# 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 then 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 β-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

## Figure Listing

# Table Listing