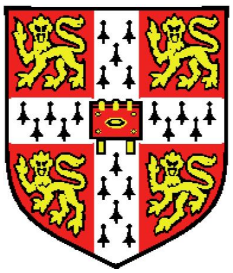
The role of protein interactions in evolution and disease



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