Protein Domains: New methods for detection and evolutionary analysis

Lachlan James Murray Coin
Magdalene College
Cambridge

January 2005

A dissertation submitted for
the degree of Doctor of Philosophy
at the University of Cambridge

Preface

The work presented in this dissertation was carried out at the Sanger Institute in Cambridge between October 2001 and December 2004. 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. No part of this dissertation nor anything substantially the same has been or is being submitted for any qualification at any other university.

Summary

Protein domains are the structural, functional and evolutionary units of proteins. A useful way to predict the function and structure of a new protein coding gene, or one which is poorly genetically or biochemically characterised, is to identify the domain architecture on the basis of the amino acid sequence, and then infer function and structure from other proteins with similar domain architectures.

The first part of my thesis concerns improving techniques for identification of protein domains from amino acid sequence. I investigate the application of language modelling techniques from speech recognition to integrate contextual information into domain prediction. This takes advantage of the observation that certain combinations of domains are more likely to occur than others. I also investigate using knowledge of the species in which the protein occurs to improve domain prediction, and develop an integrated model of species and domain context. Lastly, I investigate the degree to which protein domains can be identified on alignments of homologous proteins, rather than on the sequences taken individually. This method relies on the development of models of evolution which reflect the structural and functional constraints of conserved sites in the protein domain and using these models to calculate the likelihood that the given protein cluster has been evolving within these structural and functional constraints. I have tested each of these approaches on proteins of known structure, and demonstrated improvements in domain identification in each case.

The second part of my thesis concerns using annotated protein domains to understand the evolution of gene families. I look for cases in which the gene family unambiguously contains a particular protein domain, but also contains proteins which are diverging away from the domain. Using evolutionary models developed in the first part of my thesis which reflect functional/structural constraints at conserved sites, I develop a technique for scoring the degree to which evolution along a branch in the gene tree is constrained by the need to

maintain the structure and function of the protein, and conversely, the likelihood that it is not evolving under these constraints. I have used this approach as the basis of a test for pseudogenes, which has been tested against standard methods for identifying pseudogenes on the manual annotation of human chromosome six. I have also used this approach to develop a test for positive selection, and characterised positive selection in several gene families.

Acknowledgements

Firstly I thank my supervisor, Richard Durbin, for his advice, support and encouragement. I also thank Alex Bateman for useful advice and support. I thank Ewan Birney for many useful suggestions and research directions, particularly with regard to pseudogene detection. Many thanks also to Nick Goldman and Simon Whelan for stimulating discussions and critique of parts of this work. Thanks to Ian Korf, Rob Finn, Sam Griffiths-Jones, Corin Yeats, Mhairi Marshall, Ashwin Hajarnavis, David Carter, Mark Minichiello, Irmtraud Meyer, Marc Sohrmann, Thomas Down, Kevin Howe, Andy Futreal and Mike Stratton as well as all of the other members of the Wellcome Trust Genome campus.

I am grateful to Magdalene College, Cambridge, for a Leslie Wilson studentship and to the Cambridge Australia Trust for a scholarship.

Contents

Summary **iii**

1 Introduction **1**

1.1 Protein Domains 3

1.2 Sequence based protein domain detection 7

1.3 Models of sequence evolution 16

1.3.1 Probabilistic models of sequence evolution 17

1.3.2 Models of residue substitution 22

1.3.3 Likelihood calculation 26

1.4 Outline of thesis 27

2 Enhanced Domain Detection Using Approaches From Speech Recognition 29

2.1 Statistical Speech Recognition Techniques 30

2.2 Patterns of domain occurrence and co-occurrence 35

2.3 Methods: Application to protein domain detection 39

2.3.1 Formulation 40

2.3.2 Context model and smoothing strategy 41

2.3.3 Context score of a domain in a protein with fixed context 43

2.3.4 Dynamic programming algorithm 43

2.3.5 Incorporating the sequence score threshold 44

2.3.6 Variable length Markov model 45

2.3.7 Incorporating Pfam clans 46

2.3.8 Significance scores 46

2.3.9 Implementation 49

2.4 Results 50
2.4.1 SCOP test 51
2.4.2 Pfam scan 59
2.5 Discussion 69

3 Enhanced Domain Recognition Using Phylogeny 71
3.1 Algorithm 74
3.1.1 Phylogenetic profile HMM 74
3.1.2 Using the phylogenetic profile HMM 83
3.2 Results 85
3.3 Conclusion 92

4 Using protein domains to identify pseudogenes and positive selection 97
4.1 Pseudogenes 99
4.2 Positive selection 100
4.3 Algorithm 101
4.3.1 Allowing for a single frame-shifted nucleotide sequence 111
4.3.2 Restricting the size of the input cluster 111
4.3.3 Calculating PSILC scores for internal nodes 112
4.4 Results: Vega pseudogene test set 112
4.4.1 Test data 112
4.5 Results: detection of positive selection 116
4.5.1 Analysis of selective pressures on APOBEC/AID enzymes 118
4.5.2 Analysis of selective pressures on Abalone lysin protein 125
4.6 Results: Global scan for pseudogenes and positive selection 133
4.7 Discussion 145

5 Conclusion 149

Bibliography 170