6 Summary and Conclusions

At the start of this thesis I set out to find and describe novel protein domains that are of significant interest to biology in general. Through the identification of these sequence families I hoped to identify novel biological processes, elucidate previously unsuspected mechanisms in known processes, and to further characterise well studied proteins.

I chose to primarily study the proteins of Prokaryotes since there are many more sequenced prokaryotic genomes than eukaryotic genomes; this reduces any bias towards well studied families and allows increased focus on domains of general interest. However, the searches were not carried out exclusively in bacteria, and several domains that occur in Eukaryotes were identified - for instance the SCP domain and the Dabb domain.

The domain hunt methods I employed were generally aimed at identifying manageable (around 200) numbers of targets that may represent novel protein domains. One of the principal methods for identifying domains was based around identifying repeats within proteins; the second was based on clustering proteins of 100 residues or less in length.

Identification of sequence families allowed the construction of a multiple sequence alignment. Through these, several sources of information were related and interpreted, allowing the generation of new knowledge without recourse to laboratory experimentation. Data for each family was extracted from the published literature, from computational predictions - for example secondary structure predictions - and

from web sources. From the multiple sequence alignment itself, conserved regions and residues can be identified and correlated with previous observations. As an example the most conserved patch in the PepSY domain contains the loss-of-function mutation identified by Braun et al, 2000.

In total 41 domains, repeats, motifs and families are presented in this thesis and another 54 were identified but are not discussed. Most of these domains appear to be ligand-binding domains, or structural, with very few novel enzymes being uncovered. Many of them are also found on the bacterial cell surface. These are of particular interest, not only because of their involvement in pathogenicity, host interactions and antibiotic resistance, but also because of the difficulty in examining the structures of cell membrane associated proteins. By identifying the structural units it becomes simpler to excise them from the surrounding protein and solve their structure through crystallography or NMR; it also becomes possible to identify them in proteins that may be more amenable to laboratory-based investigation. Such an approach was successfully carried out by Wilson, Matsushita *et al*. (2003) to solve the structure of the PPC domain, and could work well with domains like He_PIG.

As has been found in the case of the PASTA domain, creating an alignment that links several species together can allow deeper interpretation of species specific information. In this case of PASTA I found that there appeared to be correlation between the number of PASTA-containing proteins in a genome and the different morphological types exhibited by that species. Specifically if a species has more than one PASTA-containing penicillin-binding protein (pPBP) or PASTA-containing serine/threonine protein kinase (pPSTK) then it will display more than one cell

morphology. *Streptomyces coelicolor* has three cell morphologies and three pPSTKs. This leads to the prediction that the Bacillus cereus group, which includes *Bacillus anthracis*, may have one more cell shape than other closely related Bacillales.

Other insights have been made into mechanisms of biofilm formation - the PepSY domain - and the immune system evasion methods of *Chlamydophila abortus* and *Tropheryma whipplei*. Also, entirely novel processes have been uncovered, paving the way for new areas of research. As an example the short repeats found in the Theileria in chapter 5.4 have not been described before and initial characterisation suggests that they are the result of a process unlike anything previously described in the literature. Similarly the SCP domain has now been strongly implicated in the establishment of cell polarity and copper ion chelation, creating the framework for investigating the details of its function.

Whilst the merits of domain hunting have been long known, the work in this Thesis demonstrates that there is still much to be learned from this level of protein analysis. Certainly by no means have all the biologically important domains been identified, characterised and modelled, and it appears that many of the interesting ones remain. I have also shown that the multiple sequence alignment is an extremely powerful tool for relating the information from different proteins that were analysed in isolation and without consideration of the rest of the family.

I suggest that not only is the detailed analysis of homology a useful task to carry out subsequent to experimental work, but that experimental work should be guided by the multiple sequence alignment. By this I mean that the identification of a domain

family's global characteristics should be used in directing laboratory-based experimentation. From the alignment it should possible to identify the biggest gaps in our description of a family and hence carry out the most useful experiment on the most effective example protein for further characterisation of the family. This approach will generate more information about a greater number of proteins that the current piecemeal approach that underlies many protein investigations. Whilst there is still clearly a place for the investigation of specific proteins in a specific context, the genomic age introduces a new paradigm. For instance further study into the bacterial secretin family should be guided by consideration of its, at least, 16 different domain architectures rather than the current organism-based approach.

Comparative analysis is already a powerful approach, and is increasing in power as more genomes are sequenced and individual families become better represented. However, there is also a need for improved automatic detection of novel domains. In Pfam there are over 7000 sequence families, and yet only around 50% of amino acid residues in UniProt are accounted for; furthermore many of these families may cover more than one domain, or even include partial domains. In order to speed up the rate of discovery, tools for splitting the known families into their subcomponents and for the *ab initio* prediction of domains are essential. Such tools would be useful for driving the type of semi-automatic investigations I have been carrying out as well as enabling effective analyses of large numbers of proteins; a task that will increase in import as the rate of genome sequencing accelerates.