

# Biological Investigations through Sequence Analysis

**Corin Yeats**

Submitted for Degree

Doctor of Philosophy, University of Cambridge

October 2004

Sidney Sussex College, University of Cambridge

& The Wellcome Trust Sanger Institute

## **Acknowledgements**

Many people have contributed, both accidentally and intentionally, both positively and negatively, to helping me get to here. Finding myself at the Sanger Institute with three years and 250 pages behind me and on the desk in front of me, has been as much from the efforts of others as from my own actions. I have always tried to approach life with an open mind and with a constant desire for learning; and I have tried to take lessons from both the positive and negative. And for instilling this attitude, and for providing many lessons in both of the above, I would like to thank my parents. They have made all things possible and given me the security to explore freely. Thank you.

Next up, from school: Chas, Andy

-and much love to everyone I've ever met with the name Oury -

Nick, Hamish. What can I say? We were there and we left, and it could have been very different. One word: Excellent. And in the nearly ten years since, second word: Excellent.

Anna, thank you, you've been wonderful.

And of course there are the people who have contributed directly to my work and learning. First and foremost my supervisor Alex Bateman has been an inspiration, giving me enough room to learn and putting in far more hours into my education than I had any right to expect. The whole of the Pfam group are superb and I wish them all much future success (is one paper in the top ten most cited enough?!). And I'd like to thank everyone I've collaborated with -especially Steve Bentley.

And in no particular order: Ali M, Ali W, Amy, {Bob, Mike, Barney), Ben D-J, Ben M, Ben S, Big Al, Billy, Buttercuts, Cath, Charlie C, Charlie T, Chris, Dan B, Dave U, Doug, Iffy, Jim, Jude, Matt & Anne (the antithesis of nuisance neighbours), Mike C, music & musicians everywhere, Nicola, Nikki, Tim, Waseem, Wee Al, Will, *et al.*

## **Declaration**

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

## Summary

The examination of three dimensional protein structures has revealed that most proteins are made up from modular building blocks. These blocks normally form stable globular structures, and carry particular functions - e.g. catalytic properties - and hence have been termed 'domains'. Domains can be considered both the functional and evolutionary units from which proteins are formed. It has also been demonstrated that if two protein amino acid sequences show significant similarity, then their structures also display similarity.

I have sought to take advantage of the huge amount of sequence data that is being generated by the current wave of genome sequencing projects to identify novel domains and build alignments of homologous sequences. These alignments provide a powerful means to integrate multiple sources of data and hence enable the derivation of novel biological knowledge without recourse to further laboratory experimentation.

Novel domains identified include: the PASTA domain, a  $\beta$ -lactam antibiotic binding domain, with various roles in eubacterial cell wall growth and maintenance; the eubacterial BON domain, a probable phospholipid membrane binding domain, with roles in osmotic shock protection and mechanosensitive channel function; the PepSY domain, which is likely to inhibit eubacterial M4 peptidases but is also found in archaea, and is possibly important in microbe-microbe interactions as well as self-protection and Bacillales sporulation.

## Contents Listing

<b>Acknowledgements</b>	...	...	...	...	...	...	...	...	<b>ii</b>
<b>Summary</b>	...	...	...	...	...	...	...	...	<b>iv</b>
<b>1 Overview</b>	...	...	...	...	...	...	...	...	<b>1</b>
<b>1.1 Aim</b>	...	...	...	...	...	...	...	...	<b>1</b>
<b>1.2 Background</b>	...	...	...	...	...	...	...	...	<b>1</b>
<b>1.3 Protein Domains, Repeats, Motifs and Families</b>	...	...	...	...	...	...	...	...	<b>6</b>
<b>1.4 Characteristic Properties of a Protein Domain</b>	...	...	...	...	...	...	...	...	<b>9</b>
<b>1.5 The Limitations and Difficulties of Domain Hunting</b>	...	...	...	...	...	...	...	...	<b>13</b>
1.5.1 Domain Boundary Identification	...	...	...	...	...	...	...	...	13
1.5.2 The Stepping Stone Phenomenon	...	...	...	...	...	...	...	...	15
1.5.3 Replication of Experiments	...	...	...	...	...	...	...	...	18
<b>1.6 Tools</b>	...	...	...	...	...	...	...	...	<b>21</b>
1.6.1 Search Software	...	...	...	...	...	...	...	...	21
1.6.2 Alignment Software	...	...	...	...	...	...	...	...	25
1.6.3 Databases	...	...	...	...	...	...	...	...	31
1.6.4 Structural Collections and Classifications	...	...	...	...	...	...	...	...	37
1.6.5 Presenting Domain Architectures and Alignments	...	...	...	...	...	...	...	...	40
<b>2 Identifying Novel Domains</b>	...	...	...	...	...	...	...	...	<b>44</b>
<b>2.1 Domain Hunt Methods</b>	...	...	...	...	...	...	...	...	<b>44</b>
2.1.1 Introduction	...	...	...	...	...	...	...	...	44
2.1.2 Details of Methods	...	...	...	...	...	...	...	...	45
<b>2.2 Domain Hunting in <i>Streptomyces coelicolor</i></b>	...	...	...	...	...	...	...	...	<b>52</b>
2.2.1 Introduction to <i>Streptomyces coelicolor</i> - a Complex Prokaryote	...	...	...	...	...	...	...	...	52
2.2.2 Methods	...	...	...	...	...	...	...	...	54
2.2.3 Summary of Results	...	...	...	...	...	...	...	...	54
2.2.4 Notes on Table of All Identified Novel Domains	...	...	...	...	...	...	...	...	57

<b>2.3 Descriptions of Novel Domains</b>	...	...	...	...	...	...	<b>57</b>
<i>HA</i>	...	...	...	...	...	...	57
<i>BTAD</i>	...	...	...	...	...	...	62
<i>ALF</i>	...	...	...	...	...	...	65
<i>SPDY</i>	...	...	...	...	...	...	68
<i>PASTA</i>	...	...	...	...	...	...	70
<i>HHE</i>	...	...	...	...	...	...	73
<i>PPC</i>	...	...	...	...	...	...	77
<i>FMN_bind</i>	...	...	...	...	...	...	81
<i>MbtH</i>	...	...	...	...	...	...	84
<b>2.4 Significantly Extended Families</b>	...	...	...	...	...	...	<b>86</b>
<i>SCP</i>	...	...	...	...	...	...	86
<i>FG-GAP</i>	...	...	...	...	...	...	92
<b>2.5 Concluding Comments</b>	...	...	...	...	...	...	<b>92</b>
<b>3 Multi-genome Domain hunting</b>	...	...	...	...	...	...	<b>95</b>
<b>3.1 Rationale</b>	...	...	...	...	...	...	<b>95</b>
<b>3.2 Results</b>	...	...	...	...	...	...	<b>97</b>
3.2.1 Summary of Results	...	...	...	...	...	...	97
3.2.2 Notes on Table of All Identified Novel Domains	...	...	...	...	...	...	98
<b>3.3 Descriptions of Novel Domains</b>	...	...	...	...	...	...	<b>98</b>
3.3.1 Domains Identified From Repeats	...	...	...	...	...	...	100
<i>PepSY</i>	...	...	...	...	...	...	100
<i>Gate</i>	...	...	...	...	...	...	100
<i>STN</i>	...	...	...	...	...	...	102
<i>Secretin_N</i>	...	...	...	...	...	...	105
<i>Secretin_N_2</i>	...	...	...	...	...	...	108
<i>Reg_prop</i>	...	...	...	...	...	...	110
<i>Y_Y_Y</i>	...	...	...	...	...	...	113
<i>DUF1533</i>	...	...	...	...	...	...	115
<i>Coat_X</i>	...	...	...	...	...	...	115
<i>Cleaved_adhesin</i>	...	...	...	...	...	...	118
<i>FIVAR</i>	...	...	...	...	...	...	118
<i>FlaE</i>	...	...	...	...	...	...	127
<i>Glug</i>	...	...	...	...	...	...	132
3.3.2 Domains Found Through Small Protein Clustering	...	...	...	...	...	...	132
<i>Coat_F</i>	...	...	...	...	...	...	132
<i>CTnDOT_TraJ</i>	...	...	...	...	...	...	132
<i>Dabb</i>	...	...	...	...	...	...	135

<i>Nif11</i>	...	...	...	...	...	...	...	142
<b>3.4 Other Potential Uses</b>	...	...	...	...	...	...	...	<b>142</b>
<b>4 Detailed Investigations of Individual Domains</b>	...	...	...	...	...	...	...	<b>145</b>
<b>4.1 The PASTA Domain: A <math>\beta</math>-lactam-Binding Domain</b>	...	...	...	...	...	...	...	<b>145</b>
4.1.1 Background	...	...	...	...	...	...	...	145
4.1.2 Searching for PASTA	...	...	...	...	...	...	...	147
4.1.3 Structure of PASTA	...	...	...	...	...	...	...	150
4.1.4 Roles of PASTA	...	...	...	...	...	...	...	151
4.1.5 PASTA and Cell Morphology	...	...	...	...	...	...	...	153
4.1.6 The PASTA Domain as an Antibiotic Target	...	...	...	...	...	...	...	154
4.1.7 Subsequent Research	...	...	...	...	...	...	...	156
<b>4.2 The BON Domain: A Putative Membrane Binding Domain</b>	...	...	...	...	...	...	...	<b>157</b>
4.2.1 Identification of the Conserved Regions	...	...	...	...	...	...	...	158
4.2.2 OsmY Comprises Two BON Domains	...	...	...	...	...	...	...	159
4.2.3 Other BON-containing Proteins	...	...	...	...	...	...	...	161
4.2.4 Phylectic Distribution	...	...	...	...	...	...	...	163
<b>4.3 The PepSY Domain: A Putative Regulator of Peptidase Activity</b>	...	...	...	...	...	...	...	<b>164</b>
4.3.1 Background to the M4 Peptidases	...	...	...	...	...	...	...	165
4.3.2 PepSY Domain Identification	...	...	...	...	...	...	...	166
4.3.3 Description of the PepSY Domain	...	...	...	...	...	...	...	166
4.3.4 Domain Architecture of the M4 Propeptide	...	...	...	...	...	...	...	170
4.3.5 Species Distribution of PepSY	...	...	...	...	...	...	...	170
4.3.6 PepSY Family Characteristics	...	...	...	...	...	...	...	171
4.3.7 PepSY Domains are Likely to be Inhibitors	...	...	...	...	...	...	...	173

4.3.8 The Biological Role of PepSY	...	...	...	...	...	...	...	174
<b>4.4 Peptidase_A24 - the Prepilin Peptidase</b>	...	...	...	...	...	...	...	<b>175</b>
<b>5 Contributions to Genome Annotation Projects</b>	...	...	...	...	...	...	...	<b>180</b>
<b>5.1 Tropheryma whipplei</b>	...	...	...	...	...	...	...	<b>180</b>
5.1.1 Background	...	...	...	...	...	...	...	180
5.1.2 The WiSP Protein Family	...	...	...	...	...	...	...	181
5.1.3 The WiSP Domains	...	...	...	...	...	...	...	183
<i>WND</i>	...	...	...	...	...	...	...	183
<i>CCD</i>	...	...	...	...	...	...	...	183
<i>He_PIG</i>	...	...	...	...	...	...	...	187
5.1.4 Implications for the Immune System	...	...	...	...	...	...	...	194
<b>5.2 Burkholderia pseudomallei</b>	...	...	...	...	...	...	...	<b>194</b>
5.2.1 Background	...	...	...	...	...	...	...	194
5.2.2 Novel Domains	...	...	...	...	...	...	...	195
<i>SCPU</i>	...	...	...	...	...	...	...	195
<i>BTP</i>	...	...	...	...	...	...	...	197
<i>PHB_acc</i>	...	...	...	...	...	...	...	199
<i>The Repetitive <math>\beta</math>-helix Surface Structure Superfamily</i>	..	..	..	..	..	..	..	199
<b>5.3 Chlamydomophila abortus</b>	...	...	...	...	...	...	...	<b>205</b>
5.3.1 Background	...	...	...	...	...	...	...	205
5.3.2 The Chlamydial Polymorphic Membrane Protein	...	...	...	...	...	...	...	206
<i>ChlamPMP_M</i>	...	...	...	...	...	...	...	206
<b>5.4 Theileria annulata</b>	...	...	...	...	...	...	...	<b>211</b>
5.4.1 Background	...	...	...	...	...	...	...	211
5.4.2 FAINT	...	...	...	...	...	...	...	212
5.4.3 The TASR Repeat Families	...	...	...	...	...	...	...	217
<b>6 Conclusions</b>	...	...	...	...	...	...	...	<b>223</b>

<b>Bibliography</b>	...	...	...	...	...	...	...	...	227
---------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----

<b>Appendix A: List of All Domains in this Thesis</b>	...	...	...	...	...	...	...	...	255
---	-----	-----	-----	-----	-----	-----	-----	-----	-----

## Figure Listing

<b>Figure 1.1:</b> Growth of the UniProt database	...	...	...	...	...	...	...	...	5
<b>Figure 1.2:</b> Examples of different protein structure types	...	...	...	...	...	...	...	...	8
<b>Figure 1.3:</b> Graph displaying the average size of domains recognised by Pfam	...	...	...	...	...	...	...	...	11
<b>Figure 1.4:</b> A common protein clustering error caused by multidomain proteins	...	...	...	...	...	...	...	...	14
<b>Figure 1.5:</b> Graph of the average percent identity for each Pfam family	...	...	...	...	...	...	...	...	16
<b>Figure 1.6:</b> The iterative search methodology	...	...	...	...	...	...	...	...	18
<b>Figure 1.7:</b> Example of the Dotter output	...	...	...	...	...	...	...	...	26
<b>Figure 1.8:</b> A simple example of a "bad" alignment compared to a "good" alignment.	...	...	...	...	...	...	...	...	31
<b>Figure 1.9:</b> Example Pfam family page - the PASTA domain	...	...	...	...	...	...	...	...	33
<b>Figure 1.10:</b> Key to the architecture figures and some example architectures	...	...	...	...	...	...	...	...	43
<b>Figure 2.1:</b> General method for identifying novel domains	...	...	...	...	...	...	...	...	49
<b>Figure 2.2:</b> HA example alignment	...	...	...	...	...	...	...	...	58
<b>Figure 2.3:</b> HA example architectures	...	...	...	...	...	...	...	...	59

<b>Figure 2.4:</b> BTAD example alignment	...	...	...	...	...	63
<b>Figure 2.5:</b> BTAD example architectures	...	...	...	...	...	64
<b>Figure 2.6:</b> ALF alignment, architectures, and genome context	...	...	...	...	...	67
<b>Figure 2.7:</b> SPDY alignment and architectures	...	...	...	...	...	69
<b>Figure 2.8:</b> Evidence for the presence of a mobile DNA element	...	...	...	...	...	71
<b>Figure 2.9:</b> Alignment of the original HHE domains and predicted secondary structure against a Hemerythrin domain and known structure	...	...	...	...	...	74
<b>Figure 2.10:</b> HHE architectures	...	...	...	...	...	75
<b>Figure 2.11:</b> PPC alignment along with predicted and known secondary structure	...	...	...	...	...	78
<b>Figure 2.12:</b> PPC domain architectures	...	...	...	...	...	79
<b>Figure 2.13:</b> FMN_Bind alignment	...	...	...	...	...	82
<b>Figure 2.14:</b> FMN_bind domain architectures	...	...	...	...	...	83
<b>Figure 2.15:</b> Alignment and architectures for the MbtH domain	...	...	...	...	...	85
<b>Figure 2.16:</b> SCP domain alignment	...	...	...	...	...	88
<b>Figure 2.17:</b> SCP domain architectures	...	...	...	...	...	89
<b>Figure 3.1:</b> Simplified taxonomic tree of bacteria investigated in the multigenome hunt	...	...	...	...	...	96
<b>Figure 3.2:</b> Alignment and architectures for the Gate domain	...	...	...	...	...	101
<b>Figure 3.3:</b> STN alignment and architectures	...	...	...	...	...	103
<b>Figure 3.4:</b> Secretin_N example alignment	...	...	...	...	...	106
<b>Figure 3.5:</b> Secretin_N example architectures	...	...	...	...	...	107
<b>Figure 3.6:</b> Secretin_N_2 alignment and architectures	...	...	...	...	...	109
<b>Figure 3.7:</b> Example Reg_prop alignment	...	...	...	...	...	111

<b>Figure 3.8:</b> Example Reg_prop architectures	...	...	...	...	112
<b>Figure 3.9:</b> Y_Y_Y alignment and architectures	...	...	...	...	114
<b>Figure 3.10:</b> DUF1533 alignment and architectures	...	...	...	...	116
<b>Figure 3.11:</b> Coat_X alignment and architectures	...	...	...	...	117
<b>Figure 3.12:</b> Cleaved_adhesin alignment and architectures	...	...	...	...	119-120
<b>Figure 3.13:</b> Example FIVAR alignment	...	...	...	...	122
<b>Figure 3.14:</b> Example FIVAR architectures	...	...	...	...	123-125
<b>Figure 3.15:</b> Example FlaE alignment and architectures	...	...	...	...	129
<b>Figure 3.16:</b> Example Glug repeat alignment	...	...	...	...	130
<b>Figure 3.17:</b> Example Glug repeat architectures	...	...	...	...	131
<b>Figure 3.18:</b> Example Coat_F alignment and architectures	...	...	...	...	133
<b>Figure 3.19:</b> Example CTnDOT_TraJ alignment and architectures	...	...	...	...	134
<b>Figure 3.20:</b> Example Dabb alignment and architectures	...	...	...	...	136-137
<b>Figure 3.21:</b> Structure of the Dabb barrel	...	...	...	...	139
<b>Figure 3.22:</b> Example Nif11 architectures and alignment	...	...	...	...	142
<b>Figure 4.1:</b> Example PASTA alignment	...	...	...	...	148
<b>Figure 4.2:</b> Example PASTA domain architectures	...	...	...	...	149
<b>Figure 4.3:</b> Stereo view of the two PASTA domains of <i>Streptococcus pneumoniae</i> PBP2X	...	...	...	...	150
<b>Figure 4.4:</b> Distribution of resistance mutations in the PASTA domains of PBP2X	...	...	...	...	155
<b>Figure 4.5:</b> Example BON domain alignment	...	...	...	...	160
<b>Figure 4.6:</b> Example BON domain architectures	...	...	...	...	162
<b>Figure 4.7:</b> Example PepSY domain alignment	...	...	...	...	167
<b>Figure 4.8:</b> FTP motif example alignment	...	...	...	...	168

<b>Figure 4.9:</b> PepSY_TM example alignment	...	...	...	...	169
<b>Figure 4.10:</b> Example PepSY domain architectures	...	...	...	...	172
<b>Figure 4.11:</b> Peptidase_A24 example alignment and architectures	...	...	...	...	177
<b>Figure 5.1:</b> Example WiSP family architectures	...	...	...	...	184
<b>Figure 5.2:</b> Example WND alignment	...	...	...	...	185-186
<b>Figure 5.3:</b> CCD example alignment	...	...	...	...	188
<b>Figure 5.4:</b> Example He_PIG architectures	...	...	...	...	189
<b>Figure 5.5:</b> He_PIG example alignment	...	...	...	...	190
<b>Figure 5.6:</b> N-J Tree of all He_PIG domains from <i>Tropheryma whipplei</i>	...	...	...	...	193
<b>Figure 5.7:</b> SCPU example alignment and architectures	...	...	...	...	196
<b>Figure 5.8:</b> BTP example alignment and architectures	...	...	...	...	198
<b>Figure 5.9:</b> PHB_acc example alignment and architectures	...	...	...	...	200
<b>Figure 5.10:</b> Example Chlam_PMP alignment with related sequences	...	...	...	...	202
<b>Figure 5.11:</b> Fil_haemagg and Ice_nucleation example architectures	...	...	...	...	203
<b>Figure 5.12:</b> The <i>Chlamydomonas abortus</i> polymorphic membrane protein family	...	...	...	...	207
<b>Figure 5.13:</b> ChlamPMP_M example alignment	...	...	...	...	209-210
<b>Figure 5.14:</b> FAINT example alignment	...	...	...	...	213-214
<b>Figure 5.15:</b> FAINT example architectures	...	...	...	...	216
<b>Figure 5.16:</b> Neighbour-Joining tree of the choline kinases of <i>Theileria parva</i> and <i>Theileria annulata</i>	...	...	...	...	218
<b>Figure 5.17:</b> Example alignments of the TASR short repeats of <i>T. annulata</i> (A) and <i>T. parva</i> (B)	...	...	...	...	219
<b>Figure 5.18:</b> <i>Theileria annulata</i> TASR_1 example architectures	...	...	...	...	220

## Table Listing

<b>Table 1.1:</b> Results from Edgar's (2004) comparison of MAFFT, T-Coffee and Clustalw ... .. 30	...	...	...	...	...	...	...	...	...
<b>Table 1.2:</b> The ClustalX colouring scheme ... .. 42	...	...	...	...	...	...	...	...	...
<b>Table 2.1:</b> Table of all novel domains identified in <i>S. coelicolor</i> ... .. 56	...	...	...	...	...	...	...	...	...
<b>Table 3.1:</b> Table of all novel domains identified in the multigenome hunt ... .. 99	...	...	...	...	...	...	...	...	...
<b>Table 5.1:</b> Statistics testing whether the overlap between the <i>T. parva</i> TASR_2-containing proteins and the <i>T. annulata</i> TASR_1 proteins is by chance ... .. 221	...	...	...	...	...	...	...	...	...