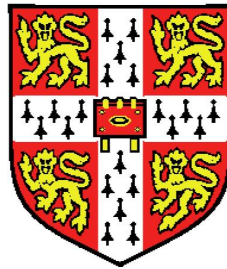


# The role of protein interactions in evolution and disease



Benjamin Schuster-Böckler

Selwyn College

University of Cambridge

A dissertation submitted for the degree of

*Doctor of Philosophy*

September 2008

---

In loving memory of Margarethe and Erwin Gregor.

*“Humility and knowledge are the origins of wisdom.”*

## **Declaration**

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between March 2005 and October 2008. 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

The network of interactions between proteins is the scaffold that shapes the properties of every living cell. Whether it is enzymatic pathways or cascades of signal transduction, most processes rely on the ability of proteins to recognise and bind each other. New experimental techniques have fuelled interest in these networks, leading to a rapid increase in available data on protein interactions from various species.

In the first part of this thesis, I investigate to what extent networks of protein interactions are mediated by conserved regions in proteins, generally called domains. I make use of a set of domain pairs which have been shown to interact in 3-dimensional structures. By analysing the frequency of co-occurrence of these domain pairs in networks of protein interactions from five different species, I show that some domain pairs form reusable recognition modules, while others are confined to a specific protein pair. Overall, the number of known protein interactions that contain a domain pair with known structure is small. This underlines the necessity to resolve more structures of interacting proteins. Finally, I observe a large overlap in the domain pairs present in different species, suggesting many recognition modules are ancient in origin.

In the second part of my thesis, I combine sequence analysis techniques to investigate the impact of protein interactions on human diseases. I make use of the detailed information provided by 3-dimensional structures to

---

identify interacting residues within known protein domains. I then use hidden Markov models to search for structurally corresponding residues in proteins that cause genetic diseases. I identify cases where these structurally corresponding residues have been reported to cause Mendelian disorders, such as an Ile to Val substitution in the dimerisation interface of the H-Twist transcription factor leading to Baller-Gerold syndrome. I report 1428 mutations which potentially affect a protein interaction. This corresponds to  $\approx 4\%$  of all known single-residue mutations.

I found that mutations in interaction interfaces frequently cause dominant phenotypes. I subsequently discovered that many dosage sensitive genes related to human disease are members of protein complexes. From the analysis of recently published data of gene expression and structural variation between individuals it emerges that members of protein complexes exhibit lower expressional noise than the rest of the genome and that variation of gene copy-number between individuals has a measurable effect on dosage. I show that this effect causes negative selection against large scale copy-number variations in dosage sensitive genes, such as members of protein complexes.

---

## Acknowledgements

First and foremost, I am greatly indebted to my supervisor Alex Bateman for patience, guidance and ongoing support over the last three years. Secondly, I want to thank Robert Finn not only for creating the databases and tools which form the basis of most of this work but also for countless fruitful discussions that greatly helped me to put a sense of direction into this work.

I also wish to thank all current and previous members of the Pfam, Rfam and Merops teams. I am grateful for the many helpful suggestions by my Thesis Committee (Sarah Teichmann and Anton Enright). Many people have contributed in various ways to this work: Donald Conrad, Gavin Wright, Ben Lehner, Manolis Dermitzakis, Matt Hurles, Christine Bird to name just a few. Cara Woodwark, Marija Buljan and Rafael Najmanovich kindly read a first draft of this manuscript and provided many helpful comments.

Finally, I want to thank Roberta for putting up with me, and for supporting me over the years, especially in the last few months.

This work was made possible by generous financial support from the Wellcome Trust.

# Contents

Summary . . . . .	iii
Acknowledgements . . . . .	v
<b>1 Introduction</b>	<b>1</b>
1.1 Protein Interactions . . . . .	2
1.1.1 Methods to detect protein interactions . . . . .	6
1.1.1.1 Affinity purification based methods . . . . .	6
1.1.1.2 The yeast-two-hybrid approach . . . . .	9
1.1.1.3 Literature Curation . . . . .	11
1.1.1.4 X-ray crystallography . . . . .	12
1.1.1.5 Other methods . . . . .	16
1.1.2 Error rate and coverage . . . . .	17
1.1.3 Protein Interaction Databases . . . . .	18
1.1.4 Interactomics - The science of networks . . . . .	20
1.2 Genetic variation . . . . .	22
1.2.1 Types and causes of mutations . . . . .	22
1.2.2 Human variation . . . . .	24
1.2.3 Variation in healthy individuals . . . . .	25
1.2.3.1 Genetic diseases . . . . .	26
1.3 Protein Domains and the Pfam database . . . . .	28



---

1.3.1	<i>i</i> Pfam . . . . .	30
1.4	Outline of this thesis . . . . .	34
<b>2</b>	<b>Distribution and evolution of interacting domains</b>	<b>35</b>
2.1	Introduction . . . . .	35
2.2	Methods . . . . .	37
2.2.1	Protein interaction data . . . . .	37
2.2.2	Filtering . . . . .	39
2.2.3	Species . . . . .	39
2.2.4	<i>i</i> Pfam . . . . .	40
2.2.5	Prediction of crystal contacts . . . . .	41
2.2.6	Random Networks . . . . .	43
2.2.7	P-values . . . . .	44
2.3	Results . . . . .	44
2.3.1	Coverage of <i>i</i> Pfam domain pairs on different interactomes . . . . .	44
2.3.2	Domain pair frequency within interaction networks . . . . .	46
2.3.3	Promiscuous domain pairs . . . . .	48
2.3.4	Domain co-occurrences . . . . .	50
2.3.5	<i>i</i> Pfam domain pairs in stable complexes of <i>S. cerevisiae</i> . . . . .	52
2.3.6	<i>i</i> Pfam domain pair conservation between species . . . . .	53
2.3.7	Predicting the total number of <i>i</i> Pfam domain pairs in nature . . . . .	57
2.4	Discussion . . . . .	59
2.4.1	Many domain–domain interfaces remain to be resolved . . . . .	59
2.4.2	<i>i</i> Pfam domain pairs can act as modules . . . . .	60
2.4.3	<i>i</i> Pfam domain pairs are conserved during evolution . . . . .	61
<b>3</b>	<b>Disease mutations in interaction interfaces</b>	<b>63</b>
3.1	Introduction . . . . .	63

---

3.2	Materials and Methods . . . . .	66
3.2.1	Disease Mutations . . . . .	66
3.2.2	<i>i</i> Pfam . . . . .	67
3.2.3	Predicting crystal contacts . . . . .	67
3.2.4	Homology Detection and Alignment . . . . .	69
3.2.5	Residue prevalence . . . . .	69
3.2.6	Alanine Scanning Database . . . . .	71
3.2.7	Compiling the curated set of interaction-related mutations . . . . .	72
3.2.8	Statistical Analysis . . . . .	72
3.2.9	Graphics . . . . .	72
3.3	Results . . . . .	72
3.3.1	Prediction algorithm . . . . .	72
3.3.2	Prediction accuracy . . . . .	73
3.3.2.1	Percent sequence identity with structural template . . . . .	73
3.3.2.2	Prevalence of mutated residues . . . . .	75
3.3.2.3	Cross validation results . . . . .	75
3.3.2.4	ASEdb results . . . . .	77
3.3.3	Application to Disease Mutations . . . . .	81
3.3.4	Properties of mutations in interaction interfaces . . . . .	83
3.3.4.1	Curated set of interaction-related mutations . . . . .	83
3.3.4.2	Classification according to function . . . . .	84
3.3.4.3	Mode of inheritance . . . . .	85
3.3.4.4	Residue frequency . . . . .	85
3.3.5	Examples of putative interaction-related mutations . . . . .	86
3.3.6	2-Methyl-3-Hydroxybutyryl-CoA Dehydrogenase Deficiency [OMIM: #300438] . . . . .	86
3.3.6.1	GrisCELLI syndrome, type 2 [OMIM: #607624] . . . . .	88

3.3.6.2	ACTH deficiency [OMIM: #201400] . . . . .	91
3.3.6.3	Baller-Gerold Syndrome [OMIM: #218600] . . . . .	91
3.4	Discussion . . . . .	94
3.4.1	Accuracy of interacting residue prediction . . . . .	94
3.4.2	Disease causing interacting residues occur frequently . . . . .	95
3.4.3	Enrichment for dominant mutations . . . . .	96
<b>4</b>	<b>Protein complexes, dosage sensitivity and copy-number variations</b>	<b>97</b>
4.1	Introduction . . . . .	97
4.2	Methods . . . . .	103
4.2.1	Gene identifiers . . . . .	103
4.2.2	Mammalian protein complexes . . . . .	103
4.2.3	Interaction and complex data . . . . .	105
4.2.4	Set of dosage sensitive genes . . . . .	106
4.2.5	Expression profiles . . . . .	106
4.2.6	Correlation computation . . . . .	109
4.2.7	Copy-number variations . . . . .	109
4.2.8	Segmental duplications . . . . .	111
4.2.9	Gene Ontology analysis . . . . .	111
4.2.10	Identification of paralogs . . . . .	111
4.2.11	Analysis of selection pressure . . . . .	112
4.2.12	P-Values . . . . .	112
4.3	Results . . . . .	113
4.3.1	Effects of CNVs on gene expression . . . . .	113
4.3.2	Limited expressional noise of protein-complex genes . . . . .	116
4.3.3	Dosage sensitive genes and CNVs . . . . .	120
4.3.4	Compositional bias of copy-number varied genes . . . . .	121

4.4	Discussion . . . . .	123
4.4.1	Protein complexes are sensitive to alterations in gene expression	123
4.4.2	CNVs affect expression levels of contained genes . . . . .	124
4.4.3	CNVs as the source of recent duplications . . . . .	126
4.4.4	Dosage sensitivity and negative selection on CNVs . . . . .	127
<b>5</b>	<b>Concluding Remarks</b>	<b>128</b>
	<b>References</b>	<b>131</b>
	<b>Appendices</b>	
	<b>A</b>	
	<b>B</b>	
	<b>C</b>	
	<b>D</b>	
	<b>E</b>	
	<b>F</b>	
	<b>G</b>	
	<b>H</b>	
	<b>I</b>	
	<b>J</b>	