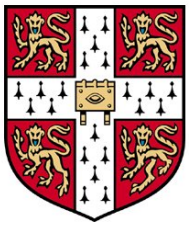
# Mechanisms of change in protein architecture



Marija Buljan

Trinity College

University of Cambridge

A dissertation submitted for the degree of

*Doctor of Philosophy*

September 2010

"*Well begun is half done.*"

Aristotle

# Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between May 2006 and September 2010. This dissertation is the result of my own work. No part of this dissertation or anything substantially the same has been or is being submitted for any qualification at any other university.

# Summary

Proteins are the basic building blocks and functional units in all living organisms. Moreover, differences between species can frequently be explained with differences in their protein complements. Importantly, proteins are often composed of segments, i.e. domains that have a certain level of evolutionary, structural and/or functional independence. The majority of proteins in nature contain two or more domains, and an individual domain can often occur in combinations with different domain partners.

In the first part of my thesis, I traced the history of animal gene families and the proteins these genes encode. By this means, I was able to infer events where changes in protein domain architectures took place. This showed that both insertions and deletions of single copy domains preferentially occur at protein termini, but also that changes are more likely to occur after gene duplication than organism speciation. Finally, domains that were most frequently gained were the ones that are related to an increase in organismal complexity, thus underlining the important role of domain shuffling in animal evolution.

In the second part of my thesis, I focused on a set of high confidence domain gain events and investigated the evidence for molecular mechanisms that caused these domain gains. In agreement with observations from the first part - that changes preferentially occur at the termini - I have found that the strongest contribution to gains of novel domains in proteins comes from gene fusion through the joining of exons from adjacent genes into a novel gene unit. Two other mechanisms that have been suggested to play a major role in the evolution of animal proteins, retroposition and middle insertions through intronic recombination, have a smaller role in comparison to gene fusions. Since the majority of these domain gains are again observed after gene duplication, this suggests a powerful mechanism for neofunctionalization after gene duplication.

Finally, in the last part of my thesis, I address a mechanism that increases the number and variety of proteins in an organism – alternative splicing. In particular, I investigate the functional consequences of tissue-specific alternative splicing events. I found that tissue-specific splicing tends to affect exons that encode protein regions without defined secondary or tertiary structure. Importantly, it is known that these disordered regions frequently play a role in protein interactions. In agreement with this, I observed significant enrichment of tissue-specifically encoded protein segments in disordered binding peptides and posttranslationally modified sites. A possible result of the finely regulated alternative splicing of these segments is a tissue-specific rewiring of protein network. In conclusion, both alternative splicing and domain shuffling can increase proteome diversity. However, a protein with a new function can often directly or indirectly shape the functions of other proteins in its environment.

# Acknowledgments

During my PhD, I was lucky to meet many people that I will remember as inspiring scientists and great persons. In these acknowledgments, I wish to say thanks to those individuals that most influenced the work that I deliver here. I am very grateful to my supervisor Alex Bateman for his support during my PhD, for teaching me how to approach science critically and with attention to details and for providing me a valuable feedback on everything that is written in this thesis. My thesis committee meetings were of immense value for my progress in the PhD. Madan Babu, Richard Durbin, Avril Coghlan and Manolis Dermitzakis, my thesis committee members, contributed with great ideas and most helpful criticism. I am also indebted to Avril for her help at the beginning of my PhD when I was starting with phylogenetic analyses. Madan gave valuable contribution to the work in Chapter 4, which is also a joint project with him, and I am very grateful to him for that. The Xfam group members were extremely helpful during my PhD, providing advice and support when I needed it. I am particularly grateful to Benjamin Schuster-Bockler, Cara Woodwark, Lars Barquist, Paul Gardner and Rob Finn. I also need to thank other students in my year, who were always available for discussions, gave valuable feedback so many times and provided excellent PhD environment. I need to specially thank there Matias Piipari and Leo Parts for their help. Finally, Neil Rawlings, Paul Gardner and Penny Coggill kindly proofread the thesis chapters and provided the most useful comments.

# Contents

1. Introduction 1

1.1. Characterization of functional elements in proteins....3

1.1.1. Protein domains....3

1.1.2. Disordered protein regions....7

1.1.3. Sites of posttranslational modification.... 11

1.2. Protein evolution....12

1.2.1. Domain shuffling....14

1.2.2. Mechanisms for formation of novel genes....18

1.2.3. Gene duplication and protein evolution....24

1.2.4. Evolutionarily related proteins....28

1.3. Protein isoforms of the same gene....29

1.4. Outline of the thesis....33

1.5. Bibliography....34

2. Evolution of multidomain proteins 45

2.1 Introduction....45

2.2 Methods....50

2.2.1 Analysis of TreeFam families....50

2.2.2 Assignment of domains to proteins with refinement....50

2.2.3 Domain gains and losses....51

2.3 Results....53

2.3.1 Phlyogenetic trees can guide refinement of domain assignments....53

2.3.2 Single copy domains are predominantly gained and lost at protein termini....57

2.3.3 Gains and losses of domains in repeats....61

2.3.4 Changes in domain architectures preferentially occur after gene duplications....65

2.3.5 Effect of domain gains on the evolution of protein function.....67
2.3.6 Estimate of domain gain and loss events strongly depends on the input parameters.....69
2.4 Discussion.....71
2.4.1 Confidence in the comparison of domain architectures.....71
2.4.2 Molecular mechanisms and evolutionary selection shape the evolution of domain architectures.....72
2.4.3 Set of confident domain gain or loss events.....76
2.5 Bibliography.....77

3. Mechanisms of domain gain in animal proteins 80
3.1 Introduction.....80
3.2 Methods.....86
3.2.1 Assignment of domains to proteins with refinement.....86
3.2.2 Exclusion of possible false domain gain calls.....86
3.2.3 Parsing trees.....87
3.2.4 Intron-exon structures of genes.....91
3.2.5 Positions of gained domains.....91
3.2.6 Genomic origin of the inserted domain.....92
3.3 Results.....94
3.3.1 Set of high confidence domain gain events.....94
3.3.2 Characteristics of the high confidence domain gain events.....95
3.3.3 Characteristics of the medium confidence domain gain events.....97
3.3.4 Supporting evidence for the representative transcripts.....99
3.3.5 Donor genes of the gained domains.....101
3.3.6 Investigation of cellular mechanisms that caused domain gain events.....102
3.3.6.1 Retroposition as a mechanism of domain gain.....102
3.3.6.2 Joining of adjacent genes as a mechanism of domain gain.....105
3.3.6.4 Insertion of exons into ancestral introns as a mechanism of domain gain.....112
3.3.6.4 Exonisation of previously non-coding sequences as a mechanism of domain gain.....113
3.3.7 Domain gains most frequently occur after gene duplications.....115

3.3.8 Gained domains do not have their origin in the adjacent genes...119
3.3.9 Domain gain events affect cellular regulatory networks.....119
3.4 Discussion.....123
3.4.1 Scope of the study.....123
3.4.2 Approach for obtaining the set of confident domain gain events.....124
3.4.3 Mechanisms of domain gain.....125
3.4.4 Domain gains were assisted by recombination events.....128
3.4.5 Different trends in domain gains in different lineages and at different time points during evolution.....130
3.4.6 Functional implications of domain gain events.....131
3.5 Bibliography.....132

4. Protein products of tissue-specific alternative splicing 138
4.1 Introduction.....138
4.2 Methods.....142
4.2.1 Sets of tissue-specific, cassette and constitutive exons.....142
4.2.2 Enrichment of genes with specific function in the set of tissue-specific exons.....143
4.2.3 Prediction of disordered protein residues.....144
4.2.4 Prediction of functional residues.....144
4.2.5 Conservation of exons in the three different datasets.....145
4.2.6 Significance of observed trends.....145
4.2.7 Comparison of MEK1 and MEK2 protein sequences.....146
4.2.8 Enrichment of known disease genes in the set of tissue-specific exons.....146
4.2.9 Disorder signatures in the protein products of the p73 gene.....147
4.3 Results.....147
4.3.1 Sets of exons with different expression profiles.....147
4.3.2 Tissue-specific exons are enriched in disordered residues.....149
4.3.3 Functional residues in disordered segments encoded by tissue-specific exons.....151
4.3.4 Distribution of functional residues in the control sets of cassette and constitutive exons.....154

4.3.5 Disordered residues encoded by tissue-specific exons are highly conserved....156
4.3.6 Genes with tissue-specifically regulated exons have an important function in organism development and survival....160
4.3.7 Alternative isoforms of the gene p73....165
4.3.8 Tissue-specific splicing and protein domains....167
4.4 Discussion....170
4.4.1 Evolution and function of alternative splicing....170
4.4.2 Unstructured functional residues direct isoform-specific networks....172
4.4.3 Examples for the role of disordered protein segments in signal transduction....175
4.4.4 Genes with tissue-specific isoforms and disease development...178
4.5 Bibliography....180

5. Concluding remarks 186

Appendices 190
Appendix A....191
Appendix B....194
Appendix C....204