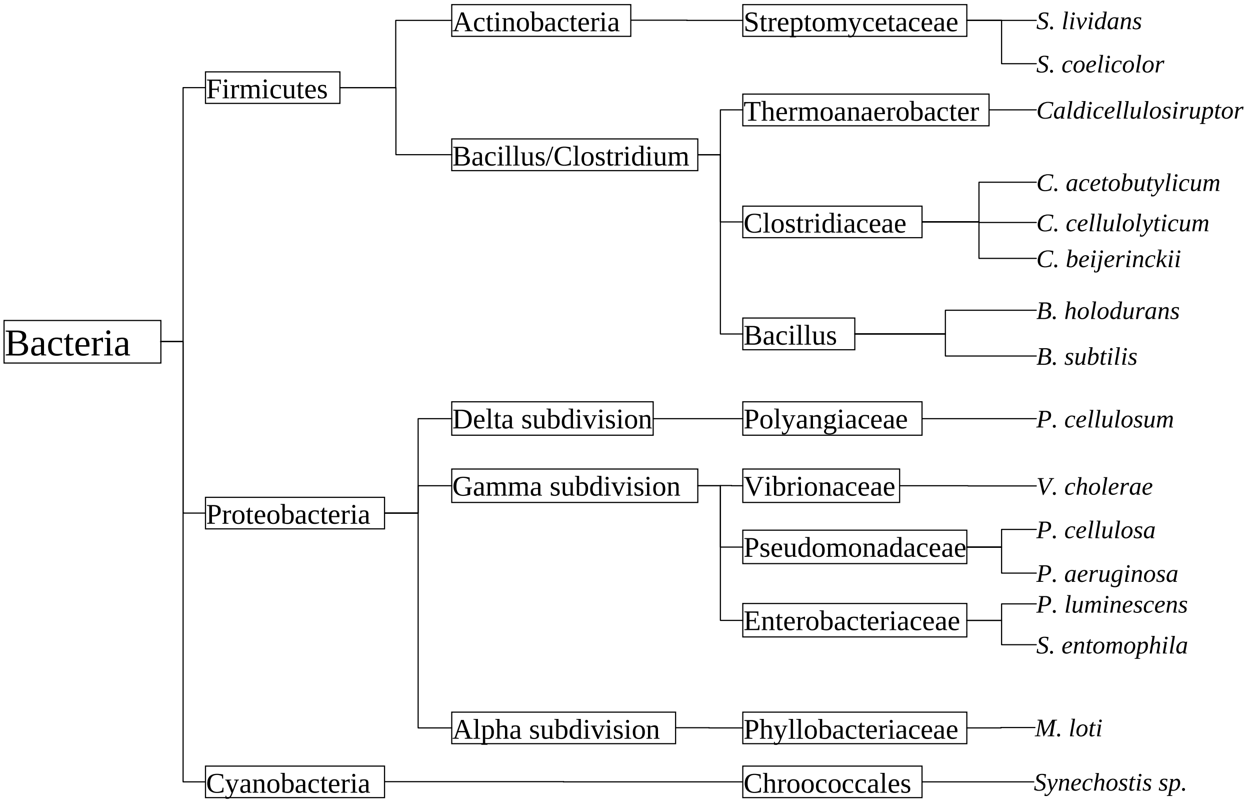


Figure 13
Species tree showing the distribution of FG-GAP proteins in bacteria according to Pfam 7.5. The broad distribution indicates that more thorough searching may find them to be ubiquitous.

to assume a β-propeller conformation. The occurrence of this repeat as sets of four or five tandem copies casts doubt on this (e.g. Swiss:ITA2_DROME), as they are normally six or seven bladed [3]. However the hemopexin repeat (PF00045) has been seen as a four-bladed propeller e.g. as in mammalian blood serum haemopexin glycosylated-native protein (PDB:1qjs), so perhaps FG-GAP repeats might be more structurally similar to these repeats.

Conclusions

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*. To manually investigate every single protein is an immensely time-consuming enterprise, and it would not be possible to add significant annotation to many of the families built. However fully automatic methods of family building lack precision, and the automated production of detailed annotation is currently not feasible. Hence we employed a combination in an LHF ("low-hanging fruit") process in order to concentrate on potentially the most interesting observations.

To underline the speed of this approach there are 204 copies of the novel domains listed in Table 1 in *S. coelicolor*, 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 very 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, 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. We hypothesise that the HA domains bind DNA, most likely telomere-specific structures, based on secondary structural similarities to the Myb-like DNA-binding (PF00249) domain. 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 the extension of the SCP domain into prokaryotes does appear to be significant. The substantial differences in sequence conservation suggest that the prokaryotic versions are not simply the product of lateral transfers, but are of ancient origin. The lack of conservation of the cysteines, after which the domain was originally named, suggests that they are not functionally important but are involved in stabilizing the protein over the greater distances involved in eukaryotic signaling. In contrast the conservation of three of the four proposed active site residues confirms that these are the functionally significant residues. The apparent importance of SCPs in eukaryotes suggests that these domains will prove to be similarly important in bacteria.

It is important to recognize when basing future work on bioinformatic studies such as this one, that the results are sets of hypotheses rather than true descriptions. This does not detract from the success of such approaches. Previously a researcher investigating an HHE-containing protein would have known little about it apart from the sequence; now three strong candidates for the functional or active site residues are clear and a putative function (cation-binding) assigned that can be tested. 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.

Supplementary Information

S1: Architecture diagrams for all HA, BTAD, SPDY, PASTA and HHE domain-containing proteins in *S. coelicolor*, as well as other proteins referred to in the text. Architectures are based on data from Pfam and SMART. Domain names are as given in Pfam and SMART. Small orange boxes at the N-termini indicate signal peptide sequences. TM indicates transmembrane regions.

S2: Architecture diagrams for all PPC, FMN_bind and MbtH domain-containing proteins in *S. coelicolor*, as well as other proteins referred to in the text. Architectures are based on data from Pfam and SMART. Small orange boxes at the N-termini indicate signal peptide sequences. TM indicates transmembrane regions.

Author's Contributions

CY carried was involved in all aspects of the work. SB was involved in annotation of the families, providing *S. coelicolor*-specific information and supervising writing of the manuscript. AB designed the search methodologies and provided advice on family annotation as well as supervising writing of the manuscript.

Correction

Subsequent to submission of this manuscript it came to the authors' attention that SCP domains have previously been described in prokaryotes [53]; so this section should be considered as a formal report of the prokaryotic version rather than an initial observation.

Abbreviations

Pfam accession numbers are indicated by '(PF#####)', # represents a numeral.

Acknowledgments

The Pfam group for help with software use and interaction with the Pfam database. Robert Finn for his previous work on the PASTA domain.

References

1. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H and Harper D **Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2).** *Nature* 2002, **417:**141-147
2. Bateman A and Birney E **Searching databases to find protein domain organization.** *Advances in Protein Chemistry, Vol 54* 2000, **54:**137-157
3. Murzin AG **Structural Principles for the Propeller Assembly of Beta-Sheets – the Preference for 7-Fold Symmetry.** *Proteins-Structure Function and Genetics* 1992, **14:**191-201
4. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25:**3389-3402
5. Pearson WR **Rapid and Sensitive Sequence Comparison with Fastp and Fasta.** *Methods in Enzymology* 1990, **183:**63-98
6. Bult CJ, White O, Olsen GJ, Zhou LX, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA and Gocayne JD **Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.** *Science* 1996, **273:**1058-1073
7. Galperin MY and Koonin EV **Sources of systematic error in functional annotation of genomes: domain rearrangement, nonorthologous gene displacement and operon disruption.** *In Silico Biology* 1998, **1:**55-67

8. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417:**399-403
9. Ponting CP and Russell RR **The natural history of protein domains.** *Annual Review of Biophysics and Biomolecular Structure* 2002, **31:**45-71
10. Sonnhammer ELL and Kahn D **Modular Arrangement of Proteins as Inferred from Analysis of Homology.** *Protein Science* 1994, **3:**482-492
11. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL **The Pfam Protein Families Database.** *Nucleic Acids Research* 2002, **30:**276-280
12. Ponting CP, Mott R, Bork P and Copley RR **Novel protein domains and repeats in Drosophila melanogaster: Insights into structure, function, and evolution.** *Genome Research* 2001, **11:**1996-2008
13. Wootton JC and Federhen S **Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases.** *Computers & Chemistry* 1993, **17:**149-163
14. Mott R 2000, [http://www.well.ox.ac.uk/rmott/ARIADNE/prospero.shtml]
15. Eddy SR **HMMER: Profile hidden Markov models for biological sequence analysis.** 2001,
16. Bairoch A and Apweiler R **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Research* 2000, **28:**45-48
17. Notredame C, Higgins DG and Heringa J **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *Journal of Molecular Biology* 2000, **302:**205-217
18. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP and Bork P **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Research* 2002, **30:**242-244
19. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K and Bairoch A **The PROSITE database, its status in 2002.** *Nucleic Acids Research* 2002, **30:**235-238
20. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti T, Corpet F and Croning MDR **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Research* 2001, **29:**37-40
21. Thompson JD, Higgins DG and Gibson TJ **Clustal-W – Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.** *Nucleic Acids Research* 1994, **22:**4673-4680
22. Nielsen H, Engelbrecht J, Brunak S and von Heijne G **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Engineering* 1997, **10:**1-6
23. Krogh A, Larsson B, von Heijne G and Sonnhammer ELL **Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.** *Journal of Molecular Biology* 2001, **305:**567-580
24. Lupas A, Vandyke M and Stock J **Predicting Coiled Coils from Protein Sequences.** *Science* 1991, **252:**1162-1164
25. Rost B **PHD: Predicting one-dimensional protein structure by profile-based neural networks.** *In: Computer Methods for Macromolecular Sequence Analysis* 1996, **266:**525-539
26. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA and Barrell B **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945
27. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T and Berry K **Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39.** *Nucleic Acids Research* 2000, **28:**1397-1406
28. Bey SJ, Tsou MF, Huang CH, Yang CC and Chen CW **The homologous terminal sequence of the Streptomyces lividans chromosome and SLP2 plasmid.** *Microbiology-Uk* 2000, **146:**911-922
29. Aravind L, Dixit VM and Koonin EV **The domains of death: evolution of the apoptosis machinery.** *Trends in Biochemical Sciences* 1999, **24:**47-53
30. Floriano B and Bibb M **afsR is a pleiotropic but conditionally required regulatory gene for antibiotic production in Streptomyces coelicolor A3(2).** *Molecular Microbiology* 1996, **21:**385-396
31. Umeyama T, Lee PC and Horinouchi S **Protein serine/threonine kinases in signal transduction for secondary metabolism and morphogenesis in Streptomyces.** *Applied Microbiology and Biotechnology* 2002, **59:**419-425
32. Sheldon PJ, Busarow SB and Hutchinson CR **Mapping the DNA-binding domain and target sequences of the *Streptomyces peucitis* daunorubicin biosynthesis regulatory protein DnrI.** *Molecular Microbiology* 2002, **44:**449-460
33. Yeats C, Finn RD and Bateman A **The PASTA domain: a beta-lactam-binding domain.** *Trends in Biochemical Sciences* 2002, **27:**438-440
34. Vollack K and Zumft WG **Nitric oxide signalling and transcriptional control of denitrification genes in *Pseudomonas stutzeri*.** *Journal of Bacteriology* 2001, **183:**2516-2526
35. Pohlmann A, Cramm R, Schmeiz K and Friedrich B **A novel NO-repsonding regulator controls the reduction of nitric oxide in *Ralstonia eutropha*.** *Molecular Microbiology* 2000, **38:**626-38
36. Brunskill EW, de Jonge BLM and Bayles KW **The *Staphylococcus aureus* scdA gene: a novel locus that affects cell division and morphogenesis.** *Microbiology* 1997, **38:**626-638
37. Rawlings ND, O'Brien E and Barrett AJ **MEROPS: the protease database.** *Nucleic Acids Research* 2002, **30:**343-346
38. Barquera B, Hase CC and Gennis RB **Expression and mutagenesis of the NqrC subunit of the NQR respiratory Na+ pump from Vibrio cholerae with covalently attached FMN.** *Febs Letters* 2001, **492:**45-49
39. Hayashi M, Nakayama Y, Yasui M, Maeda M, Furuishi K and Unemoto T **FMN is covalently attached to a threonine residue in the NqrB and NqrC subunits of Na(+)-translocating NADH-quinone reductase from *Vibrio alginolyticus*.** *FEBS Letters* 2001, **488:**5-8
40. Jouanneau Y, Jeong H-S, Hugo N, Meye C and Willison JC **Overexpression in *Escherichia coli* of the rnf genes from *Rhodobacter capsulatus* – characterisation of two membrane-bound iron-sulfur proteins.** *European Journal of Biochemistry* 1998, **251:**54-64
41. Lauer B, Russwurm R, Schwarz W, Kalmanczhelyi A, Bruntner C, A Rosemeier and Bormann C **Molecular characterisation of co-transcribed genes from *Streptomyces tendae* Tu901 involved in the biosynthesis of the peptidyl moiety and assembly of the peptidyl nucleoside antibiotic nikkomycin.** *Molecular and General Genetics* 2001, **264:**662-673
42. Galm U, Schima J, Fiedler HP, Schimdt J, Li SM and Heide L **Cloning and analysis of the simocyclonine biosynthetic gene cluster of *Streptomyces antibioticus* Tu 6040.** *Archives of Microbiology* 2002, **178:**102-114
43. Wang ZX, Li SM and Heide L **Identification of the coumermycin A1 biosynthetic gene cluster of *Streptomyces rishiriensis* DSM 40489.** *Antimicrobial Agents and Chemotherapy* 2000, **44:**3040-3048
44. Crosa JH and Walsh CT **Genetics and assembly line enzymology of siderophore biosynthesis in bacteria.** *Microbiology and Molecular Biology Reviews* 2002, **66:**223-249
45. Szyperski T, Fernandez C, Mumenthaler C and Wuthrich K **Structure comparison of a human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system.** *Proceedings of the National Academy of Sciences* 1998, **95:**2262-2266
46. Maeda T, Nishida J and Nakanishi Y **Expression pattern, subcellular localization and structure-function relationship of rat Tpx-1, a spermatogenic cell adhesion molecule responsible for association with Sertoli cells.** *Development Growth & Differentiation* 1999, **41:**715-722
47. Roberts KP, Ensrud KM and Hamilton DW **A comparative analysis of expression and processing of the rat epididymal fluid and sperm-bound forms of proteins D and E.** *Biology of Reproduction* 2002, **67:**525-533
48. Yamakawa T, Miyata S, Oqawa N, Koshikawa N, Yasumitsu H, Kanamori T and Miyazaki K **cDNA cloning of a novel trypsin inhibitor with similarity to pathogenesis-related proteins and its frequent expression in human brain cancer cells.** *Biochimica et Biophysica Acta* 1998, **1395:**202-208
49. Asmann YW, Kosari F, Wang K, Cheville JC and Vasmatzis G **Identification of differentially expressed genes in normal and malignant prostrate by electronic profiling of expressed sequence tags.** *Cancer Research* 2002, **62:**3308-3314

50. Olson JH, Xiang X, Ziegert T, Kittelson A, Rawls A, Bieber AL and Chandler DE **Allurin, a 21-kDa sperm chemoattractant from *Xenopus* egg jelly, is related to mammalian sperm-binding proteins.** *Proceedings of the National Academy of Sciences* 2001, **98:**11205-11210
51. Moyle M, Foster DL, McGrath DE, Brown SM, Laroche Y, De Meutter J, Stanssens P, Bogowitz CA, Fried VA and Ely JA **A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18.** *Journal of Biological Chemistry* 1994, **269:**10008-10015
52. Brown RL, Haley TL, West KA and Crabb JW **Pseudochetoxin: A peptide blocker of cyclic nucleotide-gated channels.** *Proceedings of the National Academy of Sciences* 1999, **96:**754-759
53. Subramanian G, Koonin EV and Aravind L **Comparative genome analysis of the pathogenic spirochetes Borrelia burgdorferi and Treponema pallidum.** *Infection and Immunity* 2000, **68:**1633-1648
54. Koonin EV, Wolf YI and Aravind L **Protein fold recognition using sequence profiles and its application in structural genomics.** *In: Advances in Protein Chemistry* 2000, **54:**245-275
55. Galperin MY, Gaidenko TA, Mulkidjanian AY, Nakano M and Price CW **MHYT, a new integral membrane sensor domain.** *Fems Microbiology Letters* 2001, **205:**17-23
56. Qi Y, Pei JM and Grishin NV **C-terminal domain of gyrase A is predicted to have a beta-propeller structure.** *Proteins-Structure Function and Genetics* 2002, **47:**258-264