



2809077481

REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author JONSSON
Páll Freyr

COPYRIGHT

This is a thesis accepted for a Higher Degree of the University of London. It is an unpublished typescript and the copyright is held by the author. All persons consulting the thesis must read and abide by the Copyright Declaration below.

COPYRIGHT DECLARATION

I recognise that the copyright of the above-described thesis rests with the author and that no quotation from it or information derived from it may be published without the prior written consent of the author.

LOAN

Theses may not be lent to individuals, but the University Library may lend a copy to approved libraries within the United Kingdom, for consultation solely on the premises of those libraries. Application should be made to: The Theses Section, University of London Library, Senate House, Malet Street, London WC1E 7HU.

REPRODUCTION

University of London theses may not be reproduced without explicit written permission from the University of London Library. Enquiries should be addressed to the Theses Section of the Library. Regulations concerning reproduction vary according to the date of acceptance of the thesis and are listed below as guidelines.

- A. Before 1962. Permission granted only upon the prior written consent of the author. (The University Library will provide addresses where possible).
- B. 1962 - 1974. In many cases the author has agreed to permit copying upon completion of a Copyright Declaration.
- C. 1975 - 1988. Most theses may be copied upon completion of a Copyright Declaration.
- D. 1989 onwards. Most theses may be copied.

This thesis comes within category D.

☐

This copy has been deposited in the Library of UCL

☐

This copy has been deposited in the University of London Library, Senate House, Malet Street, London WC1E 7HU.

Computational Analysis of Protein-Protein Interaction Networks

Páll Freyr Jónsson

August 2006

Biomolecular Modelling Laboratory,
Cancer Research UK London Research Institute
and
Department of Biochemistry and Molecular Biology,
University College London

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
at the University of London.

UMI Number: U592937

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U592937

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I, Páll Freyr Jónsson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Protein-protein interactions play a crucial role in all biological systems and an increasing emphasis has been placed on identifying the full repertoire of interacting proteins in cellular systems. Recent developments have enabled large-scale screening of protein interactions, which has yielded extensive information on protein-protein interactions. These efforts have been complemented by a number of methods aimed at predicting interactions *in silico*, based on a variety of factors, ranging from sequence to structural features.

This work explores the theme of protein-protein interactions, starting with the molecular aspect of proteins, leading on to predicting interaction partners and, at the top level, examining genome-scale protein-protein interaction networks. On the molecular level, the sequence and structural details of proteins were examined, particularly focusing on the location of intron-exon boundaries in relation to protein interfaces. In addition, a homology-based method for predicting protein-protein interactions was developed, along with a scoring function for estimating the confidence of the prediction. Large-scale protein networks or 'interactomes' for key species were constructed, followed by a validation of the scoring function which confirmed its usefulness as an indicator of prediction reliability.

The value of the predicted interactomes was demonstrated by two separate studies. First, the overall topology of the human interactome was examined and the network properties of cancer-related proteins compared to non-cancer proteins. Cancer-related proteins were shown to display network characteristics that differed markedly from non-cancer proteins. The second study was aimed at identified key proteins likely to be implicated in cancer metastasis. This was done by mapping gene expression data from highly metastatic rat cell-lines onto the rat interactome. A cluster analysis of the data revealed distinct, tightly interconnected protein communities that play a role in metastasis.

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible. First of all, I remain very grateful to Paul Bates for taking me on board his team to work on this challenging project and for his supervision. It is also a pleasure to thank all the members of the Biomolecular Modelling Laboratory at the London Research Institute for all the time we had together in the last few years. I extend my gratitude to Bruno, who encouraged me during our collaboration in the initial stage of my studies. My work would not have been possible without the help from Paul Fitzjohn, who frequently rescued me with his technical skills and was an endless source of solutions to every computational problem thinkable. Many thanks go to Chris for all the good discussions we had at the Seven Stars—and not to mention for teaching me how to typeset with Latex. Marc, Marcin, Alex, Katie and Raphael were also fantastic—it was a pleasure to work with you all.

My appreciation goes to Daniel Zicha and Tamara Cavanna in the Light Microscopy lab for the fruitful collaborations, and I am, of course, especially indebted to Cancer Research UK and its supporters for the funding and making my research possible. Furthermore, I owe a great deal to Jón Bragi Bjarnason og Sigríður Ólafsdóttir, whose professional enthusiasm has continued to motivate me since my undergraduate days at the University of Iceland.

I cannot end without thanking my entire family, including Roland, whose constant encouragement and support has made everything possible. It is to you that I dedicate this work.

Contents

Abstract	3
Acknowledgements	4
Contents	5
List of Figures	9
List of Tables	11
List of Abbreviations	12
Peer-reviewed publications	14
1 Introduction	15
1.1 The systems view of biology	16
1.2 Biological databases	19
1.2.1 Protein sequence depositories	19
1.2.2 The Protein Data Bank	20
1.2.3 Structural classification	20
1.2.4 ASTRAL	23
1.2.5 OMIM	23
1.2.6 Gene Ontology	24
1.2.7 Interaction data	24
1.3 Methods for detecting protein interactions	25
1.3.1 Structure-based methods	26
1.3.2 Biochemical methods	26
1.4 Global experimental analyses	27
1.4.1 Genome screening by yeast two-hybrid	27

1.4.2	Quality assessment of high-throughput data	29
1.5	Properties of networks	31
1.6	Towards predicting interactions	34
1.6.1	Structure-based prediction	35
1.6.2	Structure-independent prediction	36
1.7	Overview of the thesis	43
2	Exon arrangement and protein structure	45
2.1	Gene expression	46
2.2	Origin of introns	47
2.2.1	Splicing	48
2.3	Protein structure	51
2.3.1	Primary structure	51
2.3.2	Secondary structures	52
2.3.3	Tertiary and quaternary structure	53
2.4	Methods	53
2.4.1	Data	53
2.4.2	Assignment of introns to protein sequences	54
2.4.3	Secondary structure assignment	54
2.4.4	Structural conservation and IEBs	55
2.4.5	Domain contacts from PDB structures	55
2.4.6	Distinguishing obligomers from transient complexes	56
2.4.7	Sequence conservation assessment	58
2.5	Results and Discussion	60
2.5.1	Secondary structure at intron-exon boundaries	60
2.5.2	Local structural variability at intron-exon boundaries	62
2.5.3	Packing of exons using structural alignments	64
2.5.4	Intron-exon boundaries and their relation to domain interfaces	64
2.5.5	Intron-exon boundaries and sequence conservation	70
2.5.6	Splice variation in interfaces	72
2.6	Conclusions	73
3	<i>In silico</i> construction of interactomes	77
3.1	Predicting protein interactions	77
3.2	Relational databases	78

3.2.1	One-to-one relationships	79
3.2.2	One-to-many relationships	80
3.2.3	Many-to-many relationships	80
3.2.4	Normalisation and table design	80
3.3	Methods	83
3.3.1	Prediction of protein-protein interactions	83
3.3.2	ROC curve analysis	84
3.3.3	PIP database schema	88
3.3.4	Dynamic interaction maps	92
3.4	Results and Discussion	92
3.4.1	PIP server	92
3.4.2	Network validation	97
3.4.3	Network properties	99
3.5	Conclusions	102
4	Global topological features of cancer proteins	104
4.1	Introduction	104
4.1.1	Community structure in networks	105
4.2	Methods	107
4.2.1	Clustering of interconnected proteins	107
4.2.2	Data sets	107
4.3	Results and Discussion	108
4.3.1	Network properties of cancer proteins	108
4.3.2	Conclusions	120
5	Clustering of metastasis-related proteins	123
5.1	Introduction	123
5.1.1	Gene expression analysis	123
5.1.2	Network analysis of expression data	124
5.1.3	Metastasis	126
5.2	Methods	127
5.2.1	Validation	127
5.2.2	Microarray expression data for metastatic rat cells . . .	128
5.2.3	Creation of networks around differentially expressed genes	129
5.3	Results and Discussion	130

5.3.1	Validation of the scoring function	130
5.3.2	Identification of metastatic communities	134
5.3.3	Network view of gene expression	140
5.4	Conclusions	142
6	Concluding remarks	143
6.1	Intron-exon boundaries	143
6.2	Predicting interactions	144
6.3	Interactome analysis	145
6.4	Future directions	145
A	Statistical methods	147
A.1	Comparative statistics	147
A.2	Statistics for frequency data	149
A.3	Association statistics	149
B	A detailed view of the human cancer communities	152
C	KEGG pathway information for human protein communities	158
D	A closer view of the metastasis-related communities	169
E	Annotations for the metastasis-related proteins	174
F	Internet resources	184
	Bibliography	185

List of Figures

1.1	Network models	15
1.2	Developments in system-wide biology	17
1.3	Protein-protein interaction network	18
1.4	Structures in the PDB	21
1.5	Experimental evidence of interactions	30
1.6	Network models	33
1.7	Principles of SVMs	37
1.8	Sequence similarity vs. interaction similarity	41
2.1	Eukaryotic protein expression	46
2.2	Exon splicing	49
2.3	Variable splicing patterns	51
2.4	The structure of α -helices and β -sheets	52
2.5	Solvent accessible surface area	55
2.6	Detecting domain-level interactions	56
2.7	Detailed domain contact information	57
2.8	Location of IEBs in secondary structure elements	61
2.9	Structural conservation of IEB residues	63
2.10	Packing arrangements at intron-exon boundaries	65
2.11	53BP1 complexed with tumour-suppressor protein p53	67
2.12	p53 Tumour Suppressor alignments	68
2.13	p53 Binding Protein alignments	69
2.14	Frequency of IEBs by conservation groups	71
2.15	Exon conservation in protein complex interfaces	72
2.16	Exon arrangement in trypsin	75
3.1	Inferring interactions by homology	85
3.2	Table design for the PIP database	89

3.3	PIP query interface	93
3.4	PIP results page	94
3.5	PIP results page (cont.)	95
3.6	The interactive protein map	96
3.7	ROC-curve for the interaction prediction	98
3.8	Connectivity distributions: human and rat	101
3.9	Connectivity distributions: fission yeast	102
4.1	Community structure	106
4.2	Connectivity of cancer proteins	109
4.3	k -clique cluster analysis on the human interactome	115
4.4	Cellular roles of protein communities at $k = 6$	117
4.5	Community sizes of cancer and non-cancer proteins	118
4.6	Network centrality	119
4.7	The network aspect of HSPCA	122
5.1	High-density oligonucleotide microarrays	125
5.2	Bit scores as a function of sequence identity	132
5.3	Score distribution in the rat interactome	133
5.4	Identifying protein communities by cluster analysis	136
5.5	The intracellular signalling cascade	139

List of Tables

1.1	Large-scale Y2H maps	28
1.2	Database servers for predicted interactions	35
2.1	Secondary structure of IEB residues	62
2.2	Frequencies of IEBs in interfaces of proteins	70
2.3	Frequencies of splice variants in interfaces.	73
3.1	Experimental sources for building the interactome	86
3.2	The model organisms used for prediction	87
3.3	Sensitivity, specificity and PPV	97
3.4	The number of predicted interactions	99
4.1	Connectivity of cancer proteins	108
4.2	Connectivity at variable score cut-offs	110
4.3	Connectivity of somatic and germline mutated cancer genes	111
4.4	The 20 most frequently observed protein domains	113
4.5	Number of communities at different k -values	114
4.6	Multiple community membership distribution	116
4.7	Biological processes and connectivity	120
5.1	The metastatic cascade	126
5.2	Gene ontology cellular compartments	128
5.3	Protein communities at different clustering threshold values	130
5.4	Distribution of protein-protein interaction scores	131
5.5	Domain frequency within the clustered communities	137
5.6	The connectivity of up- and down-regulated proteins	141

List of Abbreviations

ASA	Accessible Surface Area
BIND	Biomolecular Interaction Network Database
BLAST	Basic Local Alignment Search Tool
CATH	Class, Architecture, Topology and Homologous superfamily
cDNA	Complementary DNA
CGI	Common Gateway Interface
DBI	Database Interface
DBMS	Database Management System
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic Acid
DSSP	Dictionary of Protein Secondary Structures
E-value	Expectation Value
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
GO	Gene Ontology
GUI	Graphical User Interface
HMM	Hidden Markov Model
HPRD	Human Protein Reference Database
HTML	Hypertext Markup Language
IEB	Intron-Exon Boundary
IP	Immunoprecipitation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MINT	Molecular INTeraction Database
MIPS	Munic Information Center for Protein Sequences
mRNA	Messenger RNA
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information

NMR	Nuclear Magnetic Resonance
OMIM	Online Mendelian Inheritance in Man
ORF	Open Reading Frame
PDB	Protein Data Bank
PIP	Potential Interactions of Proteins
PIR	Protein Information Resource
PPI	Protein-protein interaction
PSI	Proteomic Standards Initiative
RMSD	Root Mean Squared Deviation
RNA	Ribonucleic Acid
SCOP	Structural Classification of Proteins
SGC	Structural Genomics Consortium
siRNA	short interfering RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoproteins
SSE	Secondary Structure Element
SVM	Support Vector Machine
SQL	Structured Query Language
TAP	Tandem Affinity Purification
tRNA	Transfer RNA
TrEMBL	Translation of coding sequences from the EMBL database
URL	Uniform Resource Locator
WWW	World Wide Web
XML	Extensible Markup Language
Y2H	Yeast two-hybrid

Peer-reviewed publications

A chronological list of publications published in peer-reviewed journals during the PhD project:

- Contreras-Moreira, B.* , Jonsson, P. F.* and Bates, P. A. (2003). Structural context of exons in protein domains: implications for protein modelling and design. *Journal of Molecular Biology*, **333**:1057–1071.
* Joint first authors.
- Jonsson, P. F., Cavanna, T., Zicha, D. and Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, **7**:2.
- Jonsson, P. F. and Bates, P. A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**:2291–2297.

Chapter 1

Introduction

The world of proteins with its fascinating diversity and functionality has been the source of much interest over the decades. Proteins, whose name originates from the Greek *protas* meaning ‘of primary importance’, are the building blocks of life. They are polymers of amino acids and serve in a variety of ways, ranging from having structural and mechanical roles in the cell, to catalysing chemical reactions and transporting cellular signals. Proteins are therefore essential to all living cells.

In most organisms, proteins are encoded in strands of deoxyribonucleic acid, or DNA. This forms the basis of the ‘central dogma’ in biology, proposed by Francis Crick (1970), which states that information is transferred from DNA to proteins in an irreversible process. The ‘sequence hypothesis’ that is derived from the central dogma outlines the information transfer of coded genetic information: from DNA that is transcribed to ribonucleic acid, RNA, and in turn translated to proteins (see Figure 1.1).

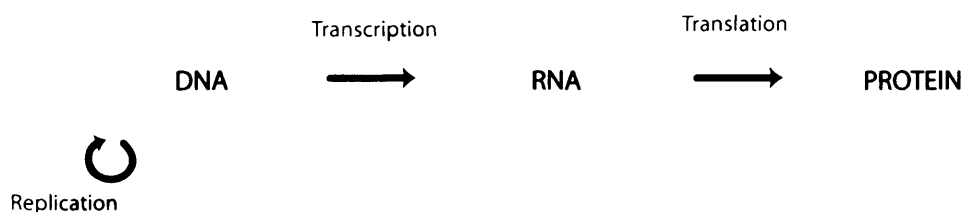


Figure 1.1: The sequence hypothesis that is derived from the central dogma of biology represents the general flow of information that is coded in DNA. The pathway can be summarised in a simplified manner, with the genetic information originating in DNA, which is then transcribed to RNA, from which proteins are translated.

Since the idea was first put forward, it has developed to reflect the added knowledge of biological systems that has accumulated over the years. For instance, it has become apparent that the DNA \rightarrow RNA transfer can be reversed in a process that makes use of reverse transcriptase, and that some RNA is non-coding, and can be spliced out via constitutive or alternative splicing. Nevertheless the underlying idea of the sequence hypothesis is still valid and is the foundation on which our understanding of molecular biology is built.

1.1 The systems view of biology

The recent developments in genome sequencing have resulted in an extraordinary amount of genomic data that provides a wealth of information; currently, just under 370 prokaryotic and 22 eukaryotic genomes have been completely sequenced (NCBI Genome Project, 2006). The sequence data has enabled gene prediction for most of the genomes and this, in turn, allows scientists to interpret this data in terms of the gene products—the proteins.

Even with the genes and the corresponding proteins established, there is still a long way to go until a full understanding of the functional aspects of the proteins in a given organism is achieved. A great emphasis has therefore been put on unravelling the functional features of proteins through large-scale screenings of protein interactions, building on the idea that a protein's function can be inferred by looking at the proteins with which it interacts. Moreover, protein-protein interactions are crucial for mediating most biological processes: tens of thousands of proteins and other macromolecules are, at any given time, expressed in a typical cell, and each will participate in a number of interactions with other proteins (Alberts *et al.*, 2002). A comprehensive determination of all protein-protein interactions that take place in an organism is therefore needed to provide an understanding of the molecular basis of the cellular function.

As technology advances, the emphasis of research efforts has moved from *genomics* to *proteomics*. The term proteomics refers to the study of the complete set of proteins that is expressed by the entire genome in a cell. The advances in proteomic study were largely driven by methods for large-scale protein separation and identification (such the combination of chro-

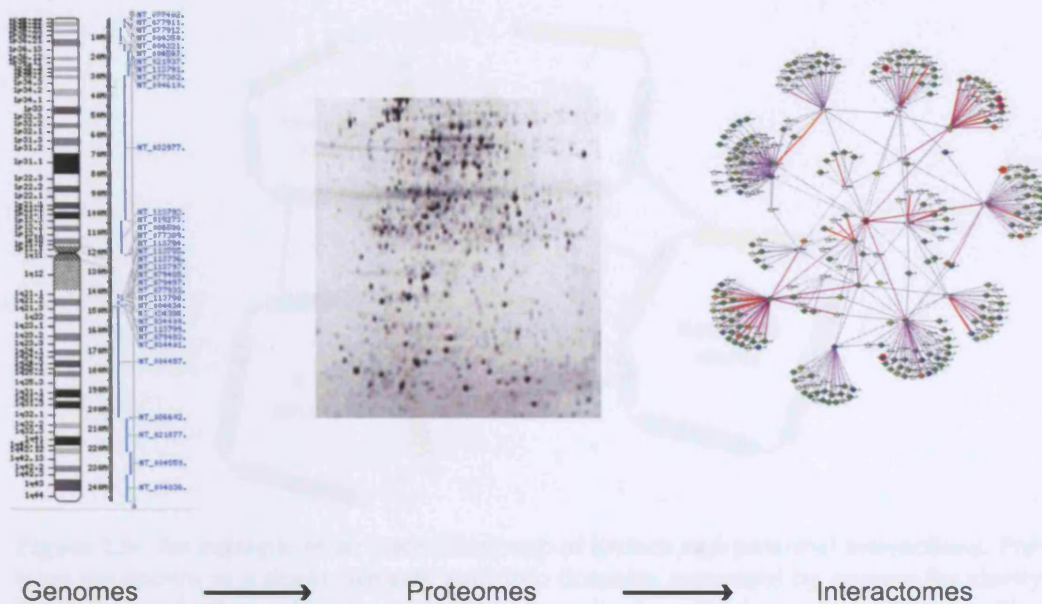


Figure 1.2: The advances in genome sequencing have brought research emphasis from genomics to proteomics, which focuses on identification of the entire set of expressed proteins under a given cellular condition. More recently, system wide approach for examining the complex networks of protein interactions, or interactomics, has gained momentum as a way towards understanding biological systems.

matography with mass spectrometry, see Lee and Lee (2004) for a review). Once the participating proteins have been identified, the next logical step is to examine the context in which the protein exists. This is the subject of *protein interactomics*, the study of the complete set of physical interactions mediated by all proteins of an organism (see Figure 1.2). The emergence of interactomics has been facilitated by another technological leap which has enabled large-scale and high-throughput mapping of protein interactions, where yeast two-hybrid (Y2H) is one of the major methods (see Section 1.3). As this has progressed, scientists are increasingly examining cellular processes in terms of networks of protein-protein interactions, rather than following the more traditional reductionist view of functions brought about through simple molecular pathways. This approach is especially appropriate when complicated diseases, such as cancer, are being investigated. As more and more signalling pathways are being identified, it has become apparent that cellular regulation is achieved through networks of signals, rather than simple linear pathways (Weng *et al.*, 1999; Hornberg *et al.*, 2006).

In spite of ever increasing data on protein interactions, there are still

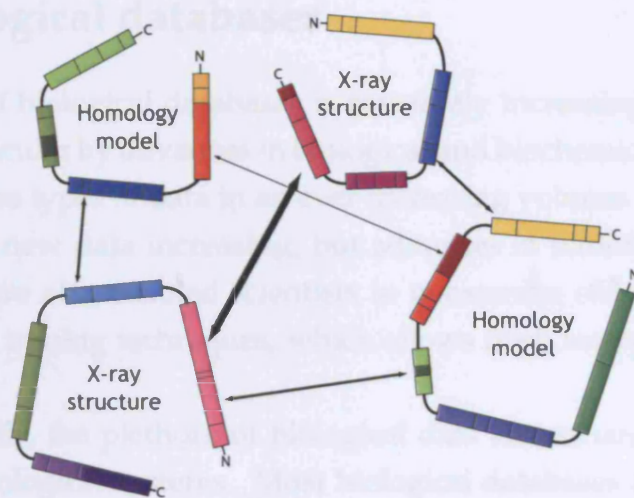


Figure 1.3: An example of an interaction map of known and potential interactions. Proteins are shown in a linear manner, split into domains separated by spacers for clarity. Domains are further split into exons and alternatively spliced exons are shown with a darker shade of domain colour. Proteins, whose structures have not been solved but for which interaction data exists (e.g. from Y2H studies), are labelled as “models”, whereas proteins of known structure are labelled “X-ray”. The strength of association is represented by the arrow width; crystallographic data having the strongest association and data involving protein models weaker association.

gaps in our knowledge of genome-wide interactome maps, partially due to lack of coverage, and also as a result of the experimental errors associated with some of the high-throughput screening methods. Empirical data must therefore be integrated with comparative and predictive bioinformatics analyses.

Towards this end, much effort is being channelled into firstly, expressing known protein-protein interaction information in a form that is useful for interpretation and, secondly, filling in as many gaps in our knowledge as possible, by considering the evolutionary aspects of proteins, sequence and function. Figure 1.3 shows the principle of one interpretation of the idea, depicting a hypothetical contact map between a few proteins in a eukaryotic genome. It shows a mixture of known and predicted data, the ratio of which will continually change as experimental data accumulates.

1.2 Biological databases

The number of biological databases is constantly increasing. This trend is fuelled in particular by advances in biological and biochemical research that produce diverse types of data in an ever increasing volume. Not only is the availability of new data increasing, but advances in bioinformatics in the last decade have also enabled scientists to re-examine old data, for example using data mining techniques, which allows fresh interpretation of old discoveries.

Undoubtedly, the plethora of biological data has enhanced our understanding of biological systems. Most biological databases available to the public are accessible on the Internet and they cover aspects such as DNA, RNA and protein sequences, gene mutation and disease information, protein structures, enzymatic activity and ligand binding, to name a few. Before proceeding further it is appropriate to introduce the data sources that were particularly valuable in this work.

1.2.1 Protein sequence depositories

The amount of sequence data has expanded enormously in the last decade. The largest amount of amino acid data is now produced by translation of nucleic acid sequences, whose numbers have been on the increase as ever more genomic sequencing projects are completed. Presently, sequence data and associated annotations are stored in several databases run by separate organisations. The Swiss Institute of Bioinformatics has collaborated with the European Bioinformatics Institute (EBI) to construct a database of annotated protein sequences, named Swiss-Prot (Boeckmann *et al.*, 2003). Swiss-Prot is a curated protein sequence database of low redundancy that includes diverse information associated with its sequences (such as description of protein functions, domain structure, post-translational modification, alternative splicing, etc.). TrEMBL is the name of a computer-annotated supplement to Swiss-Prot and it contains sequences not yet ready to be included in the main database. The Protein Information Resource, or PIR, (Wu *et al.*, 2003) is another high-quality database established by the National Biomedical Research Foundation in 1984. These three databases have now joined forces under the name UniProt (Wu *et al.*, 2006), and function as a central

repository of protein sequence and function.

Another important source for protein sequence information, and one that is primarily used in this thesis, is the RefSeq database (Pruitt *et al.*, 2005) which is maintained by the National Center for Biotechnology Information (NCBI). RefSeq provides annotated sequence information on both nucleotide and protein sequences. Its main strength lies in the emphasis on data validation, format consistency and non-redundancy. This means that it is smaller in size than many other databases; however, the data integrity makes it a good source for genome-wide protein studies.

1.2.2 The Protein Data Bank

The Research Collaboratory for Structural Bioinformatics maintains the Protein Data Bank (PDB) (Berman *et al.*, 2000). The PDB is the primary repository of three-dimensional protein structures in the scientific community. The database was constructed by Bernstein *et al.* (1977). Initially the rate of submitted structures was very low, which reflects the limitations of the experimental methods that were used at the time. By 1986 the number of protein structures had reached 214, with an average of 23 structures being added to the database yearly. With advances in technology, the rate of submission has increased exponentially, as can be seen in Figure 1.4, and in July 2006 the PDB contained 37,873 structures in total. This trend is expected to continue, boosted by the efforts of the Structural Genomics Consortium (SGC) (Williamson, 2000), which is filling in the gaps of the protein structure space by targeting a large number of proteins relevant to human health and disease. The PDB contains structures solved primarily by two methods: X-ray crystallography and Nucleic Magnetic Resonance. Lately there has been an addition of another method, electron microscopy, which has to date yielded 127 structures, particularly large macromolecular structures.

1.2.3 Structural classification

The Structural Classification of Proteins (SCOP) database is a manually curated database that orders proteins of known structure according to their evolutionary and structural relationships (Murzin *et al.*, 1995). SCOP describes protein structures in the basic units of domains, which are evolution-

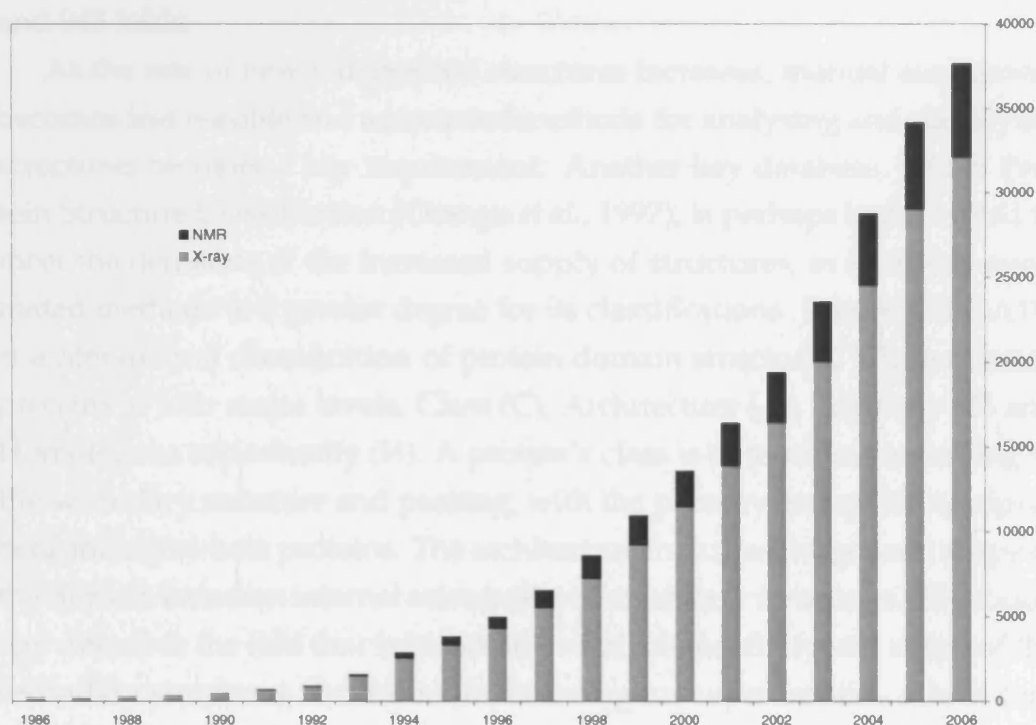


Figure 1.4: The number of structures in the Protein Data Bank over the last twenty years. The structures are divided into the two main experimental methods by which they are solved: X-ray crystallography and NMR. The third method (electron microscopy) has yielded limited number of data to date (127 structures in 2006).

ary units of protein structure and can exist either on their own or as building blocks in multi-domain proteins. The domains in SCOP are hierarchically classified into increasingly descriptive groups, starting with the most general class description (α -helix proteins, β -sheet proteins, etc.) and further subclassifications into folds, superfamilies and finally into families. The first official SCOP release in 1995 comprised 3,179 protein domains grouped into 498 families, 366 superfamilies and 279 folds. The latest version of the database (release 1.69, July 2005) contains 2,845 families, 1,539 superfamilies and 945 folds.

As the rate of newly deposited structures increases, manual assignment becomes less feasible and automated methods for analysing and classifying structures becomes a key requirement. Another key database, CATH Protein Structure Classification (Orengo *et al.*, 1997), is perhaps better suited to meet the demands of the increased supply of structures, as it utilises automated methods to a greater degree for its classifications. Like SCOP, CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). A protein's class is determined according to the secondary structure and packing, with the primary groups being alpha, beta and alpha-beta proteins. The architecture indicates the general shape of the domain based on internal orientation of secondary structures. The topology describes the fold that is based on overall connectivity and shape of the secondary structures. Furthermore, homologous superfamilies, which contain proteins thought to share a common ancestor, get their classification based on sequence and structural similarity. Domains within each homologous superfamily are further classified into sequence families based on their sequence identity overlap. In July 2006 CATH contained 40 individual architecture types, 1,110 types of topology, 2,147 homologous superfamilies and 7,841 sequence families.

Another database that documents protein families and domains is Pfam (Bateman *et al.*, 2004). This database contains multiple sequence alignments representing each family, along with domain architecture, functional annotation, literature reference and database links for each family. Each family is assigned two sets of multiple alignments. One is the seed alignment that contains a small number of representative sequences and the other is the

full alignment that contains all members in the database that can be detected. In addition, each family contains profile hidden Markov models (profile HMMs) (see e.g. Eddy (1996)) for detecting domains in sequences not already covered in its alignments.

An offshoot of Pfam is the iPfam database, which is a resource that describes domain-domain interactions that are observed in PDB entries. It maps the parent's database domain definitions onto protein structures. When multiple domains occur in a single structure, the domains are analysed to see if they form interactions. Version 20.0 of iPfam contains 3,019 domain-domain interactions of which 25% are intrachain and 75% are inter-chain interactions.

1.2.4 ASTRAL

The ASTRAL database (Chandonia *et al.*, 2004a) complements and is partially derived from the SCOP database. The database provides sequence-level information on the structures in the PDB by incorporating domain information from SCOP. PDB files were primarily designed for containing structural detail of proteins. The implications are that extracting sequence details of domains from PDB files is often cumbersome and automatic sequence retrieval from these files can be inaccurate. ASTRAL was provided as a solution to these problems. Another feature of ASTRAL, and perhaps the most useful one, is the classification of proteins into groups of homologues, which brings together structures that share similarity at the sequence level.

1.2.5 OMIM

Of particular interest to scientists involved in cancer research is OMIM—'Online Mendelian Inheritance in Man'—which is compiled by Johns Hopkins University (Hamosh *et al.*, 2005). OMIM is a catalogue of human genes and genetic disorders with links to both literature and sequences. It first appeared in a printed form in 1966 and has since been curated by physicians and scientists and this mixture of contributors makes it a valuable source for unravelling the complex relationship between genes and disease.

1.2.6 Gene Ontology

The Gene Ontology (GO) project is a collaborative effort to define a set of controlled and unified vocabulary for genes and gene products in a wide variety of organisms (Gene Ontology Consortium, 2006). The project was initiated in 1998 and initially included data on *Drosophila melanogaster*, *Saccharomyces cerevisiae* and the mouse genomes. Since then several plant, animal and microbial genomes have been added. The structure of the ontologies is hierarchical; the top level contains general descriptions of the gene products in terms of their associated properties, i.e. 'biological processes', 'cellular components' and 'molecular functions'. These ontologies successively branch to form a complex web of descriptions, which have been used by scientists to help interpret complicated genome-size data sets.

1.2.7 Interaction data

Currently, there are numerous sources for direct experimental evidence of protein-protein interactions. Many cover single-species experimental data, such as the Human Protein Reference Database, HPRD (Peri *et al.*, 2003), the Munich Information Center for Protein Sequences yeast database, MIPS CYGT (Güldener *et al.*, 2005) and the CuraGen database for *Drosophila melanogaster* (Giot *et al.*, 2002), to name a few. Multiple genome databases are, however, progressively covering most single-gene data sources. An example of the available multi-species databases are BIND (Biomolecular Interaction Network Database) (Bader *et al.*, 2003) which, in addition to information on interactions, contains details on molecular complexes and pathways from various sources; BioGRID (Stark *et al.*, 2006), which includes interactions primarily from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*; MINT (Molecular INTERaction Database) (Zanzoni *et al.*, 2002), which data is mined from the scientific literature; and DIP, Database of Interacting Proteins (Xenarios *et al.*, 2002), which includes both high-throughput data from multiple sources as well as smaller high quality data sets. In June 2006 DIP contained over 55,755 interactions in 110 species, many of which were from high-throughput experiments. In spite of being some of the most comprehensive interaction databases available, the above databases contain relatively little protein-

protein interaction data from mammals. *Saccharomyces cerevisiae*, or budding yeast, was among the first species to be extensively studied for protein-protein interactions (Uetz *et al.*, 2000; Ho *et al.*, 2002), and although useful as a model organism, medical and therapeutic research often requires mammalian models for optimal results. The MIPS Mammalian Protein-Protein Interaction Database (Pagel *et al.*, 2005) addresses this need and provides well-curated interaction data from mammalian species, however it is still quite limited in size (about 1,800 interactions among 900 proteins from 10 mammalian species in June 2005).

The advances in experimental techniques have meant that it has become cheaper and quicker to perform protein-protein interaction surveys. Accordingly, data has been made available for a diverse range of species, and is often produced by individual laboratories who collect data into their own local databases. The current situation therefore calls for a comprehensive and unified data depository for experimental results that conform to standards for experimental procedures agreed by the scientific community. The Proteomic Standards Initiative, or PSI (Hermjakob *et al.*, 2004a), has been formed as a step in that direction. The PSI has proposed a standardised data format with the Molecular Interaction (MI) Extensible Markup Language (XML) with the aim of facilitating data integration. Moreover, the PSI MI makes use of controlled vocabularies or ontologies to standardise the contents of data attributes. Combined, these approaches will make comparative analyses of protein interaction data more streamlined and promote data integration. The scientific community is already seeing the results of this effort in the IntAct database, developed by the European Bioinformatics Institute, which is effectively a meta-database, containing information from many of the above mentioned data sources (Hermjakob *et al.*, 2004b).

1.3 Methods for detecting protein interactions

Experimental methods for detecting protein-protein interactions are many and varied. In general, they can be classified into two categories: structure-based methods and biochemical methods.

1.3.1 Structure-based methods

In the first category are methods that aim to elucidate the physical structure of the proteins and the complexes they make. The methods in this first category are X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy. Used in combination with protein sequencing they can be used to identify the actual physical interaction, including the protein interfaces and their makeup. An interaction documented by structural evidence obtained by one of these methods can be considered proved to the highest standard possible—the physical aspects of the interaction can be seen and interpreted by the images they provide. These methods form the backbone of structural biology studies, but as a form of surveying a large amount of data they are very limited. X-ray crystallography, which to date has yielded most structures (see Section 1.2.2), relies on the proteins under study being highly purified, folded correctly under conditions close to physiological and subsequently crystallised. For some proteins this process can take months and others have proved very difficult to crystallise.

1.3.2 Biochemical methods

The second category encompasses a variety of biochemical and genetic methods, which lend themselves better to extensive screening of interactions. Biochemical methods particularly successful for detecting protein-protein interactions include genome-wide assays by yeast two-hybrid technology (Y2H), immunoprecipitation (IP), tandem affinity purification (TAP), mass spectrometry (MS), and protein chips (see Piehler (2005) for an overview of these methods).

The yeast two-hybrid approach is of particular interest, as it has been the method that has produced by far the most interactome data and, indeed is the method that yielded 70% of the experimental data used in the network studies in this thesis (see Table 3.1 in Chapter 3). The assay was inspired by the modular structure of transcription factors in yeast, *Saccharomyces cerevisiae*. The transcription factors contain a DNA-binding domain (DBD) and a transcription activation domain (AD) and in order for transcription to take place both domains must be in close proximity. The two genes coding for the proteins to be tested for interactions are fused to transcription factor do-

mains, one to the AD (the ‘prey’) and the other to the DBD (the ‘bait’). If the bait and prey interact the transcription factor activity is reconstituted and reporter genes that have upstream binding sites for the DBD get activated. The expression of the reporter gene allows the diploid yeast cells to grow on selective media and the interaction is thereby detected. In large-scale surveys, the DBD and AD expression vectors are placed initially into different haploid yeast strains of opposite mating types. Pairs of BD and AD fused proteins can then be tested for interaction by mating the appropriate pair of yeast strains and assaying reporter activity in the resulting diploid cells (Finley and Brent, 1994; Uetz, 2002; Stanyon *et al.*, 2004; Piehler, 2005). In addition to the above approach, several improvements and variations have been reported, see for instance a review by Toby and Golemis (2001). The simplicity and relatively inexpensive nature of the technology, and the fact that the method is easily scalable for large studies, has made yeast two hybrid extremely popular. The method is not without negative aspects though, as it has relatively high false-positive and false-negative rates (Titz *et al.*, 2004). This is often attributed to misfolding of the fusion proteins; the interaction between the bait and prey has to take place in the yeast nucleus, where many proteins are not in their native physiological environment. Moreover, the method is only suitable for binary protein-protein association and cannot automatically detect multi-protein complexes.

1.4 Global experimental analyses

1.4.1 Genome screening by yeast two-hybrid

Most large-scale protein-protein interaction screens have, to date, been performed using the yeast two-hybrid method. The first genome-wide interaction map generated by the two-hybrid approach was for *Escherichia coli* bacteriophage T7 (Bartel *et al.*, 1996). This screen was later followed by comprehensive surveys of the budding yeast *Saccharomyces cerevisiae* (Uetz *et al.*, 2000; Ito *et al.*, 2001) which paved the way for subsequent genome-wide screens in a variety of species (Parrish *et al.*, 2006). The most prominent and extensive screening efforts to date are listed in Table 1.1. Initially the yeast protein maps made up the majority of data available to the sci-

entific community, however, a few years later an extensive survey of the first metazoan proteome—the fruit fly—was published by Giot *et al.* (2002). Combined with the two subsequent *Drosophila melanogaster* surveys listed in Table 1.1 over 50% of the fruit fly proteome has been mapped (Pacifico *et al.*, 2006). The generation of this map is of particular interest to those studying human disease as over half of the human disease-associated genes in the OMIM database have orthologues in *Drosophila* (Parrish *et al.*, 2006).

Table 1.1: A comparison of 12 of the most prominent large-scale Y2H maps to date, showing the relative number of proteins and interactions detected for each study. * A subset of highly confident interactions. ** Study focusing only on the Smad signalling system.

Species	Published by	Total number	
		Proteins	Interactions
Bacteriophage T7	Bartel <i>et al.</i> (1996)	55	25
Vaccinia virus	McCraith <i>et al.</i> (2000)	266	37
Herpes virus VZV	Uetz <i>et al.</i> (2006)	69	173
<i>H. pylori</i>	Rain <i>et al.</i> (2001)	261	1,280
<i>S. cerevisiae</i>	Uetz <i>et al.</i> (2000)	1,005	905
<i>S. cerevisiae</i>	Ito <i>et al.</i> (2001)	797	754
<i>C. elegans</i>	Li <i>et al.</i> (2004)	1,415	2,082
<i>D. melanogaster</i>	Giot <i>et al.</i> (2002)	7,048 (4,679*)	20,405 (4,780*)
<i>D. melanogaster</i>	Stanyon <i>et al.</i> (2004)	488	1,814
<i>D. melanogaster</i>	Formstecher <i>et al.</i> (2005)	102	2,338 (710*)
<i>P. falciparum</i>	LaCount <i>et al.</i> (2005)	1,308	2,846
<i>H. sapiens</i>	Stelzl <i>et al.</i> (2005)	1,705	3,186
<i>H. sapiens</i>	Rual <i>et al.</i> (2005)	1,549	2,754
<i>H. sapiens</i> **	Colland <i>et al.</i> (2004)	591	755

Although multicellular model organisms have been helpful as tools for studying human disease, a complete interaction map based on human proteins would be enormously beneficial. Until recently most Y2H screens involving human proteins were focused on specific diseases or pathways and the first large-scale human maps were published by Stelzl *et al.* (2005) and Rual *et al.* (2005). The two human studies have a low level of overlap (Parrish *et al.*, 2006) making up a total of about 5900 protein interactions. This, however, is only a small a fraction of the interactions thought to take place in the human cell, as the number of genes in the human genome is currently estimated at about 20,000-25,000 (International Human Genome Se-

quencing Consortium, 2004). Ramani *et al.* (2005), for instance, combined available experimental data with literature mining and estimated that each human protein is involved in approximately 15 interactions, which implies more than 375,000 interactions in the complete human protein interaction network. This figure is very likely an underestimate, owing to a number of protein products that are produced by alternative splicing, which may affect up to 40–50% of all genes (Modrek *et al.*, 2001).

1.4.2 Quality assessment of high-throughput data

False-positive results from high-throughput Y2H studies have been a considerable problem when it comes to interpreting the results of these large-scale experiments. The size of the problem has been assessed in several publications and it has even been suggested that as much as 50% of the interactions in some earlier high-throughput surveys were false-positive (Sprinzak *et al.*, 2003; Deane *et al.*, 2002). The problem with false-positive results can be defined as two fold: Firstly, the false-positive interaction can be a result of a non-specific binding in the yeast two-hybrid assay. Duplicate assays and orthogonal experimental approaches can help to determine this class of true positives. The second issue, and one that is much harder to tackle, is the question of biological significance—i.e. in the case of a true specific binding, is the interaction likely to exist *in vivo* and contribute to a cellular function?

In order to be considered reliable, interactions have to be detected by multiple screens and, ideally, by complementary experimental methods (Bader and Hogue, 2002). For large genomes, the situation is not likely to improve considerably in the immediate future, particularly given the composition of the data available in some of the larger databases. For example, by pooling the protein-protein interaction data from both DIP and MIPS Mammalian databases and examining the underlying experimental evidence it becomes apparent that most interactions are documented by one piece of experimental evidence only (see Figure 1.5). Together the databases contained about 45,200 interactions that originated from 51,200 experimental measurements. This equates to 1.13 experiments per interaction, where 93% of interactions are supported by one experimental detail only.

The two-hybrid approach is increasingly complemented by alternative

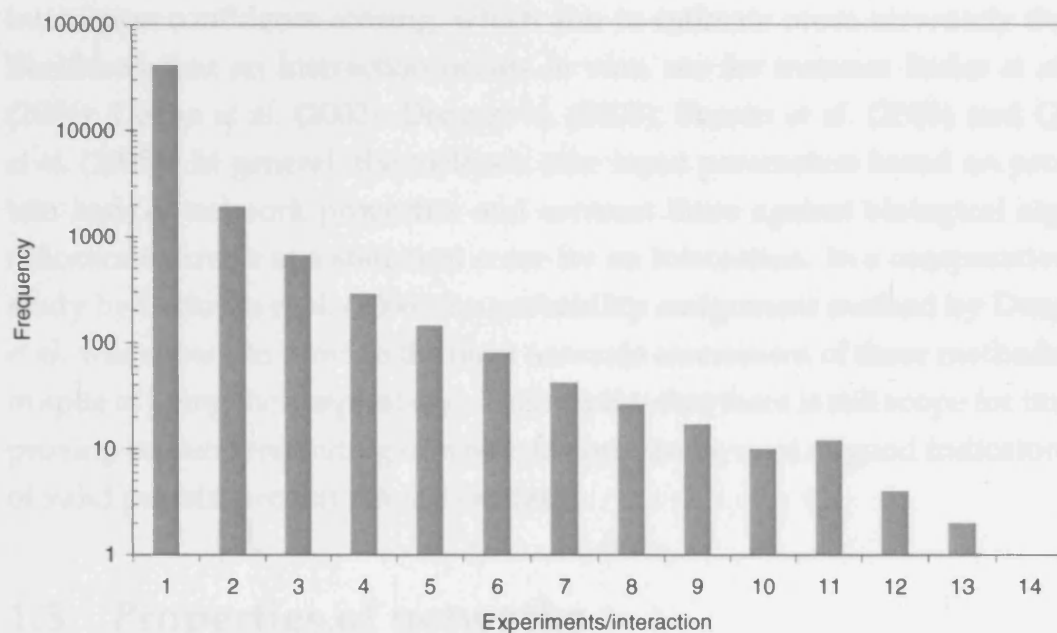


Figure 1.5: The frequency distribution of the experimental data behind documented protein-protein interactions. The abscissa shows the number of separate experiments for each documented interaction in the DIP and MIPS Mammalian Databases as at March 2005.

methods, such as mass spectroscopy, with the aim of validating the results (Bork *et al.*, 2004). Independent coimmunoprecipitation and pull-down assays, to name a few, have also been used in conjunction with yeast two-hybrid screening. These techniques are normally performed on a smaller scale, with the aim of identifying a high-confidence core set within the results. Validation of larger datasets, however, requires a bioinformatics approach for validation owing to the expensive and time consuming nature of the complementary experimental techniques.

For this purpose, several groups have proposed different methods for interaction confidence scoring, which aim to estimate more accurately the likelihood that an interaction occurs *in vivo*, see for instance Bader *et al.* (2004); Deane *et al.* (2002); Deng *et al.* (2003); Sharan *et al.* (2005) and Qi *et al.* (2005). In general, the methods take input parameters based on protein and/or network properties and contrast these against biological significance to arrive at a statistical score for an interaction. In a comparative study by Suthram *et al.* (2006) the probability assignment method by Deng *et al.* was shown to provide the most accurate assessment of these methods, in spite of being the simplest one. This implies that there is still scope for improving our understanding of which factors can be used as good indicators of valid protein-protein interaction data.

1.5 Properties of networks

Within complex systems there are numerous diverse processes that are brought about by carefully coordinated pairwise interactions. The cell is a good example of such a complex system, where functions are driven by protein-protein interactions. The increasing complexity of biological observations has motivated scientists to use models that originate from network theories to describe the behaviour of their systems. In the abstract network representation of protein behaviour, proteins are reduced to nodes that are connected to each other, whereby each connection represents a protein-protein interaction.

Network theories have a wide range of applications and they have been used to describe systems as diverse as telecommunications, social and biological networks. Various network models have been proposed and adapted

throughout the years. Network theories traditionally assumed that networks were either completely regular or completely random, with most complex physical networks being random, i.e. networks formed by nodes that are connected together by edges in a random manner. The random-network model is often termed the ‘ER-model’ as it was proposed by Erdős and Rényi who pioneered the study of the mathematical properties of random networks (Erdős and Rényi, 1959). It later appeared that many technological, social and biological systems showed signs of mixed properties and other theories were put forward, including the small-worlds theory (Watts and Strogatz, 1998) and the scale-free model (Barabási and Albert, 1999). The small-world theory gets its name from the fact that in small-world networks it is possible to connect any one node to another node through relatively few intermediate connecting nodes. Small-world architecture also implies a highly clustered architecture, such that when a node is connected to two other nodes, the latter two also tend to have a direct connection to each other. The scale-free model is a modification of the small-world model and assumes an inhomogeneous distribution of connectivity, with most nodes having few connections and a small, but significant, fraction of nodes involved in a large number of interactions (see Figure 1.6). In scale-free systems the probability of finding a node that connects to k other nodes follows a power law, i.e.

$$P(k) \sim k^{-\gamma}, \quad (1.1)$$

where $P(k)$ represents the connectivity distribution, or the probability that a chosen node has exactly k links, and γ is a degree exponent that characterises the system.

Scale-free networks have unique properties that make them robust to the random removal of nodes. This feature of robustness seems to describe biological networks well, for instance in the situation where a failure of a certain protein leads to the activation of an alternative network path to restore the original function (Albert *et al.*, 2000; Barabási and Oltvai, 2004). There is increasing evidence that protein-protein interaction networks exhibit scale-free properties. Scale-free topology has been observed in the biological context, for example in a study of cellular metabolism (Jeong *et al.*, 2000) and later from a study on the robustness of the p53 network (Dartnell *et al.*, 2005). Larger, more global proteomic studies on different species have

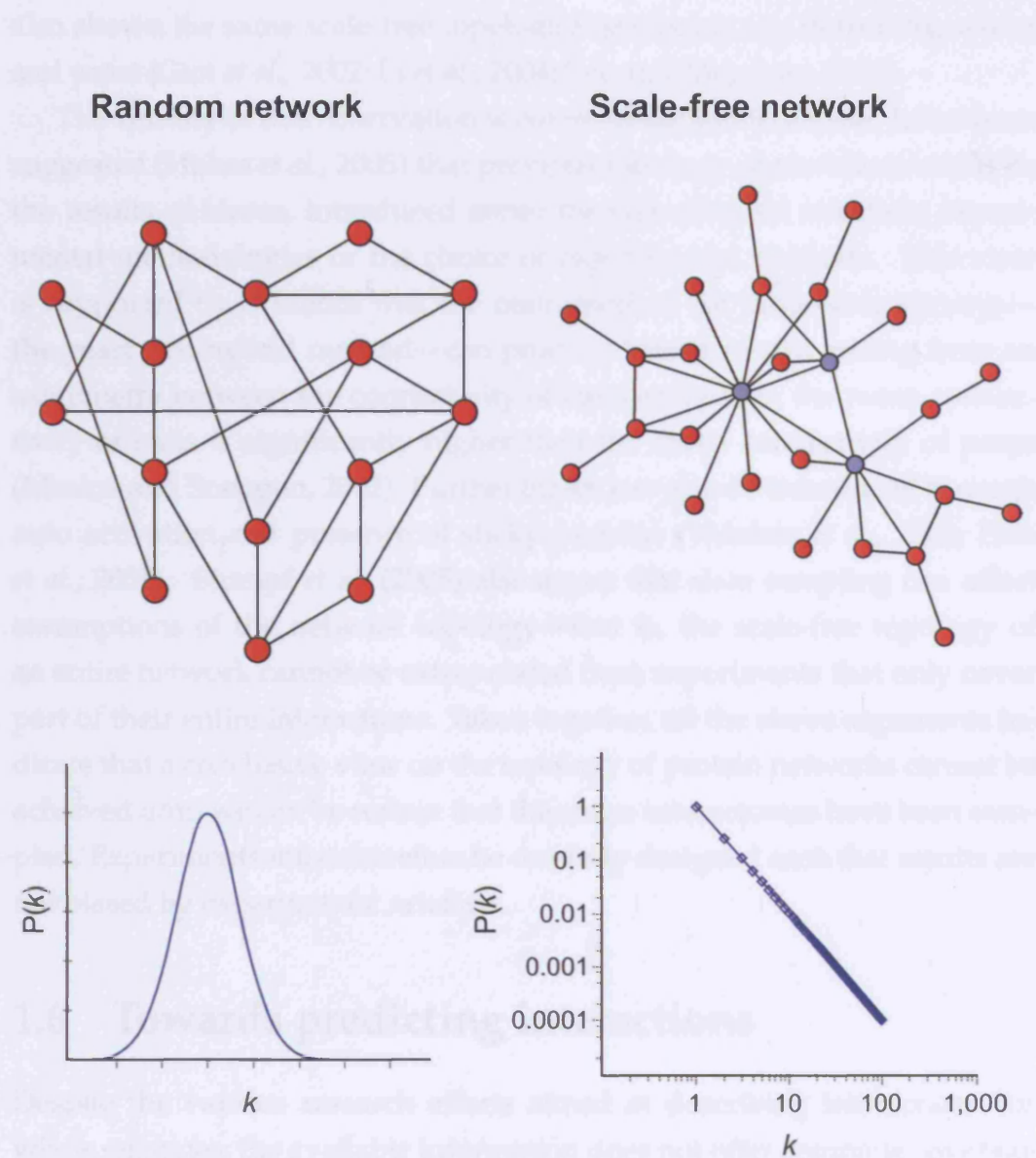


Figure 1.6: Comparison between a random (left) and a scale-free (right) network. **The random network:** The model consists of N nodes and connects each pair of nodes with probability p , which results in a graph with approximately $pN(N-1)/2$ randomly placed links. The random network follows a Poisson distribution (shown below the network), which indicates that most nodes have approximately the same number of links. **The scale-free network:** The network has a small number of highly connected nodes (shown in blue) and the probability of finding a node that shows a greater-than-average connectivity is more significant than in the random graph. The probability that a node has k links is expressed by $P(k) \sim k^{-\gamma}$, where γ is the degree exponent. The scale-free model is characterised by a power law degree distribution of connectivity as can be seen below the network. From Barabási and Oltvai (2004).

also shown the same scale-free topological structure, e.g. in fruit fly, worm and yeast (Giot *et al.*, 2002; Li *et al.*, 2004; Lee and Megeney, 2005).

The validity of this observation is currently subject to debate. It has been suggested (Hakes *et al.*, 2005) that previous topology observations could be the results of biases, introduced either through data set selection, experimental methodologies or the choice of experimental methods. This view is supported by evidence that the main method for large-scale surveys—the yeast two-hybrid method—can produce biased results arising from an asymmetry between the connectivity of baits and preys; the mean connectivity of baits is significantly higher than the mean connectivity of preys (Maslov and Sneppen, 2002). Further biases can also be introduced through auto activation and presence of sticky proteins (Vidalain *et al.*, 2004; Han *et al.*, 2005). Stumpf *et al.* (2005) also argue that data sampling can affect assumptions of the network topology—that is, the scale-free topology of an entire network cannot be extrapolated from experiments that only cover part of their entire interactome. Taken together, all the above arguments indicate that a conclusive view on the topology of protein networks cannot be achieved until we can be certain that the entire interactomes have been sampled. Experiments must therefore be carefully designed such that results are not biased by experimental artefacts.

1.6 Towards predicting interactions

Despite the various research efforts aimed at describing interactions for whole genomes, the available information does not offer complete coverage of any one genome, and in addition, owing to the lack of overlapping data within the genomes under study, has not been verifiable to a great extent. This situation has been the motivation for development of *in silico* methods for predicting individual interactions, with the aim of building complete networks for both model species and the human. Table 1.2 lists some of the databases containing predicted protein interactions. Computational techniques have been developed to predict gene and protein interactions based on *a priori* biological knowledge. These include methods based on gene fusion (Marcotte *et al.*, 1999), gene neighbourhood (Dandekar *et al.*, 1998) and phylogenetic tree similarity profiling (Pellegrini *et al.*, 1999), but the focus of

the following section is on methods that heavily base their predictions on experimental data. The approaches can be broadly divided into the following: Firstly, using structural information to infer both interaction sites and subsequently binary interactions, and secondly, non-structure based methods building on a variety of methods and data sources.

Table 1.2: Database servers available on the Internet. These databases contain protein-protein interaction data, often combining both experimental data and predicted interactions.

Name	Species	Method Database URL	Authors
OPHID	Human	Orthology http://ophid.utoronto.ca/	Brown and Jurisica (2005)
HomoMint	Human	Orthology http://mint.bio.uniroma2.it/HomoMINT/	Persico <i>et al.</i> (2005)
POINT	Human	Miscellaneous http://point.bioinformatics.tw/	Huang <i>et al.</i> (2004)
Prolinks	83 organisms	Miscellaneous http://dip.doe-mbi.ucla.edu/pronav/	Bowers <i>et al.</i> (2004)
Visant	103 organisms	Miscellaneous http://visant.bu.edu/	Hu <i>et al.</i> (2005)
STRING	179 organisms	Miscellaneous http://string.embl.de/	von Mering <i>et al.</i> (2005)
PEP	105 organisms	Miscellaneous http://cubic.bioc.columbia.edu/pep/	Carter <i>et al.</i> (2003)

1.6.1 Structure-based prediction

The structure-based approach has a longer history than the structure independent, and has been focused on characterisation of protein interfaces (see for example Jones and Thornton (1997), Valdar and Thornton (2001b) and Ofra and Rost (2003)), with the aim of subsequently using this knowledge for predicting interfaces. A large amount of work has been done to this end, particularly making use of neural networks (Fariselli *et al.*, 2002; Keil *et al.*, 2004; Hoskins *et al.*, 2006) and Support Vector Machines (SVM) (Yuan *et al.*, 2002; Kim and Park, 2003; Koike and Takagi, 2004; Bordner and Abagyan, 2005; Bradford and Westhead, 2005). SVMs (Vapnik, 1995; Burges, 1998) are computational learning systems based on statistical learn-

ing theories. SVMs have been used for various applications, such as text categorisation, hand-written character recognition, image recognition and various other analyses (Cristianini and Shawe-Taylor, 2000). They are also well suited for the binary task of discriminating between interacting and non-interacting protein pairs. SVMs perform pattern recognition by using binary classification. SVMs map their non-linear n -dimensional input space into a high dimensional feature space and seek a separating hyperplane in this space. The hyperplane is selected in such a way as to maximise its distance from the closest training samples (see Figure 1.7). In contrast with the SVM's discriminative nature, the neural networks use a probabilistic approach for constructing networks so that the individual interactions that make up the network receive a probability based on an assessment by a probability scoring function.

The machine learning methods use a variety of sequence, structure and physical properties (e.g. amino acid propensity, solvent accessibility, electrostatic potential, residue hydrophobicity, sequence conservation, surface planarity and protrusion) as input parameters for their predictions. These methods are designed to identify binding sites but do not indicate which proteins are likely to interact. The natural progression for extensive prediction of interactions would be to incorporate the interface prediction with genome-wide protein docking. Docking of proteins on a genomic scale is a huge challenge with today's technology and knowledge (Szilagyi *et al.*, 2005), although small steps have been made towards docking approximate protein structures (Tovchigrechko *et al.*, 2002), but generally, the current docking methods would need a substantial improvement of computational efficiency and reliability for this to become possible.

1.6.2 Structure-independent prediction

The second approach—using structure-independent based methods—is more feasible in the immediate term. In contrast to the above, these methods do not rely on protein structures being available for complete genomes, but take advantage of the high-throughput data already available and extrapolate to fill in the gaps where experimental knowledge is missing. This subject has expanded greatly in the last few years, particularly since the first genome-wide screening of *Saccharomyces cerevisiae*.

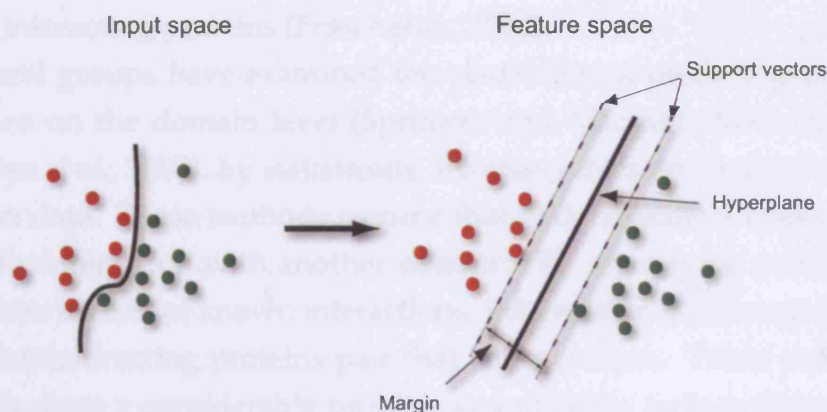


Figure 1.7: The principles of support vector machines. The original objects are transformed from the input space into a multi-dimensional feature space, using a set of mathematical functions, known as kernels. The process of rearranging the objects from input space to feature space is known as mapping. The SVM finds an optimal hyperplane separating the two classes (represented by red and green dots), so that a maximum distance is achieved between the hyperplane and the support vectors. Note that in the feature space, the classes are linearly separable which facilitates an easy solution for the separation.

Predicting protein-protein interactions based on primary sequence properties is probably one of the earliest methods suitable for large-scale predictions. Sprinzak and Margalit (2001) correlated experimental yeast two-hybrid data with sequence signatures that corresponded to protein superfamilies and used those to predict the interacting partners. A more extensive approach was taken by Bock and Gough (2001). Although their method used primary structure as the input for their predictions, the sequences were interpreted in terms of physicochemical properties (charge, hydrophobicity and surface tension). These factors were then fed into an SVM learning system to identify protein-protein binding interactions from the descriptors. The method of Bock and co-workers has since been further developed by Martin *et al.* (2005), who used 'signature products' that combine the sequences of both proteins involved in an interaction, without the need to transform sequence into the physicochemical attributes.

Gene expression has proved useful for validating high-throughput experimental data (Deane *et al.*, 2002; Bader *et al.*, 2004). Additionally, it has been suggested that gene coexpression may be a useful indicator for predicting interactions. This is based on the observation that interacting proteins appear to be coexpressed to maintain a correct chemical stoichiometry

among interacting proteins (Fraser *et al.*, 2004).

Several groups have examined the possibility of predicting protein interactions on the domain level (Sprinzak and Margalit, 2001; Deng *et al.*, 2002; Nye *et al.*, 2005), by statistically analysing domain composition of interaction data. These methods assume that each domain is responsible for a specific interaction with another domain. By identifying correlated domain pairs in a set of known interactions, information is subsequently used to predict interacting proteins pair that contain them. These probabilistic methods show a considerable promise as validation tools to filter out false positives and false negatives in large-scale data sets and may even be useful for inferring interactions based on observation from multiple species (Pagel *et al.*, 2004; Riley *et al.*, 2005; Liu *et al.*, 2005).

Furthermore, gene ontology and functional classification features are particularly useful for prediction of protein interactions, as demonstrated by Lin *et al.* (2004). They used a Bayesian network approach for prediction and examined the usefulness of data from different sources, including mRNA expression data, Gene Ontology and MIPS functional annotations, essentiality data and high-throughput experimental data. They found that annotation data from GO and MIPS (Mewes *et al.*, 2002) were better contributors for predicting the protein-protein interactions than the others.

Data integration methods

The use of machine learning for predicting protein-protein interactions has opened the door for integration of data from diverse sources in an attempt to provide more accurate results. Supervised learning methods can use both direct (experimental) biological data and indirect (e.g. annotation and ontologies) as sources for prediction, often combining fundamentally different sets of genomic information. This is done with the aim of minimising the effect of noise that often comes from the error-prone high-throughput data, and has resulted in progressively more identifications of protein-protein association (Gerstein *et al.*, 2002).

Although SVMs have proved useful for predictions based on data integration, there are several other approaches. Jansen *et al.* (2003), for instance, used a naïve Bayes classifier to predict interactions in yeast. Based on high-throughput Y2H data, in addition to mRNA coexpression, colocal-

isation from ontology and protein essentiality for survival, they predicted over 10,000 interactions and subsequently validated a number of them with tandem affinity purification. Another probabilistic model, proposed by Iosifov *et al.* (2004), used data from literature mining in combination with yeast two-hybrid experiments to predict specific domain-domain interactions, for which no experimental data was available.

Studies focusing on interactions involved in defined aspects of biological functions have also been published. Gunsalus *et al.* (2005) took an integrated approach for an investigation of the molecular machines involved in early embryogenesis in *Caenorhabditis elegans*, by integrating gene and protein network information that was generated from three types of functional relationships, i.e. protein interaction, expression profiling similarity and phenotypic profiling expressed as embryonic lethality. A more extensive study on human networks was done by Rhodes *et al.* (2005), who used a naïve Bayes probabilistic classifier to predict 40,000 interactions. The classifier was fed four types of evidence (model organism interactome data, co-expression matrices, gene ontology biological function, and domain-domain enrichment).

With integrative studies on the increase, it is important to assess the effectiveness of data integration and the limits of the current tools. One study examining this area was done by Lu *et al.* (2005). They investigated the effect of the number of features used for prediction using a naïve Bayes classifier, and found that including a large number of features does not improve prediction quality; future efforts should therefore be limited to a few (up to four) good quality features. Qi *et al.* (2006) assessed the ability of six different classifiers, including Bayes and SVMs, to predict interactions from the same data feature set. They found that a classifier named 'Random Forest' consistently ranked as one of the top two classifiers for all combinations of feature sets. The Random Forest classifier, developed by Breiman (2001), bases its predictions on a collection of decision trees and outputs a classification that is based on the mode (the most frequently occurring) of the classes output by individual trees. Qi and co-workers also assessed the relative importance of each of the data set features that were fed into the Random Forest prediction. Gene expression ranked highest in importance followed by gene ontology process and component descriptions. Fourth on the list

was TAP MS data, followed by gene ontology function terms and lastly mutant phenotype. It is noteworthy that yeast two-hybrid does not feature in this list of the most important prediction factors. Indirect information, such as GO annotation, appears to be highly important in the decision process, which can be attributed to the extensive coverage—direct experiments only covered about 20% of the protein pairs in the study.

The above mentioned studies are only a few of the many that have been published recently, collectively indicating that *in silico* data integration is a method that will continue to grow and establish itself in biological sciences.

The orthology approach

The advantages of integration of orthogonal data sources are evident by the rapid evolution of integrative approaches. The benefits also extend to the use of experimental data from different species as demonstrated by Liu *et al.* (2005), who confirmed that the integrated approach provides a more reliable inference of protein-protein interactions than an analysis from a single organism.

In a seminal publication, Chothia and Lesk (1986) demonstrated the relationship between sequence divergence and protein structure (decreasing sequence similarity results in structural divergence), which is the basis of today's homology modelling principle. More recently Aloy *et al.* (2005) demonstrated, in an analogous way, the correlation between sequence and protein interaction divergence. This means that for two pairs of interacting proteins, $A \leftrightarrow B$ and $C \leftrightarrow D$, where A shares homology with C and B with D, the structural composition of the interfaces is retained for the homologous pairs. This was found, in particular, for sequences sharing more than 30% sequence similarity and indicates that homologous proteins interact in a physically similar manner (see Figure 1.8). This observation is in agreement with the increasing evidence from other sources that protein-protein interfaces are conserved through evolution (Wuchty *et al.*, 2003; Pagel *et al.*, 2004; Rhodes *et al.*, 2005).

The concept of 'interologs'—conserved protein-protein interactions through orthology—was first proposed by Walhout *et al.* (2000) who were investigating a small number of proteins associated with vulval development in *Caenorhabditis elegans*. The concept builds on the idea that physically

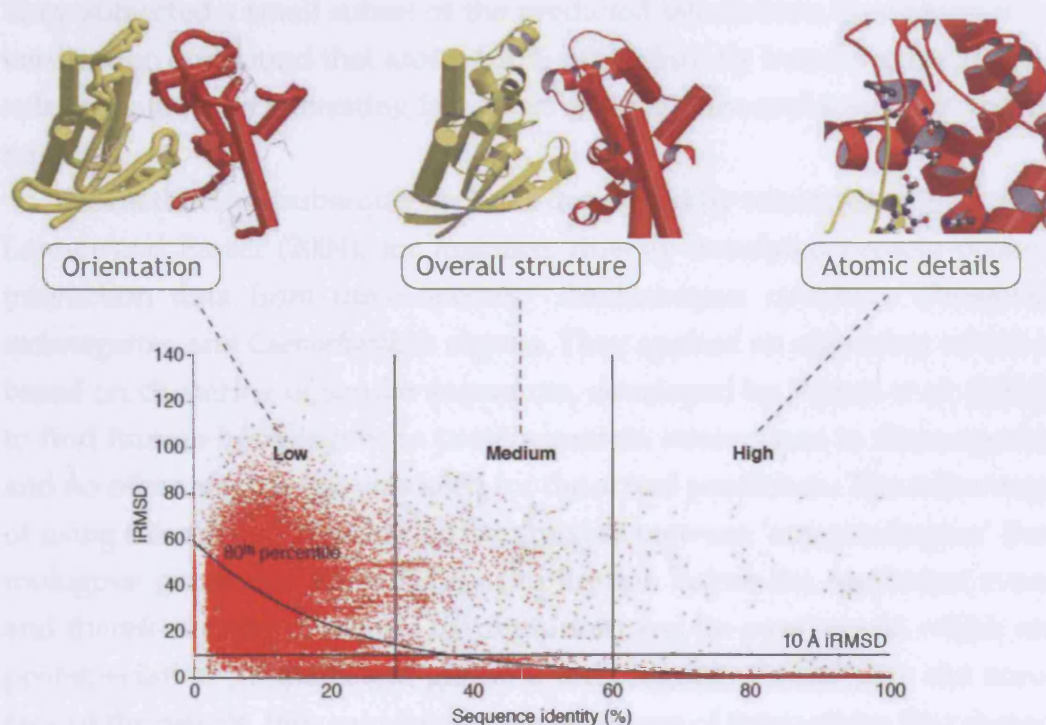


Figure 1.8: The relationship between sequence and interaction divergence shown as the structural interaction similarity (iRMSD; interaction Root Mean Square Deviation) against sequence identity (%). Above 60% sequence similarity, protein interfaces retain similar atomic details, in the range between 30–60% the overall structure of the interface is retained, and below 30% only the overall orientation may be conserved. Figure adapted from Aloy, Pichaud, and Russell (2005).

interacting proteins in one organism have coevolved so that the respective orthologues in other related organisms interact in a same manner. The idea was developed and scaled up by Matthews *et al.* (2001), who used the approach to map protein-protein interactions from yeast into *Caenorhabditis elegans* and verified them experimentally by yeast two-hybrid. Matthews and co-workers started from a sample of 1,195 pairwise interactions from yeast and identified 257 potentially orthologous interactions in the fruit fly. They subjected a small subset of the predicted interactions to experimental verification and found that around 30% of the fruit fly interactions could be substantiated—an interesting fact given that the yeast and worm are distant relatives

The method has subsequently been developed by other research groups. Lehner and Fraser (2004), for instance, directly translated protein-protein interaction data from three species: *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. They applied an algorithm which is based on clustering of similar sequences, developed by Remm *et al.* (2001), to find human homologues to protein-protein interactions in these species and no other mechanism was used for the actual prediction. The advantage of using this approach is that it distinguishes between ‘out-paralogues’ (homologous genes that appeared by duplication before the speciation event and therefore not orthologues by definition) and ‘in-paralogues’ which are post-speciation products and genuine orthologues. To estimate the accuracy of the results, they calculated the percentage of interactions that shared at least one gene ontology term. The complete network contained 71,496 interactions between 6,231 human proteins. Persico *et al.* (2005) applied the same approach for a construction of a set human interactions, but they utilised a string matching algorithm to filter out orthologous proteins whose domain organisation was not conserved and predicted 5,200 interactions, substantially fewer than Lehner and Fraser. Brown and Jurisica (2005) used similar approach to Lehner and Fraser, but used an alternative approach for identifying homologues and added more methods for the validation. They mapped *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* to the human reference frame by assigning a human homologue to the experimental data in the the four model species via the reciprocal best-hit approach (using Blast homology searches in both

directions) (Yu *et al.*, 2004), rather than the in-paralogue method used by Lehner. In addition, they incorporated two methods (domain-domain co-occurrence and gene coexpression) as well as gene orthology to identify additional evidence for part of the data. Their data contained 4,552 proteins involved in 23,889 interactions, of which 5,483 interactions remain (23%) once interactions supported by one of the three methods are removed. Brown and Jurisica compared their homology approach to Lehner and Brown's and found theirs marginally better in terms of the extent to which networks could be validated by GO terms (20.6% compared with 17.7%).

1.7 Overview of the thesis

The construction of protein-protein networks should aid our understanding of both normal and aberrant cellular functions. However, it has been shown that the quality of high throughput protein-protein detection methods is questionable and additionally only a fraction of the interactions have been mapped in many of the key species. This has served as a motivation for the work presented in this thesis. The central theme of the work is protein-protein interactions and the thesis explores this theme on different levels, starting with the molecular aspect of proteins, progressing on to predicting interaction partners and, at the top level, examining genome-scale protein-protein interaction networks.

In *Chapter 2*, the focus is on the molecular level, where the sequence and structural details of proteins are examined. In a eukaryotic genome, such as the human, the coding regions are interspersed with non-coding regions called introns. Therefore, as a first step to understanding protein-protein interactions, we look at the building blocks of eukaryotic genes and investigate whether they affect the overall structure of proteins. Moreover, the effect of intron-exon boundaries on the structure of protein interfaces are studied with the view of uncovering differences that may aid the validation or prediction of binding sites.

Chapter 3 explores the idea of predicting protein-protein interactions by expanding the 'interolog' concept. In particular, a novel scoring function to estimate the reliability of the prediction is proposed and networks of theoretical protein-protein interactions—or interactomes—are created for three

key species. The chapter investigates whether the large-scale networks that were created exhibit the small-world properties often observed in biological networks. Furthermore, the development of a database to house the interactome data is described, and an interactive web server to accompany the database is presented.

Chapter 4 contains an analysis of the topological network features of the human interactome. It asks whether proteins whose mutation can lead to cancer, exhibit a degree of topological difference compared with proteins not associated with cancer. To facilitate further the comparative analysis between cancer and non-cancer proteins, a clustering technique for highlighting parts of the network that are involved in distinct cellular processes is introduced and subsequently the domain composition of both groups are examined in an effort to shed light on the findings.

Chapter 5 takes the study to the rat interactome and explores the possibility whether the computationally-constructed interactome network might be used to help interpret microarray expression data. For this purpose, gene expression data from cell-lines that show high levels of metastasis are incorporated into the protein network. The analysis that follows explores the possibility of identifying novel interactions that may play a key role in cancer metastasis.

Chapter 6 contains an overview of the work presented in the thesis and examines the overall conclusions that can be drawn from the material in the previous chapters.

Chapter 2

Exon arrangement and protein structure

Protein networks are made up of a wide variety of individual proteins that interact and perform diverse tasks within the cell. In order to understand what dictates protein-protein interactions, it is important to recognise that the overall protein structure affects the physical properties of proteins and their functionality. More importantly, the structure of the interface—the site which comes in contact with the protein's binding partner—directly affects the function and specificity of the protein. Understanding of protein structure and interfaces is therefore important for a greater understanding of interactions and subsequently for the verification and even prediction of protein-protein interactions.

The following work looks at one aspect that may affect protein structure: namely the arrangement of exons in the genetic material, from which eukaryotic proteins are translated. It examines whether the area, where two exons meet through RNA splicing, has any special features in terms of its location in the protein structure, and furthermore whether the intron-exon boundaries (IEBs) have any correlation with the highly important protein interface. This chapter contains the results of an initial analysis on the arrangement of introns and exons in relation to protein structure, followed by an investigation into the correlation of intron-exon boundaries and protein interfaces, and finally concluding by examining the special case of alternatively spliced exons in protein interfaces.

Part of the work in this work in this chapter was done in collaboration

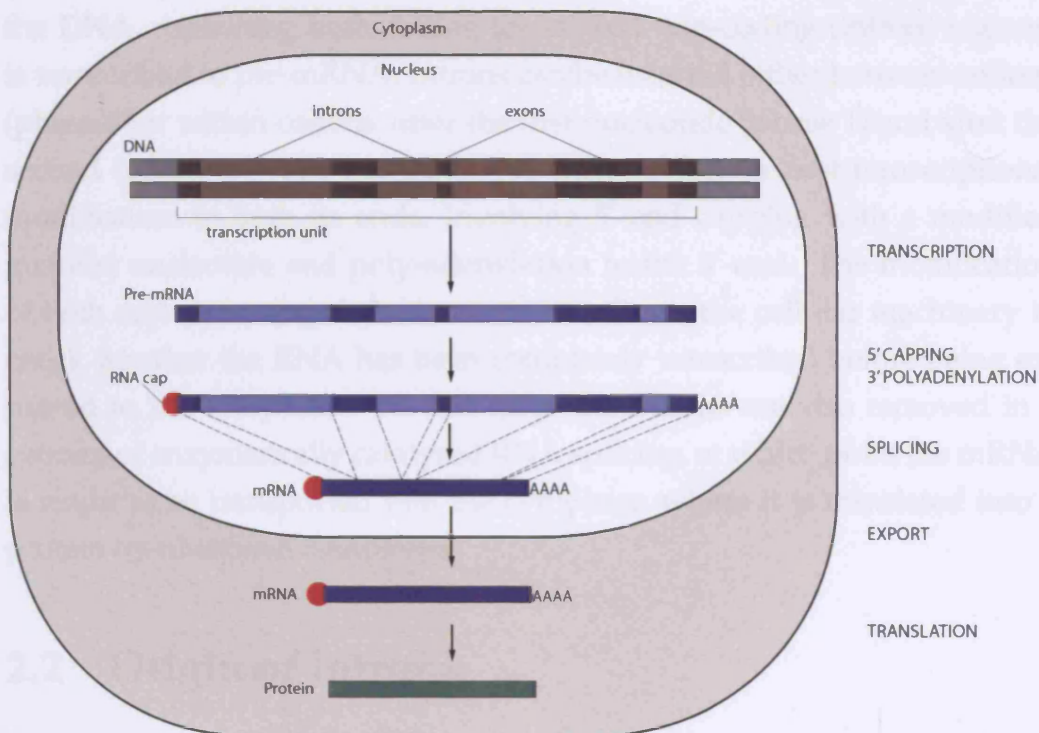


Figure 2.1: A general overview of the processes involved in eukaryotic gene expression. The main events that take place in the nucleus are transcription and post-transcriptional processes (5' capping, 3' poly-adenylation and splicing of introns). Once these events have taken place the mRNA is transported to the cytoplasm where ribosomal complexes translate the mRNA into a protein. Figure based on Alberts *et al.* (2002, pg. 315).

with Bruno Contreras-Moreira and was published in the Journal of Molecular Biology (Contreras-Moreira, Jonsson, and Bates, 2003). Contreras-Moreira and Jonsson were joint first authors of this work; Contreras-Moreira's contributions related to the modelling aspects and genetic algorithms for modelling and the author of this thesis performed the work relating the structural aspect of IEBs.

2.1 Gene expression

Before turning to protein structure, it is important to look at the biological process involved in eukaryotic gene expression. The main events are transcription from DNA to RNA, the splicing of RNA and subsequent translation into the final protein product. Figure 2.1 summarises the processes involved in eukaryotic protein expression. Initially, the transcription unit of

the DNA, containing both coding (exon) and non-coding (intron) regions, is transcribed to pre-mRNA. Introns can be inserted either between codons (phase 0) or within codons, after the first nucleotide (phase 1) and after the second (phase 2). The pre-mRNA is then subject to post-transcriptional modification to both its ends, involving 5'-end capping with a modified guanine nucleotide and poly-adenylation to the 3'-end. The modification of both ends is a control mechanism that allows the cellular machinery to check whether the RNA has been completely transcribed before being exported to the cytoplasm. At this stage the introns are also removed in a process of enzymatically catalysed RNA splicing, at which point the mRNA is ready to be transported into the cytoplasm where it is translated into a protein by ribosomal complexes.

2.2 Origin of introns

The organisation of eukaryotic DNA into exons (expressed sequences) and introns (intervening sequences) requires a more complicated transcriptional mechanism than is observed in lower organisms, such as the much simpler prokaryotes, which generally do not have introns. Some exceptions to this general observation have, nevertheless, been found in archaeobacteria and some eubacteria, where introns have been discovered (Marlene Belfort and Dalgaard, 1995). In the 30 years since they were discovered by Berget *et al.* (1977) and Chow *et al.* (1977), introns have been a source of much interest in the scientific community. What at a first glance appears to be a waste of genetic material has subsequently been shown to be a feature that efficiently maximises the coding potentials of genes. The intron-exon arrangements appear to be a way of speeding up evolution, allowing new proteins to evolve by combinations of different exons into a new protein product (Patthy, 1999a). Furthermore, the cell has a mechanism that allows alternative splicing of exons, i.e. allowing different proteins to be produced from the same gene under different conditions (Letunic *et al.*, 2002).

Introns are recognised as genomic regions involved in insertion, deletion or duplication of new exons, or even in the formation of chimeric proteins (Patthy, 1999b). For this reason, introns are potential places for insertion or deletion of fragments in proteins, and therefore possible locations for

significant changes in protein structure and function. Three theories have been put forward as an explanation of the existence of introns: the introns-early (Gilbert, 1987), introns-late (Palmer and Logsdon, 1991) and synthetic (de Souza, 2003) theories. The three theories can be summaries as follows (Stoltzfus *et al.*, 1994; Roy, 2003):

Introns early:

1. Exons are the descendants of ancient mini-genes and introns stem from the spacers between them.
2. Exons were joined together to form larger genes.
3. The splicing mechanism originates from ancient RNA.
4. Introns were lost from bacteria through evolution.

The introns late theory contradicts the above in most aspects:

Introns late:

1. Split genes originate from insertion of introns.
2. Genes encoding modern proteins developed without the participation of introns.
3. The splicing mechanism developed from fragmented self-splicing introns.
4. Spliceosomal introns were never present in the ancestors of intron-free organisms.

The earlier debate was highly polarised, but more inclusive perspectives that allow elements from both theories are now becoming increasingly common, for instance in the synthetic theory:

Synthetic theory:

1. A mixed model: some introns are ancient and others new.
2. Most introns, especially phase 1 and phase 2 are recent additions to the eukaryotic genomes (agreeing with introns-late theory).
3. Some phase 0 introns are ancient and are correlated with protein modules (agreeing with introns-early theory).

2.2.1 Splicing

The splicing of most exons takes place in the spliceosome, although some introns are capable of self splicing (Cech, 1986). RNA molecules known as

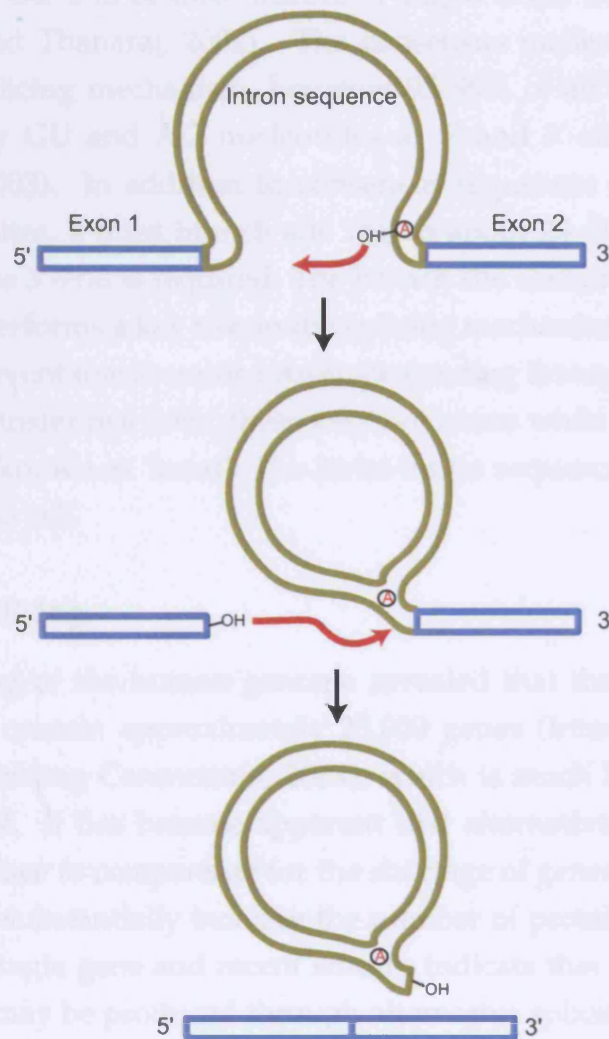


Figure 2.2: An overview of the RNA splicing mechanism. In the first step, an adenine nucleotide at the branch site, towards the 3'-end of the intron sequence (indicated red in a black circle), attacks the 5' splice site and cuts the backbone of the RNA. In the second step the cut 5'-end covalently joins the adenine nucleotide and a loop in the RNA is created. The final step takes place by the released 3'-OH end of the exon reacts with the start of the next exon, joining the two exons together and releasing the intron sequence as a 'lariat'. Adapted from Alberts *et al.* (2002, pg. 318).

snRNA (small nuclear RNAs) complex with several snRNP (small nuclear ribonucleoproteins) to form the core of the spliceosome. The spliceosome recognises consensus nucleotide sequences in the pre-mRNA that signal the beginning and the end of most introns (Padgett *et al.*, 1986; Alberts *et al.*, 2002; Clark and Thanaraj, 2002). The consensus nucleotides depend on the specific splicing mechanism, however 95–99% of all mammalian sites are marked by GU and AG nucleotides at 5' and 3'-ends (Burset *et al.*, 2000; Black, 2003). In addition to consensus sequences directly at the 5' and 3' splice sites, a third branch site region about 20–50 nucleotides upstream from the 3'-end is required. The branch site contains an adenine nucleotide that performs a key role in the splicing mechanism (see Figure 2.2). Each splicing event removes one intron, proceeding through two sequential phosphoryl-transfer reactions; these join two exons while removing the intron in a loop known as 'lariat'. The lariat intron sequence is subsequently degraded in the cell.

Alternative splicing

The sequencing of the human genome revealed that the human genome is thought to contain approximately 25,000 genes (International Human Genome Sequencing Consortium, 2004), which is much lower than previously expected. It has become apparent that alternative splicing mechanisms are in place to compensate for the shortage of genes. Alterations in a splice site can substantially increase the number of protein products stemming from a single gene and recent studies indicate that up to 74% of human proteins may be produced through alternative splicing (Johnson *et al.*, 2003). Furthermore, alternative splicing is one of the most important mechanisms in gene regulation (Stamm *et al.*, 2005), allowing different proteins to be produced in response to physiological changes in the cell. Exons that are present in all mRNA after processing are termed constitutive, but those that vary are referred to as alternatively spliced exons and they can be formed by several distinct splicing patterns (Black, 2003), shown in Figure 2.3.

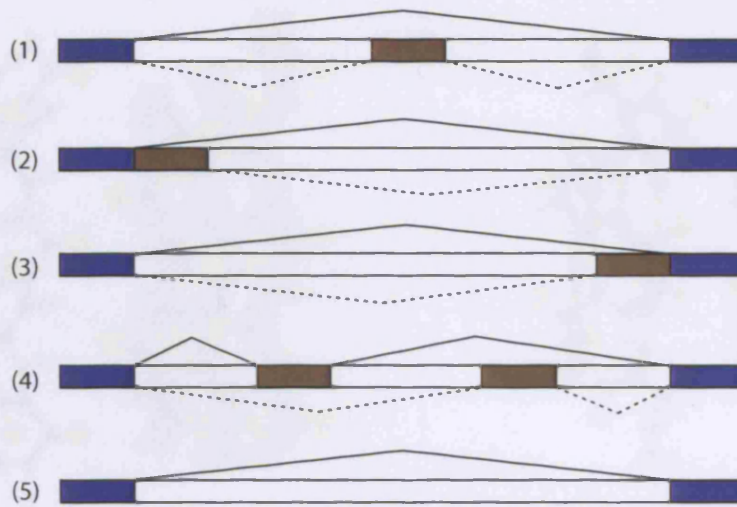


Figure 2.3: A schematic interpretation of a pre-mRNA and the possible splicing patterns. Constitutive exons present in all final mRNA are shown in blue, alternatively spliced exons in brown and introns in white. The lines joining one exon to another indicate possible splicing paths: one alternative path shown with a solid line and the other with a dotted line. (1) Exon skipping/inclusion; (2) alternative 5' splice sites; (3) alternative 3' splice sites; (4) mutually exclusive exons; (5) Intron retention. (Based on Cartegni *et al.* (2002)).

2.3 Protein structure

In the subsequent sections the intron-exon arrangement will be examined in relation to protein structure and therefore a description of the elements of protein structure is needed. Protein structure is classified into different levels: primary, secondary, tertiary and quaternary structure.

2.3.1 Primary structure

The primary protein structure is defined as the one dimensional amino acid sequence and is determined by the covalently linked amino acids in the polypeptide backbone. The structure of a protein determines its function, but what determines the structure? Anfinsen (1973) first demonstrated that the primary structure—the amino acid sequence—determines the higher degree three-dimensional structure. This can be reasoned by understanding that the polypeptide seeks to fold into a native structure that is thermodynamically the most stable in the intracellular environment (Creighton, 1993).

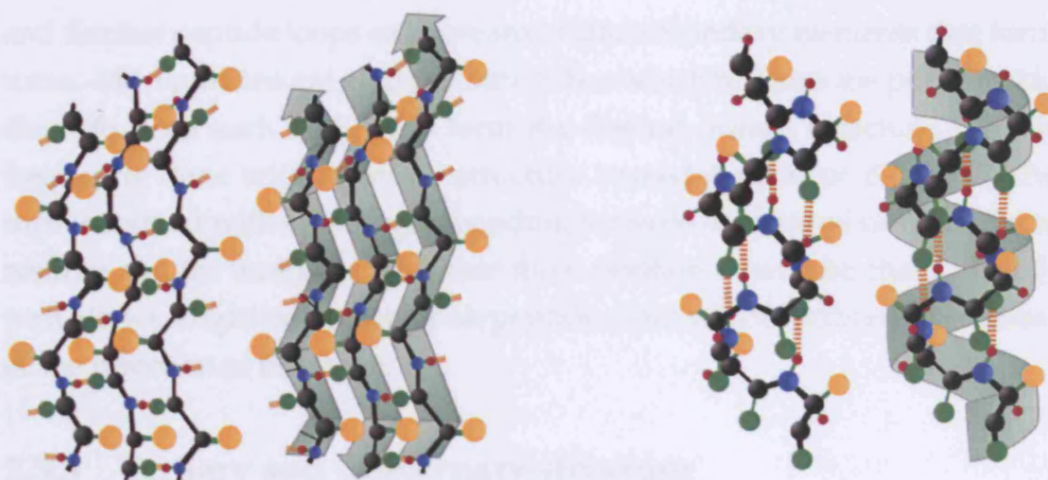


Figure 2.4: Structure of β -strands (left) and α -helices (right). The colour coding of the atoms is as follows: carbon (black), nitrogen (blue), oxygen (green), hydrogen (red) and side-chains (yellow). The polypeptide backbone (black) is formed by peptide bonds between carbon and nitrogen molecules. The structure is supported by hydrogen bonds (dotted orange lines) that extend from amine groups (NH) to carbonyl groups (CO) of nearby peptides. (From the Department of Biology, Penn State University (<http://www.bio.psu.edu>)).

2.3.2 Secondary structures

Contrary to primary structures, secondary and higher structures are chiefly held together by non-covalent forces such as hydrogen bonds, ionic bonds, van der Waals and hydrophobic interactions. Adjacent residues in polypeptide chains can form hydrogen bonding interactions between the backbone oxygens and amide hydrogens, facilitating formation of structural arrangements that are the protein's essential building elements in three-dimensional space.

The α -helices and β -sheets are the two main types of secondary structure elements (see Figure 2.4). The α -helix is a spiral-like construction, stabilised by hydrogen bonding between backbone amine and backbone carbonyl atoms. One turn of the most common helix represents 3.6 amino acid residues containing 13 atoms, which is why the α -helix is sometimes referred to as 3.6₁₃-helix. Strands, usually 5 to 10 residue long, make up β -sheets through hydrogen bonding as they are structurally unstable on their own. Adjacent β -strands can be either parallel (running in the same direction) or antiparallel (running in the opposite direction to each other).

Helices and sheets are frequently joined together by loosely structured

and flexible peptide loops or more structured secondary elements that form turns. Most proteins are globular structures, which requires the polypeptide chain to bend such that it can form the desired overall structure. This is frequently done with a simple structure termed β -turn (or β -bend). The turn is formed with a hydrogen bonding between a carbonyl oxygen of one residue and the amide of a residue three position down the chain. The β -turn allows a tight bend of the polypeptide chain which facilitates a reversal of the direction of the chain.

2.3.3 Tertiary and quaternary structure

Protein domains are examples of tertiary structure elements. The tertiary structure is formed once the secondary structure elements fold in order to assume a low energy state and a more compact three-dimensional shape. The tertiary structure is often globular as this allows the formation of large numbers of intramolecular hydrogen bonds, and reduces the solvent accessible surface. Amino acids with hydrophobic side-chains are folded in such a way that they form the core of domains whereas residues with hydrophilic side-chains are exposed to the solvent on the outside of the structure. Finally, the highest structural form, quaternary structure, is the arrangement of two or more tertiary structures, often from separate polypeptide chains, into larger complex structures.

2.4 Methods

Exons, both constitutively and alternatively spliced, and the intron-exon boundaries were examined using the following data and methods:

2.4.1 Data

The protein set used in the study of intron-exon boundaries and protein structure was composed of human and mouse proteins obtained from the Protein Data Bank (PDB), as at 22 January 2003 (Berman *et al.*, 2000). To avoid large multi-domain proteins, structures with at least 100 residues but no more than 300 were selected. Immunoglobulins and T-cell receptors were identified by sequence similarity and excluded from this data set to avoid

spliced genes. Chimeric proteins were also excluded. After excluding proteins with only one exon (about 25% of the original set), this data set contained a total of 684 PDB chains. These proteins contained, on average, 3.2 introns.

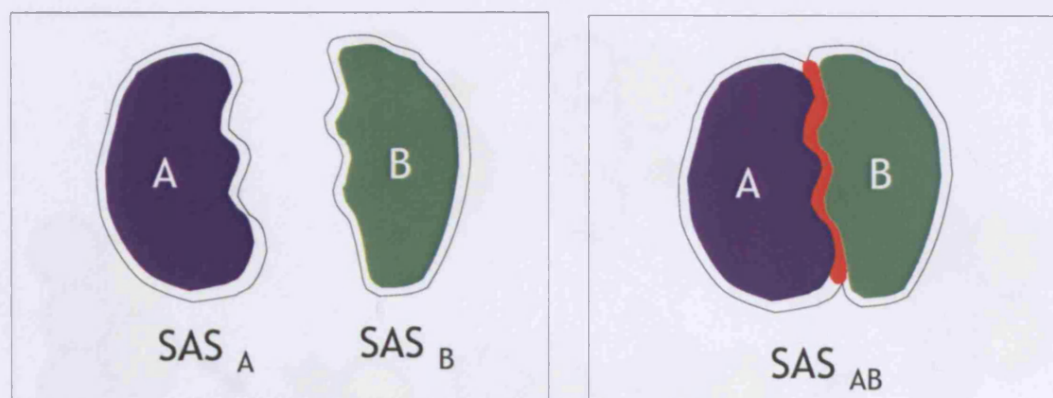
Information on variable sequence splicing in the data set was obtained from release 41.16 of the Swissprot database (Bairoch and Apweiler, 2000). Swissprot sequences for which splice variant information was available were matched against the sequences containing intron-exon boundary (IEB) information (see below). Stringent matching criteria were used, allowing two gap regions in the sequence alignments, yielding only 24 sequences where IEBs could confidently be identified in alternatively spliced areas. For the study of human-mouse homologues, human and mouse sequence pairs of sequence identity $\geq 40\%$ were extracted from the above data set, resulting in 118 pairs.

2.4.2 Assignment of introns to protein sequences

Intron-exon boundaries were assigned by mapping protein sequences to the human and murine genome assemblies (Human genome build 31 from the National Center for Biotechnology Information (Lander *et al.*, 2001) and Mouse genome release 3 from the Mouse Genome Sequencing Consortium (Waterston *et al.*, 2002)), using the BLAT server (Kent, 2002). When using amino acid sequences in this work, introns are defined as the residues corresponding to the left hand side of the boundary at the DNA level. IEBs in homologous proteins are defined as conserved if they occupy the same place in a structural alignment of those proteins. Phases of exons at IEBs were obtained by dividing the genomic position of the last DNA base of each exon by three and calculating the modulus.

2.4.3 Secondary structure assignment

Secondary-structure of proteins was assigned using the program DSSP (Kabsch and Sander, 1983). Protein structure figures were prepared using Rasmol (Sayle and Milner-White, 1995), Molscript (Kraulis, 1991) and VMD (Humphrey *et al.*, 1996).



$$\text{Interface area} = (\text{SAS}_A + \text{SAS}_B) - \text{SAS}_{AB}$$

Figure 2.5: A schematic diagram of domain A and domain B forming a complex. Solvent accessible surface area is calculated for each domain separately and for the whole complex. A change in the accessible surface area indicates a buried surface area (coloured red) and thereby interaction between the domains.

2.4.4 Structural conservation and IEBs

Proteins were aligned by structure using the program Msuper, which was developed for progressive multiple structure alignment (Gerstein and Levitt, 1996; Russell and Barton, 1992) in the Biomolecular Modelling Laboratory. The level of structural conservation, or variability, was based on Msuper's alignment scores for each column of structurally aligned residues, centering on the distance between C^β atoms, ranging from 0 (close structural alignment) to 9 (distant alignment) and '-' where structural alignment was not possible.

2.4.5 Domain contacts from PDB structures

A database of protein domain interfaces was created by the following approach: experimentally determined structures were retrieved from the PDB and each of the structures was split into domains using information from SCOP. Domain-domain contacts were detected by systematically removing one domain at a time from each protein structure. The exposed surface area was then calculated using Naccess (<http://wolf.bms.umist.ac.uk/naccess>), which estimates the accessible area when a probe with a radius of 1.4\AA (the radius of a H_2O molecule) is rolled around the van der Waals

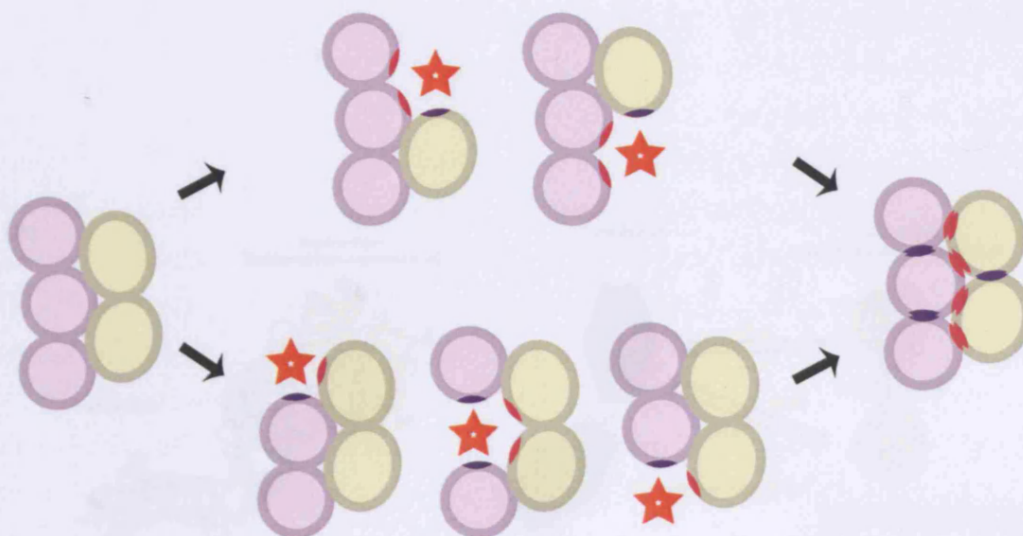


Figure 2.6: A schematic figure of two multi-domain proteins (one lilac, the other yellow) bound in a complex. The solvent accessible surface area is calculated for the complex as a whole and then one domain is removed at a time and the accessible surface area recalculated. Removing a domain exposes contact residues on remaining domains and these contacts are identified by recalculating the accessible surface area. This process is repeated for all domains in both proteins (domain removal is indicated by a star) and the domain-domain contacts recorded. Inter-chain contacts are distinguished from intra-chain contacts (shown as red and blue patches respectively).

surface of the domain (see Figure 2.5). Any change in the solvent accessible surface area (ASA) of a domain when an adjacent domain is removed indicates a buried surface area and thereby an interaction between two domains. Once this has been calculated for a domain, it is replaced on the complex and the calculation then proceeds to the next domain, see Figure 2.6 for an illustration. Residues participating in a domain-domain contact were classified as 'core' if completely buried in the interface or 'peripheral' if partially exposed to solvent. All available structures in the PDB were processed in this way and stored in a relational database (see Figure 2.7 for an example of the information recorded for a complex).

2.4.6 Distinguishing obligomers from transient complexes

When examining PDB files containing more than one polypeptide chain, it is not obvious whether chains making contact form obligatory or nonobligatory complexes. The terms were first defined by Jones and Thornton (1996)

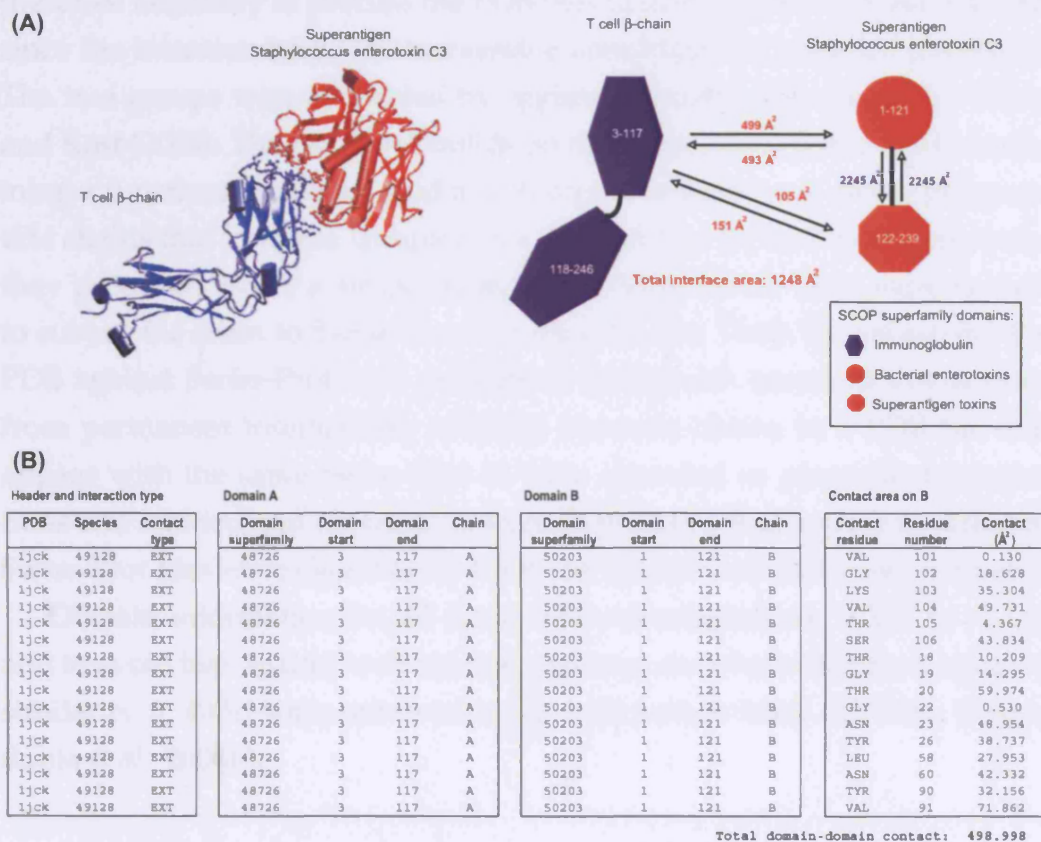


Figure 2.7: (A) The structure of a superantigen–T-cell complex (PDB ID: 1JCK) with the atomic details of contacts of both molecules shown (left) and a schematic diagram of the same complex (right), showing the domain composition and residue numbers. The arrows represent contacts between domains, showing the exposed surface on the target domain (arrow head) when the binding domain (arrow tail) is removed. Intra-chain contact is shown in blue and inter-chain contact in red. (B) An example of the database entry for the immunoglobulin–enterotoxin domain interaction. The table shows details such as the species (49128=human) and the type of interaction (EXT=inter-chain contact) as well as domain information for both interacting domains (SCOP id 48726 for Immunoglobulin and 50203 for Bacterial enterotoxins). Finally, the contact area is shown with residue details.

and refer to whether chains are necessarily and permanently bound, to form a multimeric complex that functions as a cellular machine (obligomer), or transiently bound where the proteins are functional on their own—and interaction is brought about by temporary connection (nonobligomer). It was therefore necessary to process the PDB files to differentiate between the two since the intention here was to examine nonobligatory complex interfaces. The two groups were separated by applying a method proposed by Ofran and Rost (2003). Their method builds on the observation that experts examining a functional complex tend to submit the sequences of all the polypeptide chains that form the complex in a single file to Swiss-Prot. Conversely, they proposed that if a single chain is functional on its own, experts tend to submit the chain to Swiss-Prot in a separate file. Thus, by comparing the PDB against Swiss-Prot, it is possible to distinguish transient interactions from permanent interactions: contacts between chains in a PDB file that appear with the same Swiss-Prot id were classified as permanent interactions and omitted and contacts between PDB chains that appear in different Swiss-Prot files were classified as transient interactions and taken forward.

Domain information for all chains were obtained from SCOP (v. 1.63), and to avoid bias against well studied proteins, close homologues (sequence similarity $\geq 40\%$) were removed using information from ASTRAL (Chandonia *et al.*, 2004b).

2.4.7 Sequence conservation assessment

PSI-BLAST sequence similarity searches (Altschul *et al.*, 1997) were run for each sequence in the human, mouse and rat data set against the non-redundant NCBI protein sequence database (9 November 2003). Hits with sequence coverage $\geq 80\%$ and sequence similarity $\geq 20\%$ were taken forward. PSI-BLAST parameters were set to the following values: maximum number of passes, $j = 3$; Expectation value, $e = 1.0$; E-value threshold for multipass, $h = 0.0005$; and maximum number of alignments, $b = 100,000$. Complete sequences were retrieved for each hit using Fastacmd from the BLAST package.

A maximum of 250 sequences, equally distributed through the homology hits were then aligned by CLUSTAL W (Thompson *et al.*, 1994). Intron-exon boundaries were mapped onto the alignments and sequence conserva-

tion scores were calculated for each column, using the sequence weighted matrix method proposed by Valdar and Thornton (2001b) for quantifying residue conservation in the multiple sequence alignment.

A weighted sum of all pairwise similarities between all residues in a position, i is given by the $Cons(i)$ equation:

$$Cons(i) = \frac{\sum_j^N \sum_{k>j}^N w_j w_k M(s_j(i), s_k(i))}{\sum_j^N \sum_{k>j}^N w_j w_k}, \quad (2.1)$$

where M is a normalised substitution matrix with a range of $[0,1]$; $M(a, b)$ is the element (a, b) in the matrix M , showing the similarity between amino acids a and b ; N is the number of sequences in the alignment; and $s_j(i)$ and $s_k(i)$ are the amino acids at alignment position i of sequences s_j and s_k respectively. w_j and w_k are the weights of sequence s_j and s_k , respectively and given by

$$w_j = \frac{1}{N-1} \sum_{k \neq j}^N Dist(s_j, s_k), \quad (2.2)$$

where $Dist(s_j, s_k)$ is the evolutionary distance between sequences s_j and s_k :

$$Dist(s_j, s_k) = 1 - \frac{1}{n(Aligned_{jk})} \sum_{i \in Aligned_{jk}} M(s_j, s_k) \quad (2.3)$$

and $Aligned_{jk}$ is the set of all non-gap positions in s_j or s_k , and $n(Aligned_{jk})$ is the number of such positions.

The matrix used for the above calculations provides information on the likelihood of amino acids in aligned positions. Several such matrices are available, for instance PET91, the Pairwise Exchange Table (Jones *et al.*, 1992), which was used in this work. Gaps in the PET91 substitution matrix were replaced with zeros and the matrix normalised to the range of 0-1 as described by Karlin and Brocchieri (1996):

$$M(j, k) = \frac{M'(j, k)}{\sqrt{M'(j, j) \times M'(k, k)}}, \quad (2.4)$$

where $M(j, k)$ is the normalised value of M' in position (j, k) . The weight of each sequence was calculated based on the distribution between sequence pairs derived from the normalised substitution matrix and the conservation

score for each column calculated.

The data pertaining to domains and protein interfaces were retrieved from the domain contact database described in Section 2.4.5, and mapped onto the sequence alignments for further statistical analysis (see Appendix A for a review of some of the statistical methods used in this work).

2.5 Results and Discussion

In order to assess the effect of intron-exons boundaries on domain structure, the set of 684 single-domain human and mouse protein structures from the PDB was subjected to the following investigation.

2.5.1 Secondary structure at intron-exon boundaries

An analysis was performed to compare the composition of secondary structures (1) at intron-exon boundaries and (2) away from boundaries. The results, shown on the left-hand side of Table 2.1, show a significant preference for IEBs to exist in coil regions of proteins and less inside α -helices and extended β -strand elements. This could indicate that insertion of introns into sections of ordered structure, for instance α -helices and β -sheets, is likely to affect the overall structure and function, which in return affects the fitness of proteins in natural selection terms. Also, even when boundaries occur within strands and helices, they tend to be close to the end of their secondary structure element, as shown in Figure 2.8. This is especially apparent for non-conserved IEBs in extended strands (see Figure 2.8(A)) and supports the idea that boundaries occur in less-ordered areas.

The question now arises whether the observed secondary structure biases could reflect different types of introns. Introns appearing in proteins as a result of late exon duplications and insertions have a phase class that is identical with that of the recipient intron (Patthy, 1987). An analysis of phase classes of exons and their boundaries (see the right-hand side of Table 2.1) does not indicate any correlation between the phasing of exons and the secondary structure of IEBs. This, however, does not imply that phases are not conserved in particular genes, since we are comparing many different proteins from different genes. Splice variants within proteins could

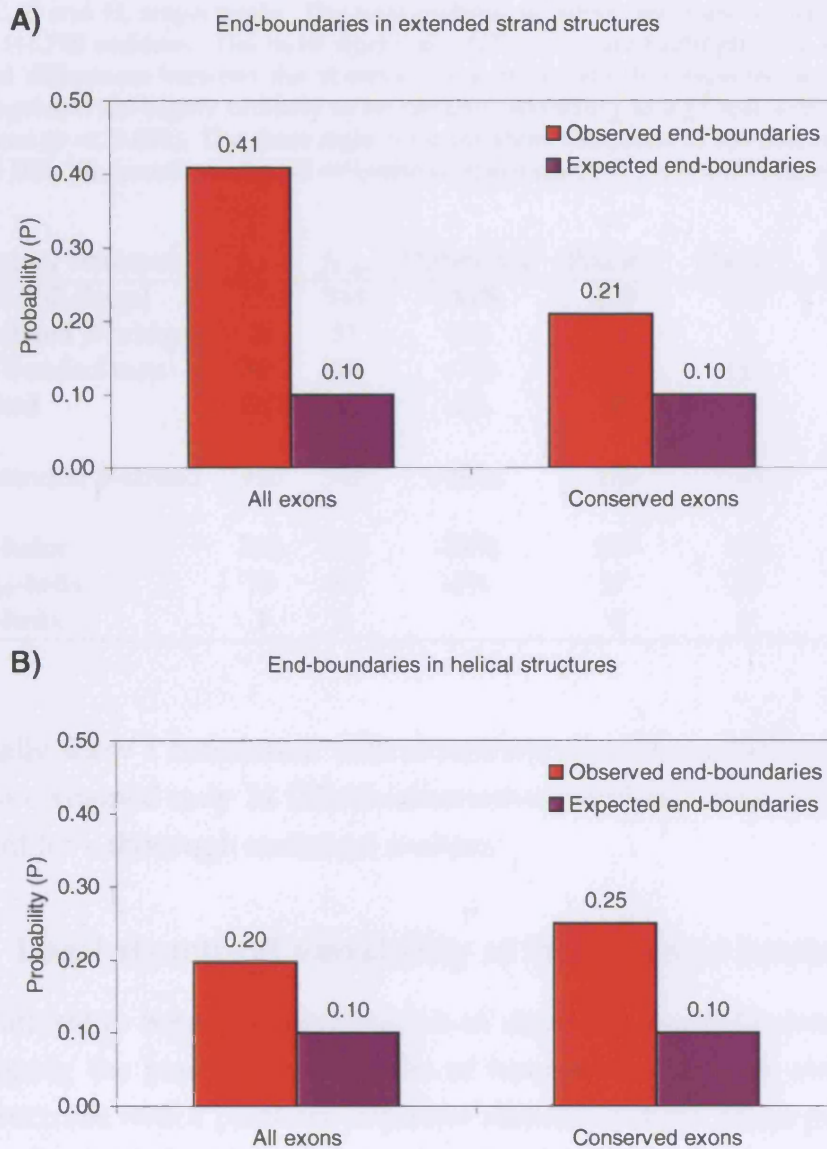


Figure 2.8: The probability of intron-exon boundaries appearing at the ends of (A) extended β -strands and (B) α -helices. Ends are defined as the first or last 5% of the secondary structure element length. Red columns show the observed frequency of boundaries in ends of secondary structure elements and purple columns show the expected frequency. The left-hand side pair of columns in each graph shows statistics for all exons, whereas the right-hand columns show statistics for a subset of exons that are conserved between human and mouse. The differences are significant according to χ^2 -test ($p \ll 0.0001$ for all exons in extended strands and helices, $N = 450$ and 579 , respectively, and $p < 0.005$ for conserved exons in extended strands and helices, $N = 62$ and 60 , respectively).

Table 2.1: Observed and expected frequencies of IEB within DSSP assigned secondary structure elements. Coil, extended strand and helical structures are identified by the letters C, E and H, respectively. The total number of intron residues is 2,447, out of a total of 116,740 residues. The most significant differences are highlighted in bold. The observed differences between the observed frequencies and the expected according to the background are highly unlikely to be random, according to a χ^2 -test with 7 degrees of freedom ($p \ll 0.001$). The three right columns show the phase of the preceding exon for each IEB. We found no overall differential distribution of IEBs with respect to exon phases.

Secondary structure	f_{Obs}	f_{Exp}	Difference	Phase 0	Phase 1	Phase 2
C - No SSE (loop)	776	544	+43%	279	262	235
C - Isolated β -bridge	29	31	-6%	10	9	10
C - H-bonded turn	308	288	+7%	106	111	91
C - Bend	260	265	-2%	90	72	98
E - extended β -strand	430	537	-20%	130	148	152
H - α -helix	570	702	-19%	199	174	197
H - 3_{10} -helix	73	80	-9%	27	22	24
H - 5-helix	1	0	–	0	1	0

potentially show a correlation with secondary structure at IEBs—however, the data contained only 24 IEBs in alternatively spliced areas, which is insufficient for a thorough statistical analysis.

2.5.2 Local structural variability at intron-exon boundaries

The relationship between conservation of structure and IEBs was studied by mapping the boundaries on pairs of homologous human and mouse PDB structures with a pairwise sequence identity $\geq 40\%$. These pairs were structurally aligned and a window of seven residues was moved along the superposition. The fitness of the alignment (a score based on the distance between C^β atoms) was recorded for each of the seven positions. The secondary structure of each position was determined using DSSP, as before, and the window scores for each of the three secondary structure elements (α -helix, β -strand and coil) were then normalised and the scores for each class of secondary structure elements at boundary positions compared with the overall expected scores.

The structure conservation of boundaries in coil regions and helices was not found to differ significantly from the expected values, see Figure 2.9.

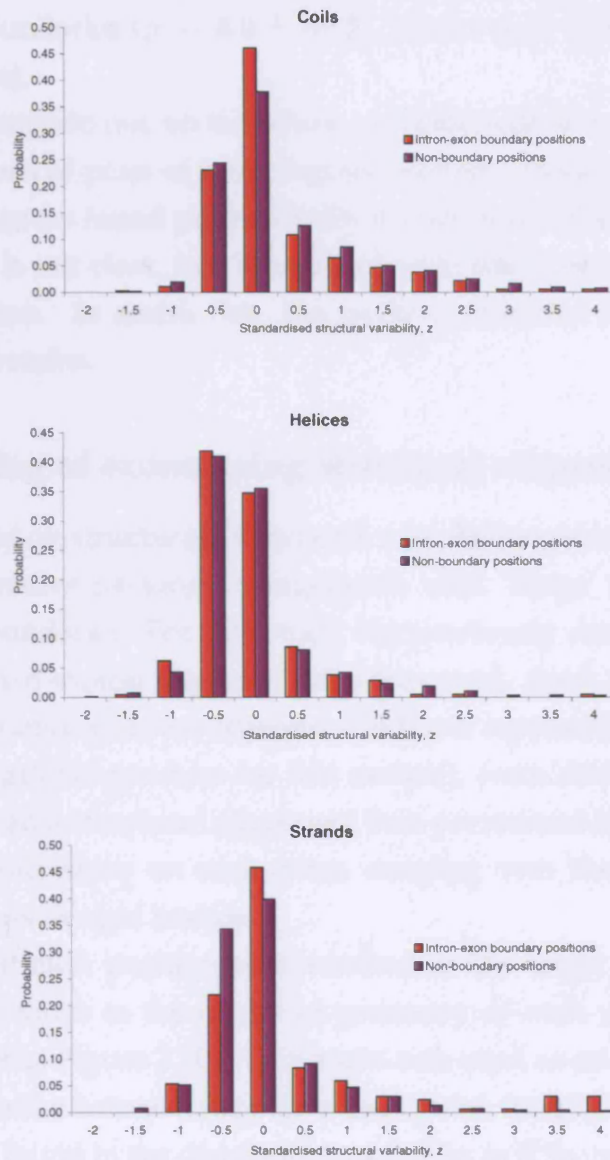


Figure 2.9: Standardised structural variability scores (standard deviations from the mean) for windows of seven residues running over the whole length of pairs of structurally aligned sequences. Windows containing IEBs (red columns) and windows not containing IEBs (purple columns) were compared, classified by the type of secondary structure; coil, α -helix or β -strand. High scores indicate more variation in structure between the two structurally aligned sequences and low scores less variation than expected in the sequence. The distributions of coils and helices were not significantly different according to an independent t-test ($p = 0.17$ and 0.25 respectively). The distribution of strands revealed, on the other hand, a small but significant difference between IEB and non-IEB residues ($p = 6.0 \times 10^{-6}$). The standardised deviate (or z-value) is given by $z = \frac{x - \mu}{\sigma}$ where x is the original structural similarity score, μ is the mean and σ is the standard deviation.

There was an increase in the structural variability of strands containing intron-exon boundaries ($p = 6.0 \times 10^{-6}$), however, it was small (0.5 standard deviations).

The boundaries do not, on the whole, coincide with the more structurally divergent regions of pairs of homologous proteins. Hence, the reason why these boundaries are found preferentially in coils and at the ends of α -helices and β -strands is not clear, but it is perceivable that it is to allow variable packing of exons. To assess this, the packing of exons was compared in homologous proteins.

2.5.3 Packing of exons using structural alignments

A method based on structural alignments was used to assess whether exons can have alternative packing arrangements with 'hinge' points located at intron-exon boundaries. For this study the previously described set of homologous human-mouse sequence pairs was used. Each pair was initially aligned by sequence and two adjacent windows, representing two exons of an average length (43 residues for this sample), were shifted along the sequence pairs, and a structural alignment then performed by superimposing the two left-hand exons on each other, carrying over the structure of the right-hand exons as rigid bodies.

Flexibility at each position was assessed as the angle between vectors from the N-terminus to the centre of geometry of each of the right hand exons (see inset in Figure 2.10). This angle was used as an indication of the structural deviation between the pair at each point. No significant difference ($p = 0.77$) was found in the distribution of angles at IEBs compared with the background distribution as shown in Figure 2.10. This would suggest either that evolution does not favour increased diversity of packing between homologous exons, or that the method that was used was not sensitive enough to pick up potential hinge points in the boundary locations.

2.5.4 Intron-exon boundaries and their relation to domain interfaces

Interaction sites of proteins have been extensively analysed, focusing on physicochemical and geometrical properties such as solvation potentials,

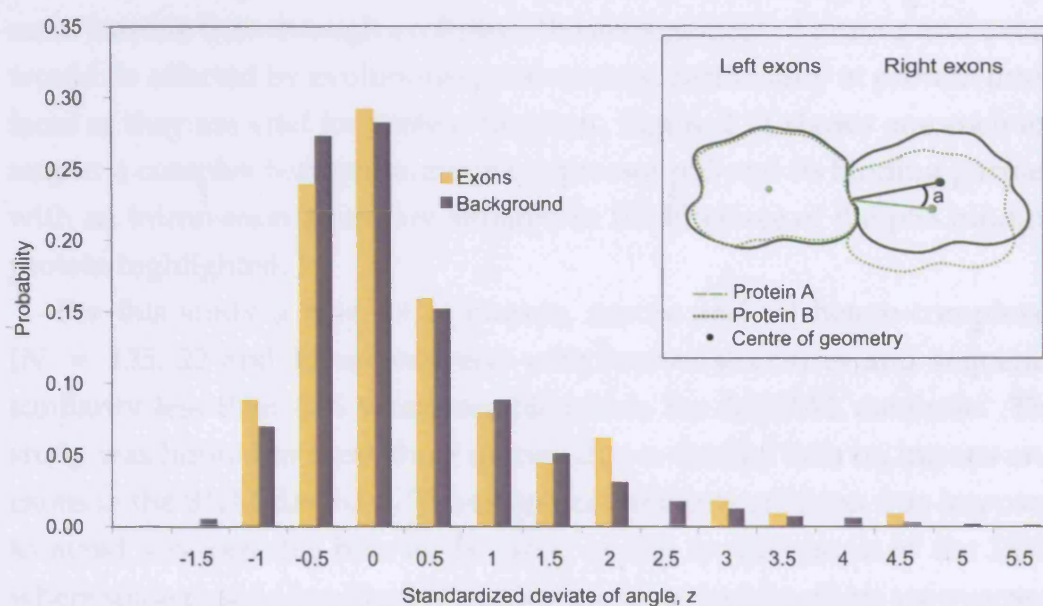


Figure 2.10: Distribution of standardised normal deviates of angles in intron-exon boundaries (yellow) and the background (grey) with a mean value of 8.1 degrees and standard deviation of 6.7 degrees. Greater z-values represent higher degree of variability between a homologous pair at a specific position. There is not a significant difference between the samples ($p = 0.77$ for an independent t-test). The insert shows a schematic diagram of the calculation on a pair of proteins consisting of two exons. Centres of geometry are depicted. By superimposing the left-hand exons and carrying over the right-hand exons as rigid bodies an angle, α , can be measured.

residue propensities, hydrophobicity, solvent accessible surface area and protrusion (see e.g. Jones and Thornton (1997); Lo Conte *et al.* (1999); Ofra and Rost (2003); Nooren and Thornton (2003)). Although the location of alternatively spliced sequence in relation to protein interaction sites has been examined (Offman *et al.*, 2004; Neverov *et al.*, 2005), the arrangement of intron-exon boundaries in interfaces has, to our knowledge, not been studied. A further investigation of intron-exon boundaries was therefore carried out, focusing on the IEBs and interfaces of protein complexes. It would seem feasible that, through evolution, the arrangement of introns and exons would be affected by evolutionary constraints, particularly at protein interfaces as they are vital for protein function. Figure 2.11 shows one such example: a complex between tumour-suppressor p53 and its binding partner, with an intron-exon boundary situated in the interface of the p53 binding protein highlighted.

For this study a new set of human, mouse and rat hetero-complexes ($N = 135$, 22 and 13 respectively) with known structures and sequence similarity less than 40% was assembled from the ASTRAL database. The study was limited to these three species due to limited data on introns and exons in the BLAT database. The sequence similarity criterion was imposed to avoid any possible bias in the data, owing to the nature of the PDB, where some protein families are underrepresented while others are overrepresented. Intron-exon boundaries were mapped onto the structure as before (see Section 2.4.2), and this information was related to interface data by using the previously constructed database of domain-domain interactions.

Figures 2.12 and 2.13 show the sequences for the human, rat and mouse p53 tumour suppressor and its binding partner p53BP1. The positions of intron-exon boundaries are relatively well conserved for these close homologues. Interface residues are indicated above the sequence alignment, demonstrating that all species contain one IEB in a protein binding site. This intron-exon boundary at residue 120 (human sequence numbering) separates exon three and four in the p53 binding protein and structurally it appears at the edge of the interface, see Figure 2.11.

The initial question to be asked was: Do intron-exon boundaries occur with higher or lower frequency at interfaces than could be expected by chance? As shown in Table 2.2, the distribution of IEBs is neither bi-

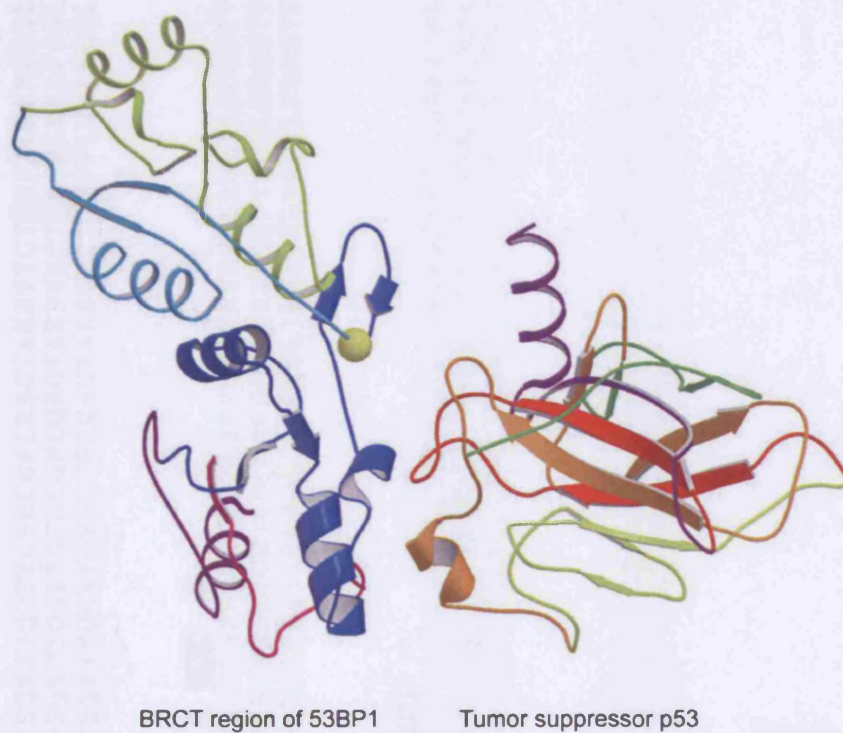


Figure 2.11: The exon structure of the tumour-suppressor protein p53 and its binding partner, 53BP1. Exons participating in the binding are shown with larger structural elements than those not involved in binding. Two exons from p53 (orange and red) and one from the BRCT region of 53BP1 (blue) form the protein-protein interface. One intron/exon boundary from 53BP1, depicted as a solid yellow sphere is located at the edge of the interface.

Human	0
Rat	...MEDSQSDMSIELPLSQETFSLWKLLPPDDILPTTATGSPNSMEDLFLPQDVAELLEGPEEALQVS.APAAQ	71
Mouse	MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILP.....SPHCMDDL LLPQDVEEFFEGPSEALRVSGAPAAQ	70
HumanSSVPSQKTYQGSYGFRLLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVD	54
Rat	EPGTEAPAPVAPASATPWPLSSSVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSISLNKLFCQLAKTCPVQLWVT	146
Mouse	DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFCQLAKTCPVQLWVS	145
Human	STPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPP	129
Rat	STPPPGTRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNPYAEYLDDRQTFRHSVVVPYEPP	221
Mouse	ATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQTFRHSVVVPYEPP	220
Human	EVGSECTTIHYNMCMSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRRDRTEENLRKK.....	198
Rat	EVGSDYTTIHYKYMCMSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRRDRTEENFRKKEHCPE	296
Mouse	EAGSEYTTIHYKYMCMSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRRDRTEENFRKKEVLCPE	295
Human	198
Rat	LPPGSAKRALPTSTSSSPQKKKPLDGEYFTLKIRGRERFEMFRELNEALELKDARAAEESGDSRAHSYPKTKK	371
Mouse	LPPGSAKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDAHATEESGDSRAHSYLKTKK	370
Human	198
Rat	GQSTSRHKKPMIKKVGPDSD	391
Mouse	GQSTSRHKKTMVKKVGPDSD	390

Figure 2.12: p53 Tumour Suppressor alignments for human, rat and mouse. Intron-exon boundaries are highlighted red in the sequences. Protein interface residue ranges (core and periphery) are shown in blue above the sequence alignment.

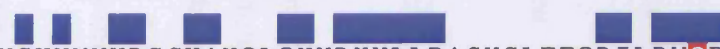
Human	L	NKTLFLGYAFL	TMATTSDKLASRSKLPDGPTGSSEEEEE	F	FLEIPPFNKQYTESQLRAGAGYILED	FNEA	Q	CNT	75
Rat	MATTSDKLASRSKLLDGPTGSSEEEEE	F	FLEIPPFNKQYTECQLRAGAGYILED	FNEA	Q	CNT	61
Mouse	L	NKTLFLGYAFL	TMATTSDKLASRSKLLDGPTGSSEEEEE	F	FLEIPPFNKQYTECQLRAGAGYILED	FNEA	Q	CNT	439
									
Human	A	YQCLLIADQHCR	TRKYFLCLASGIPC	VSHVWVHDSCHAN	QLQNYRNYLLPAGYSLEE	QRILDW	Q	PRENPFQNLK	150
Rat	A	YQCLLIADQHCR	TRKYFLCLASGIPC	VSHVWVHDSCHAN	QLQNYRNYLLPAGYSLEE	QRILDW	Q	PRENPFQNLK	136
Mouse	A	YQCLLIADQHCR	TRKYFLCLASGIPC	VSHVWVHDSCHAN	QLQNYRNYLLPAGYSLEE	QRILDW	Q	PRENPFQNLK	514
Human	V	LLVSDQQNFLEL	WSEILMTGGAASVK	QHHSSAHNK	D	IALGVFDVVVTD	PSCPASVLKCAEAL	QLPVVSQEWVI	225
Rat	V	LLVSDQQNFLEL	WSEILMTGGAASVK	QHHSSAHNK	D	IALGVFDVVVTD	PSCPASVLKCAEAL	QLPVVSQEWVI	211
Mouse	V	LLVSDQQNFLEL	WSEILMTGGAASVK	QHHSSAHNK	K	DIALGVFDVVVTD	PSCPASVLKCAEAL	QLPVVSQEWVI	589
Human	Q	CLIVGERIGFKQ	HPKYKH	DYVSH	249				
Rat	Q	CLIVGERIGFQ	QHPKYKH	DYVSH	235				
Mouse	Q	CLIVGERIGFKQ	HPKYKH	DYVSH	613				

Figure 2.13:

p53 Binding Protein 1 alignments for human, rat and mouse. Intron-exon boundaries are highlighted red in the sequences. Protein interface residue ranges (core and periphery) are shown in blue above the sequence alignment.

Table 2.2: Frequencies of IEBs in interfaces vs. outside interfaces for both core interface (residues completely buried in the interface) and total interface (core and periphery). Observed values are compared with values expected, given a random distribution of IEBs. Expected values were calculated by multiplying the total number of IEBs by the ratio of amino acids in an interface over the total number of residues.

	Core interface			Total interface		
	f_{Obs}	f_{Exp}	Difference	f_{Obs}	f_{Exp}	Difference
Human						
IEBs in interface	9	13	-30.8%	143	142	+0.7%
IEBs not in interface	627	623	+0.6%	493	494	-0.2%
Mouse						
IEBs in interface	1	2	-50.0%	26	23	+13.0%
IEBs not in interface	87	86	+1.1%	62	65	-4.6%
Rat						
IEBs in interface	0	1	-100.0%	18	18	0%
IEBs not in interface	91	90	+1.1%	73	73	0%

ased away or towards protein interfaces. This can be seen both for IEBs in the total interface, defined as core and periphery residues, ($p = 0.76$) and core interface (interface consisting of completely buried residues), although there appears to be a very slight trend for IEBs to appear away from the core, but this observation is not statistically significant ($p = 0.25$).

2.5.5 Intron-exon boundaries and sequence conservation

In previous studies, residues in protein interfaces have been shown to exhibit a higher degree of conservation compared with those outside the interface (Valdar and Thornton, 2001a; Huan-Xiang and Yibing, 2001). The findings described in the previous section indicate that IEBs are statistically unbiased towards or away from protein interfaces, but how do IEBs relate to sequence conservation in general?

To answer this question, the sequence conservation of the human, mouse and rat functional complexes was calculated as described in the Methods section. A window of 5 columns (current position ± 2 columns) was shifted alongside the multiple alignments and average conservation score calculated alongside the scores at window positions around IEBs. The mean sequence conservation score for IEB windows was 0.52 compared with 0.83 overall, which indicates that intron-exon boundaries have a preference for

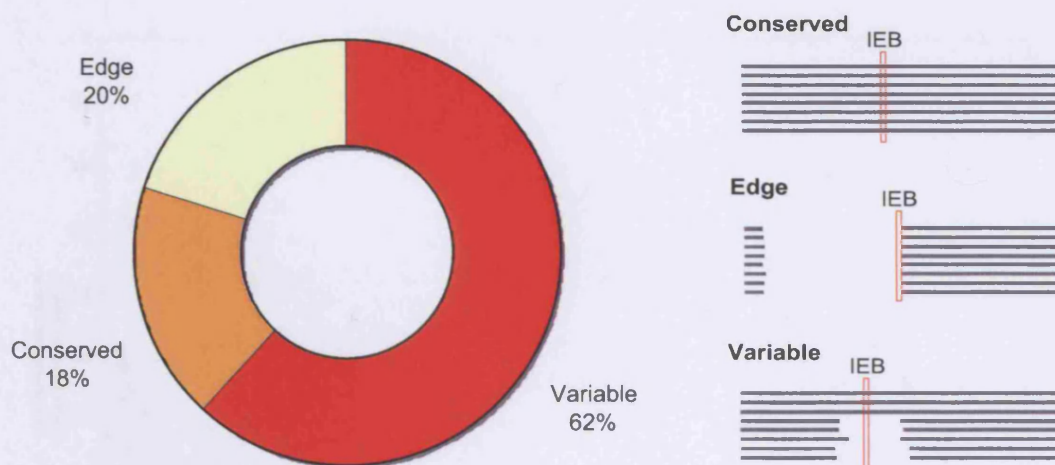


Figure 2.14: Frequency of intron-exon boundaries classified by the level of conservation in a set of homologous sequences. Up to 250 homologous sequences were aligned and the placement of IEBs was classified according to the nature of sequence conservation: conserved (no gaps in alignment), loop (some gaps) and edge (no gaps, but adjacent to gap columns).

less conserved regions of proteins.

The placements of boundaries in relation to conservation were investigated further by studying the sequence alignments. Each column in the alignments was classified in one of the three following groups: 'conserved' (no gaps in current column and no gaps in the closest two columns on either side), 'edge' (no gaps, but adjacent (± 2 columns) to a column with one or more gaps) and 'variable' (one or more gaps in current column). Using these classes, it becomes apparent that the majority of IEBs (62%) reside in variable regions of the alignment (see Figure 2.14). A further 20% of IEBs are observed in regions close to gaps, which brings the percentage of gap and near-gap IEBs to 82%. This high percentage of gap-associated IEBs agrees with the overall lower sequence conservation score in IEB locations, and also with the enrichment of IEBs in structural loops, shown previously in Table 2.1, as gap regions in sequence alignments often indicate structurally variable loop regions.

The results shown in the previous section indicated that intron-exon boundaries have no correlation with protein interfaces. It is therefore interesting to turn the attention to the exons themselves and the relationship between exon conservation and interface participation. Focusing specifically on a subset of exons that make up the protein interface, the exon interface

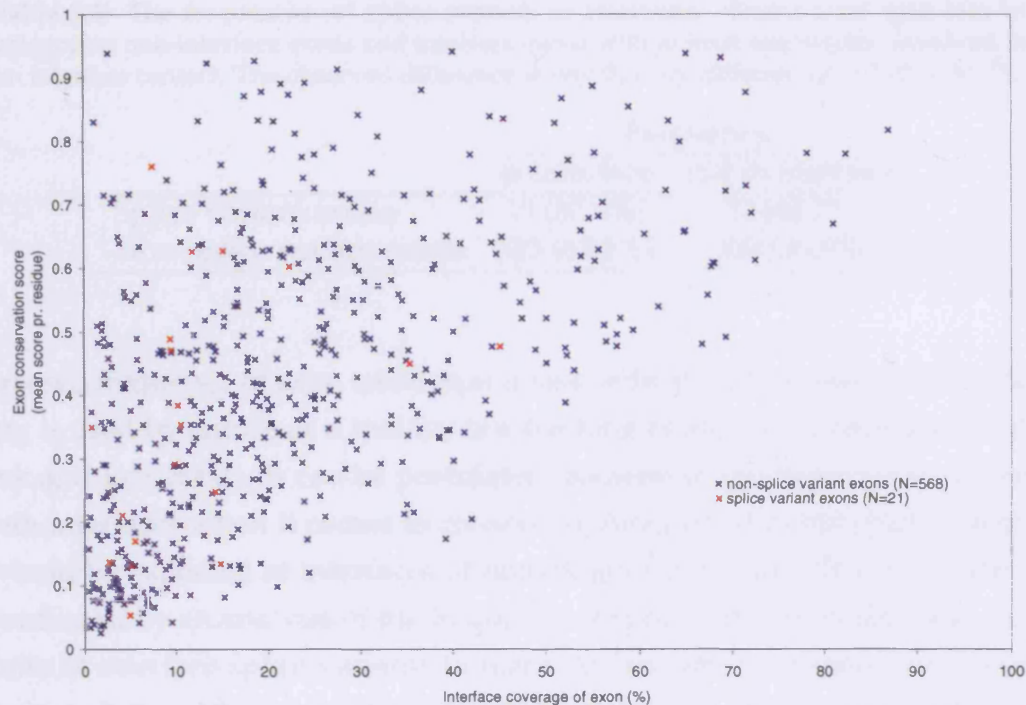


Figure 2.15: Conservation of exons in interfaces of proteins, as a function of interface coverage of the exon. Constitutively spliced exons are shown in blue and alternatively spliced ones in red. Spearman's rank coefficient for alternatively spliced variants was found to be lower than for constitutively spliced ones ($\rho_s = 0.57$, $p = 4.57 \times 10^{-4}$).

coverage (i.e. the proportion of an exon that forms an interface) was examined to establish whether a correlation to sequence conservation, as calculated from the sequence alignments, could be observed. A moderately high correlation was found between interface coverage and exon score, measured by Spearman's rank coefficient, $\rho_s = 0.81$ (see Figure 2.15). This observation is significant ($p = 6.52 \times 10^{-131}$) and agrees with the general observation that conserved residues are more likely to be found at protein interfaces (Valdar and Thornton, 2001b; Grishin and Phillips, 1994), although Caffrey *et al.* (2004) more recently showed that this correlation is weaker than previously thought.

2.5.6 Splice variation in interfaces

Intron-exon boundaries are not only sites for constitutive splicing but also alternative splicing. Alternative splicing has been associated with an in-

Table 2.3: The frequencies of splice variants in interfaces. Exons were split into two categories; non-interface exons and interface exons with at least one residue involved in an interface contact. The observed difference is significantly different ($p = 1.03 \times 10^{-4}$).

	Frequency	
	in interface	not in interface
Splice variant exons	19 (37.3%)	32 (62.7%)
Non-splice variant exons	523 (63.5%)	300 (36.5%)

creased frequency of exon creation and loss, indicating that alternative splicing is used by nature as a tool for fast-tracking evolution of proteins (Modrek and Lee, 2003). It can be postulated, because of the importance of protein interfaces when it comes to conserving function, that minimal changes would be expected at interfaces of homologous proteins. This was indeed confirmed by an analysis of the frequency of splice variants in the data. The ratio of interface splice variants to non-interface splice variants was found to be substantially lower than expected: 37.3% vs. 62.7% compared with 63.5% and 36.5% for non-splice variant exons (interface and non-interface, respectively), see Table 2.3. These results are based on a very limited set of protein structures and should be taken with some caution until more structural data becomes available to verify the findings.

2.6 Conclusions

The results presented in this chapter suggest an evolutionary feedback mechanism between natural introns and the effect they have on protein folds and interaction sites. We observed an enrichment of intron-exon boundaries in coils and the ends of secondary structure elements. This fact suggests an evolutionary model in which the position of introns is constrained by protein structure, particularly by tertiary structure contacts. Whamond and Thornton (2006) have recently proposed a reversed relationship, i.e. that a nucleotide bias at intron locations produces secondary structure bias around introns through the coding of residues such as glycine and aspartic acid, which are frequently found in coiled regions. It is not possible to exclude one or the other based on these findings, and further studies combining structural, genomic and evolutionary aspects of introns would be

needed to resolve the question. In relation to the introns early/late debate, the findings cannot rule out either theory. Some results seem to support an early origin of introns (such as secondary structure preferences), whilst others could be taken as evidence for their late origin (for example, both packing and structural variability results). Furthermore, older and newer introns were not distinguishable in the data set, since only human and mouse data were used for the analysis. Overall, the results agree with a model in which both theories are compatible (de Souza *et al.*, 1998; Fedorova and Fedorov, 2003).

In addition to the structural environment of IEBs, intron-exon arrangement was examined in the context of sequence conservation. Boundaries were shown to have a preference to exist in areas of lower conservation and therefore have a particular tendency to appear in, or near, gaps in alignments of homologous sequences. Remarkably, the intron-exon boundaries studied here do not show a tendency to avoid interfaces, be it core or total interface, although sequence conservation is generally a feature of binding sites—particularly so in core residues of binding. Indeed, exons in protein interfaces seem to be subject to conservational constraints, reflected in the observation that the larger the proportion of an exon in an interface the more highly it is conserved. Conversely, alternatively spliced exons were found to avoid protein interfaces when they were compared with constitutively spliced exons. This may stem from the deleterious effect alternative splicing may have on proteins' function, should splicing take place in an interface.

Taking an introns-early view, the arrangement of genes into exons is a way of facilitating modulation of protein function. Exons thus promote functional diversity by allowing recombination and shuffling of functional modules—leading to a quicker evolutionary adaptation. Intron-exon boundaries, particularly those surrounding alternatively spliced exons, could then be seen as hot-spots for recombination and thus less likely to appear in protein interfaces. The underrepresentation of alternatively spliced exons in interfaces would support this view.

The exon arrangement should also be looked at in a broader view, since many proteins rely on a precise arrangement of multiple exons, not just in the binding site but also in the vicinity of the interface. Any rearrangement of exons may dramatically affect protein folds and hence the function. An

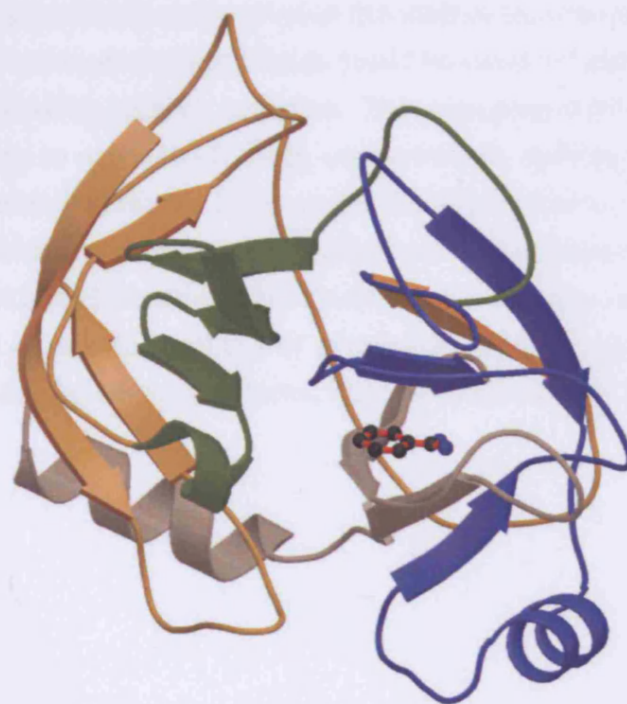


Figure 2.16: The exon structure of the human Trypsin IV (PDB code 1H4W), bound to a benzamidine inhibitor in the active site. The four exons that form the protease are shown in different colours. It is evident that the arrangement of all four exons is important for forming the active site. Figure created using Molscript (Kraulis, 1991).

example of this is the human Trypsin IV, which specificity is determined by an extended area of binding, not just the core binding site (Katona *et al.*, 2002). The four exons that constitute this enzyme are shown in Figure 2.16 in different colours. Two of the four IEBs are in close proximity ($< 7\text{\AA}$) to the catalytic site. Clearly, the binding site is the result of the precise packing of at least three exons and thus the function of the protein depends on the correct modulation of the exons. Any modification at the IEBs would therefore be likely to have a direct effect on the enzyme's function.

One of the questions put forward at the start of this chapter was whether the location of intron-exon boundaries could be used to help with validating or predicting binding sites on proteins. This was proposed on the assumption that, owing to a relatively high conservation, splicing sites would be avoided in protein interfaces. IEBs would then be usable to indicate reduced probability of binding in their area. In this work, it has been shown that this is generally not possible, although information on location of splice variants could be used as one component of predictive methods using Bayesian or SVM analysis for building networks, as they were shown to avoid protein interfaces.

Chapter 3

In silico construction of interactomes

3.1 Predicting protein interactions

The analysis of protein networks is dependent on a reliable assignment of protein-protein interactions. Protein-protein interactions are commonly studied using biochemical techniques, and several different experimental methods are currently in use. Two-hybrid screens have, to date, yielded the bulk of available data (Uetz *et al.*, 2000; Ho *et al.*, 2002); however their level of accuracy is not particularly high and they should be supported by additional evidence (Sprinzak *et al.*, 2003; Bader and Hogue, 2002). Advances in other techniques, such as tandem-affinity purification and mass spectroscopy, have also made large-scale studies increasingly feasible (Gavin *et al.*, 2002; Mann *et al.*, 2001).

A number of computational methods, either based on sequence or structural features, have been developed to complement experimental approaches to predicting protein-protein interactions (Park *et al.*, 2001; Valencia and Pazos, 2002). An increasing emphasis has been on deducing and exploring the protein-protein interaction networks that are reflected in expression data; gene networks have been inferred from gene expression data using mathematical analysis such as Bayesian regression (Bader *et al.*, 2004; Brazhnik *et al.*, 2002; Rogers and Girolami, 2005; Jansen *et al.*, 2002b). Moreover, networks have been derived by complementing gene expression data with data from different sources, such as gene ontologies, phenotypic pro-

filing and functional similarities (Jansen *et al.*, 2002a; Gunsalus *et al.*, 2005; Lu *et al.*, 2005; Rhodes *et al.*, 2005). Alternative techniques to network construction have also been taken, see for instance Cabusora *et al.* (2005), who created a map of interacting proteins based upon the principle that interacting protein modules in one organism may be fused into a single chain in another, and Calvano *et al.* (2005), who constructed the network by literature searches for information pertaining to interacting protein pairs from closely related organisms.

This chapter describes the approaches taken to predict extensive protein-protein interaction networks in three species. Predicting protein-protein interactions for entire genomes yields a great amount of data that can only be analysed effectively once it has been put into a manageable form. An efficient way is to store the data in a relational database, and this chapter describes the construction of such a database for three key organisms. However, large sets of data are not useful unless there is a way to explore and analyse them, and with this in mind an interactive web server was constructed to facilitate a visual exploration of the predicted protein interactions.

3.2 Relational databases

The amount of biological data scientist have to deal with in their studies increasingly requires systematic storage and access solutions. Databases have therefore become essential tools for biologists and biochemists, as well as bioinformaticians, who often deal with sizable aggregated data. Several different commercial Database Management Systems (DBMS) are available on the market; nevertheless the scientific community has mainly made use of the open source and freely available solutions for data management: MySQL (www.mysql.com) and PostgreSQL (www.postgresql.org). An important aspect of these DBMS is that they can be queried through a universal standard language (Structured Query Language, or SQL).

A database not only has to store data, but it must also contain information about the connection between the various aspects of the data it contains. So before designing a database for any kind of data, the relationship between various data items has to be identified. This is often not a trivial

task as there are usually several ways in which a set of data can be organised. Well designed databases are fundamental to efficient and accurate data retrieval and care must be taken to organise the structure of the database to avoid unnecessary duplication of data that can lead to inconsistency when the data is changed or deleted.

The relational model, proposed by Codd (1970), is the model on which most current database systems are based. Due to their structure, relational databases offer a fast and reliable mechanism for retrieving, updating and inserting data. The relational model is based on two main elements: the relation and the table, which differ in their nature. The tables contain the data which is divided into entities and attributes. A protein is an example of an entity and its attributes are used to provide description, such as the molecular weight, isoelectric point and so on. The attributes make up the columns in the entity table and whenever new data is added to a table (e.g. information on a new protein) a new row is inserted.

The concept of relation stems from a formal mathematical definition in set theory but effectively it describes how data in one table is related to another table. It is customary to assign an identifying column (primary key) that contains a unique identifier to each row. Relationships are achieved through the use of foreign keys that refer, or point, to primary keys in a different table, and tables are said to be joined when a reference between keys in separate tables are followed. Relationships can fall into one of the following three classifications (Atzeni *et al.*, 1999):

3.2.1 One-to-one relationships

A One-to-one relationship describes an association between two tables that are related in such a way that a single row in one table is linked to a single row in another table (i.e. a reciprocal relationship). If there are two instances of two entities (A and B) called A_i and B_i , then a one-to-one relationship exists if at all times A_i is related to no instances of entity B or one instance of entity B, and B_i is related to no instances of entity A or one instance of entity A. One-to-one relationships are infrequently used as they are easily incorporated into a single table.

3.2.2 One-to-many relationships

The most common type of relationships is the one-to-many relationship, and as a matter of fact, most databases consist solely of these relations. A One-to-many relationship denotes a link from a primary key in a single row in one table to many foreign keys (i.e. many rows) in another table. If we have instances of two entities (A and B), then a one-to-many relationship exists between two instances (A_i and B_i) if A_i is related to zero, one, or more instances of entity B and B_i is related to zero or one instance of entity A. These relationships are unidirectional.

3.2.3 Many-to-many relationships

Many-to-many relationships are also very common in real life but they pose problems in database design and are usually resolved to one-to-many relationships through an intermediary table. A many-to-many relationship exists between entities A and B if for two instances of those entities (A_i and B_i), A_i can be related to zero, one, or more instances of entity B and B_i can be related to zero, one or more instances of entity A. In database schemas, many-to-many relationships are converted by a collection of one to many relationships through a third table, often referred to as composite entity table.

3.2.4 Normalisation and table design

The quality of the database design can be assessed by rules, known as normal forms. Each normal form represents an increasingly stringent set of criteria that aim to eliminate redundancies of data, which can cause undesirable behaviour. By normalising a database, entities are separated into their own relations, which allows modification and deletion of data without disturbing entities other than the one being directly modified (Atzeni *et al.*, 1999).

First Normal Form, 1NF

First normal form is defined as 'Data stored in a two-dimensional table with no repeating groups'. A repeating group is an attribute that has more

than one value in each row, e.g.:

A relation with repeating groups

ID	Protein name	Tumour types
001	KRAS2	pancreatic, colorectal, lung, thyroid
002	BRCA2	breast, ovarian, pancreatic

The above table, once rearranged in first normal form, would appear as follows:

First normal form

ID	Protein name	Tumour types
001	KRAS2	pancreatic
001	KRAS2	colorectal
001	KRAS2	lung
001	KRAS2	thyroid
002	BRCA2	breast
002	BRCA2	ovarian
002	BRCA2	pancreatic

Second Normal Form, 2NF

In second normal form, 'The relation is in first normal form and all non-key attributes are functionally dependent on the entire primary key,' i.e. there is one relation (or table) for each entity in the 1NF relation. This can be illustrated by the following example where the relation is expressed as: relation (table) name, followed by the attributes within parenthesis:

A relation with mixed dependencies

experiments (first name, last name, date, type of experiment, protein found, reagent used, reagent's supplier)

Second normal form requires the above to be separated into independent tables, such that all the entities are dependent on their primary key (the primary key is shown underlined):

Second normal form

scientists (scientist id, first name, last name)

reagents (reagent id, supplier)

experiments (experiment id, scientist id, reagent id, date, type of experiment, protein found)

Third Normal Form, 3NF

Third normal form is defined as 'relation in second normal form and no transitive dependencies.' A transitive dependency exists when the following functional dependency pattern is observed: $A \rightarrow B$ and $B \rightarrow C$, therefore $A \rightarrow C$.

A relation with transitive dependency

reagent (item number, reagent name, distributor's name, warehouse phone number)

This relation could be normalised in third normal form as:

Third normal form

items (item number, reagent name, distributor id)

distributors (distributor id, distributor's name, warehouse phone number)

Higher degrees of normal forms exist, however a normalisation to 3NF is usually sufficient. Each degree of normalisation involves the creation of separate tables with matching primary key and foreign key. When the database is queried, it has to connect these keys up for all relevant tables in an operation called 'join'. As each join takes up extra time, the querying of a highly normalised database slows down the performance of the database. The degree to which a database is normalised is thus a balancing act between structural quality of the data and database efficiency.

3.3 Methods

3.3.1 Prediction of protein-protein interactions

Networks of interacting proteins were constructed for the human, brown rat (*Rattus norvegicus*) and fission yeast (*Schizosaccharomyces pombe*) genomes. These three species are all eukaryotes and are valuable for cancer research. The human genome is an obvious choice for human cancer studies and the rat has been chosen as it has been used as a model for cancer studies (Kerler and Rabes, 1994). The fission yeast is a model organism that is fairly simple and contains a small set of genes, but nevertheless proves useful for studies into the cell cycle (Hayles and Nurse, 2001).

In an attempt to make the networks as comprehensive as possible, the predictions were based on a large amount of experimental data from a diverse collection of species. The method of prediction was based on the principle of orthologous interactions described by Walhout *et al.* (2000) and Matthews *et al.* (2001), but in order to improve the quality and usefulness of the predictions, a procedure for assigning confidence scores to interactions was developed and implemented.

One of the benefits of the orthologous approach is the reduced noise in protein interaction data which allows predictions of interactions that would not have been detected in a single organism (Sharan *et al.*, 2005). The method identifies putative interactions based on homology to experimentally determined interactions in a range of different species. BLAST sequence similarity searches were run for the human, rat and fission yeast genomes as documented in version 9 of the Reference Sequence Database (Pruitt *et al.*, 2005) against all proteins in the DIP (Salwinski *et al.*, 2004) and MIPS Mammalian Protein-Protein Interaction databases (Pagel *et al.*, 2005). The data from both databases were sourced from the most current releases in March 2005. The putative interactions were given confidence scores based on two factors: the level of homology to proteins found experimentally to interact, and the amount of experimental data available. The confidence score, S , is given by

$$S = \sum_{i=1}^N \ln(s_{\mathbf{a}_i} s_{\mathbf{b}_i}) n \quad (3.1)$$

where s_{a_i} and s_{b_i} are sequence similarity bit scores to proteins a_i and b_i , respectively, which have experimentally been shown to interact; n is the number of experiments linking protein a_i to protein b_i ; and N is the total number of instances where the same pair of proteins is identified as interacting through different homologues (if in same species) or orthologues (if in a different species), see Figure 3.1 for further illustration of the concept. As mentioned in Section 1.4.2, two-hybrid experiments are prone to giving false-positive results. Although most of the interactions created here are derived through yeast two-hybrid links, it has been shown that confidence is higher for interactions detected in multiple independent yeast two-hybrid experiments (Bader and Hogue, 2002; Jansen *et al.*, 2002a). This fact is reflected in the additive nature of the score, where a protein interaction that shows up repeatedly in independent two-hybrid experiments gets a higher score.

The experimental data arises from several methods and the most frequent are listed in Table 3.1. The DIP and MIPS Mammalian Protein-Protein Interaction databases were selected for the reason that they contain a large amount of manually curated interaction data and do not overlap. This is an important point, as overlapping data could lead to the false assumptions, for instance implying that certain protein interactions have multiple experimental evidences, where in fact they are the same experimental observation, but only documented repeatedly in two separate databases. Such discrepancies would compromise the scoring function as it rates multiple documented interactions higher than single observations.

The interaction data included proteins from an extensive list of species, which can be seen in Table 3.2. The majority of the proteins come from three model organisms (the fruit fly, bakers yeast and worm) in addition to other species, mainly mammals and bacteria.

3.3.2 ROC curve analysis

Receiver operating characteristic (ROC) analysis is frequently used for assessing sensitivity and specificity of prediction methods. The ROC analysis builds on the outcome of the prediction, in particular on the rates of true and false positive identification of interactions. True positive interactions were sourced from the Human Protein Reference Database (HPRD), Ver-

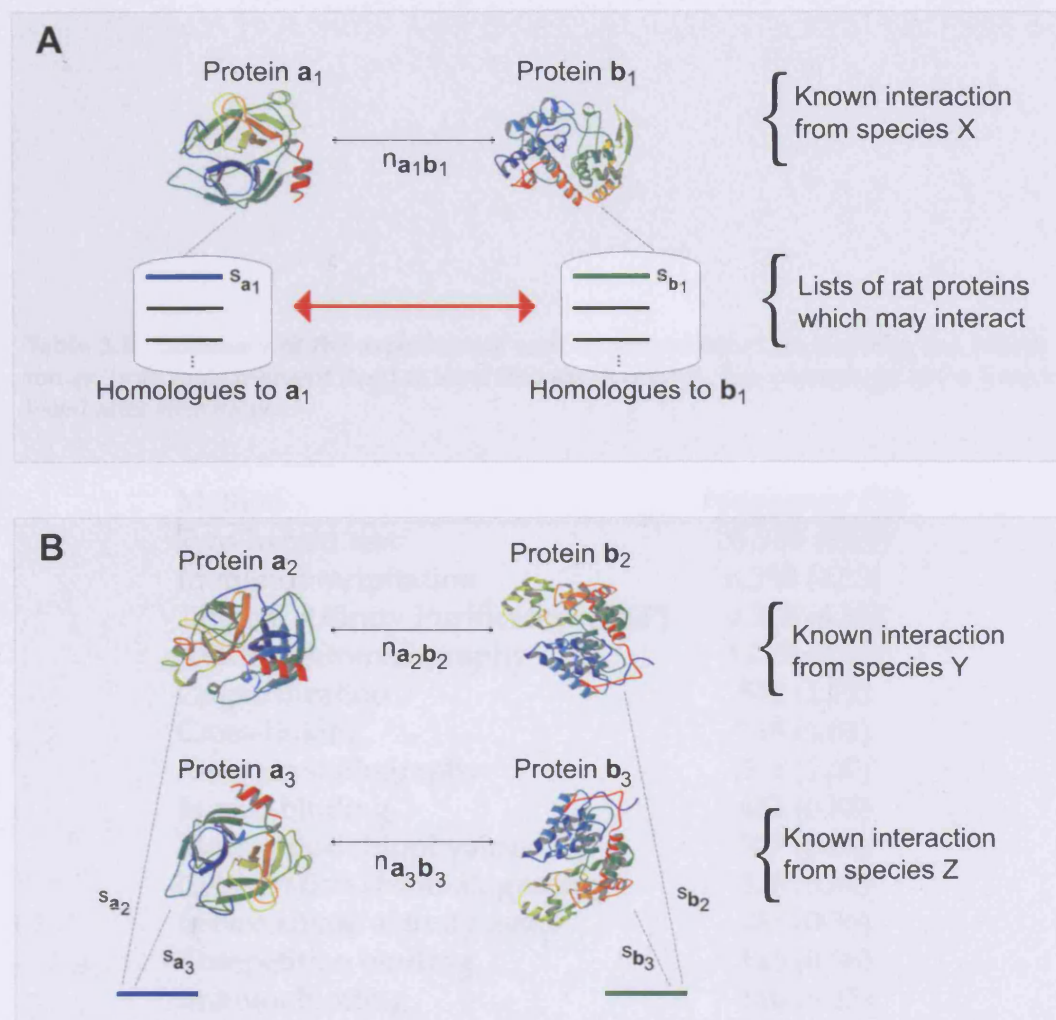


Figure 3.1: (A) Each interaction is inferred from homology to experimentally observed interactions. In this schematic, proteins a_1 and b_1 have been shown experimentally to interact in one organism, here labelled 'species X'. Lists of homologues (rat homologues in this example) are generated for both proteins, ranked by their bit score (s_{a_i} , s_{b_i} , etc.). A protein from one list may interact with a protein from the other (shown by the red arrow) and potential pairwise interactions are scored according to Equation 3.1, based on homology to the proteins involved in the known interaction. Furthermore, interactions receive a higher score if they are derived from multiple experimental sources ($n > 1$). (B) The score is additive, for instance, in the example above, the blue and green sequences are predicted to interact based on the interactions in 'species X'. If the blue and green sequences also share homology with experimentally determined protein interaction in 'species Y' and 'species Z', the process is repeated and the overall score is the sum of all pairwise scores from Equation 3.1. This additive process continues over all experimentally determined protein pairs, N , for which the rat sequences, labelled blue and green, are present.

Table 3.1: Summary of the experiments used as a foundation for building the interactomes, from most frequent (top) to least frequent (bottom). The percentage of the total is listed after each value.

Method	Frequency (%)
Two-hybrid test	35,759 (69.9)
Immunoprecipitation	6,290 (12.3)
Tandem Affinity Purification (TAP)	3,503 (6.85)
Affinity chromatography	1,070 (2.09)
Co-purification	572 (1.12)
Cross-linking	518 (1.01)
X-ray crystallography	511 (1.00)
<i>In vitro</i> binding	452 (0.88)
Biochemical/biophysical	327 (0.64)
Gel filtration chromatography	326 (0.64)
<i>In vivo</i> kinase activity assay	185 (0.36)
Competition binding	185 (0.36)
Immunoblotting	140 (0.27)
Co-sedimentation	133 (0.26)
Gel retardation assays	106 (0.21)
Native gel electrophoresis	103 (0.20)
Other	973 (1.90)

Table 3.2: Summary of the model organisms from which experimental data was sourced. The number of proteins from each organism is listed in a descending order in the right-hand side column (and corresponding percentages within parentheses). In addition, the data set included 186 proteins from 77 miscellaneous species, mostly bacteria.

Species	Number of proteins (%)
<i>Drosophila melanogaster</i>	7,052 (39.3)
<i>Saccharomyces cerevisiae</i>	4,751 (26.5)
<i>Caenorhabditis elegans</i>	2,638 (14.7)
<i>Homo sapiens</i>	1,293 (7.2)
<i>Helicobacter pylori</i>	708 (4.0)
<i>Escherichia coli</i>	545 (3.0)
<i>Mus musculus</i>	433 (2.4)
<i>Rattus norvegicus</i>	182 (1.0)
<i>Vaccinia virus</i>	38 (0.2)
<i>Bos Taurus</i>	36 (0.2)
Other (77 entries)	186 (1.3)

sion 13 (Peri *et al.*, 2003). The HPRD documents protein-protein interactions that have been manually curated from the literature by expert biologists which makes it suitable as a gold standard positive set. Selection of negative examples for the gold standard negative set is more difficult; unlike positive interactions, it is rare to find validated reports of non-interacting proteins, especially not on a large scale. Following the approach of Rhodes *et al.* (2005), interactions were classified as false positive if interacting proteins had been documented in incompatible cellular compartments (plasma membrane proteins interacting with nuclear proteins), as annotated in the December 2005 release of the Gene Ontology (GO) Consortium database (Ashburner *et al.*, 2000). Ben-Hur and Noble (2006) have highlighted the possibility that this approach for selecting negative samples can lead to biased estimates of accuracy. However, it is unlikely to affect accuracy estimates of the predictions here as they are based on experimental data, rather than being based on machine learning algorithms which rely on accurate definitions of true positive and negative samples for their training.

The receiver operating characteristic analysis uses the rates of true and false positive identification of interactions (TP and FP, respectively). The

relationship between these values, along with their counterparts, the true negatives and false negatives (TN and FN), can be represented by the following:

$$\text{True positive fraction (TPF)} = P(1|I+) = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{True negative fraction (TNF)} = P(0|I-) = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{False positive fraction (FPF)} = P(1|I-) = \text{FP}/(\text{FP} + \text{TN})$$

$$\text{False negative fraction (FNF)} = P(0|I+) = \text{FN}/(\text{FN} + \text{TP})$$

In each case the probability of predicted interaction or absence of interaction ('1' and '0') is calculated, both for the cases where a real interaction has been proved (I+) or disproved (I-). The TPF equals the sensitivity and TNF the specificity. Furthermore, the positive predictive value, PPV, where $\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$, indicates the likelihood of true interaction, given a positive prediction.

The utility of the ROC curve is best appreciated once the area under the curve (AUC) is calculated. Hanley and McNeil (1982) demonstrated that the AUC is a general measure of the probability that a predicted interaction is a genuine interaction and furthermore, the quality of the curve can be measured by the curve standard error (SE), which is given by:

$$\text{SE} = \sqrt{\frac{A(1 - A) + (n_p - 1)(Q_1 - A^2) + (n_n - 1)(Q_2 - A^2)}{n_n n_p}}, \quad (3.2)$$

where A is the area under the curve; n_n and n_p are the numbers of true negatives and true positives respectively; $Q_1 = A/(2 - A)$; and $Q_2 = 2A^2/(1 + A)$. The SE is a useful indicator whether the sample size is large enough in order for the ROC curve to be accurate—an increased sample size will yield lower standard error.

3.3.3 PIP database schema

The large amount of data involved in the project necessitated a structured and efficient approach to data management. This was achieved with the implementation of a MySQL relational database which was named 'PIP: Potential Interactions of Proteins'. The database can be queried in different ways: from the command line, by using a scripting language, or through a

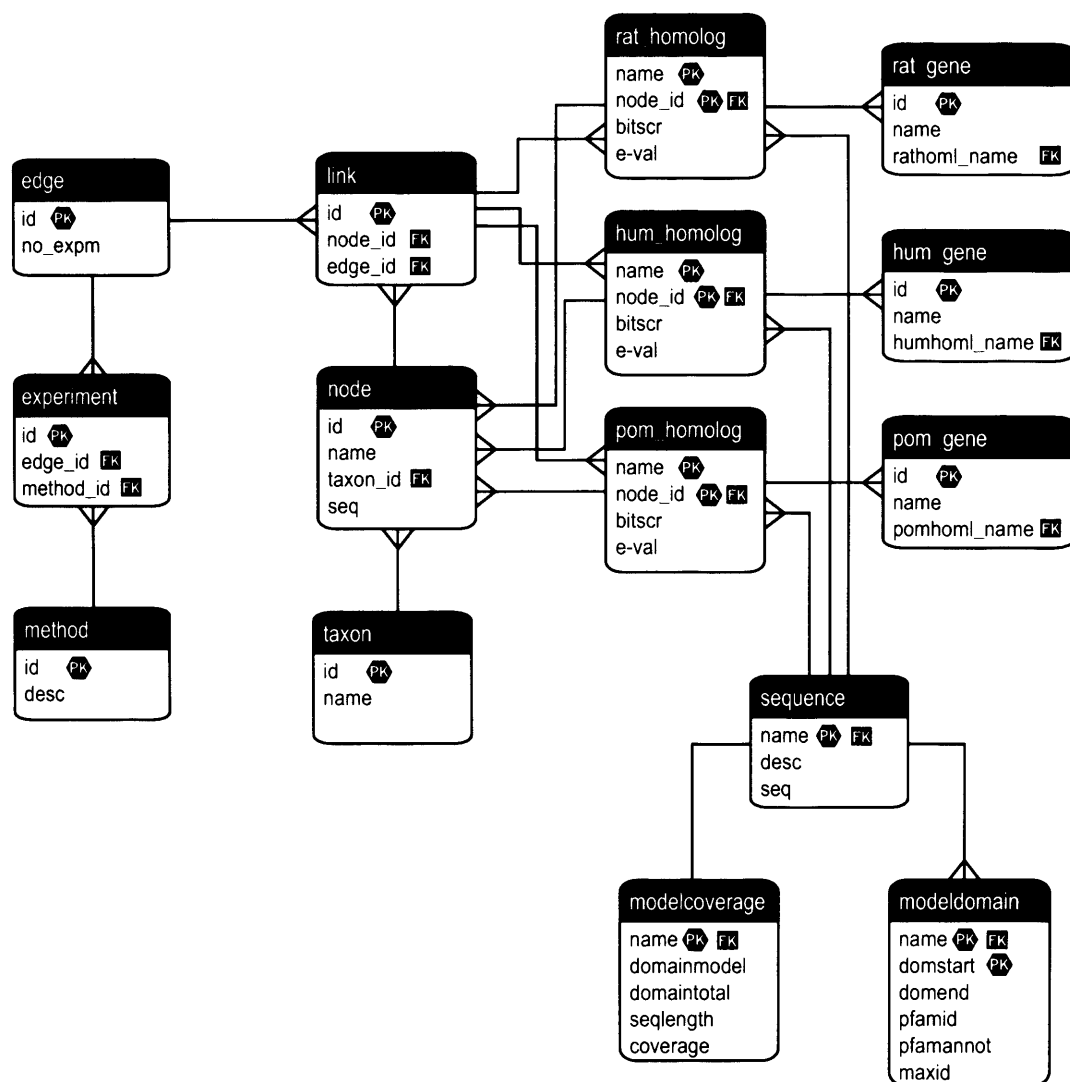


Figure 3.2: Database tables used for the PIP database. The relationship between tables is shown by conventional symbols: one-to-one (—), one-to-many (—<) and many-to-one (>—). Primary and foreign keys (PK and FK, respectively) are also indicated.

web interface. Most of the data analysis was done with Perl scripting after implementation of a Database Interface (DBI) module, which provides the basic abstraction layer for working with databases.

The basic layout of PIP's tables is shown in Figure 3.2. The tables were designed to conform to the third normal form in most aspects, but some compromises were made to speed up query response time. The database comprises 45,154 separate experimental evidences (edges) based on 59 biochemical methods and involving 17,926 proteins (nodes). The numbers of homologues to the proteins involved in the experimental data set differ between species; the database documents 737,000 entries for human, 594,000 for rat and 101,000 for fission yeast homologues.

The following examples outline the sets of queries involved in exploring the potential interaction partners of a protein, for example in the rat genome. Assuming the RefSeq identification for the query protein is known, the first step is to establish whether the protein shares homology to any of the proteins in the experimental set:

```
SELECT rat_homolog.bitscr, node.name AS node.name, taxon.name AS node.species, edge.no_expm,
edge.id AS source, node.id AS node1 FROM rat_homolog, edge, LEFT JOIN link ON (link.edge_id
= edge.id AND link.node_id = node.id), LEFT JOIN node ON rat_homolog.node_id = node.id,
LEFT JOIN taxon ON node.taxon_id = taxon.id WHERE rat_homolog.name = '[query sequence id]'
ORDER BY rat_homolog.bitscr DESC;
```

The above query finds the experimental homologues along with annotations as shown below:

bitscr	node.name	node.species	no_expm	source	node1
328.1760	LCK_HUMAN	Homo sapiens	3	MIPSM:P30530_P06239	17396
328.1760	LCK_HUMAN	Homo sapiens	2	MIPSM:P28907_P06239	17396
327.0200	PIR:TVHUA	Homo sapiens	1	DIP:25E	5742
326.6350	ABL1_MOUSE	Mus musculus	3	MIPSM:P04629_P00520	17907
322.7830	SR64_DROME	Drosophila melanogaster	1	DIP:40003E	3428

The next step involves an interactive query to get the interaction partners of the nodes in the experimental data set and then retrieving their rat homologues along with associated information and bit scores. The following queries are for the first result line in the table above:

```
SELECT node.id as node2, FROM node, link WHERE node.id = link.node_id AND link.edge_id =
'MIPSM:P30530_P06239' AND link.node_id != '17396';
```

which gives the interacting node:

```
+-----+
| node2 |
+-----+
| 17414 |
+-----+
```

and once the interacting node has been found the details of the homologous rat sequences to the partner node can be listed (ranked by score):

```
SELECT node.name AS node.name, sequence.name AS rathomolog.id, rat_homolog.bitscr,
sequence.desc, rat_gene.name AS gene FROM node LEFT JOIN rathomolog ON node.id =
rathomolog.node_id, LEFT JOIN ratseq ON ratseq.name = rathomolog.ratseq_name,
LEFT JOIN sequence ON sequence.name = rat_homolog.name, LEFT JOIN rat_gene ON
rat_gene.rathomol_name = rathomol.name WHERE node.id = '17414'
ORDER BY rathomolog.bitscr DESC;
```

which yields:

node.name	rathomolog.id	bitscr	descr	gene
UFO_HUMAN	XP_218346	1483.39	similar to rat Axl shortform [Rattus norvegicus]	Axl
UFO_HUMAN	NP_058788	624.39	TYRO3 protein tyrosine kinase 3 [Rattus norvegicus]	Tyro3
UFO_HUMAN	NP_075232	614.76	MERTK [Rattus norvegicus]	Mertk
UFO_HUMAN	NP_113705	235.34	met proto-oncogene [Rattus norvegicus]	Met
UFO_HUMAN	XP_347256	235.34	similar to met proto-onco [Rattus norvegicus]	LOC368085

This process is repeated for all the homologues to the initial query sequence, all data collected and the scores calculated. Once this process is completed the results are displayed, ranked by the interaction score, i.e. the most confident hits listed first.

The interactome data is available through the PIP (Potential Interactions of Proteins) web server at <http://bmm.cancerresearchuk.org/servers/pip/>. This is an interactive server, allowing scientists to submit queries for the three species. It is built on Perl scripts that access the database and perform the necessary task to process the data and finally display the results via a Common Gateway Interface (CGI). The combination of CGI and Perl is a well established approach for generating web interfaces for number crunching, querying databases, generating customised graphics and perform any other server-side tasks.

3.3.4 Dynamic interaction maps

In addition to textual information, the web server displays graphs of the protein-protein interactions that are predicted for the query sequence. Graph drawing and the study of algorithms to display graphs is a well established field and has been the subject of a number of publications, see e.g. Di Battista *et al.* (1994) and Kaufmann and Wagner (2001). Graph drawing involves mapping the data, usually on a two-dimensional plane utilising cost functions to optimise the arrangement of edges and nodes such that it can more easily be interpreted. The web server utilises the program Graphviz (<http://www.graphviz.org>) for visualisation of the interaction networks. Graphviz was originally intended for graph drawing in telecommunication networks and software engineering. It builds undirected graphs by running an iterative solver to find low energy configurations for the arrangement of the nodes and edges, and is implemented in a plain text graph description language named 'DOT' (Gansner *et al.*, 1993).

3.4 Results and Discussion

3.4.1 PIP server

The server comprises an SQL database and a web interface, allowing the user to enter either the sequence or the gene name of the protein of interest (see Figure 3.3). The database is then queried and potential binding partners are listed, ranked by the score as shown in Figures 3.4 and 3.5. The results are also presented graphically with the aid of the software package Graphviz. Further queries can be made by clicking on a protein in the map, and expanding the network as needed (see Figure 3.6).

To help with analysis and validation of the networks, further details on each protein are available through external links. Information on the domain structure and the facility to build homology models are also provided, with links to the Domain Fishing (Contreras-Moreira and Bates, 2002) and 3D-JIGSAW (Bates *et al.*, 2001) modelling servers. These features have been included to assist the user in understanding the structural compositions of the proteins involved in the interactions and can be used to confirm association or highlight a certain interaction as an unlikely one.

Potential Interactions of Proteins

Welcome to PIP, a webserver for potential protein-protein interactions of human, rat and fission yeast proteins. It predicts interactions, which are derived from homology with experimentally known protein-protein interactions from various species.

Select the species which you would like to investigate and then enter the details of your protein below. The server will go through its databases and reply with possible interaction partners of your protein, ranked by a confidence score.

The screenshot shows the PIP web interface. At the top, there are three tabs for species selection: **Schizosaccharomyces pombe** (selected), **Rattus norvegicus**, and **Homo sapiens**. Below the tabs, there are two input options: (A) Gene name (e.g. *cdc2*): with a text input field, and (B) Protein sequence: with a larger text area. Below these is a checkbox labeled "Lower detection threshold to get more hits:" which is currently unchecked. At the bottom left of the form are "Submit" and "Reset" buttons. At the bottom of the page, there are five links: [Help & information](#), [Cite PIP](#), [Contact us](#), [Disclaimer](#), and [Biomolecular Modelling Laboratory](#). The Cancer Research UK logo is visible in the bottom right corner.

Figure 3.3: The PIP web interface. This simple form allows the user to enter a sequence or a gene name for one of the three species: human, rat and fission yeast. The data are submitted to a server which processes the query and returns an interactive results page.

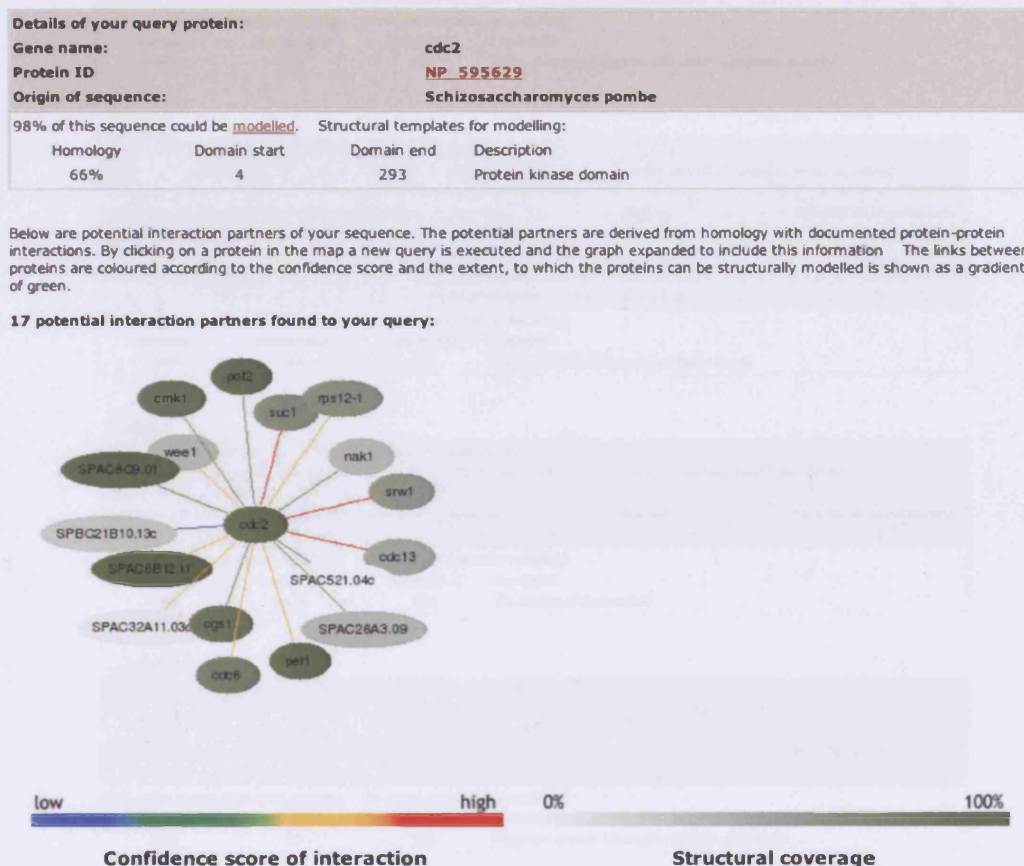


Figure 3.4: A view of the top of a PIP results page. Details of the query protein are listed and potential interaction partners are shown on a graph below. The confidence of each interaction is shown by a colour gradient that ranges from blue (low scoring interactions) to red (high scoring interactions). The proteins are also coloured in a shade of green according to how much of the sequence can be structurally modelled using homology techniques.

cdc13			
Score	Protein ID	Gene name	Description
89.07	NP_595171	cdc13	g2/mitotic-specific cyclin [Schizosaccharomyces pombe]
Homologous experimental interaction documented in:			
	Species	Source	Number of experiments
	Schizosaccharomyces pombe	DIP:1022E	1
	Mus musculus	DIP:40034E	1
	Homo sapiens	DIP:546E	1
	Homo sapiens	DIP:1121E	1
	Homo sapiens	DIP:1119E	1
	Homo sapiens	DIP:1120E	3
51% of this sequence could be modelled. Structural templates for modelling:			
Homology	Domain start	Domain end	Description
48%	206	452	Cyclin, N-terminal domain plus Cyclin, C-terminal domain

suc1			
Score	Protein ID	Gene name	Description
85.96	NP_595431	suc1	cyclin-dependent kinases regulatory subunit [Schizosaccharomyces pombe]
Homologous experimental interaction documented in:			
	Species	Source	Number of experiments
	Homo sapiens	DIP:1010E	1
	Homo sapiens	DIP:1012E	1
	Saccharomyces cerevisiae	DIP:1013E	1
	Saccharomyces cerevisiae	DIP:2258E	5
72% of this sequence could be modelled. Structural templates for modelling:			
Homology	Domain start	Domain end	Description
100%	22	102	Cyclin-dependent kinase regulatory subunit

srw1			
Score	Protein ID	Gene name	Description
59.36	NP_594674	srw1	WD domain containing srw1 protein [Schizosaccharomyces pombe]
Homologous experimental interaction documented in:			
	Species	Source	Number of experiments
	Saccharomyces cerevisiae	DIP:11729E	5
64% of this sequence could be modelled. Structural templates for modelling:			
Homology	Domain start	Domain end	Description
23%	175	531	WD domain, G-beta repeat

rps12-1			
Score	Protein ID	Gene name	Description
39.03	NP_587869	rps12-1	40s ribosomal protein s12 [Schizosaccharomyces pombe]
Homologous experimental interaction documented in:			
	Species	Source	Number of experiments
	Homo sapiens	DIP:40128E	4
66% of this sequence could be modelled. Structural templates for modelling:			
Homology	Domain start	Domain end	Description
36%	32	126	Ribosomal protein L7Ae/L30e/S12e/Gad45 family

wee1			
Score	Protein ID	Gene name	Description
36.52	NP_587933	wee1	mitosis inhibitor protein kinase wee1 [Schizosaccharomyces pombe]
Homologous experimental interaction documented in:			
	Species	Source	Number of experiments
	Schizosaccharomyces pombe	DIP:1122E	1
	Saccharomyces cerevisiae	DIP:44393E	2
35% of this sequence could be modelled. Structural templates for modelling:			
Homology	Domain start	Domain end	Description
30%	566	869	Protein kinase domain

Figure 3.5: Details regarding the interacting partners are displayed further down the results page. The confidence score and a brief description for each of the proteins are shown. A link to the RefSeq protein record of each protein provides an easy way to retrieve further annotations. The experimental data behind each prediction is also shown, along with a link to the source (DIP or MIPS). The last bit of detail contains information on the domain composition of the proteins.

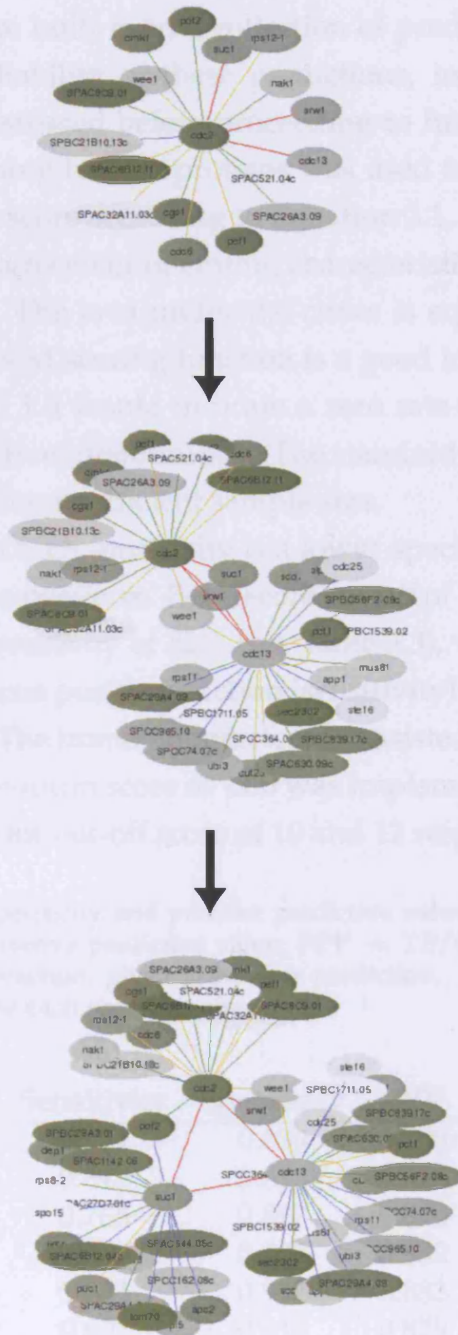


Figure 3.6: The interactive protein map. By clicking on a node, a new query is started and the protein-protein interaction map is redrawn with the results of the query. The top of the illustration shows the interaction partners of *cdc2*. By clicking on *cdc13* the interaction partners of *cdc13* are added to the map (middle) followed by clicking on *suc1* which expands the network even further (bottom).

3.4.2 Network validation

The interactomes were built from a collection of predicted protein-protein interactions. The reliability of these predictions, in terms of sensitivity and specificity, was assessed before proceeding to further analyses. Gold-standard data containing human proteins was used for this analysis. Each interaction received a score according to Equation 3.1, and the scoring function was assessed by a receiver operating characteristic curve (ROC-curve), shown in Figure 3.7. The area under the curve is equal to 0.89, which illustrates that the derived scoring function is a good indicator of prediction reliability—an area of 1.0 would indicate a zero rate of false negative and 100% rate of true positive identification. The standard error for the graph is 0.00078, which indicates a sufficient sample size.

The trade-off for a high sensitivity is a lower specificity, which means a higher fraction of false positives. For a score cut-off of 11.0 we obtained sensitivity of 85% and specificity of 82% (see Table 3.3), which is a reasonable balance between the true positive fraction (sensitivity) and the true negative fraction (specificity). The human interactome consisted of 108,113 binary interactions when a minimum score of 11.0 was implemented. This compares to 196,213 and 66,944 for cut-off score of 10 and 12 respectively.

Table 3.3: Sensitivity, specificity and positive predictive value (PPV) as a function of the cut-off score. The positive predictive value, $PPV = TP/(TP + FP)$, indicates the likelihood of a true interaction, given a positive prediction. Also shown is the total number of interactions for each score cut-off.

Score cut-off	Sensitivity	Specificity	PPV	Interactions
10	0.921	0.639	0.708	196,213
11	0.849	0.816	0.810	108,113
12	0.765	0.887	0.870	66,944
13	0.686	0.908	0.882	53,947
14	0.641	0.915	0.883	51,034
15	0.607	0.916	0.879	50,378

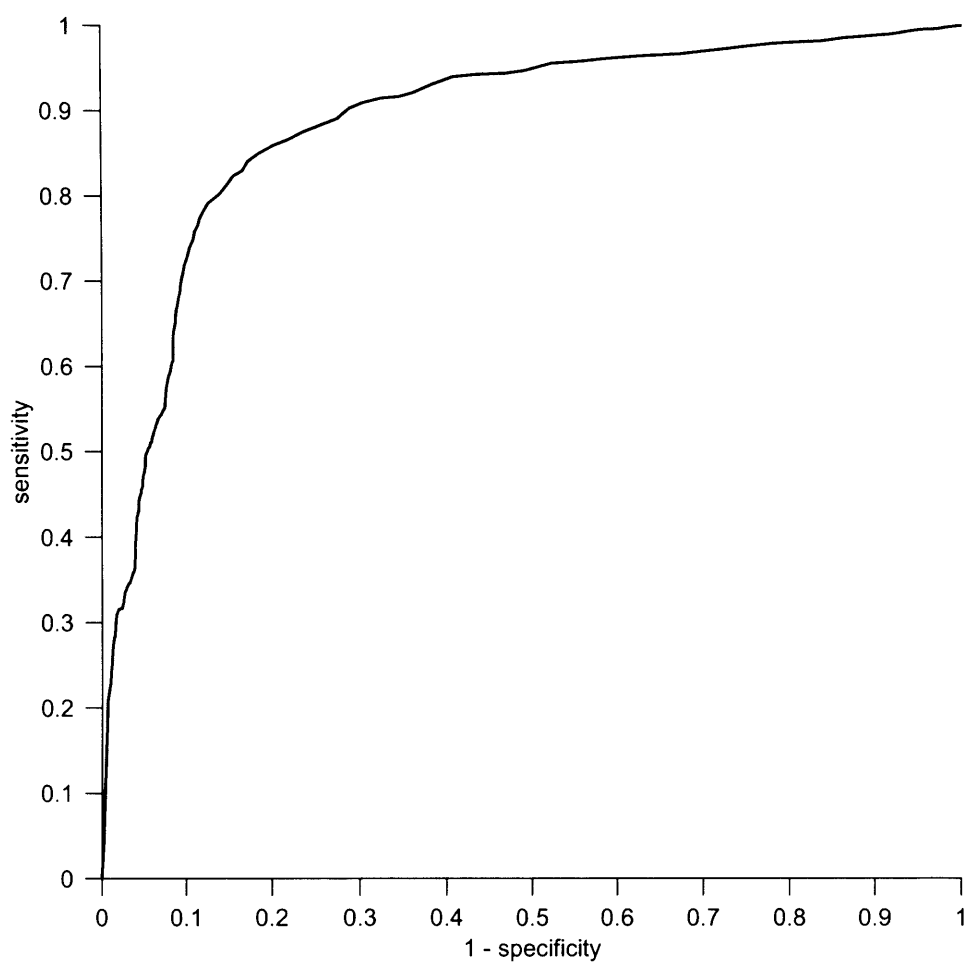


Figure 3.7: Curve of receiver operating characteristics (ROC) at different cut-off points of scores. The area under the curve is 0.89.

3.4.3 Network properties

The predicted interactomes for the three species consist of individual interactions, each of which has been assigned a confidence score as previously described. The size of each interactome therefore varies as the cut-off for the minimum confidence score is changed. Table 3.4 illustrates this for the three genomes. The selection of a score cut-off involves balancing the number of false interactions, which increases as the cut-off is lowered, against the number of undetected interactions, which increases as the cut-off is set higher (see Section 3.4.2 on network validation).

Table 3.4: The number of predicted interactions for the human, mouse and fission yeast genomes, as a function of the confidence score cut-off.

Score cut-off	Number of predicted interactions		
	Human	Rat	Fission yeast
10	196,213	151,905	30,762
11	108,113	82,621	19,314
12	66,944	51,552	12,612
13	53,947	42,179	9,755
14	51,034	40,333	9,256
15	50,378	39,909	9,103
20	38,867	30,348	6,405
30	19,155	14,223	2,978
40	11,228	8,170	1,714
50	7,518	5,442	1,047

As discussed in Chapter 1, there has been a debate regarding the structure of biological networks. The general consensus seems to be that biological networks display scale-free topology, where relatively few proteins have a high connectivity (Albert *et al.*, 2000), although the subject will remain under debate until more complete and reliable interactome data become available. Of particular concern is the fact that biological data studied to date is biased because of the limited sampling, and therefore the topology observation for the data sets so far cannot be extrapolated to whole genomes (Han *et al.*, 2005). It is however of interest to examine the network structure of the interactomes that have been constructed by the method described here. As can be seen in Figures 3.8 and 3.9 the interactome data for both the mam-

malian networks and the fission yeast show scale-free properties, i.e. the distribution of interaction frequency (degree distribution) can be fitted to a power-law curve, $P(k) = k^{-\gamma}$, where k is the degree of connectivity of a protein and $P(k)$ is the probability of observing a protein with a k connectivity. The human data shows a least-square fit R^2 -value of 0.93 and the rat and fission yeast equivalents are 0.91 and 0.94, respectively.

The appearance of scale-free properties in the predicted networks is in agreement with the same observation in a number of networks (see Chapter 1). Two main theories have been put forward to explain the scale-free nature that is frequently observed in biological networks. The first one is based on the combination of growth and ‘preferential attachment’, which stipulates that networks are the result of a growth process, during which new nodes join the system over an extended period. In the network growth, nodes prefer to connect to nodes that already have many links in a process termed ‘preferential attachment’ (Barabási and Albert, 1999). In biological systems this would be achieved through gene duplication and divergence (Pastor-Satorras *et al.*, 2003; Amoutzias *et al.*, 2004; van Noort *et al.*, 2004). The second theory, proposed more recently by Deeds *et al.* (2006), is non-evolutionary and suggests that the observed biological interactions are influenced by non-specific interactions, e.g. they show that desolvation of surface residues is a physical factor in protein-protein interactions and can affect the results from yeast two-hybrid systems and subsequently influence the observed topology of the network. As the networks described here are largely based on experimental data obtained by yeast two-hybrid techniques, the scale-free properties observed here could therefore be an artefact of the hybridisation method. In addition to this, an extra layer of uncertainty is encountered as the networks are not directly observed but inferred by computational methods. It is therefore not possible to confirm the above observation with absolute confidence—until larger portions of genomes are surveyed with alternative experimental methods the question of the scale-free nature of the protein-protein interaction networks will remain.

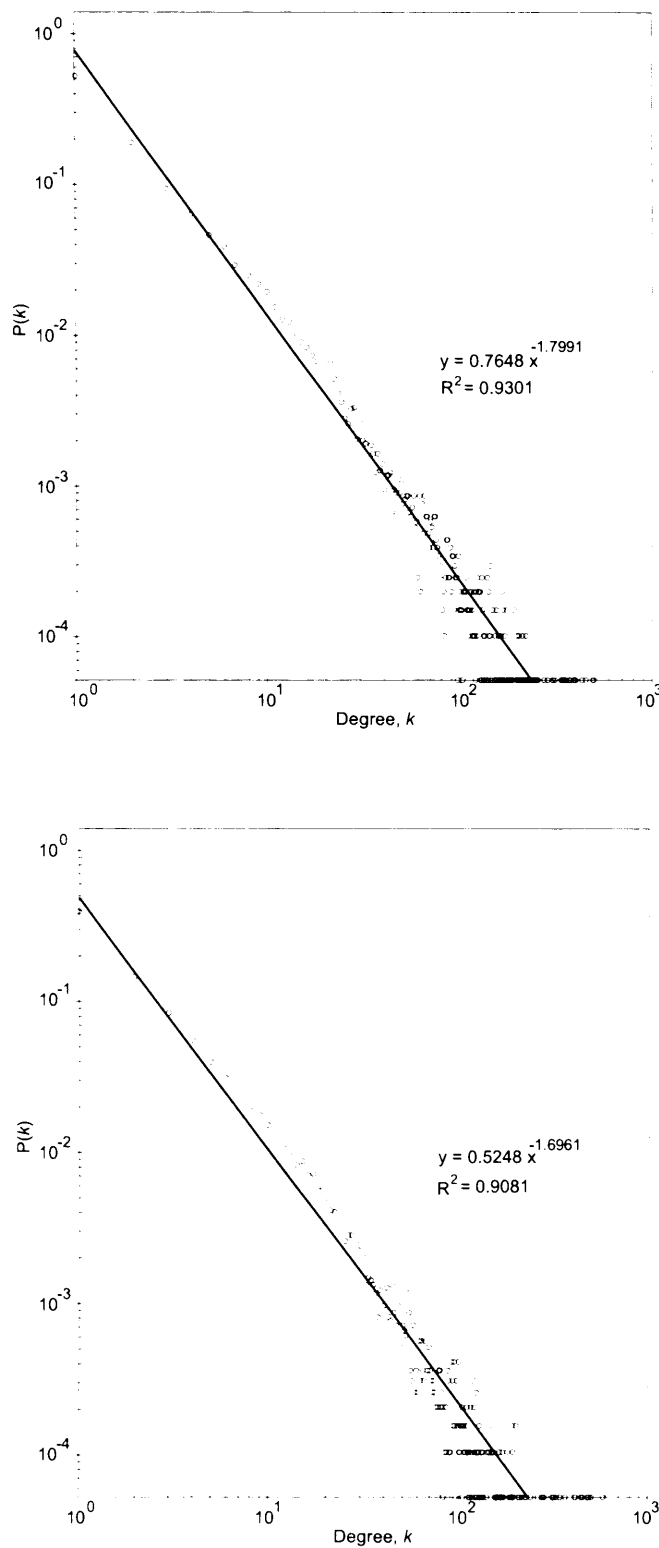


Figure 3.8: The connectivity (degree) distributions for the human (top) and rat (bottom) interactomes. Networks were built with interactions with a confidence score of 11.0 or above. The node degree k is represented on the x-axis and the probability of observing nodes with a particular k is represented on the y-axis. Both axes are logarithmic and the data exhibits a power law distribution shown by the fitted lines.

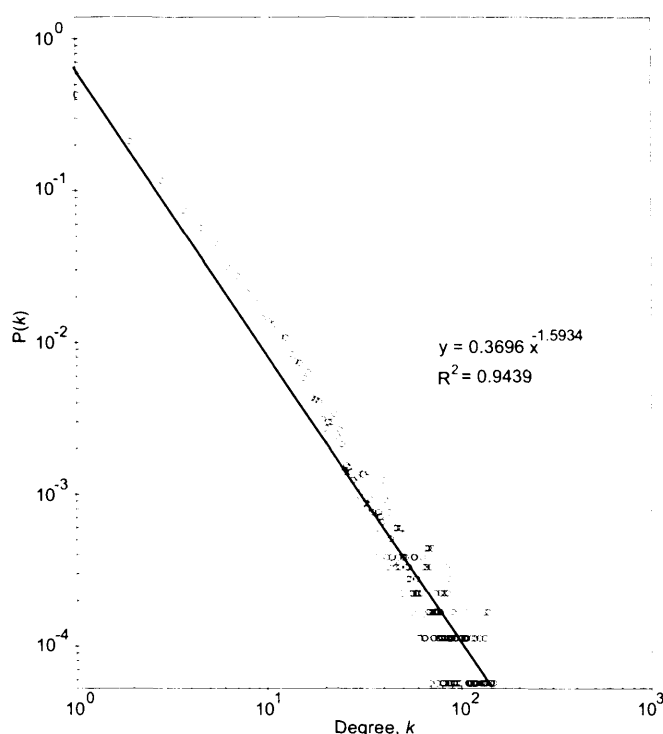


Figure 3.9: The connectivity (degree) distributions for the *Schizosaccharomyces pombe* interactome. Refer to Figure 3.8 for further description.

3.5 Conclusions

The method of using homologous protein interaction data to infer protein-protein information could be particularly useful for proteins for which there is no definite binding partner information. By expanding the networks around the protein of interest one can identify possible binding partners. The server has been implemented in such a way that it can be updated as new experimental data becomes available, thereby increasing the confidence in the predicted interactions. Furthermore it can be extended to include networks for other species, in addition to the human, rat and fission yeast networks.

The web server allows scientists to explore focused subsets of the protein-protein networks, and identify potential targets for further experimental study, but once data has been generated on a genome-wide level perhaps more interesting and revealing questions can be asked. The possibility of mapping extra attributes—such as gene ontology and expression data, to name a few—to the network opens up many interesting avenues

for network studies. This is the subject of the following two chapters, where firstly the topology of the human network is examined with a special focus on cancer-related proteins and secondly, where the rat interactome is explored in combination with microarray data from metastatic cell-lines.

Chapter 4

Global topological features of cancer proteins

4.1 Introduction

The availability of high-throughput experimental data has allowed construction of increasingly comprehensive protein-protein interaction networks (see Chapter 1). The structure, or topology, of such networks not only sheds light on the complex cellular mechanisms and processes, but also gives insight into evolutionary aspects of the proteins involved (Jeong *et al.*, 2001; Fraser *et al.*, 2002; Bu *et al.*, 2003; Wuchty, 2004; Calvano *et al.*, 2005). Charting interaction maps of entire genomes is undoubtedly useful for improved understanding of cellular function, especially once they have been integrated with a wider collection of biological data. It is now possible to map a number of different attributes, or data types, onto interactomes. Examples of this are microarray expression data (Ideker *et al.*, 2002; Sohler *et al.*, 2004; de Lichtenberg *et al.*, 2005), gene ontology (Formstecher *et al.*, 2005), structural information (Dunker *et al.*, 2005) and domain composition (Riley *et al.*, 2005; Wuchty and Almaas, 2005). This chapter continues on this theme and investigates how combining cancer-related information with a comprehensive protein network can yield useful information on the general nature of cancer proteins in the context of their location in the network. For this purpose, the study makes use of the predicted human interactome that was constructed in the previous chapter.

In a recent publication, Wachi *et al.* (2005) reported increased interaction

connectivity of proteins which were differentially expressed in lung cancer tissues. However, a comprehensive survey of the 'social nature' or interaction characteristics of all known predisposing human cancer genes has not previously been attempted. In this study the network properties of a large set of proteins known to be susceptible to mutations leading to cancer (Futreal *et al.*, 2004) was examined. In addition to statistical analysis of the network properties of cancer proteins, a clustering method was utilised to highlight proteins in centrally connected hubs that form the backbone of the interactome.

4.1.1 Community structure in networks

A network can be represented as a set of points (vertices or nodes), joined in pairs by lines, or edges. Most networks show some level of inhomogeneity, that is, the nodes are not evenly distributed throughout the network but form areas where the density is higher than the average density (see Figure 4.1). This forms the basis of the idea of a 'community structure' within networks (Newman, 2004). Strong signs of community structure can be found in networks of diverse origin, including computer networks (Albert *et al.*, 1999), social systems (such as social networks (Dodds *et al.*, 2003) and scientific collaborations (Newman, 2001)), and biological phenomena (for instance epidemiology (Heath, 2005) and biochemical networks (Holme *et al.*, 2003; Guimera and Nunes Amaral, 2005)). Communities are of interest because neighbouring and well-interconnected nodes that make up a network community are likely to share some similarity and, in the case of protein networks, may participate in similar cellular functions (Guimera and Nunes Amaral, 2005).

Finding the communities, or the clusters, is important in order to understand the network's internal structure. A large number of clustering algorithms have been developed for this purpose, based on many different approaches for distinguishing the clusters. The clustering criteria essentially fit into three categories, based on the network properties they use for their task: compactness, connectedness and spatial separation (Handl *et al.*, 2005). Methods that aim to identify clusters based on their compactness look at the global distribution of nodes and identify those that appear more densely located compared with the general distribution (MacQueen, 1976;

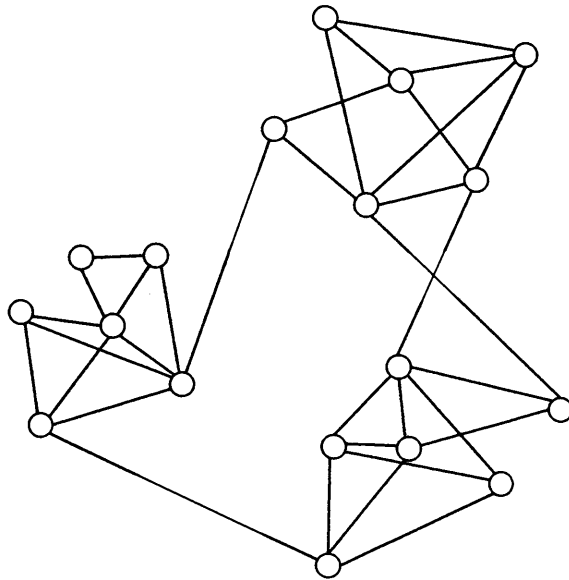


Figure 4.1: Community structure within a network. In many natural networks the nodes appear in groups or communities (shown by grey shading). Within a community there are many edges, whereas only a small number of edges connect nodes of different groups.

McLachlan and Krishnan, 1997; Kohonen, 2001). These methods are usually effective for spherical or well-separated clusters, but may not detect asymmetric or unusually shaped clusters. Basing clustering on connectedness of nodes is particularly suitable for identifying oddly shaped clusters, as the clustering is based on more local network properties, i.e. distances between individual nodes (Griths *et al.*, 1984). This approach, however, is less practical where there is little separation between the clusters. The third methodology, spatial separation, is usually combined with additional methods as it often fails to find a meaningful separation in complicated networks (Rayward-Smith *et al.*, 1996). Most of these algorithms separate the network into non-overlapping clusters, assigning each node to a specific cluster, that is they perform ‘hard clustering’, although methods that allow overlapping between clusters are increasingly getting more attention (Reichardt and Bornholdt, 2004; Gfeller *et al.*, 2005).

4.2 Methods

4.2.1 Clustering of interconnected proteins

The network of interactions was analysed using clique percolation clustering (Adamcsek *et al.*, 2006), which locates maximal complete subgraphs (k -cliques) in the networks and then identifies ‘communities’ by carrying out standard component analysis of the clique-clique overlap. In this context, the variable k is defined as the number of nodes in the subgraph and a k -clique community is defined as the union of all k -cliques that can be reached from each other through a series of adjacent k -cliques, where cliques sharing $k - 1$ nodes are defined as adjacent. The community definition is based on the observation that a typical member in a community is linked to many, but not necessarily all other nodes in the community. In other words, a community can be regarded as a union of smaller, complete, fully-connected subgraphs that share nodes. The resulting communities are allowed to exhibit a degree of overlap, which is particularly useful, as such methods have been shown to be more suitable for identifying nodes of central importance in biological networks compared with non-overlapping clustering algorithms (Wuchty and Almaas, 2005; Palla *et al.*, 2005).

4.2.2 Data sets

Information on cancer genes was obtained from a comprehensive census of human cancer genes (Futreal *et al.*, 2004). The data, 346 genes encoding 509 protein isoforms, were mapped on to the protein-protein interaction network. For the study of domain interactions, statistics relating to the frequency of domain-domain interactions were obtained from version 19.0 of PFAM (Bateman *et al.*, 2004). And lastly, the analysis of functional aspects of the networks involved classification of proteins into cellular processes according to information from release 37.0 of the KEGG Pathway Database (Kanehisa *et al.*, 2006) and June 2006 release of the Gene Ontology database.

4.3 Results and Discussion

4.3.1 Network properties of cancer proteins

The construction of a validated human protein-protein interaction network allows an in-depth analysis of individual proteins in the context of their surroundings. Here the network topographies of human cancer proteins were examined with the aim of uncovering intrinsic properties that distinguish proteins prone to cancerous mutations from those that are not.

Connectivity

The number of interaction partners for each protein in the network was calculated, using a cut-off score of 11.0 (see Section 3.4.2). Statistics were accumulated for two groups: proteins classified as cancer proteins, and those that were not linked to cancer. Cancer proteins were shown to have twice as many interaction partners as non-cancer proteins, with 23.4 ± 1.9 ($n = 439$) and 11.4 ± 0.2 ($n = 16,600$) interaction partners respectively (mean values \pm standard error of the mean, see Figure 4.2). Table 4.1 shows a more detailed breakdown of the interaction frequency. Cancer proteins are under-represented in the category of the least connected proteins, but show the reverse trend in all other categories. This trend is highly statistically significant ($p = 5 \times 10^{-34}$ for a χ^2 -test).

Table 4.1: The observed interaction frequency distribution of cancer proteins compared with the expected distribution in the genome-wide network.

Interactions	f_{Obs}	f_{Exp}	Difference (%)
1-10	215	318	-32
11-20	99	69	43
21-30	48	21	134
31-40	12	8	42
≥ 40	65	23	185
Sum	439	439	

Because each interaction in our predicted protein-protein interaction map receives a confidence score, it is possible to test the robustness of the

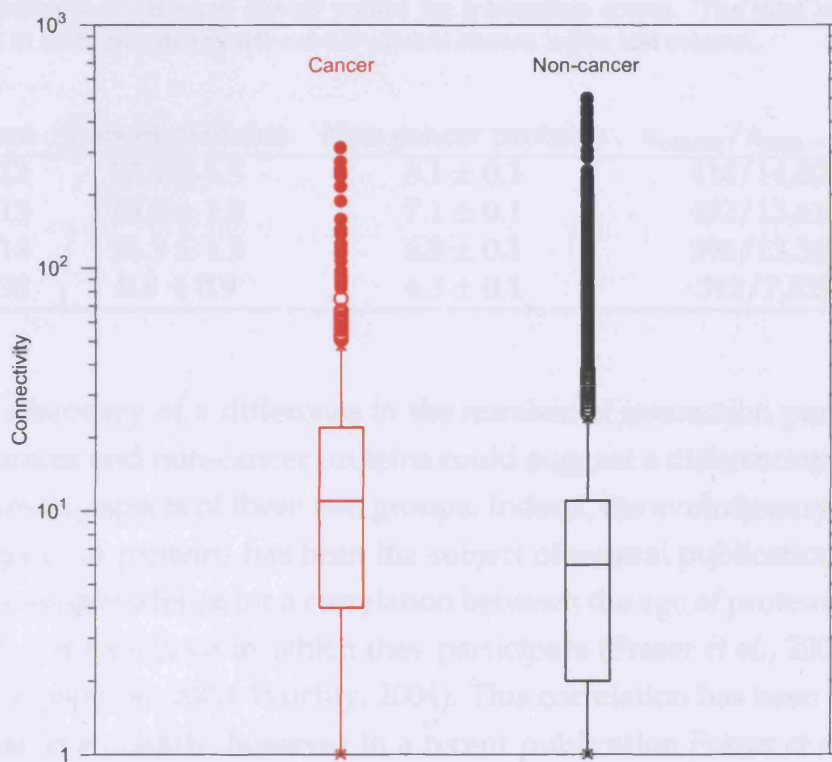


Figure 4.2: The number of interactions in which cancer proteins participate (red), compared with the number of interactions of non-cancer proteins (black). Cancer proteins are, on average, involved in 23.4 interactions, whereas non-cancer proteins are involved in 11.4. The centre of the box is the median and the box spans from first to third quartiles (the inter-quartile range). The whiskers extend to the furthest point within 1.5 times the inter-quartile range. Beyond the whiskers, all outliers are shown, in open circles up to a distance of 3 times the inter-quartile range and closed beyond that.

above observation. Increasing the cut-off score limits the amount of interaction data, but the confidence of each interaction is increased. Cancer proteins consistently showed twice as many interaction partners as non-cancer proteins, when the score cut-off was increased, indicating that the observation is robust (see Table 4.2).

Table 4.2: The mean connectivity (\pm standard error of the mean) of cancer and non-cancer proteins at different cut-off values for interaction scores. The total number of proteins in both groups at each cut-off value is shown in the last column.

Score	Cancer proteins	Non-cancer proteins	$n_{\text{cancer}}/n_{\text{non-cancer}}$
≥ 12	17.9 ± 1.5	8.1 ± 0.1	414/14,501
≥ 13	15.0 ± 1.3	7.1 ± 0.1	402/13,610
≥ 14	14.3 ± 1.3	6.9 ± 0.1	396/13,369
≥ 30	8.8 ± 0.9	4.3 ± 0.1	312/7,829

The discovery of a difference in the number of interaction partners between cancer and non-cancer proteins could suggest a differentiation in the evolutionary aspects of these two groups. Indeed, the evolutionary rate and age of genes or proteins has been the subject of several publications, showing increasing evidence for a correlation between the age of proteins and the number of interactions in which they participate (Fraser *et al.*, 2002; Eisenberg and Levanon, 2003; Wuchty, 2004). This correlation has been disputed by Jordan *et al.* (2003), however in a recent publication Fraser *et al.* (2003), and subsequently, Saeed and Deane (2006), have stated that a weak correlation does exist, although it is dependent on the completeness and quality of the data set under study. Wuchty (2004) and Saeed and Deane (2006) have, however, found a strong correlation (Spearman's coefficient = 0.97 and 0.98, respectively) between the age of a protein (calculated as Excess Retention) and connectivity. The Excess Retention calculation is based on the observation of orthologues in related genomes and as such can indicate gene age but does not indicate evolutionary rate through changes in the genomes by events such as gene loss or horizontal gene transfer. Thus the results presented here indicate that proteins whose mutation results in a detrimental change of function that leads to cancer may generally be older than the non-cancer proteins.

Jeong *et al.* (2001) showed that the most highly connected proteins in

yeast are also the ones that are phenotypically most important, and critical for the survival of the organism. Furthermore, it has been shown that toxicity-modulating proteins exhibit a greater number of interactions (Said *et al.*, 2004). It follows that the increased connectivity of cancer proteins, shown here, suggests that they play a central role in the protein network.

Cancers result from the accumulation of inherited and somatic mutations in oncogenes and tumour suppressor genes. With this in mind, it is of interest to distinguish between somatic and germline mutations that result in cancer. Germline mutations can be passed between generations, whereas somatic mutations are not passed on to offspring. This study shows a modest but statistically significant difference ($p = 0.002$) in the distribution of interaction partners of somatically mutated vs. germline mutated proteins. About two-thirds of the proteins showing somatic mutations interact with a relatively low number of proteins (twenty interaction partners or less), whereas just over half of germline mutated proteins fell in to the same grouping (see Table 4.3). This may indicate somewhat younger mean evolutionary age of the somatically mutated cancer proteins compared with germline, reflecting the fact that evolutionary selection pressure affects germline mutated proteins.

Table 4.3: Connectivity ratio of somatic and germline mutated cancer proteins (number of proteins in parentheses). The observed difference is statistically significant according to a χ^2 -test ($p = 0.002$).

Interactions	Somatic	Germline	Difference (%)
1-20	0.68 (175)	0.53 (30)	28.88
≥ 20	0.32 (83)	0.47 (27)	-32.08
<i>Sum</i>	1.00 (258)	1.00 (57)	

Protein domain frequencies

Comparing the structural and domain composition of cancer proteins against domain propensities of the general network reveals a functional diversity that differs from that of proteins in general. Table 4.4 lists the twenty most frequent domains found in the cancer protein population. Most of those domains appear with a higher frequency than expected. Many of the

proteins, whose frequency is increased compared with the expected values, appear to be of nuclear origin and have a functionality that particularly focuses on DNA regulation and repair, such as the Zinc-finger, PHD-finger, BRCT and Paired-box domains.

The concern that the apparent increased connectivity of cancer proteins is a result of a bias in the protein network needed to be addressed. As described, the construction of the human interaction network builds on experimental data and it could be argued that cancer proteins, having been under particular scrutiny by the scientific community, have been studied in greater detail which could explain the higher number of known interaction partners for cancer proteins. However, this appears not to be the case and is more likely to be a consequence of domain usage. By using interaction frequencies for domain-domain interactions from the PFAM database, the propensity of highly-promiscuous domains, in terms of the variety of different domains with which they interact, was found to be higher in the cancer population, with 22.2% of domains highly promiscuous compared with 6.9% of the non-cancer population (the number of unique domains, $n = 212$ and $n = 4042$ respectively).

Some of these domains ($p_{\text{promiscuity}} < 0.005$) appear in the list of the most frequently observed cancer domains (Table 4.4), and are preceded with the symbol '★'. This is based on a statistical analysis of interaction frequencies of PFAM domains in general, which conform to a Probability Density Function:

$$f(x, \lambda) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases} \quad (4.1)$$

where x is the domain interaction frequency and λ is the rate constant, derived from the mean value of interactions per domain (0.56), as observed in the PFAM data.

There still remains the argument that the PFAM domain-domain interaction data could be biased towards cancer protein studies. Since PFAM is derived from the Protein Data Bank (PDB) it is important to establish whether there is a bias within this primary structural database. Less than 1% of the PDB was found to contain structural information on the cancer proteins studied here, or their close homologues (E-value $\leq 1 \times 10^{-5}$, sequence identity $\geq 70\%$), thereby indicating no obvious bias towards cancer

Table 4.4: The most frequently observed domains in human cancer proteins, listed in a descending order. Their frequency is compared against the expected frequency derived from a network-wide survey. The $p_{\text{promiscuity}}$ -value shows the probability of observing a domain with higher interaction promiscuity as calculated by a Probability Density Function based on the PFAM domain population. Domains preceded with \star show statistically higher-than-expected interaction promiscuity ($p_{\text{promiscuity}} \leq 0.005$) in terms of the number of different domains with which they interact. Domains without any documented PFAM interactions receive a $p_{\text{promiscuity}}$ -value of 1.

	PFAM id	Domain annotation	f_{Obs}	f_{Exp}	n -fold difference	$p_{\text{promiscuity}}$
\star	PF00047	Immunoglobulin domain	102	37	2.76	7.93×10^{-11}
	PF00096	Zinc-finger, C2H2 type	51	35	1.46	0.0279
\star	PF00069	Protein kinase domain	51	14	3.64	3.70×10^{-13}
	PF00628	PHD-finger	35	4	8.75	1
\star	PF00097	Zinc-finger, C3HC4 type	28	8	3.50	4.67×10^{-3}
	PF00855	PWWP domain	21	1	21.00	1
\star	PF00046	Homeobox domain	17	7	2.43	7.80×10^{-4}
	PF00643	Zinc-finger, C3HC4 type plus B-box	16	3	5.33	1
\star	PF00036	EF hand	16	8	2.00	1.33×10^{-11}
\star	PF00008	EGF-like domain	13	20	0.65	1.33×10^{-11}
	PF00533	BRCA1 C Terminus (BRCT) domain	12	1	12.00	0.0279
\star	PF00010	Helix-loop-helix DNA-binding domain	11	3	3.67	4.67×10^{-3}
	PF00292	Paired-box domain	10	1	10.00	0.167
\star	PF00856	SET-domain	10	2	5.00	4.67×10^{-3}
\star	PF00076	RNA recognition motif. (RRM, RBD, or RNP)	10	9	1.11	4.67×10^{-3}
\star	PF00020	TNFR/NGFR cysteine-rich region	9	1	9.00	7.80×10^{-4}
\star	PF00018	SH3-domain	8	7	1.14	6.09×10^{-7}
	PF00439	Bromodomain	8	2	4.00	0.167
	PF00531	Death-domain	7	1	7.00	0.167
\star	PF00178	Ets-domain	7	15	0.47	7.80×10^{-4}

proteins.

Cluster analysis of the human interactome

Clustering methods have previously been shown to be useful in identifying protein interactions that take place within the same cellular process (Palla *et al.*, 2005). This can be attributed to the fact that subnetworks of proteins involved in a defined cellular process are more heavily interconnected by direct protein interactions than would be expected by chance (Jeong *et al.*, 2001; Gunsalus *et al.*, 2005).

The *k*-clique clustering method was applied repeatedly at different *k*-cluster values. Figure 4.3 demonstrates the concept of clustering, showing a subsection of the human interaction map, on which cancer proteins have been highlighted. A low *k*-value yields a large number of extensive communities which overlap to a high degree. Conversely, raising the *k*-value results in identification of fewer and more distinct protein communities which contain highly interconnected proteins (Table 4.5). Interestingly, even though cluster sizes decrease with increasing *k*-value, the proportion of cancer genes identified in the protein communities increases, indicating the enrichment of cancer proteins in the most tightly connected communities.

Table 4.5: The number of protein communities in the entire human interactome, identified by *k*-clique analysis at different *k*-values. The number of non-cancer and cancer proteins in the communities at each *k*-value is listed on the right-hand side of the table

<i>k</i> -value	Communities	Non-cancer proteins	Cancer proteins (%)
3	222	8870	334 (3.6)
4	189	4245	234 (5.2)
5	98	1918	117 (5.8)
6	37	764	53 (6.5)
7	19	325	28 (7.9)
8	9	193	14 (6.8)

An example of the protein communities identified by this method is shown in Figure 4.4. The communities contain proteins involved in a diverse range of protein functions and are either self contained or connected

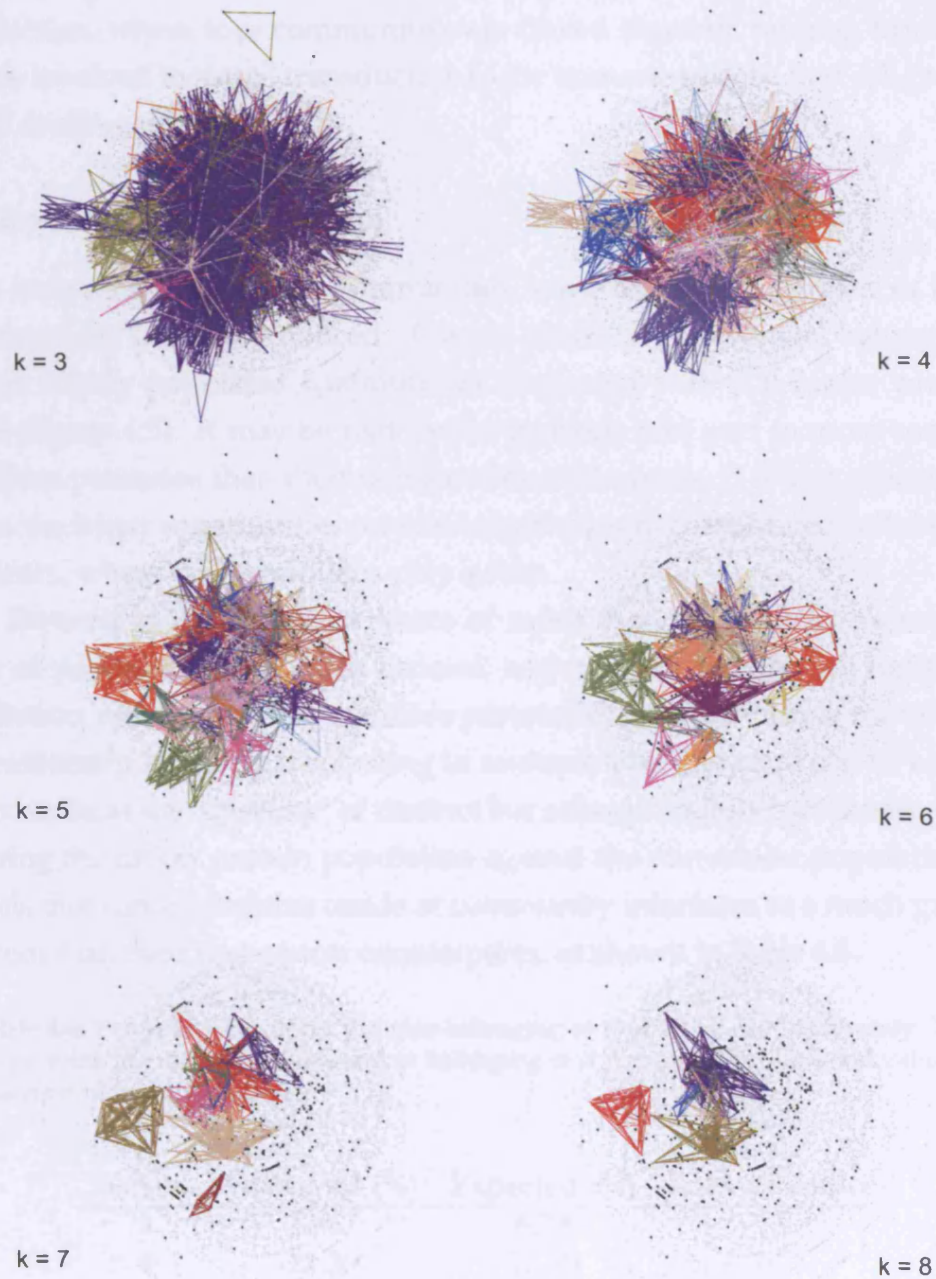


Figure 4.3: A subsection of the human interactome, focusing on protein networks around cancer proteins. Each community is shown in a different colour. Cancer proteins are depicted as dark triangles and non-cancer proteins as grey circles. Protein communities resulting from cluster analyses at different k -values are shown. Interactions within communities of highly inter-connected proteins are highlighted in different colours.

to other protein communities. In the latter case, cancer proteins can be seen mediating interactions between different pathways, such as in the largest collection, where four communities are linked together, ranging from proteins involved in signal transduction to the immune system, and cell growth and death regulation.

Community size and overlap

On examining the protein communities, some interesting differences in the community size were noticed. Cancer proteins, on average, belonged to more highly populated communities compared with non-cancer proteins (see Figure 4.5). It may be that cancer proteins take part in more complex cellular processes than their non-cancer counterparts. It is also conceivable that the larger communities contain larger or more complicated cellular machinery, where cancer proteins play a role.

Proteins identified as members of more than one protein community are of particular interest. In general, each protein community represents a distinct cellular process; therefore proteins that have multiple community membership may be participating in multiple processes and can be considered to be at the ‘interface’ of distinct but adjacent cellular processes. Comparing the cancer protein population against the non-cancer population reveals that cancer proteins reside at community interfaces to a much greater extent than their non-cancer counterparts, as shown in Table 4.6.

Table 4.6: Percentage of cancer proteins belonging to more than one community (based on proteins identified by clustering as belonging to a community). Expected value was based on non-cancer proteins.

<i>k</i> -value	Observed (%)	Expected (%)	Fold difference
3	12.67	8.38	1.5
4	21.39	12.38	1.7
5	12.37	9.96	1.2
6	17.07	13.67	1.2
7	17.39	7.26	2.4
8	7.69	2.66	2.9

While connectivity gives an indication of a protein’s importance, it is possible to further classify the topological role of highly connected proteins

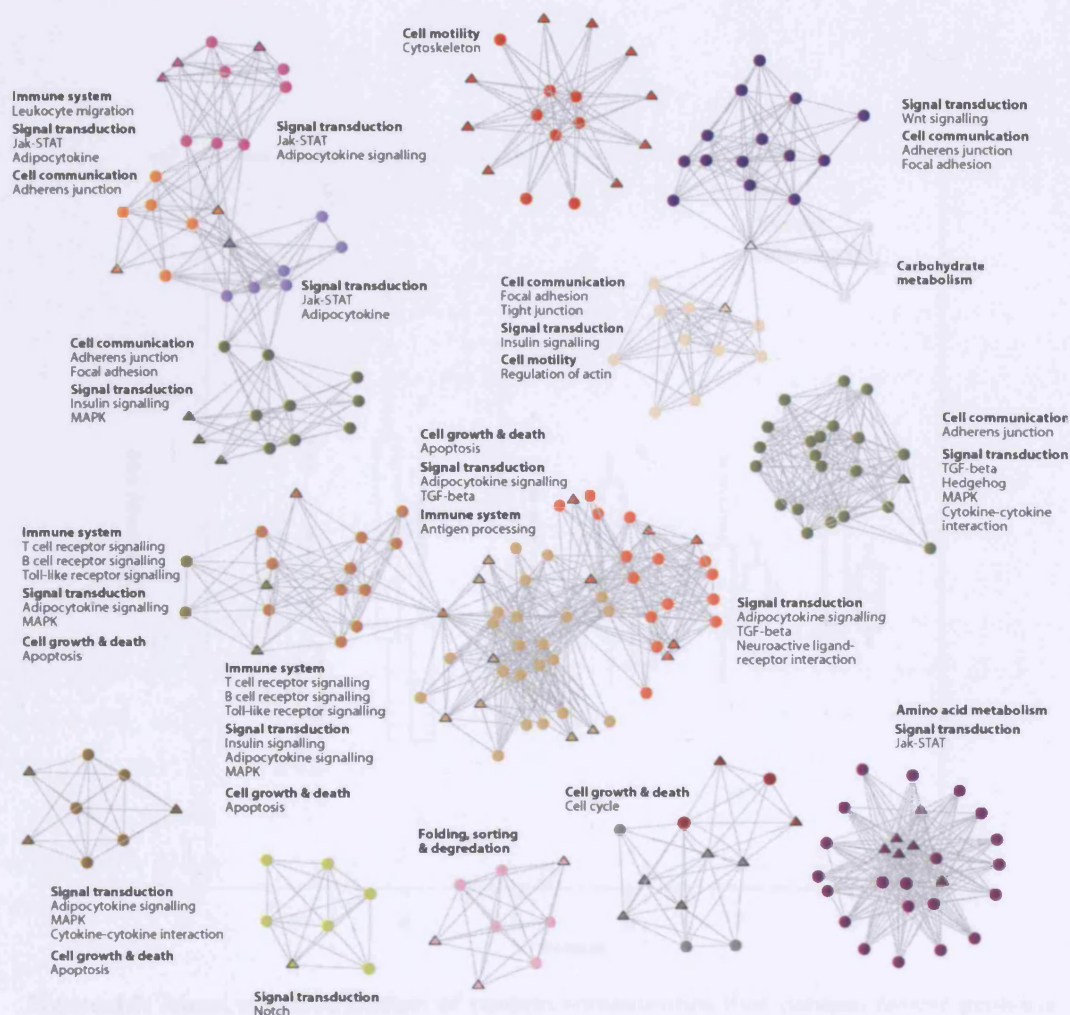


Figure 4.4: A more detailed description of the protein communities identified by k -clique cluster analysis ($k = 6$) in Figure 4.3. Each community is distinctly coloured, with cancer proteins shown as triangles. The main functional classes of each cluster (in bold) and individual pathways, as defined in the KEGG database, are listed alongside each community. Note that proteins can be members of more than one community, but the figure shows only one community assignment for each protein. Appendix C lists KEGG pathway classifications for these communities.

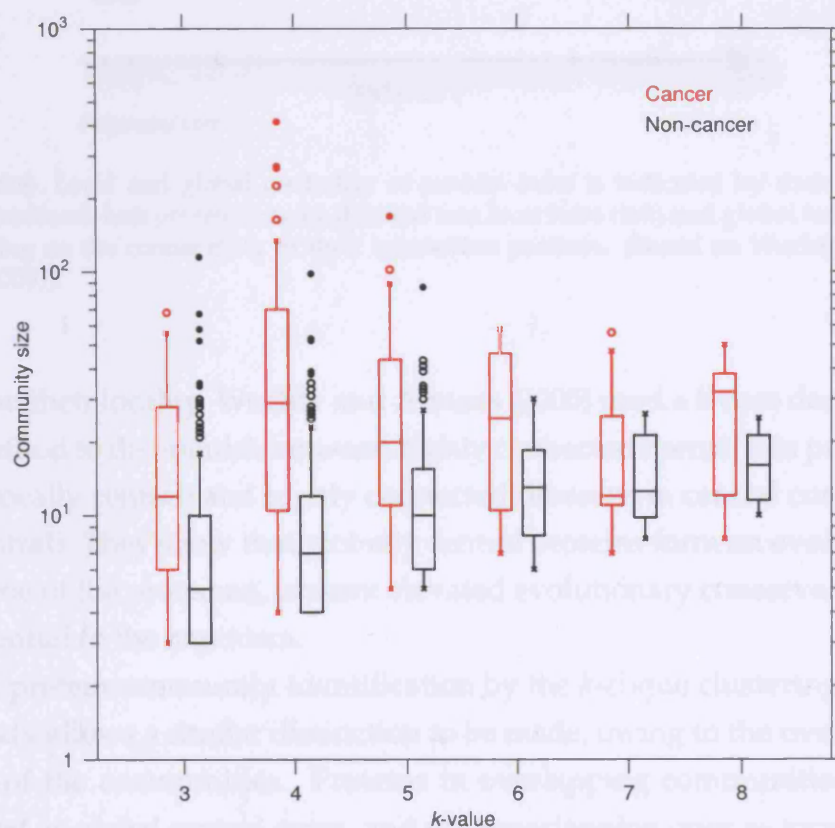


Figure 4.5: Mean size distribution of protein communities that contain cancer proteins compared with those containing non-cancer proteins. Distributions are classified according to clustering k -value, with cancer communities in red and non-cancer in black. The difference between cancer and non-cancer groups is statistically significant, according to Wilcoxon rank sum tests, for k -values 3, 4, 5 ($p < 0.005$), and 6 ($p < 0.05$).

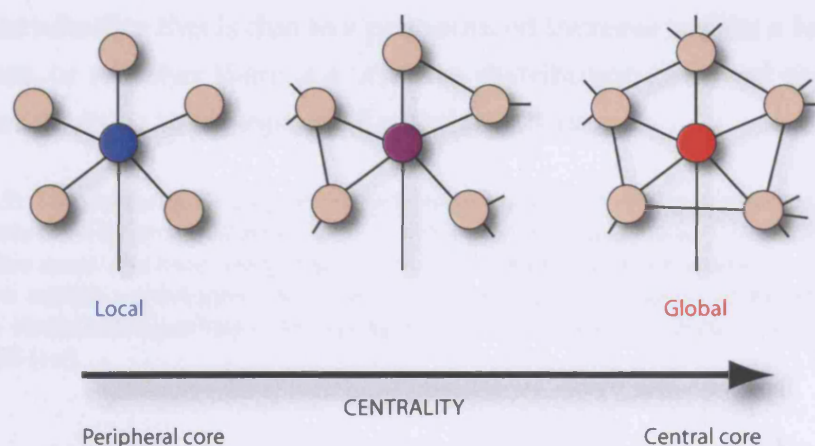


Figure 4.6: Local and global centrality of protein hubs is indicated by their network neighbourhood: hub proteins can be divided into local hubs (left) and global hubs (right) depending on the connectivity of their interaction partners. (Based on Wuchty and Almaas (2005)).

based on their locality. Wuchty and Almaas (2005) used a k -core decomposition method to distinguish between highly connected domains in peripheral cores (locally central) and highly connected domains in central cores (globally central). They show that globally central proteins form an evolutionary backbone of the proteome, present elevated evolutionary conservation, and are essential to the organism.

The protein community identification by the k -clique clustering used in this study allows a similar distinction to be made, owing to the overlapping nature of the communities. Proteins in overlapping communities can be classified as global central cores, and non-overlapping ones as local central cores (see Figure 4.6 for illustration of the concept). The results in Table 4.6 thus emphasise the key role of cancer proteins, which is reflected in their protein-network topology. This observation supports and extends the findings of Wachi *et al.* (2005), who showed that differentially expressed proteins in squamous cell carcinoma of the lung tend to be global hubs.

Cellular processes

Overall, the above findings reveal a topological distinction of cancer proteins that is primarily displayed in an increased interaction frequency compared with non-cancer proteins. In the light of this, it is interesting to in-

investigate whether this is due to a pronounced increase within a few cellular processes, or whether there is a uniform distribution of raised connectivity of cancer proteins in all aspects of cellular function.

Table 4.7: The connectivity of proteins in primary GO biological process categories. The mean connectivity and standard error of the mean for cancer and non-cancer proteins in the five most common categories. The total number of observations for each group is shown within parentheses. Stars after the percentage difference of the mean values indicate statistical significance according to Wilcoxon rank sum tests, $p < 0.05$ (*) and $p < 0.005$ (**).

GO process class	Connectivity		%
	Cancer proteins	Non-cancer proteins	
Physiological process	25.4±2.6 (264)	13.3±0.3 (7657)	+91.0 **
Cellular process	27.3±2.8 (261)	14.0±0.3 (7509)	+94.3 **
Development	14.1±1.9 (73)	12.6±0.7 (1069)	+11.7 *
Regulation of biol. process	16.1±4.5 (61)	11.3±0.7 (1018)	+43.0
Response to stimulus	35.9±8.0 (40)	12.2±0.7 (1076)	+193 **

Functional description could be assigned to 404 cancer and 10,757 non-cancer proteins by using Gene Ontology annotation. Owing to the possibility of multiple annotations for a single protein, the total number of functional descriptions was 714 and 18,711 respectively. The majority of the proteins can be assigned to five key biological processes: response to stimulus, cellular processes, physiological processes, regulation of biological processes and development. Cancer proteins are more highly connected in all these categories compared with non-cancer proteins (see Table 4.7), with the stimulus response groups and cellular processes (encompassing processes that are carried out at the cellular level, e.g. cell adhesion, communication, signal transduction, differentiation, recognition) showing the greatest enrichment of highly connected cancer proteins.

4.3.2 Conclusions

In summary, the work presented here has shown that human proteins involved in cancer exhibit a network topology which is different from that of proteins not documented as being mutated in cancer. The observation is based on the study of a global protein-protein interaction network, constructed by a homology-based method, which we have shown to be capa-

ble of accurately predicting protein-protein interactions. The most striking property of cancer proteins is the increased frequency of interactions. This observation may indicate an underlying evolutionary pressure to which cancer genes, as genes of central importance, are subjected.

The k -clique clustering algorithm allows us to explore protein-protein connectivity in a more informative way than is possible just by looking at the interaction frequency of each protein. Its feature of overlapping protein communities allows us to distinguish between central and peripheral hubs of highly connecting proteins, revealing proteins forming the backbone of the proteome. The fact that an enrichment of cancer proteins is observed in this group indicates the central role of these proteins. The domain composition of cancer proteins may indicate why this is the case: it appears that cancer proteins contain a high ratio of highly promiscuous domains, in terms of the number of different proteins with which they interact.

An example of a protein showing the above features is the HSPCA (Hsp90) heat shock protein 90kDa 1 alpha, which is shown as a white triangle with a red border in a central position in Figure 4.7. Heat shock proteins are important for cellular homeostasis. Their function is primarily assisting protein folding and transport under physiological conditions, but during stress they can promote refolding of damaged proteins. HSPCA has been linked to various processes of carcinogenesis, in particular in relation to breast cancer (Teng *et al.*, 2004), and the protein has been shown to have both prognostic and therapeutic implications (Romanucci *et al.*, 2006). The diverse role of HSPCA is reflected in its network position and number of interactions. It contains a promiscuous ($p_{\text{promiscuity}} = 4.7 \times 10^{-10}$) ATPase-like domain that facilitates a number of interactions—linking together protein communities involved in signal transduction and apoptosis. Owing to a change in its binding affinity, tumour HSPCA is exclusively found in multi-chaperone complexes with high ATPase activity, whereas HSPCA from normal tissues is in a uncomplexed, latent state (Kamal *et al.*, 2003).

The results presented here provide first insights into the global network properties of cancer proteins and can be used to guide experiments towards regions of the interactome likely to modulate cellular processes involved in cancer. Further studies, however, are required to resolve the evolutionary aspect of these findings fully.

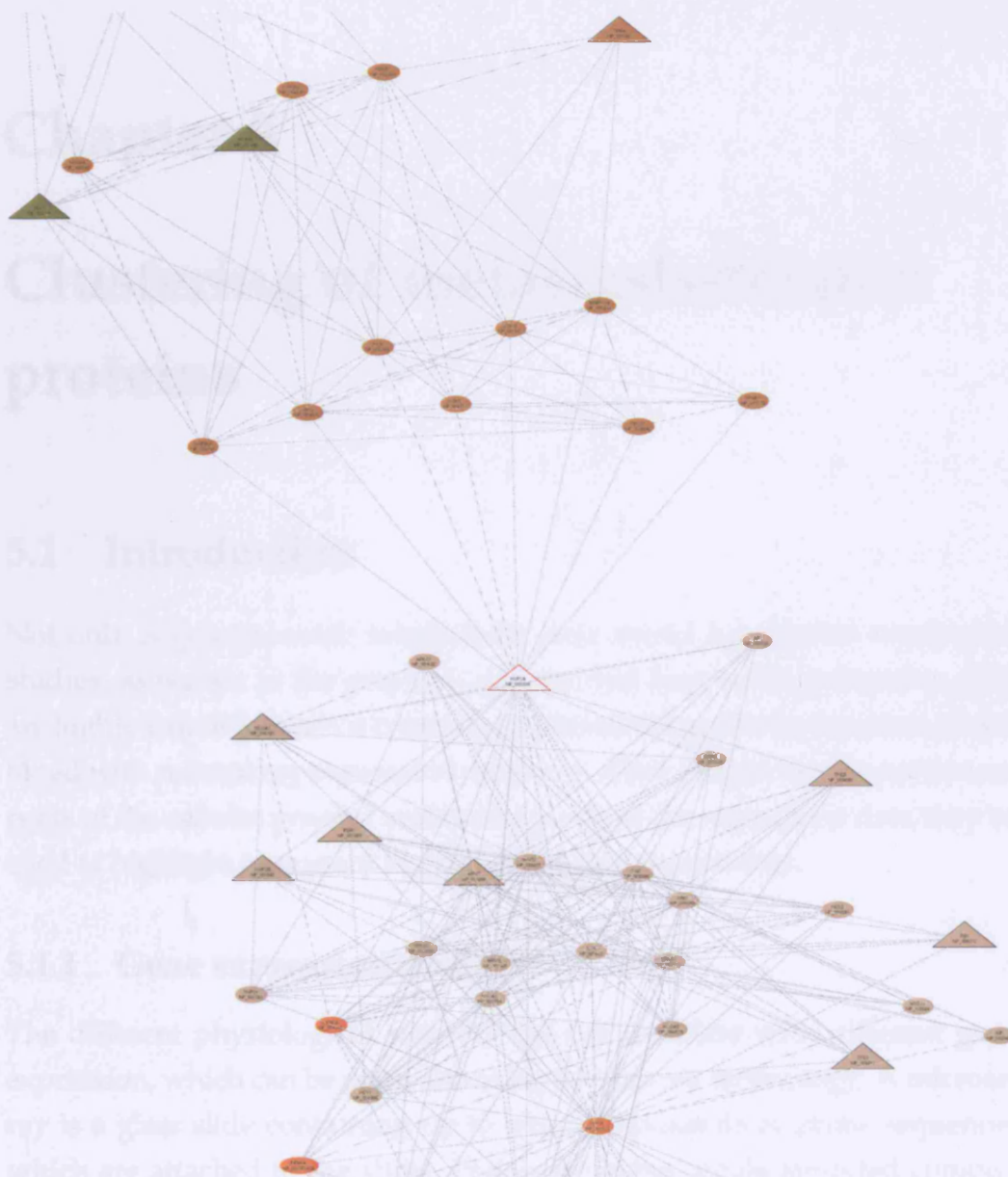


Figure 4.7: A view of a part of the largest network in Figure 4.4, centred around HSPCA (Hsp90) heat shock protein 1 alpha. Each community is coloured in a distinct colour as per the colour scheme in the previous figure, except for HSPCA, which is shown as a white triangle with a red border. Proteins are labelled with gene name and RefSeq identifiers. In addition, cancer proteins are shown as triangles and non-cancer as circles. Refer to Appendix B for the complete community collection.

Chapter 5

Clustering of metastasis-related proteins

5.1 Introduction

Not only is genome-wide interactome data useful for general topological studies, as shown in the previous chapter, but large-scale interaction data are highly valuable when it comes to understanding cellular function. Combined with microarray expression data they allow insight into important aspects of the cellular process under study, where the expression data may be used to highlight important regions of the protein network.

5.1.1 Gene expression analysis

The different physiological states of the cell correlate with different gene expression, which can be analysed using microarray technology. A microarray is a glass slide containing up to tens of thousands of probe sequences which are attached to the slide. Probes are either single-stranded complementary DNA (cDNA) or oligonucleotides ranging from 25–100 base pairs. Target RNA is extracted from the samples of interest, reverse transcribed into cDNA, labelled with fluorescent dye and then hybridised to the array. The fluorescent intensity of a spot indicates the amount of RNA expressed in the sample (see Figure 5.1 for an explanation of the approach). It is therefore possible to identify genes involved in a certain disease by comparing the expression level of an abnormal sample to that of a normal cell line. Tens

of thousands of transcripts can be detected and quantified simultaneously which makes the technology ideal for large scale analyses.

5.1.2 Network analysis of expression data

Expression patterns can be uncovered from large-scale microarray data by systematically grouping genes with the help of clustering methods. Co-clustering of genes can indicate that the genes in question have a similar function or that they participate in the same cellular process (Eisen *et al.*, 1998; Niehrs and Pollet, 1999). Nevertheless, microarray experiments typically yield hundreds of significantly differentially-expressed genes, making it difficult to draw biological conclusions. Furthermore, although microarray experiments can show correlations between co-expressed genes, they do not reveal the exact protein interaction mechanism.

Previous studies have mapped expression data of different systems onto experimentally-based networks. Ideker *et al.* (2002) used gene expression changes in response to perturbation to highlight clusters within a yeast network, and Sohler *et al.* (2004) made use of statistical analysis to highlight significant sub-clusters, also within a yeast network. Moreover, the dynamic aspect of yeast networks has been highlighted by de Lichtenberg and co-workers (de Lichtenberg *et al.*, 2005), who combined temporal cell cycle expression data with protein-protein interaction networks.

Here the multi-genome homology approach described in Chapter 3 was used to construct an interactome for *Rattus norvegicus*, the brown rat. The method, in contrast to the methodologies described above, goes beyond data integration and is therefore able to create a more extensive protein interaction network and has the added benefit of assigning a confidence score to each interaction. The scoring function was further validated and expression data on tumour progression resulting in rat sarcomas with high metastatic potential were subsequently mapped onto the interactome, creating protein networks around key proteins involved in the metastatic process.

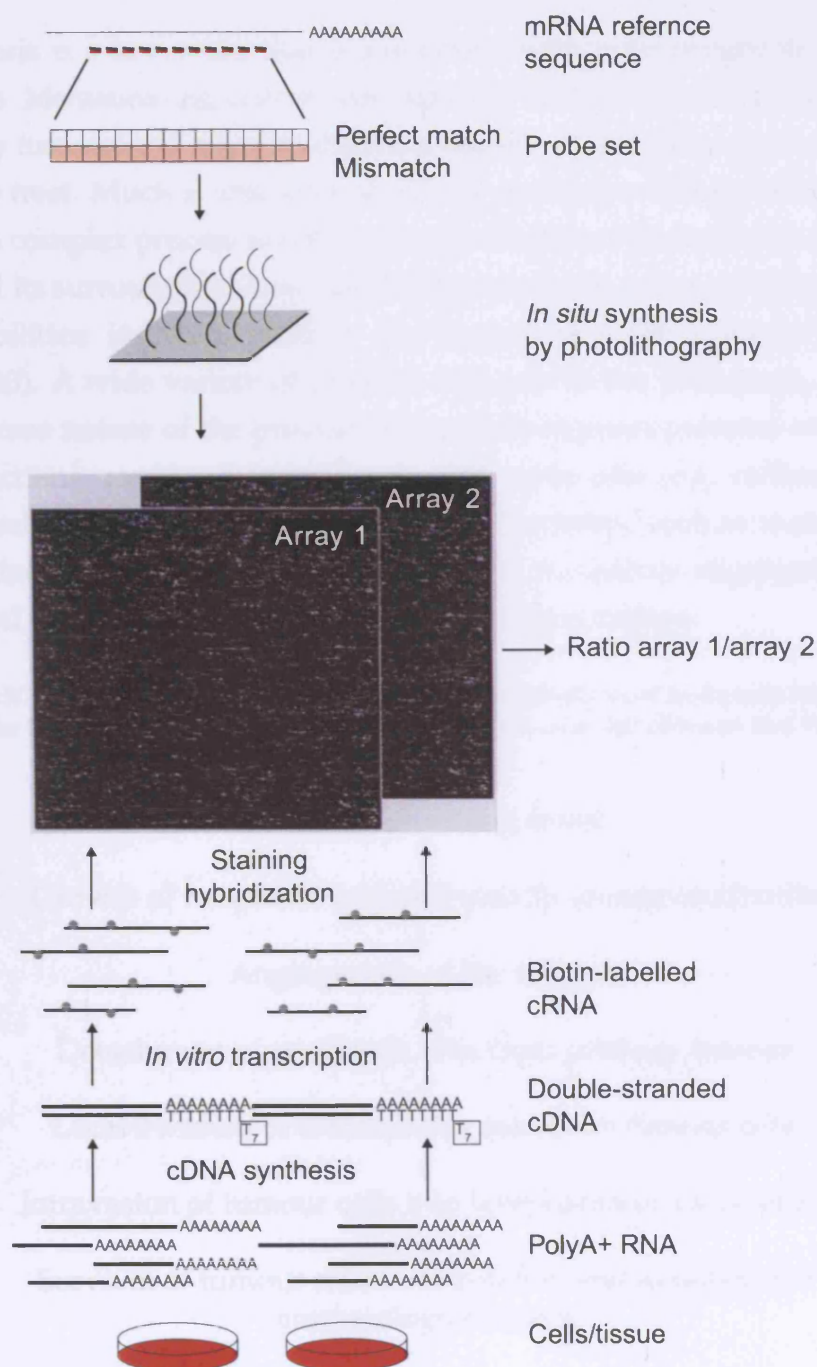


Figure 5.1: A schematic overview of the processes involved in the preparation of high-density oligonucleotide microarrays, such as the Affymetrix microarrays. Sequences of short oligonucleotides (typically 25 bases) are chosen from the mRNA reference sequence of each gene. The sequences are typically selected to represent the most unique part of the transcript. From Schulze and Downward (2001).

5.1.3 Metastasis

Metastasis is a key event that is associated with poor prognosis in cancer patients. Metastasising cancer cells have the ability to break away from the primary tumour and move to different organs, making the cancer more difficult to treat. Much is unknown about the molecular biology of metastasis, as it is a complex process involving series of interactions between the cancer cell and its surroundings that lead to the cancerous cells acquiring two separate abilities: increased motility and invasiveness (Bogenrieder and Herlyn, 2003). A wide variety of proteins take part in the metastasis, reflecting the diverse nature of the process. Metastasis requires proteins with different functions, ranging from cell adhesion molecules (e.g. cadherins, integrins and immunoglobulins) to proteolytic proteins, such as matrix metalloproteinases and serine/cysteine proteases, to various angiogenic factors involved in angiogenesis and growth inducing cytokines.

Table 5.1: The metastatic cascade. Only a very small fraction of malignant tumour cells that enter the bloodstream will produce a tumour at a new site (Ahmad and Hart, 1997).

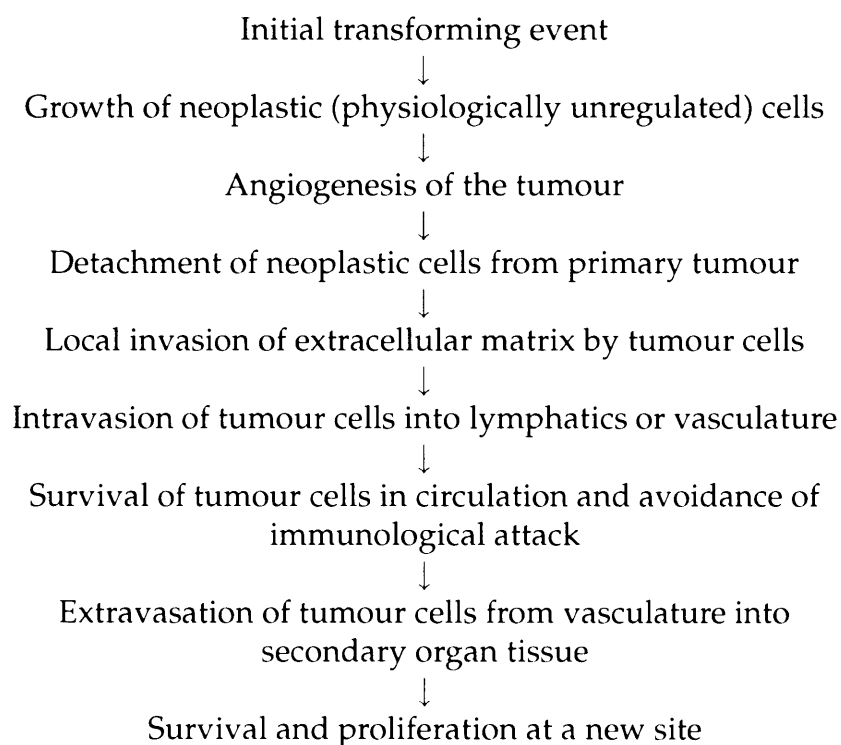


Table 5.1 illustrates the cascading metastatic process. Metastatic cancer

cells travel away from the primary tumour through the blood vessels. In order for metastasis to take place through the blood vessels, cells in the primary tumour must go through several separate steps: Firstly, they need to migrate to a region of capillaries. Then they digest through the basal lamina surrounding the capillary and slip between the endothelial cells to enter the bloodstream; a process known as intravasation. Once in the blood stream, they must attach to endothelial cells and then, by the process of extravasation, move through from the blood vessel through the epithelial cells, digesting the basal lamina and settle in the extracellular matrix.

5.2 Methods

5.2.1 Validation

The rat interactome was created using the procedure described in Chapter 3. As a complementary approach to the ROC-curve validation described there, additional tests were conducted to assess the effectiveness of the scoring function—this time using data relating to the brown rat.

For this purpose, a gold standard data set of transient rat protein complexes from the RCSB Protein Data Bank (Berman *et al.*, 2000) was constructed. The data set was assembled as follows: Protein chains with high sequence homology (sequences detected by BLAST searches with expectation value $< 1 \times 10^{-10}$) to *Rattus norvegicus* were considered. Obligomeric complexes (where multimeric protein chains are permanently bound and essential for the complex function) were distinguished from transient ones (where protein chains may be bound to a complex but may also act as a separate functional protein on its own), by applying the method described in Section 2.4.6. Only transient complexes were included in the gold data set, which was composed of 377 binary chain interactions.

Cellular localisation of proteins was obtained from the Gene Ontology Consortium (Ashburner *et al.*, 2000). Each of the proteins identified by the cluster analysis was placed in a basic cellular localisation class as per Table 5.2. Protein pairs predicted to interact were considered co-localised if they were found in compatible cellular compartments.

Table 5.2: A simplified representation of gene ontology cellular compartments. Protein accessibility between compartments is represented by ones and zeros: the former indicates the possibility of interaction between respective compartments and the latter excludes any interactions.

	Extra-cellular	Intra-cellular	Cyto-plasm	Nuc-leus	Mitochon-drion	Membrane
Extracellular	1	0	0	0	0	1
Intracellular	0	1	1	1	1	1
Cytoplasm	0	1	1	0	0	1
Nucleus	0	1	0	1	0	1
Mitochondrion	0	1	0	0	1	0
Membrane	1	1	1	1	0	1

5.2.2 Microarray expression data for metastatic rat cells

All experimental work regarding gene expression on the metastatic rat cell lines was carried out by Tamara Cavanna in the Microscopy laboratory at Cancer Research UK London Research Institute and she is acknowledged as a contributor to this work.

To investigate genes that may be important in the development of metastases, a rat sarcoma model was used, in which the cell populations K2, T15, A297 and A311 have 0, 40, 90 and 100% incidence of metastasis, respectively. Affymetrix microarray analysis was performed on the four cell populations and the primary tumours that formed when the cells were injected subcutaneously into rats. All experiments were performed in triplicate, using Affymetrix rat 230A GeneChip oligonucleotide arrays (http://www.affymetrix.com/support/technical/datasheets/rat230_datasheet.pdf).

Total RNA was extracted from each sample and used to prepare biotinylated target RNA; 10 μ g of RNA was used to generate first-strand cDNA by using a T7-linked oligo(dT) primer. After second-strand synthesis, in vitro transcription was performed with biotinylated UTP and CTP (Enzo Diagnostics), resulting in approximately 100-fold amplification of RNA. A complete description of the procedures is included in The Paterson Institute's Affymetrix GeneChip systems protocols (http://bioinf.picr.man.ac.uk/mbcf/downloads/GeneChip_Target_Prep_Protocol-CR-UK_v2.pdf).

The target cRNA generated from each sample was processed as per

the manufacturer's recommendation using an Affymetrix GeneChip Instrument System (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). Briefly, spike controls were added to 10 μ g fragmented cDNA before overnight hybridisation, arrays were washed and stained with streptavidin-phycoerythrin, and scanned on an Affymetrix GeneChip scanner. The procedure is further described in The Paterson Institute's RNA Hybridisation protocols (http://bioinf.picr.man.ac.uk/mbcf/downloads/GeneChip_Hyb_Wash_Scan_Protocol-CR-UK_v2.pdf).

The median fluorescence intensity value of each GeneChip was calculated and used to normalise the chips. Gene expression was considered in terms of fold-changes between non-metastatic and the median of the three metastatic samples.

RNA consensus sequences for each oligonucleotide probe were mapped to RefSeq rat protein sequences using Blastx, which converts a nucleotide query sequence into protein sequences in all six reading frames.

5.2.3 Creation of networks around differentially expressed genes

Rat genes that were overexpressed or underexpressed more than four-fold were used as starting points ($n = 100$). Networks were expanded two generations out from the starting points using protein-protein interactions whose S -score value from Equation 3.1 was 10 or higher (see Section 5.3.1 for the basis of the score cut-off selection). The resulting 10,628 interactions were then analysed using k -clique clustering.

Table 5.3 shows the number of individual protein communities for different k -values. Thirty-seven communities were identified for $k = 4$, i.e. setting the subgraph size threshold to a minimum of four. Selecting the k -value is a balancing act; the higher the k -value, the smaller and more internally connected the communities become, but less connection is observed between communities. The k -value was selected after observing that at $k = 4$, reasonably large communities were formed. Proteins which shared sequence identity higher than 40% within each community were merged together such that they appeared as single nodes on the protein map. These merged nodes inherited all the interactions from the individual proteins before the

merging process. This was done to correct for any possible redundancies caused by the homology-based method for predicting protein interactions and there was negligible change in the protein networks as a result of this.

Table 5.3: The number of protein communities vary as the k -threshold value for clustering is changed. The table shows the total number of separate protein communities for each k -value.

k -threshold	Number of communities
3	145
4	37
5	12
6	8
7	2
8	1
9	1
10	1
11	1

5.3 Results and Discussion

Networks of interacting proteins were constructed automatically for the entire rat (*Rattus norvegicus*) genome using the approach described in Chapter 3. The number of individual interactions was reduced from 325,087 to 151,049, when the scoring function was applied to filter out low-quality data, and was further cut down by a clustering method aimed at identifying key interconnected network nodes.

5.3.1 Validation of the scoring function

The scoring function was validated on the human data set in Chapter 3, however here it is further verified, using structural and biological data that is available for the rat system.

Selection of cut-off value for the scoring function

The network construction is based on homology to known interactions and it is therefore imperative to ascertain the minimum level of homology

whereby the structural and functional similarity of the interacting proteins is retained. As mentioned in Chapter 1, pairs of interacting proteins can be considered structurally similar if their sequence identity is no lower than 30% (Aloy *et al.*, 2005). The method here utilises Blast bit scores as a component in the scoring function and so the relationship between bit scores and sequence identity needed to be tested. At the 30% sequence identity level, the bit score ranges linearly from 86–177 (see Figure 5.2) which, according to Equation 3.1, yields minimum interaction scores ranging from 9 to 10. The minimum score for interactions was set at 10 to minimise possibilities of false positive results arising from low homology.

Identification of highly reliable interactions

Many methods for detecting protein-protein interactions can yield either false positive or false negative results, nevertheless X-ray crystal structures of complexed proteins can be considered to be a gold standard for proof. The validity of the scoring of interactions was established by examining the score distribution of proteins in two separate groups: the rat protein interactions in the gold standard set that have either been crystallised together in a complex or have a very high homology to one that has been, and interactions without any crystallographic evidence, i.e. those that do not interact or have not been proved to do so by crystallography.

The interactions in the gold standard data set, identified by X-ray crystallography, were found to score higher than those without crystallographic evidence, with median scores 128 and 16 respectively and mean scores 443 ± 34 for the gold standard set and 364 ± 1 for those without crystallographic evidence (see Figure 5.3). This difference was significant ($p < 2.2 \times 10^{-16}$) according to a Wilcoxon rank sum test, indicating that true interactions score higher than those whose association has not been confirmed

Table 5.4: Interaction score distribution for complexes confirmed by X-ray crystallography ($n = 377$) compared with the scores of all (genome-wide) predicted interactions.

Interaction score, S	X-ray complexes (%)	Genome-wide (%)
0 – 10	6.4	43.2
> 10	93.6	56.8

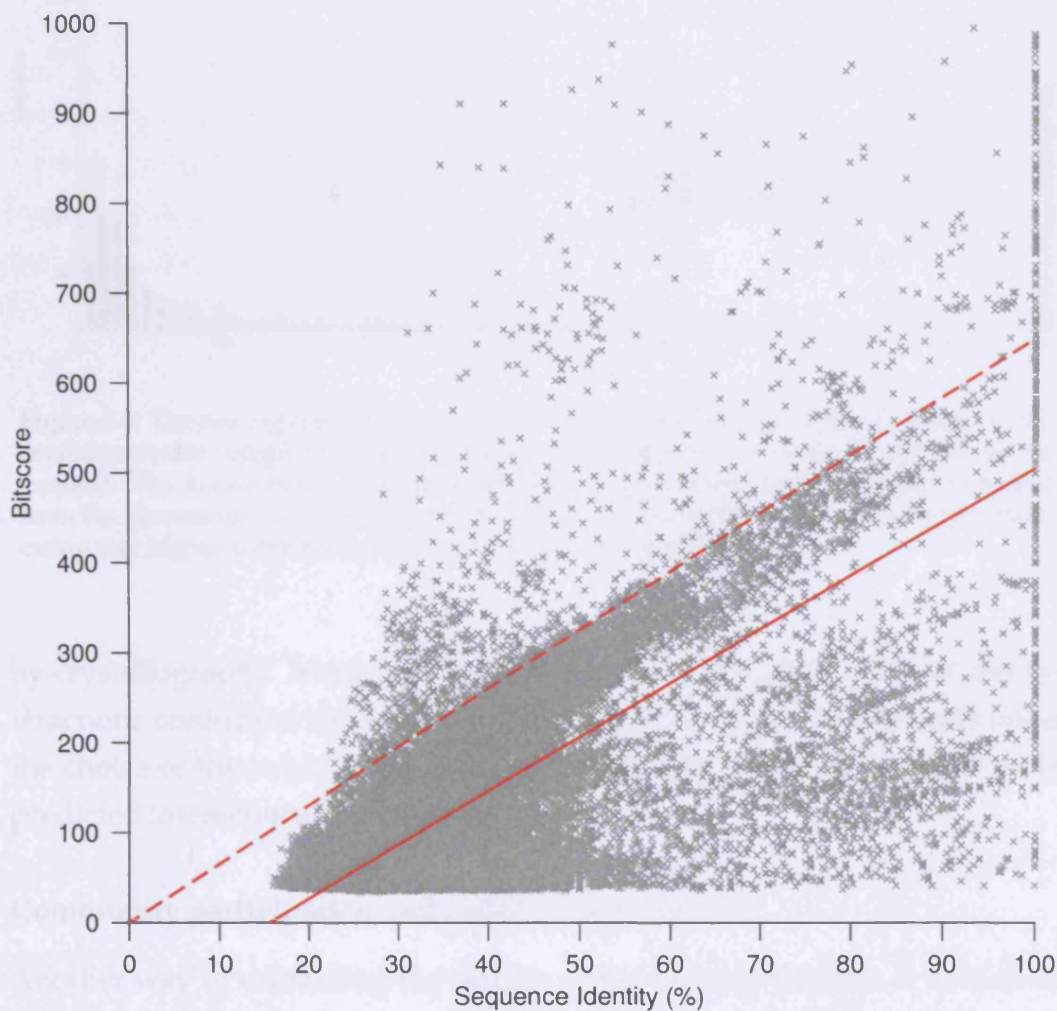


Figure 5.2: The sequence identity and bit score of each hit when proteins in the interaction data were queried against the rat genome. The solid red line shows the best linear fit to the data and shown in dotted red is a line, starting at the origin, which contains 97% of the data in the area below it. Reading from these lines at 30% sequence identity gives bit scores of 86 and 177, respectively, yielding interaction scores of 9 and 10 when inserted into Equation 3.1. To ensure a stringent criterion for the minimum interaction score the higher value was selected as a cut-off score.

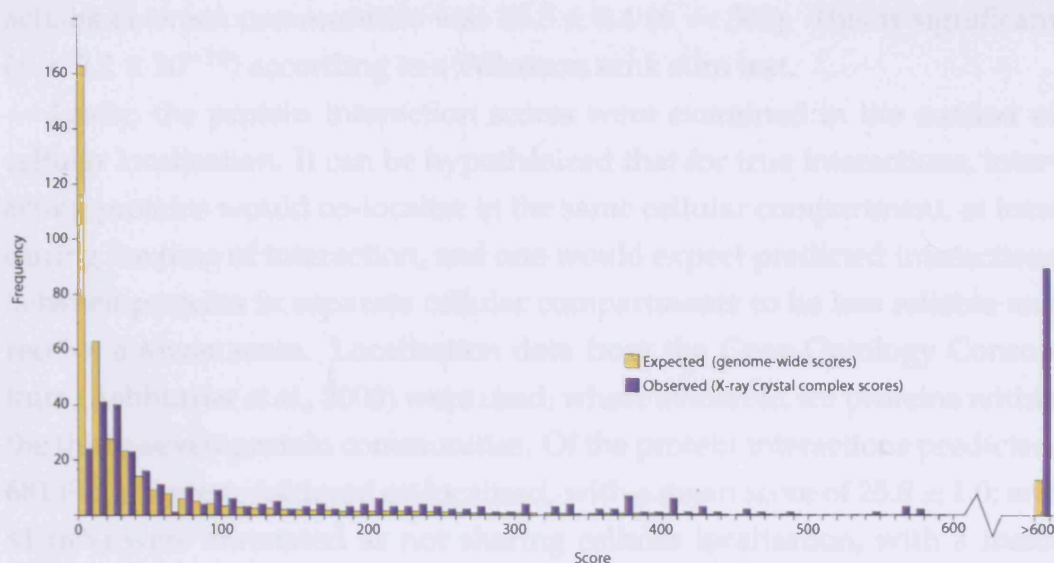


Figure 5.3: The scoring function was assessed by investigating the scores of binary X-ray crystal structure complexes, compared with the distribution of scores for the whole rat genome. The X-ray crystal structures received a higher interaction score than expected from the genome-wide distribution, with median scores of 128 and 16 respectively, indicating that higher-scoring interactions are less likely to be false.

by crystallography. Moreover, as shown in Table 5.4, about 94% of the interactions confirmed by X-ray crystallography score above 10, reaffirming the choice of the cut-off score, whereas just under half of all genome-wide predicted interactions score 10 and lower.

Community participation and cellular localisation

Another way of estimating the quality of the scoring function is to look at proteins participating in the same cellular process and compare them with proteins that are not thought to interact directly in a pathway. The clique percolation method, which was described in the previous chapter, was used to identify communities within the network that show high interconnectivity. This yielded 37 communities of tightly interconnected proteins that will be described later. One can assume that interactions within communities are more likely to be true than interactions between communities, i.e. higher scores would be expected for intra-community interactions (Palla *et al.*, 2005). This was found to be true; the mean score for interactions within a community was 25.2 ± 0.5 ($n = 2038$) and the mean score for inter-

actions between communities was 13.5 ± 0.4 ($n = 502$). This is significant ($p < 2.2 \times 10^{-16}$) according to a Wilcoxon rank sum test.

Lastly, the protein interaction scores were examined in the context of cellular localisation. It can be hypothesized that for true interactions, interacting proteins would co-localise in the same cellular compartment, at least during the time of interaction, and one would expect predicted interactions between proteins in separate cellular compartments to be less reliable and receive a lower score. Localisation data from the Gene Ontology Consortium (Ashburner *et al.*, 2000) were used, where available, for proteins within the thirty-seven protein communities. Of the protein interactions predicted, 681 (94%) were considered co-localised, with a mean score of 25.8 ± 1.0 ; and 41 (6%) were annotated as not sharing cellular localisation, with a mean score of 13.1 ± 1.1 . A Wilcoxon test shows that this score difference is statistically significant ($p = 4.87 \times 10^{-9}$).

Collectively, the results presented in this section further substantiate the validity of the scoring function and indicate that the score cut-off used for the generation of the networks in this study is appropriate.

5.3.2 Identification of metastatic communities

The metastatic process involves a complex network of cascading protein-protein interactions which have to be unravelled if an effective treatment is to be developed. Here an attempt is made to examine these networks by combining expression data with network information. As a starting point, data from a microarray analysis of cell lines with different metastatic potentials was used (see Methods). The highest up- and down-regulated genes (≥ 4 -fold up- or down-expression) were singled out and protein networks constructed around these, extending two generations from the starting point, i.e. initially including proteins that interact directly with the originating protein and then going on to include the proteins that interact with them. This subset of the rat interactome contained 10,628 interactions.

A cluster analysis was then performed to highlight areas in the protein networks that are involved in the metastatic process. The clustering is based on the previously described clique percolation method which distinguishes communities of highly interconnected proteins that make up the essential structural units of the networks. Palla *et al.* (2005) have shown that clique

clustering analysis is a powerful tool to identify communities of proteins participating in the same cellular processes. Furthermore, it has been shown that subnetworks of proteins involved in a defined cellular process are more heavily interconnected by direct protein interaction than would be expected by chance (Gunsalus *et al.*, 2005). Highly connected proteins are also more likely to be essential to cellular processes (Jeong *et al.*, 2001).

The clustering method identified 37 highly interconnected communities, containing 313 proteins involved in 1,094 interactions (Figure 5.4). The majority of the communities have been associated with cancer and metastasis. Some show a degree of overlap and are linked, the most prominent link running through the centre of the figure and containing 17 communities linked in a chain-like manner, however others are not linked, for example, the transcription regulation, which consists of only four proteins. A more detailed description of this graph can be found in Appendix D and an annotated list of all the participants is included in Appendix E.

An initial analysis of the structural- and functional composition of the networks was performed using Domain Fishing (Contreras-Moreira and Bates, 2002), which assigns structural domains to sequences based on homology to known domains. When comparing the domain composition of the communities to domain frequencies of the whole rat genome a bias was observed towards classes of domains found in proteins involved in cytoskeletal structures, cell motility and cell-signalling (see Table 5.5). All but one of the most frequent domains are overrepresented when compared with the genome-wide distribution; only immunoglobulin domains appear less frequently. Spectrin repeat domains, which top the table, are found in proteins involved in cytoskeletal structure, such as spectrin, α -actinin and dystrophin. They are known to bind to calponin homology domains, which are found in both cytoskeletal and signal transduction proteins. The IQ calmodulin-binding domains work as Ca^{2+} switches for myosin which is involved in cell motility and chemotaxis. Furthermore, protein kinase domains, SH2 and SH3 domains and protein-tyrosine phosphatase domains participate in signal transduction and are known to interact. These categories of domains, and associated functions and interactions, are all of interest in the context of cancer metastasis.

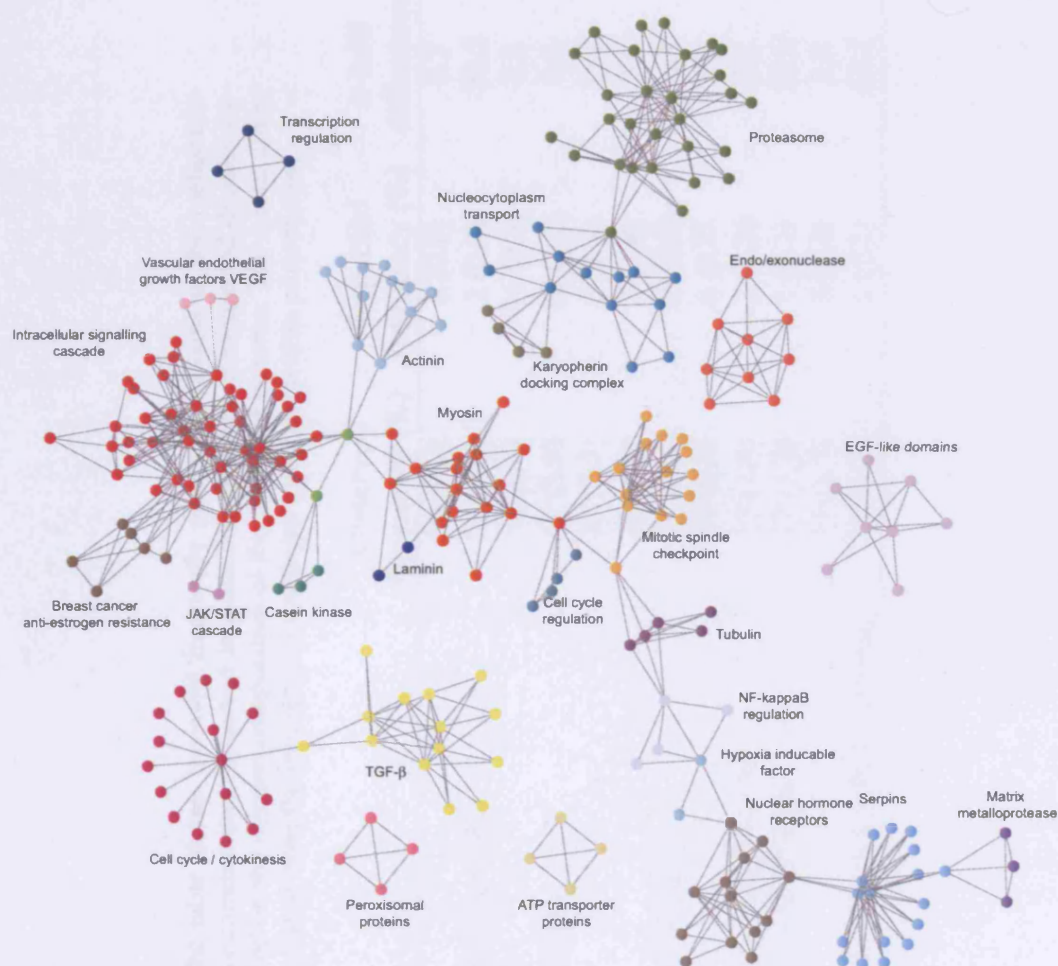


Figure 5.4: The communities identified by k -clique analysis performed on the predicted genome-wide rat protein network. The communities are distinguished by different colours and labelled by the overall function or the dominating protein class. Note that proteins, particularly at community edges, can belong to more than two communities, although this is not shown.

Table 5.5: The table shows the most frequently observed domains in the metastasis-related cluster communities (observed frequencies) alongside the expected domain frequencies, based on the domain composition of the whole rat genome. The n -fold difference was calculated from the frequency percentages (numbers within parentheses).

Domain	Observed frequency (%)	Expected frequency (%)	n -fold difference
Spectrin repeat	56 (6.9)	6 (0.7)	8.3
IQ calmodulin-binding motif	54 (6.6)	2 (0.2)	26.5
EGF-like domain	52 (6.4)	16 (2.0)	2.2
Protein kinase domain	47 (5.8)	12 (1.4)	3.0
SH2 domain	27 (3.3)	2 (0.3)	11.7
EF hand	25 (3.1)	7 (0.8)	2.6
Immunoglobulin domain	21 (2.6)	35 (4.3)	-0.4
SH3 domain	20 (2.4)	6 (0.7)	2.6
Calponin homology (CH) domain	13 (1.6)	2 (0.3)	5.4
Proteasome A-type and B-type	12 (1.5)	1 (0.1)	20.0
LIM domain	11 (1.3)	3 (0.4)	2.7
Transforming growth factor β -like domain	10 (1.2)	1 (0.1)	11.2

The intracellular signalling cascade

It is not the aim here to explore every member of each community—the automatic identification of metastatic-related protein communities is the primary focus. However, the value of the approach will be illustrated by describing a key section of the regulation pathway. The intracellular signalling cascade constitutes the head of a chain of communities (Figure 5.4), and as such warrants a closer investigation.

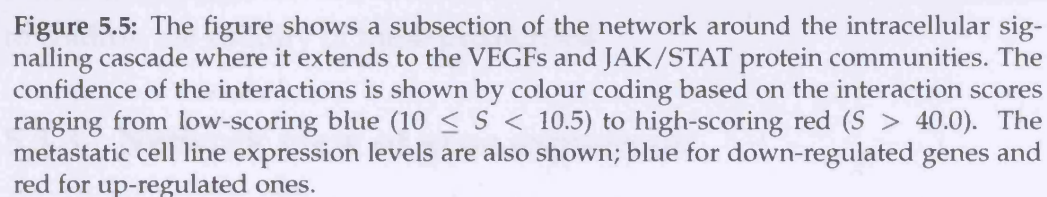
Figure 5.5 shows a detailed view of some of the interactions within that community, focused on the intersection with the vascular endothelial growth factors (VEGFs) and the JAK/STAT group. Many of the interactions in this network have been established either in rat or in other species; others have not been previously demonstrated and it may be proposed that these might have a role in the context of the surrounding proteins.

Three separate groups of proteins are distinguishable: vascular endothelial growth factors (Vegfa, Vegfc, Figf) and the receptor (Kdr), which play a principle role in tumour progression and angiogenesis (Ferrara *et al.*, 2003) and which have also been associated with tumour metastasis (Hirakawa *et al.*, 2005); insulin-like growth factors and receptors (Igf1, Igf1r and Grb 7/14); and JAK/STAT proteins (JAK2, STAT5b).

The figure shows the three ligands, Vegfa and Vegfc and Figf, at different levels of expression, all of which can bind to the kinase insert domain protein receptor Kdr, a VEGF receptor, which in turn induces mitogenesis and differentiation of vascular endothelial cells (Takahashi *et al.*, 1999).

The interaction between Kdr and Socs1, an SH2 domain-containing suppressor of cytokine signalling 1, is plausible as Kdr has a tyrosine protein kinase domain which in a mouse homologue has been shown to interact with Socs1 (Bourette *et al.*, 2001). Furthermore, up-regulation of Socs1 has been linked with the suppression of cytokine signalling and the JAK/STAT inflammatory signalling pathway (Alexander and Hilton, 2004; Park *et al.*, 2003; Ali *et al.*, 2003), which is shown here further down the network; here also, Socs1 is up-regulated and JAK/STAT down-regulated.

The proposed Ptpn11-Lck interaction is based on orthology to an interaction between Ptprc and Lck in mouse. Ptpn11 and Ptprc both have tyrosine specific protein phosphatase activity. Ptpn11 is phosphorylated by tyrosine protein kinases, contains two SH2 domains and therefore could be phos-



phorylated by Lck.

Higher up the network are the insulin-like growth factor 1 and its receptor (Igf1 and Igf1r, respectively) which are highly implicated in different cancers (Furukawa *et al.*, 2005; Hofmann and García-Echeverría, 2005; All-Ericsson *et al.*, 2002). The insulin-like growth factors are involved in several cellular processes, such as regulation of proliferation, migration, survival, size control, and differentiation (LeRoith *et al.*, 1995; Yenush and White, 1997; Massagué and Czech, 1982; Ullrich *et al.*, 1986). Igf1r is overexpressed in most malignant tumours, where it functions as an anti-apoptotic agent by enhancing cell survival. Igf1 has also been shown to enhance adhesion and motility of cancer cells (Dunn *et al.*, 1998; Andre *et al.*, 2004); however, the exact role of Igf1r in the metastatic process has not been established. The network shown here suggests a link between the insulin-like growth factor receptor and the vascular endothelial growth factors through the highly up-regulated phospholipase delta 4 (Plcd4) and phospholipase gamma 1/2 (Plcg 1/2). The Plcg 1/2 and Igf1r interaction is based on the fact that the phospholipase has been shown to interact with an insulin receptor, a close homologue of the insulin-like receptor.

Another distinguishing feature in the network is the highly down-regulated protein tyrosine phosphatase (Ptpn13). It has been reported that a protein tyrosine phosphatase, Ptp61F, negatively regulates the JAK/STAT pathway in *Drosophila melanogaster* (Müller *et al.*, 2005). The networks suggest that the signalling protein tyrosine phosphatase, Ptpn13, may act on the JAK/STAT pathway similarly, through the dephosphorylation of the growth hormone receptor Ghr.

The few examples shown here illustrate the value of the approach in terms of revealing potential pathways and interactions that play a part in cancer metastasis, nevertheless further experimental work will be needed to confirm the validity of these predictions.

5.3.3 Network view of gene expression

Extracting meaningful information from microarray expression data is often difficult, especially when looking at a complex process involving a large number of genes and unknown mechanisms. Clustering of genes may be of use when trying to find genes in a common pathway and genes with related

function, but this often has limitations, such as in identifying negative feedback loops (Armstrong and van de Wiel, 2004). Furthermore, even if key proteins are highlighted through microarray analysis, the expression data rarely reveals all proteins involved in a particular pathway.

Examining the distribution of up- and down-regulated proteins in the context of their neighbours shows that this is indeed the case for the protein networks shown in Figure 5.4. The metastatic expression data was mapped onto the networks and the frequency of pairs of interacting genes was examined, based on the pairwise expression (up-up, down-down and up-/down-regulated pairs). The results, in Table 5.6, indicate that if expression data from the network was randomly redistributed, the probability of observing two up-regulated proteins interacting with each other is about the same as the observed probability. That is, up-regulated proteins do not have a trend of directly interacting with each other, but are interlinked through either neutrally expressed or down-regulated proteins. Moreover, down-regulated proteins are much less likely to interact with each other than expected, demonstrating the benefit of projecting the expression data onto already built networks, as clustering similarly expressed genes and assigning to the same pathway would not be effective.

Table 5.6: Observed and expected frequencies of pairwise protein interactions, categorised by their expression: N-N (non-expressed protein interacting with non-expressed protein), U-U (up-regulated protein interacting with up-regulated protein), D-D (down-regulated protein interacting with down-regulated protein) and U-D (up-regulated interacting with down-regulated). For the purpose of the classification, up-regulated proteins are those up-regulated more than 20% and down-regulated proteins down-regulated more than 20%. Expected values were calculated based on a random distribution of the expression data on the network ($p < 0.001$ for a χ^2 -test).

	Observed	Expected	n -fold difference
N-N	8	5	1.5
U-U	121	109	1.1
D-D	17	41	0.4
U-D	71	67	1.1

5.4 Conclusions

Expression data has previously been put into a network context, for example by incorporating gene ontology data (Jansen *et al.*, 2002a) and protein interactions (Segal *et al.*, 2003), but here the networks were generated first and the expression data merged with the networks before being subjected to clustering analysis. This approach allows bypassing of some of the obstacles involved in traditional microarray analysis, such as clustering of gene expression patterns; as demonstrated here, interactions of up-up and down-down regulated genes are not necessarily co-localised.

Subnetworks around highly up- and down-regulated genes were initially selected to focus on the parts of the genome-wide interaction network relevant to metastasis. The clique method was then used to further highlight hubs of highly interconnected protein communities within the networks. This allows examination of the most complex parts of the network but as a result simple linear pathways do not get included. Although this general approach of combining gene expression with rat interactome data has shown some useful results, there remain some shortcomings. Most importantly, transient protein-protein interactions are unlikely to be captured by the approach. This is a direct consequence of transient not being as well documented as non-transient interactions. Moreover, the method cannot distinguish between true positives and false positives for which there is limited experimental data. These problems will be alleviated as more high-throughput proteomic studies are completed. The system-level approach taken here is a particularly appealing way to gain an understanding of complex biological processes, such as metastasis. Although not discussed here in great detail, several interesting groups of interactions have been highlighted as potentially important players in the metastatic process. Further dissection of these is the subject of ongoing studies and consequently to be confirmed experimentally.

Chapter 6

Concluding remarks

The work presented in this thesis is a study of proteins and their interactions, ranging from the most fundamental aspects—protein structure—through to predicting individual interactions and assembling large networks of interactions in as complete a way as possible.

6.1 Intron-exon boundaries

We employed statistical analysis to investigate the locations of intron-exon boundaries in protein structures with the aim of discovering the structural aspect of IEB location. We also asked whether knowing the location of IEBs may be used as one of the indicators for identifying interaction sites. The analysis revealed an enrichment of IEBs in coiled regions of protein structure as well as at the ends of α -helices and β -strands. This was also confirmed in an analysis of sequence conservation: IEBs were shown to appear with a higher frequency in areas of low sequence conservation, which is often a feature of the loosely structured coil regions. Looking more closely at this phenomenon from a protein-interaction perspective, one would expect the boundaries also to have a tendency to stay away from protein interfaces, which often show signs of increased conservation. This could not be demonstrated for the boundaries of constitutively spliced exons; however we observed that alternatively spliced boundaries tend to steer clear of interface areas. Nevertheless, in the case of the alternatively spliced exons, the observation is based on a very limited data set and it would be prudent to re-examine this issue once more IEB data becomes available. Until then, it

is not possible to say whether IEB information is useful as a parameter for prediction of protein-protein interactions.

6.2 Predicting interactions

In order to predict protein-protein interactions we developed a structure-independent algorithm that is suitable for large-scale predictions. The approach is based on the idea of conservation of interactions through sequence conservation, and as such is suitable for predicting *de novo* interactions in an organism that has not been subject to extensive experimental studies. We improved the previously described interolog-based methods by introducing a simple scoring function, which takes into account the level of homology to the experimentally detected interactions. An additional feature of the scoring function is the evidence weighting: predictions based on repeatedly experimentally-confirmed interactions receive a higher score. Furthermore, the accuracy of the scoring function was confirmed by a ROC-curve analysis. In spite of this, there is scope for improvement. In particular, the scoring could take view of the data source and assign a higher score to predictions based on small core-data that is deemed highly reliable. Furthermore, assigning different weights to the various experimental methods may also improve the accuracy by awarding higher confidence to data based on highly reliable experiments, such as X-ray crystallography, while giving less reliable high-throughput methods a lower rating. The weight assigning, however, raises the question of how to accurately rate the relative quality of one experiment against another.

The uses of a predicted interactome are mainly threefold: firstly, it allows further exploration of known pathways; secondly the predictions can be used as an aid for designing new experiments by indicating likely interactions based on orthologous interactions; and finally, the interactome information is well-suited for integration with additional datasets with the aim of discovering the mechanisms behind a particular disease. This approach, however, is not without its limitations. As the predictions are solely based on experimental data, any gaps in experimental screening will leave gaps in the predicted interactomes and no attempt was made here to estimate the prediction coverage for the three species whose interactomes were

constructed.

6.3 Interactome analysis

The study of interactomes is increasingly providing valuable information on biological systems. We demonstrated the usefulness of the predicted interactomes by conducting two separate studies. We first examined the overall topology of the human interactome and compared the network properties of two groups of proteins: proteins related to cancer and those not documented as being implicated in cancer. We revealed that cancer proteins show an increase in the number of proteins with which they interact. They also appeared to participate in central hubs rather than peripheral ones, mirroring their greater centrality and participation in networks that form the backbone of the proteome. Moreover, we showed that cancer proteins contain a high ratio of highly promiscuous structural domains, that is, domains with a high propensity for mediating protein interactions. These observations may indicate an underlying evolutionary distinction between the two groups of proteins, reflecting the central roles of proteins, whose mutations lead to cancer.

The second study identified key protein communities and potential interactions of proteins likely to be implicated in cancer metastasis. This was done by mapping gene expression data from highly metastatic rat cell-lines onto the rat interactome and subjecting the data to cluster analysis. The cluster analysis revealed distinct, tightly interconnected protein communities that play a role in metastasis. The results indicate that the combination of microarray expression data with protein network information is a powerful way to shed light on biological mechanisms. This is an approach that can be extended to other species in conjunction with expression data relating to different biological states or diseases, although any predicted interactions would have to be confirmed experimentally.

6.4 Future directions

In recent years a great deal of emphasis has been put on studying biological phenomena as a system of complex molecular networks. The success of this

approach has been driven by the development of high-throughput methods that allow screening of a large portion, or even the entire set, of proteins in an organism. This trend will no doubt continue, with the possibility of the human interactome network being studied on a proteome-scale. Large-scale data sets will need to be increasingly integrated with the aim of improving the quality of the data (Vidal, 2005). Additionally, interaction data sets will need to be supplemented by orthogonal experimental and computational approaches to increase the proportion of highly confident interactions (Ge *et al.*, 2003). To this end, experimental methods such as X-ray crystallography, TAP and MS will, without a doubt, provide valuable addition to Y2H data. Improvements in computational methods, such as homology-based modelling and docking may also provide additional means to validate and extend the experimentally determined maps.

Another development, which we may see sooner, is the application of flux and systems approaches to investigate cellular pathways and disease states. This may be achieved by making use of temporal expression data in an attempt to distinguish between active and inactive paths in the network (de Lichtenberg *et al.*, 2005; Zhao *et al.*, 2006), leading to a dynamic network construction in contrast to the static approach taken in this thesis. Moreover, the relative expression levels of neighbouring proteins may prove an important consideration, when protein networks are to be subsequently modulated in conjunction with disease analysis, for example by targeting the expression of a particular gene by short interfering RNA (siRNA) (Karagiannis and El-Osta, 2005). Adding these extra dimensions to interactome data inevitably increases the depth of the analysis, bringing us closer to the situation where we are able to use mathematical models to characterise, simulate and elucidate the mechanisms underlying both normal and abnormal cellular function.

Appendix A

Statistical methods

Much of the analysis in this work relies on statistical tests for interpretation of the significance of the observations. This appendix describes the statistical tests that were used in this work.

The distribution of data can either be parametric, i.e. normally-distributed, or non-parametric, in which case the data is not normally distributed. This needs to be taken into consideration when the statistical method is selected, as some methods assume normal distribution for a correct interpretation.

A.1 Comparative statistics

An important task in data analysis is to compare two or more sets of data to determine whether one set is significantly different from the others or if they are essentially the same, i.e. coming from the same population. The most common statistical tests are the t -test, which assumes normal distribution, and the Wilcoxon rank sum test, which does not require any assumptions in regard to the data distribution.

Student's t -test

The most common comparative statistical test is the t -test, which is used when there are two sets of continuous and normally-distributed data to compare. Each data set is characterised by its mean, standard deviation and number of data points. The t -test is used to support or reject a null hy-

pothesis (H0) that the means of the two normally distributed populations are equal, effectively indicating whether the two sets of samples are equally distributed or not. The test can be either paired (e.g. when the same population is measured twice to compare a measurement of some sort of intervention), or independent (for instance when individuals are randomly assigned into two groups).

The approach for the unpaired case is as follows (Armitage and Berry, 1994). Suppose \bar{x}_1 and \bar{x}_2 are the means of the two samples, whose size is n_1 and n_2 , respectively. The variance, s^2 , of the two samples is first calculated by:

$$s^2 = \frac{\sum_{(1)}(x - \bar{x}_1)^2 + \sum_{(2)}(x - \bar{x}_2)^2}{n_1 + n_2 - 2}. \quad (\text{A.1})$$

The standard error of the difference of the two means, is given by

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, \quad (\text{A.2})$$

and to test the null hypothesis that $\mu_1 = \mu_2$, the t -statistics is calculated by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{SE}(\bar{x}_1 - \bar{x}_2)}. \quad (\text{A.3})$$

Using information on the degrees of freedom ($n_1 + n_2 - 2$), and comparing to a statistical table containing details of the t -distribution, the probability (p) of observing a greater t value can be obtained. This p -value represents the likelihood that the null-hypothesis is wrongly rejected, that is, assuming that the distributions are different when they are not.

Wilcoxon rank sum test

The Wilcoxon rank sum, also known as the Mann-Whitney U test, is non-parametric and is used in place of a two sample t -test when the populations being compared are not normal. It requires independent random samples of sizes n_1 and n_2 . The test is simple and consists of combining the two samples into one sample of size $n_1 + n_2$, sorting the result, assigning ranks to the sorted values and then letting U_1 and U_2 be the sum of the ranks for the observations in the first and second samples, respectively. If the two

populations have the same distribution then the sum of the ranks of the first sample and those in the second sample should be close to the same value. Statistical p -values can then be obtained from Wilcoxon two-sample tables by comparison of the U -values.

A.2 Statistics for frequency data

Frequency data requires a different approach to most other data, as it is not possible to calculate the mean, standard deviation or do a t -test on this kind of data. A common test for this type of data is the χ^2 -test.

χ^2 -test

The χ^2 -test is used to investigate association between frequency data in two separate groups. Each group can have counts in two or more categories and the frequency data are set out in a contingency table. In this thesis it has mainly been used to compare categorical frequencies in a subset of data (the observed values), to the general frequencies in the complete set (the expected values). The test is calculated by

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \quad (\text{A.4})$$

for all observed (O) and expected (E) categories. The null-hypothesis, that states that the observed and expected groups are the same, is then either accepted or rejected based on comparison with a χ^2 -distribution table that gives the p -value.

A.3 Association statistics

A common task in data analysis is to investigate the correlation between variables. Correlation indicates whether the two variables vary together. A positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa. Unrelated variables would then get

a value around zero. The most common tests for correlation are the Pearson's product-moment correlation for parametric data and Spearman's rank correlation for non-parametric data.

Pearson's product-moment correlation

The Pearson product moment correlation coefficient is a dimensionless index that ranges from -1.0 to 1.0, inclusive, and reflects the extent of a linear relationship between two data sets, X and Y , that are represented by n measurements, (x, y) . It is given by

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}, \quad (\text{A.5})$$

and generally speaking, the strength of correlation can be classified according to the following levels:

0.9	to	1.0	very high correlation
0.7	to	0.9	high correlation
0.5	to	0.7	moderate correlation
0.3	to	0.5	weak correlation
0.0	to	0.3	little, if any no correlation

Spearman's rank correlation

The Spearman's rank correlation, ρ_s is more suitable compared to Pearson's if there is any uncertainty regarding the distribution of data, as it does not assume normal distribution and is also suitable for ranked data. It gives as much information as the Pearson correlation coefficient and is valid under a wider range of circumstances (Altman, 1991).

The Spearman's rank correlation coefficient is defined by

$$\rho_s = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}}, \quad (\text{A.6})$$

where x and y are the statistical rank numbers for the two groups of data being tested. Moreover, the significance of the correlation can be tested by

the Spearman t -statistic

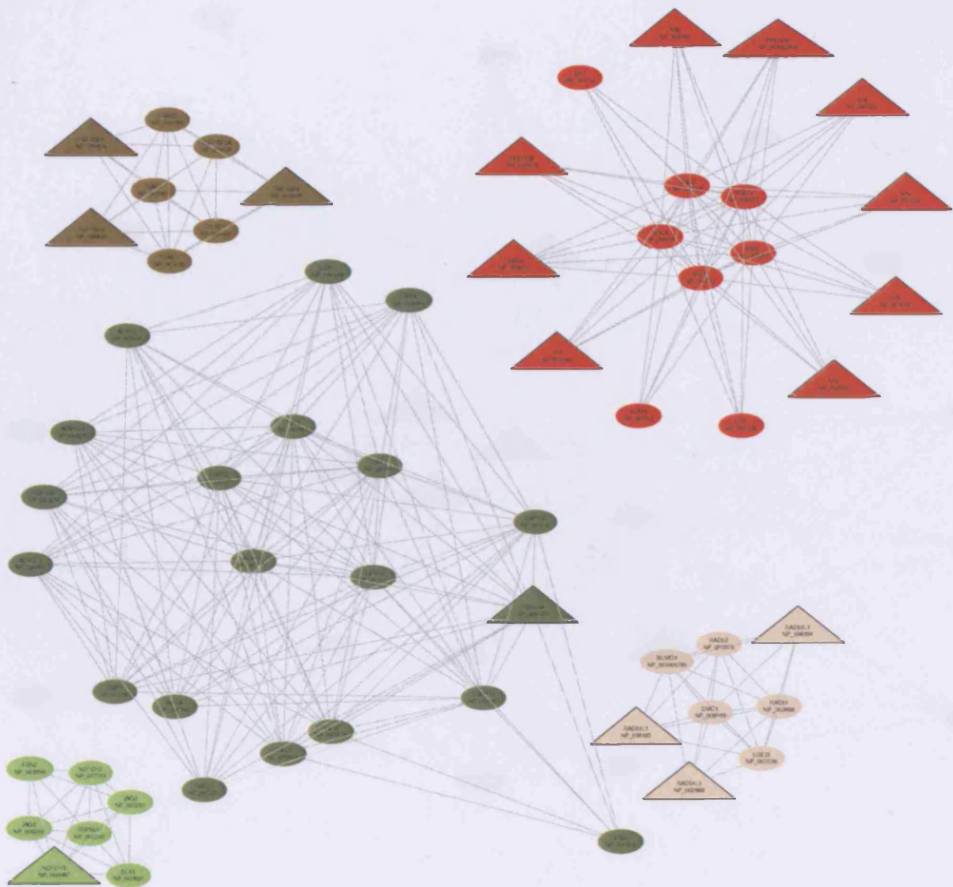
$$t = \sqrt{\frac{n-2}{1-\rho_s^2}}, \quad (\text{A.7})$$

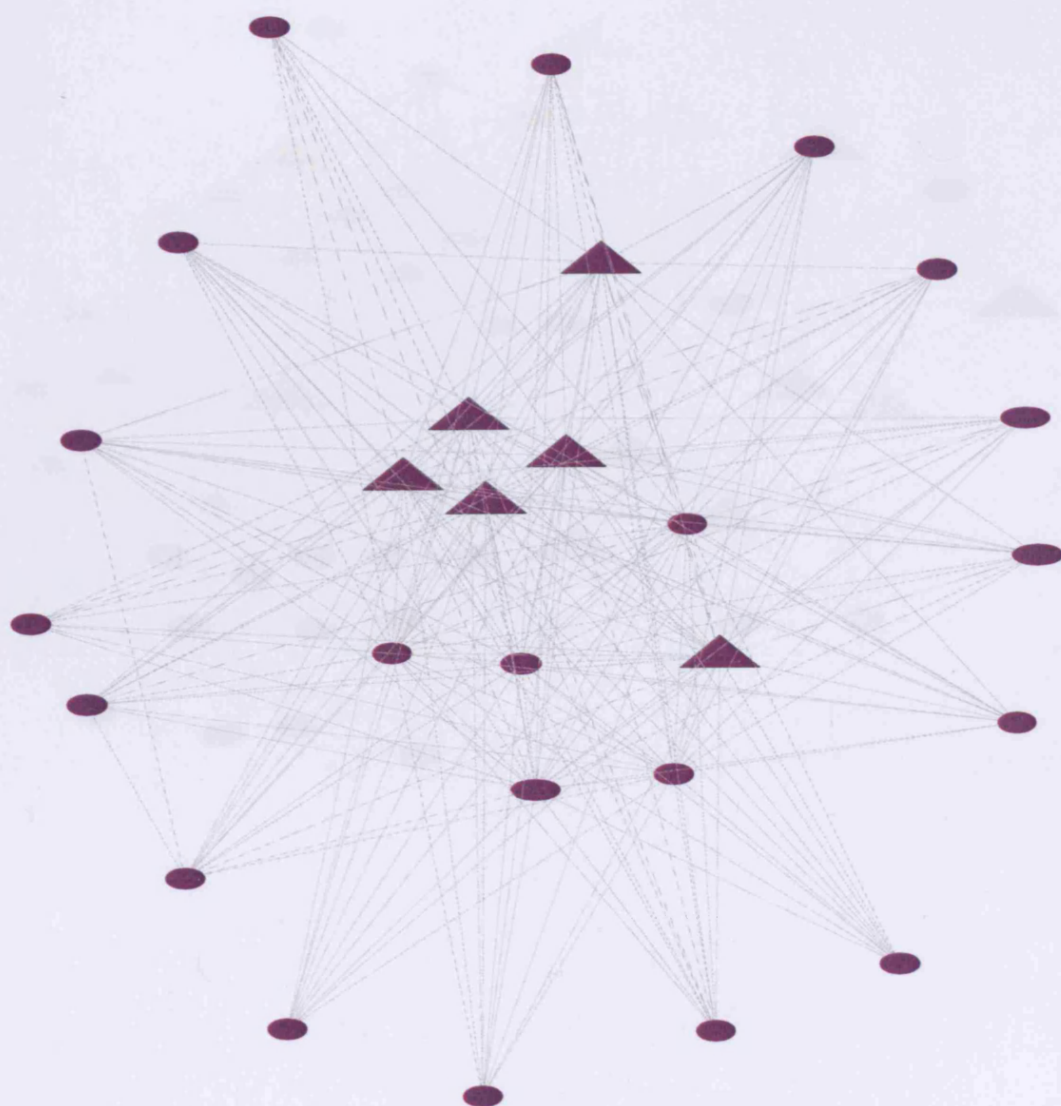
which alongside the degrees of freedom allows the estimation of the p -value.

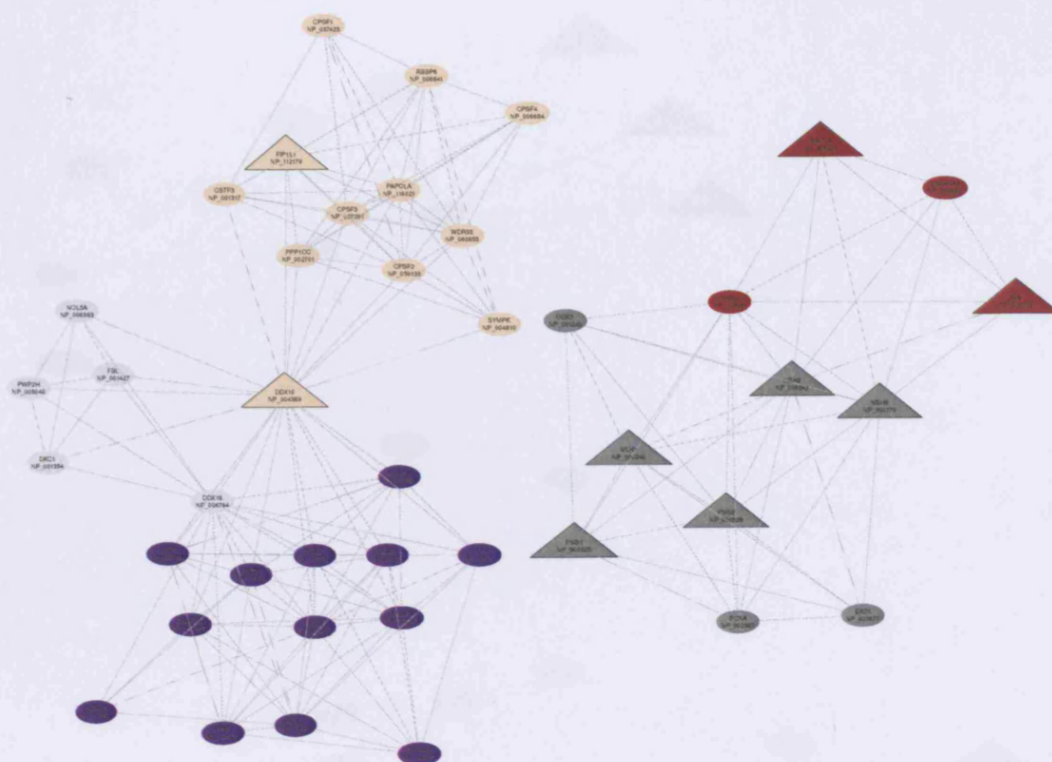
Appendix B

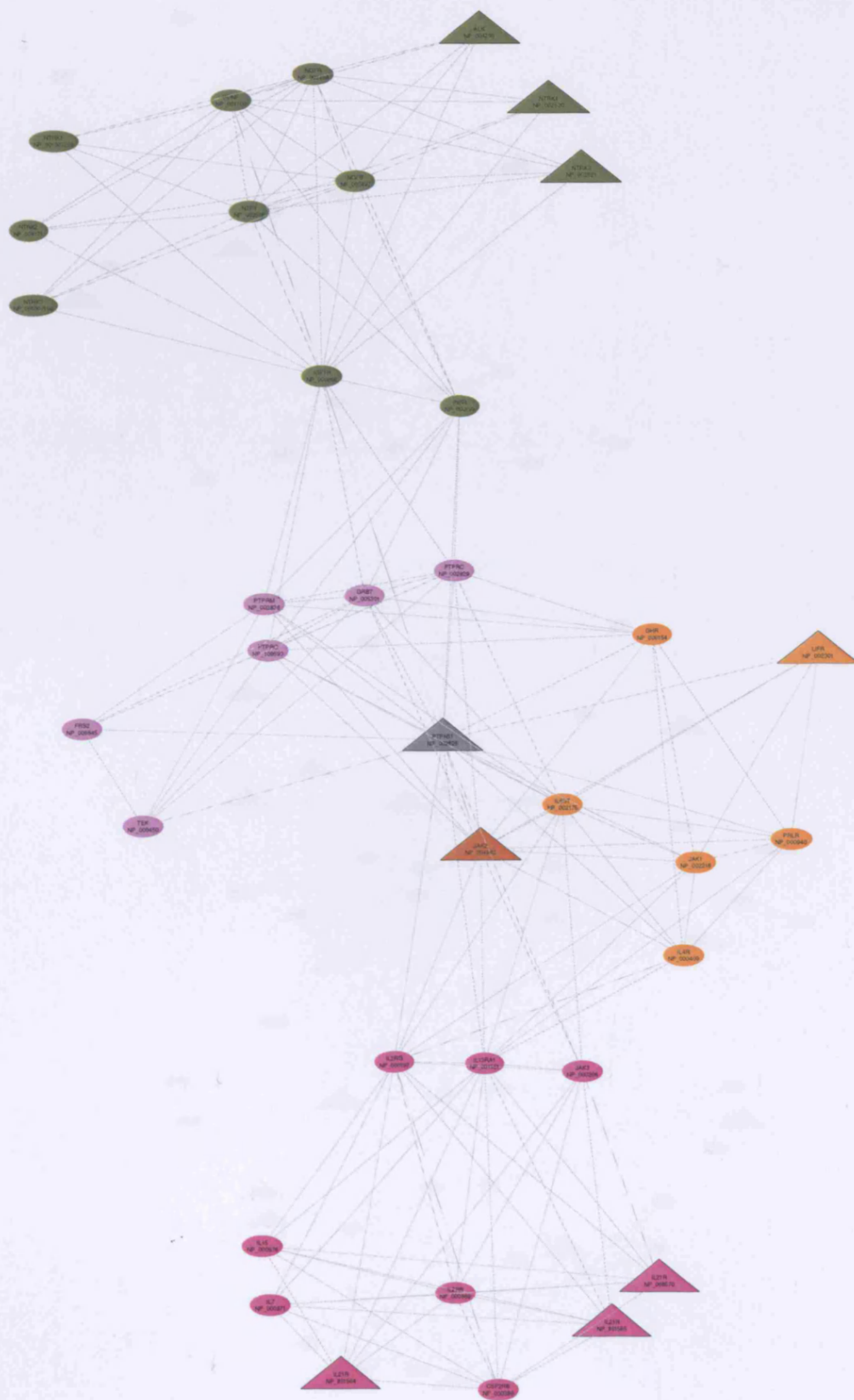
A detailed view of the human cancer communities

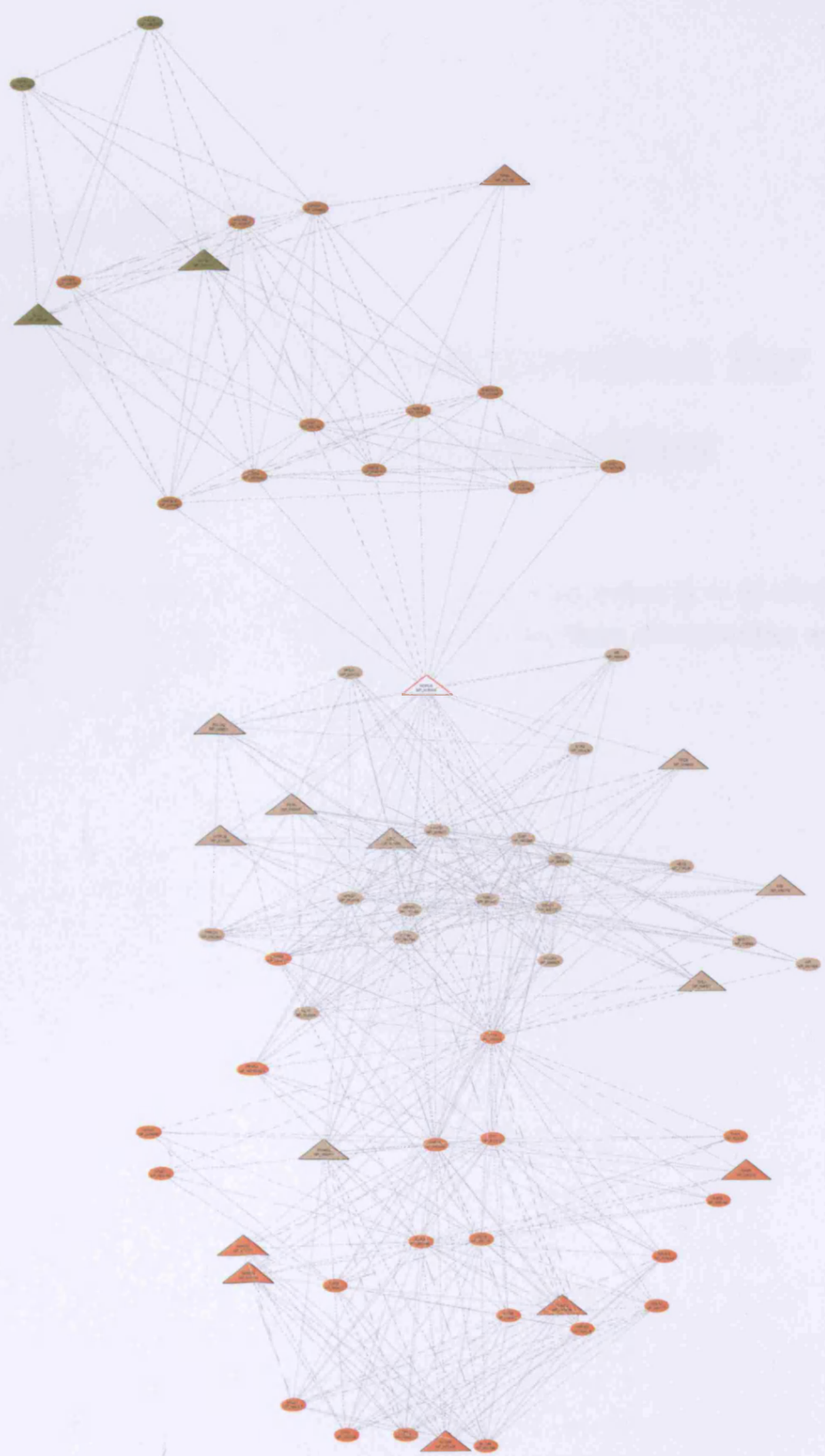
A detailed graphical view of the cancer communities at k -clustering value 6. Each community is coloured in a distinct colour as per the colour scheme in Figure 4.4. Cancer proteins are shown as triangles and non-cancer as circles. Proteins are labelled with gene name and RefSeq identifiers — in case of multiple isoforms, the gene name will appear more than once with each isoform showing a separate RefSeq identifier.











Appendix C

KEGG pathway information for human protein communities

KEGG pathway information for the protein communities ($k = 6$) identified by network clustering. The interaction maps for these communities are detailed in Appendix B.

Community 0 ($k = 6$)

NP_001004426	PLA2G6		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Nervous system	Long-term depression
	METABOLIC	Lipid Metabolism	Glycerophospholipid metabolism
	METABOLIC	Lipid Metabolism	Arachidonic acid metabolism
NP_001547	METABOLIC	Lipid Metabolism	Linoleic acid metabolism
	IKBKB		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
NP_002493	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Insulin signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
	REGULATORY	Metabolic Disorders	Type II diabetes mellitus
NP_002493	NFKB2		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
NP_003630	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
	IKBKG		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
NP_003945	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Immune System	
	REGULATORY	Immune System	
NP_003989	MAP3K14		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	T cell receptor signaling pathway
	NFKB1		
NP_004547	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
NP_005195	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
	NFKBIE		
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_005339	MAP3K8		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	HSPCA		
	REGULATORY	Immune System	Antigen processing and presentation
NP_006590	NFAT5		
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Development	Axon guidance
	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Immune System	T cell receptor signaling pathway
NP_037386	REGULATORY	Immune System	B cell receptor signaling pathway
	TBK1		
	REGULATORY	Immune System	Toll-like receptor signaling pathway
NP_054721	IKBKE		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	NFKBIA		
	REGULATORY	Cell Growth and Death	Apoptosis
NP_065390	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Immune System	Adipocytokine signaling pathway
	REGULATORY	Signal Transduction	

Community 1 ($k = 6$)

NP_002107	HILA-A		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Immune System	Antigen processing and presentation

NP_002110	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_002111	HLA-DOA		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_002112	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_002113	HLA-DQB1		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_002114	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_002115	HLA-DQA1		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_002116	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_005505	HLA-DRB1		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_005506	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_061984	HLA-DRB5		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_064440	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_068818	HLA-B		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_072049	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291032	HLA-DRA		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291033	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291034	HLA-DQA2		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291035	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291036	HLA-DRB4		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291037	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291038	HLA-DRB3		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291039	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291040	HLA-DPA1		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291041	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_291042	HLA-DQB2		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
NP_291043	REGULATORY	Immune System	Antigen processing and presentation
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus

Community 2 (k = 6)

NP_000011	ACVRI1		
NP_000470	REGULATORY	Signal Transduction	IGF-beta signaling pathway
	AMH		
NP_000548	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	IGF-beta signaling pathway
NP_000549	GDF5		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001001557	REGULATORY	Signal Transduction	IGF-beta signaling pathway
	GDF6		
NP_001001558	REGULATORY	Signal Transduction	IGF-beta signaling pathway

NP_001096	ACVR1		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001097	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	ACVR2B		
NP_001191	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001193	BMP2		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001194	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001195	BMP4		
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
NP_001196	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	BMPRII		
NP_001197	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001198	BMPRII		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001199	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	BMPRII		
NP_001200	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001201	ACVR2		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001202	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	BMP6		
NP_001203	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001204	BMP7		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_001205	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_001206	BMP8B		
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
NP_001207	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_002183	INHBA		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_002184	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	INHBB		
NP_003229	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_003230	TGFB2		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_003231	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Cell Growth and Death	Cell cycle
NP_003232	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	TGFB3		
NP_003233	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_003234	REGULATORY	Cell Growth and Death	Cell cycle
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_003235	LEFTY2		
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_004293	ACVR1B		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_004320	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_004321	REGULATORY	Cell Communication	Adherens junction
	BMPRIA		
NP_004322	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_004603	TGFBRI		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_005529	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_060525	INHBC		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_060526	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	NODAL		
NP_065434	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	AMHR2		
NP_066277	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_066278	LEFTY1		
	REGULATORY	Signal Transduction	TGF-beta signaling pathway

NP_066551	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	BMP5		
NP_113667	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_660302	INHBE	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
NP_861525	ACVR1C		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_861525	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	REGULATORY	Cell Communication	Adherens junction
XP_208545	BMP8A		
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
XP_208545	LOC283155	Signal Transduction	TGF-beta signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
Community 3 ($k = 6$)			
NP_000205	JAG1		
NP_002217	REGULATORY	Signal Transduction	Notch signaling pathway
	JAG2		
NP_005609	REGULATORY	Signal Transduction	Notch signaling pathway
	DLL1		
NP_058637	REGULATORY	Signal Transduction	Notch signaling pathway
	DLL3		
NP_077719	REGULATORY	Signal Transduction	Notch signaling pathway
	NOTCH2		
NP_077719	REGULATORY	Signal Transduction	Notch signaling pathway
Community 4 ($k = 6$)			
NP_002070	GOT1		
NP_002070	METABOLIC	Amino Acid Metabolism	Glutamate metabolism
	METABOLIC	Amino Acid Metabolism	Alanine and aspartate metabolism
NP_002070	METABOLIC	Amino Acid Metabolism	Cysteine metabolism
	METABOLIC	Amino Acid Metabolism	Arginine and proline metabolism
NP_002070	METABOLIC	Amino Acid Metabolism	Tyrosine metabolism
	METABOLIC	Amino Acid Metabolism	Phenylalanine metabolism
NP_002070	METABOLIC	Amino Acid Metabolism	Phenylalanine, tyrosine and tryptophan biosynthesis
	METABOLIC	Biosynthesis of Secondary Metabolites	Novobiocin biosynthesis
NP_002070	METABOLIC	Energy Metabolism	Carbon fixation
	METABOLIC	Biosynthesis of Secondary Metabolites	Alkaloid biosynthesis I
Community 5 ($k = 6$)			
NP_000810	GART		
NP_000810	METABOLIC	Nucleotide Metabolism	Purine metabolism
	METABOLIC	Metabolism of Cofactors and Vitamins	One carbon pool by folate
NP_002583	PCNA		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_031396	RFC5		
	METABOLIC	Nucleotide Metabolism	Purine metabolism
NP_031396	METABOLIC	Nucleotide Metabolism	Pyrimidine metabolism
Community 6 ($k = 6$)			
NP_002701	PPP1CC		
NP_002701	REGULATORY	Cell Communication	Focal adhesion
	REGULATORY	Nervous system	Long-term potentiation
NP_002701	REGULATORY	Cell Motility	Regulation of actin cytoskeleton
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_004810	SYMPK		
	REGULATORY	Cell Communication	Tight junction
Community 8 ($k = 6$)			
NP_000391	FRCC2		
NP_000391	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis
NP_002583	PCNA		
	REGULATORY	Cell Growth and Death	Cell cycle
Community 9 ($k = 6$)			
NP_037466	GMPPB		
NP_037466	METABOLIC	Carbohydrate Metabolism	Fructose and mannose metabolism
	GMPPA		

	METABOLIC	Carbohydrate Metabolism	Fructose and mannose metabolism
Community 10 ($k = 6$)			
NP_000312	RB1		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_000448	HNF4A		
	REGULATORY	Metabolic Disorders	Maturity onset diabetes of the young
NP_001169	ARNTL		
	REGULATORY	Behavior	Circadian rhythm
NP_002509	NPAS2		
	REGULATORY	Behavior	Circadian rhythm
NP_002607	PER1		
	REGULATORY	Behavior	Circadian rhythm
NP_004889	CLOCK		
	REGULATORY	Behavior	Circadian rhythm
NP_005339	HSPCA		
	REGULATORY	Immune System	Antigen processing and presentation
NP_058515	PER3		
	REGULATORY	Behavior	Circadian rhythm
NP_064568	ARNTL2		
	REGULATORY	Behavior	Circadian rhythm
NP_612482	SP1		
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
Community 11 ($k = 6$)			
Community 12 ($k = 6$)			
NP_001245	CDC6		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_002379	MCM3		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_002543	ORC4L		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_002544	ORC5L		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_003494	CDC7		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_003495	CDC45L		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_004144	ORC1L		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_004517	MCM2		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_005905	MCM4		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_005906	MCM6		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_005907	MCM7		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_006181	ORC2L		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_006730	MCM5		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_055136	ORC6L		
	REGULATORY	Cell Growth and Death	Cell cycle
Community 13 ($k = 6$)			
NP_001550	IL12RB2		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_001833	CNTFR		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_002175	IL6ST		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_002218	JAK1		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_002825	PTPN11		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity

NP_004503	REGULATORY	Immune System	Leukocyte transendothelial migration
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_004963	IL11RA	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	JAK2	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway

Community 14 ($k = 6$)

NP_001009552	PPP2CB		
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	REGULATORY	Cell Communication	Tight junction
	REGULATORY	Nervous system	Long-term depression

Community 15 ($k = 6$)

NP_001009552	PPP2CB		
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	TGF-beta signaling pathway
	REGULATORY	Cell Communication	Tight junction
	REGULATORY	Nervous system	Long-term depression

Community 18 ($k = 6$)

NP_005709	ARPC4		
	REGULATORY	Cell Motility	Regulation of actin cytoskeleton
NP_005711	ARPC1B		
	REGULATORY	Cell Motility	Regulation of actin cytoskeleton
NP_006400	ARPC1A		
	REGULATORY	Cell Motility	Regulation of actin cytoskeleton

Community 19 ($k = 6$)

NP_000167	NR3C1		
	REGULATORY	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction
NP_000448	HNF4A		
	REGULATORY	Metabolic Disorders	Maturity onset diabetes of the young
NP_000452	THRB		
	REGULATORY	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction
NP_002126	NR4A1		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002948	RXRA		
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_004124	HNF4G		
	REGULATORY	Metabolic Disorders	Maturity onset diabetes of the young
NP_068811	RXRβ		
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_612482	SP1		
	REGULATORY	Signal Transduction	TGF-beta signaling pathway

Community 20 ($k = 6$)

NP_001887	CSNK2A2		
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Cell Communication	Tight junction
NP_006764	DDX18		
	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
NP_076977	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis
	DDX54		
	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis

Community 23 ($k = 6$)

NP_001310	CSNK1G2		
	REGULATORY	Signal Transduction	Phosphatidylinositol signaling system
	METABOLIC	Carbohydrate Metabolism	Inositol phosphate metabolism
	METABOLIC	Xenobiotics Biodegradation and Metabolism	Benzoate degradation via CoA ligation
NP_002721	METABOLIC	Metabolism of Cofactors and Vitamins	Nicotinate and nicotinamide metabolism
	PRKACA		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	Calcium signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Signal Transduction	Wnt signaling pathway

NP_002722	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term potentiation
	REGULATORY	Signal Transduction	Insulin signaling pathway
	PRKACB		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	Calcium signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
NP_002723	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term potentiation
	REGULATORY	Signal Transduction	Insulin signaling pathway
	PRKACG		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	Calcium signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Cell Communication	Gap junction
NP_002751	REGULATORY	Nervous system	Long-term potentiation
	REGULATORY	Signal Transduction	Insulin signaling pathway
	PRKY		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	Calcium signaling pathway
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term potentiation
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_004148	PRKAR2A		
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_005035	PRKX		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	Calcium signaling pathway
	REGULATORY	Signal Transduction	Wnt signaling pathway
	REGULATORY	Signal Transduction	Hedgehog signaling pathway
	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term potentiation
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_006249	PRKG1		
	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term depression
NP_006250	PRKG2		
	REGULATORY	Cell Communication	Gap junction
	REGULATORY	Nervous system	Long-term depression
XP_496112	LOC440332		
	REGULATORY	Cell Communication	Focal adhesion
	REGULATORY	Signal Transduction	Insulin signaling pathway

Community 24 ($k = 6$)

NP_000154	GHR		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	REGULATORY	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction
NP_002175	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	IL6ST		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_002823	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	PTPN7		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002825	PTPN11		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Immune System	Leukocyte transendothelial migration
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_002828	PTPRB		
	REGULATORY	Cell Communication	Adherens junction
NP_002829	PTPRC		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Immune System	T cell receptor signaling pathway
NP_002831	PTPRF		

	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_002834	PTPRJ		
	REGULATORY	Cell Communication	Adherens junction
NP_002836	PTPRM		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
NP_002838	PTPRN2		
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
NP_004963	JAK2		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
Community 25 ($k = 6$)			
NP_000199	INSR		
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Signal Transduction	Insulin signaling pathway
	REGULATORY	Metabolic Disorders	Type II diabetes mellitus
	REGULATORY	Neurodegenerative Disorders	Dentatorubropallidoluysian atrophy (DRPLA)
NP_000866	IGFIR		
	REGULATORY	Cell Communication	Focal adhesion
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Nervous system	Long-term depression
NP_002823	PTPN7		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002825	PTPN11		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Immune System	Leukocyte transendothelial migration
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_002828	PTPRB		
	REGULATORY	Cell Communication	Adherens junction
NP_002829	PTPRC		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Immune System	T cell receptor signaling pathway
NP_002831	PTPRF		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_002834	PTPRJ		
	REGULATORY	Cell Communication	Adherens junction
NP_002836	PTPRM		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
NP_002838	PTPRN2		
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus
Community 26 ($k = 6$)			
NP_002823	PTPN7		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002825	PTPN11		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Immune System	Leukocyte transendothelial migration
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_002828	PTPRB		
	REGULATORY	Cell Communication	Adherens junction
NP_002829	PTPRC		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Immune System	T cell receptor signaling pathway
NP_002831	PTPRF		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
	REGULATORY	Signal Transduction	Insulin signaling pathway
NP_002834	PTPRJ		
	REGULATORY	Cell Communication	Adherens junction
NP_002836	PTPRM		
	REGULATORY	Signaling Molecules and Interaction	Cell adhesion molecules (CAMs)
	REGULATORY	Cell Communication	Adherens junction
NP_002838	PTPRN2		
	REGULATORY	Metabolic Disorders	Type I diabetes mellitus

Community 27 ($k = 6$)

NP_000391	ERCC2		
	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
NP_002740	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis
	MAPK7		
NP_004126	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Communication	Gap junction
NP_005521	IDH3G		
	METABOLIC	Carbohydrate Metabolism	Citrate cycle (TCA cycle)
NP_006494	IDH3A		
	METABOLIC	Carbohydrate Metabolism	Citrate cycle (TCA cycle)
NP_008830	PSMC4		
	IDH3B		
	METABOLIC	Carbohydrate Metabolism	Citrate cycle (TCA cycle)

Community 28 ($k = 6$)

NP_000154	GHR		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_000940	REGULATORY	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_002175	PRLR		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_002218	REGULATORY	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_002825	IL6ST		
	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
NP_004963	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	JAK1		
NP_004963	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_004963	PTPN11		
	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
NP_004963	REGULATORY	Immune System	Natural killer cell mediated cytotoxicity
	REGULATORY	Immune System	Leukocyte transendothelial migration
NP_004963	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
	JAK2		
NP_004963	REGULATORY	Signal Transduction	Jak-STAT signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway

Community 29 ($k = 6$)

NP_000199	INSR		
	REGULATORY	Cell Communication	Adherens junction
NP_000866	REGULATORY	Signal Transduction	Insulin signaling pathway
	REGULATORY	Metabolic Disorders	Type II diabetes mellitus
NP_001007793	REGULATORY	Neurodegenerative Disorders	Dentatorubropallidoluysian atrophy (DRPLA)
	IGF1R		
NP_001700	REGULATORY	Cell Communication	Focal adhesion
	REGULATORY	Cell Communication	Adherens junction
NP_002497	REGULATORY	Nervous system	Long-term depression
	NTRK1		
NP_002497	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
NP_002497	BDNF		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002497	REGULATORY	Neurodegenerative Disorders	Huntington's disease
NP_002497	NGFB		
	REGULATORY	Signal Transduction	MAPK signaling pathway
NP_002498	REGULATORY	Cell Growth and Death	Apoptosis
	NGFR		
NP_002518	REGULATORY	Signaling Molecules and Interaction	Cytokine-cytokine receptor interaction
	NTF3		
NP_002518	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Signal Transduction	MAPK signaling pathway

Community 30 ($k = 6$)

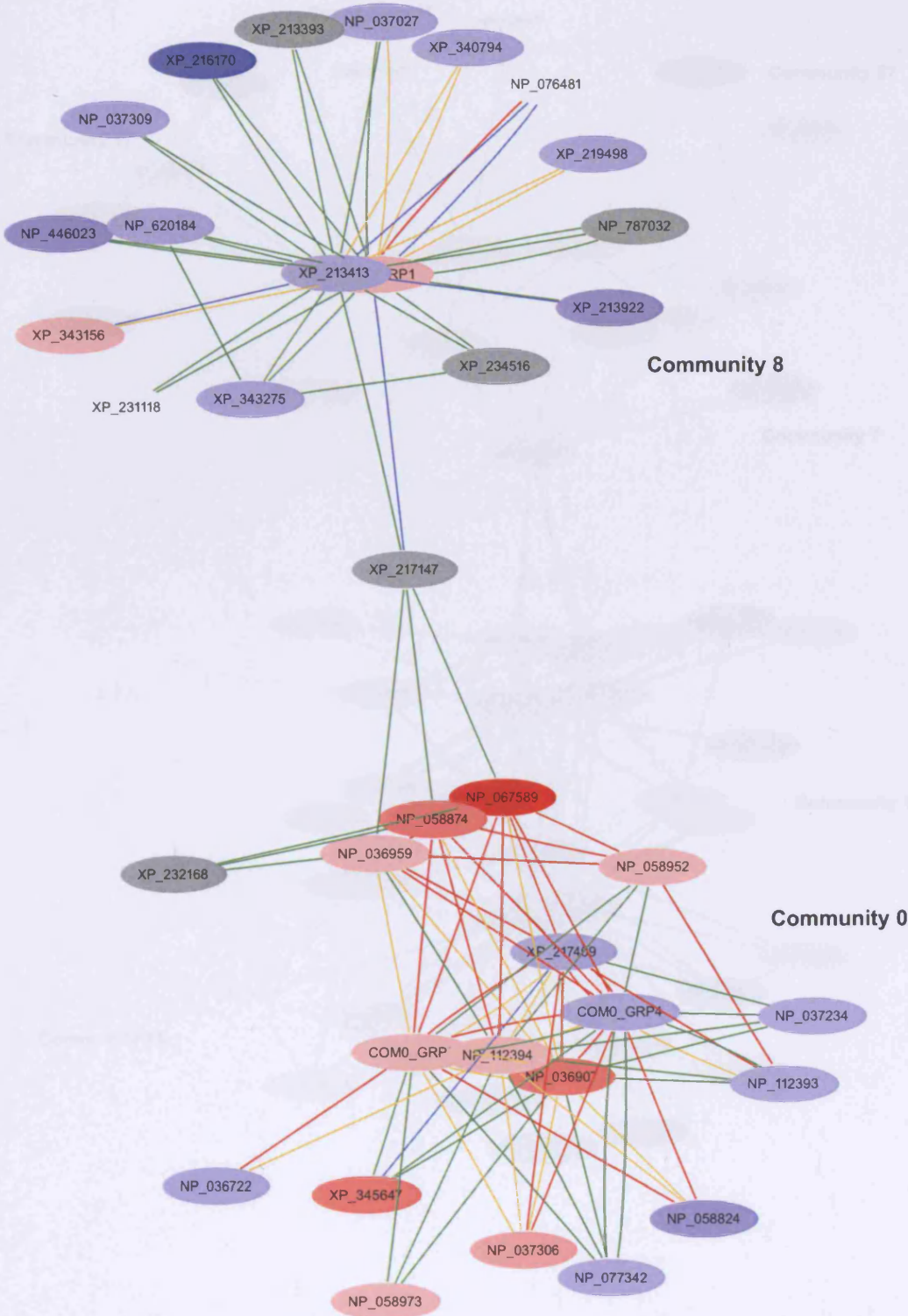
NP_003061	SMARCA2		
	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
NP_003592	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis
	SMARCA5		
NP_003592	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis

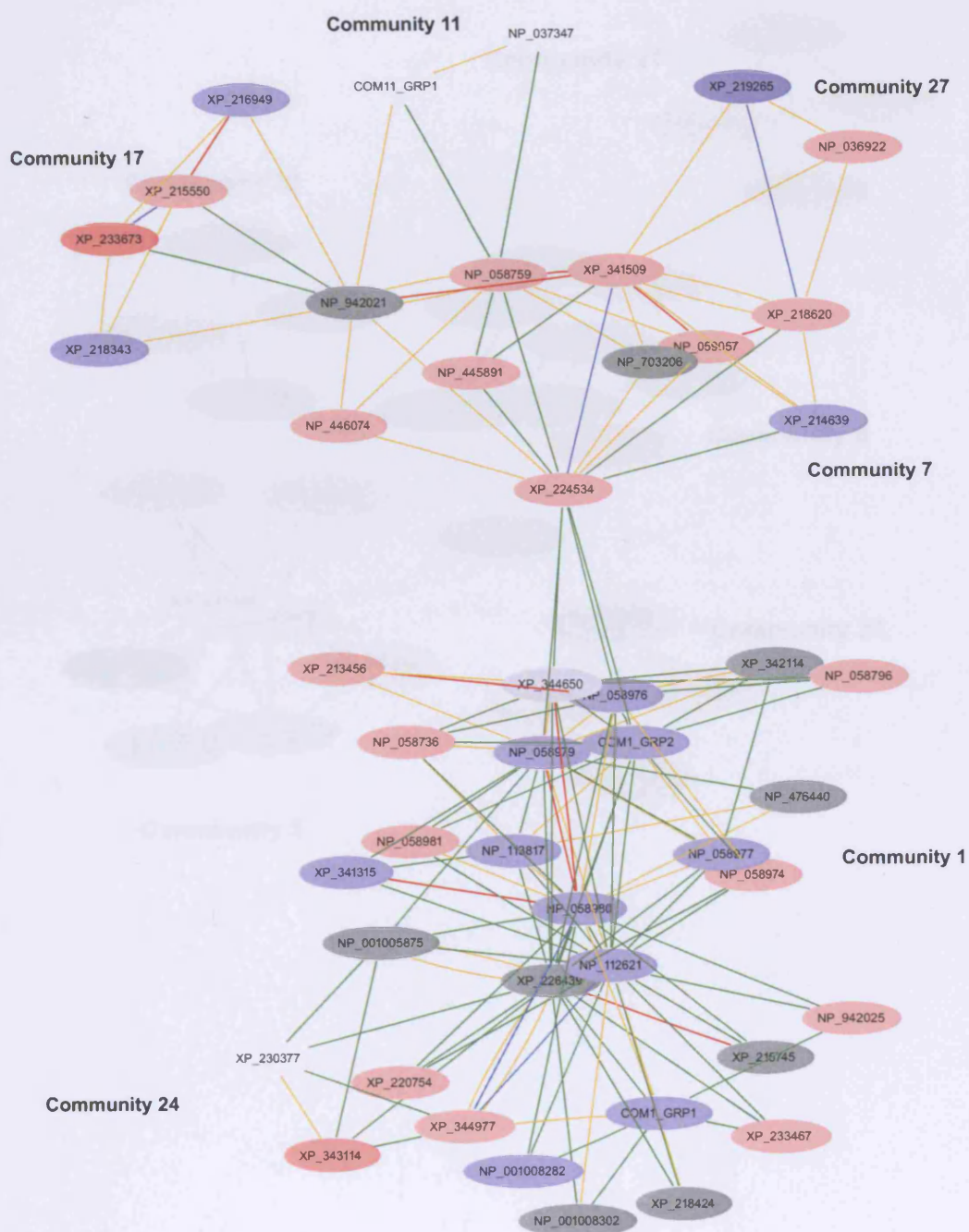
Community 33 ($k = 6$)			
NP_006297	SMC1L1		
	REGULATORY	Cell Growth and Death	Cell cycle
NP_683515	SMC1L2		
	REGULATORY	Cell Growth and Death	Cell cycle
Community 34 ($k = 6$)			
NP_006764	DDX18		
	METABOLIC	Carbohydrate Metabolism	Starch and sucrose metabolism
	METABOLIC	Metabolism of Cofactors and Vitamins	Folate biosynthesis
Community 35 ($k = 6$)			
NP_001896	CTPS		
	METABOLIC	Nucleotide Metabolism	Pyrimidine metabolism
NP_062831	CTPS2		
	METABOLIC	Nucleotide Metabolism	Pyrimidine metabolism
Community 36 ($k = 6$)			
NP_002493	NFKB2		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_003989	NFKB1		
	REGULATORY	Signal Transduction	MAPK signaling pathway
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_004547	NFKBIE		
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway
NP_065390	NFKBIA		
	REGULATORY	Cell Growth and Death	Apoptosis
	REGULATORY	Immune System	Toll-like receptor signaling pathway
	REGULATORY	Immune System	T cell receptor signaling pathway
	REGULATORY	Immune System	B cell receptor signaling pathway
	REGULATORY	Signal Transduction	Adipocytokine signaling pathway

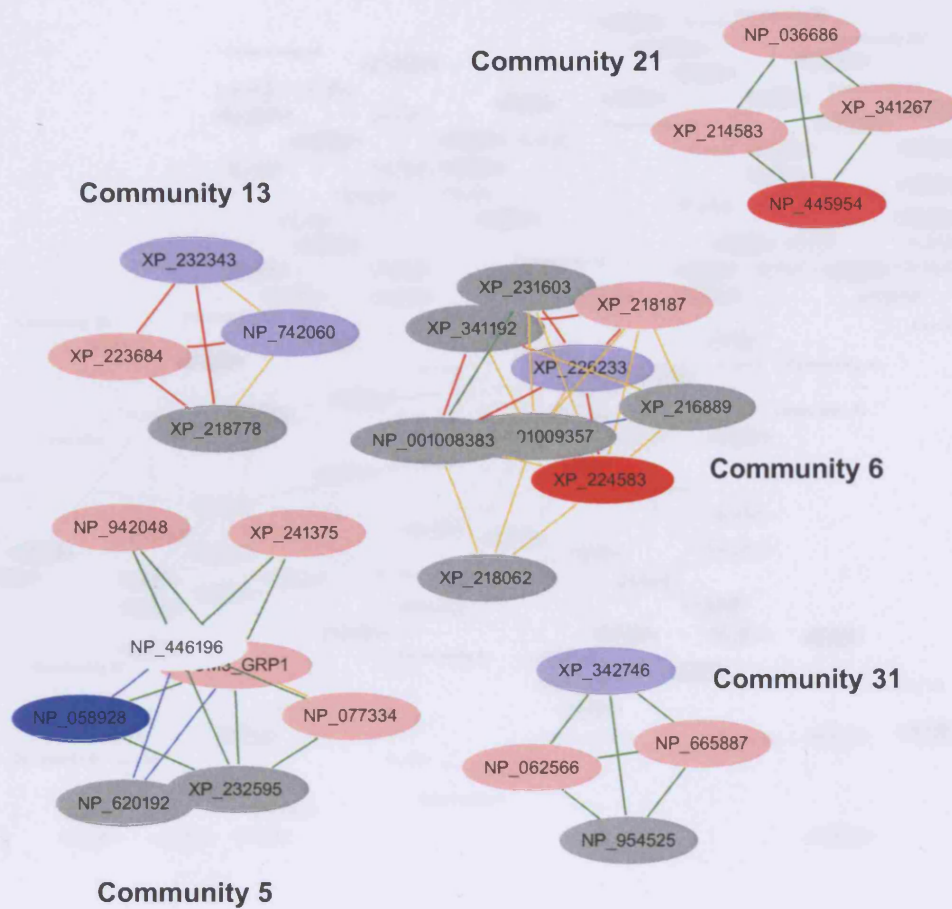
Appendix D

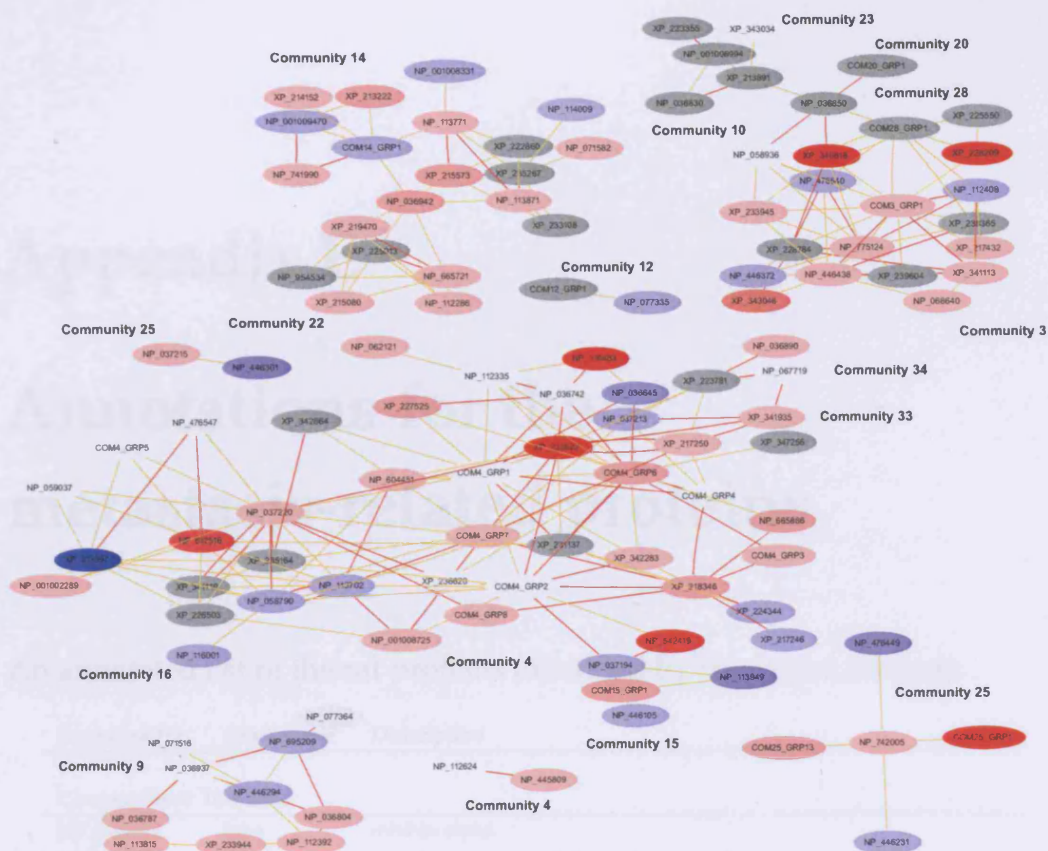
A closer view of the metastasis-related communities

A detailed description of the metastasis-related protein networks identified by cluster analysis. The protein nodes are labelled with RefSeq identifiers. The weakest links (blue and green) were removed from the last graph to improve legibility. Annotations relating to these proteins are given in Appendix E and the approximate location of the corresponding communities is indicated.









Appendix E

Annotations for the metastasis-related proteins

An annotated list of the rat proteins identified by the clique analysis.

Sequence ID	Gene name	Description
Community 0: TGF-beta		
NP_036722	Inha	inhibin alpha
NP_036907	Tgfb1	transforming growth factor, beta receptor 1
NP_036959	Bmp4	bone morphogenetic protein 4
NP_037234	Fkbp1a	FK506-binding protein 1a
NP_037239	Bmp6	bone morphogenetic protein 6
NP_037306	Tgfb3	transforming growth factor, beta 3
NP_058824	Inhba	inhibin beta A
NP_058874	Bmp2	bone morphogenetic protein 2
NP_058952	Tgfb3	transforming growth factor, beta receptor 3
NP_058973	Ap1b1	adaptor protein complex AP-1, beta 1 subunit
NP_059052	Hpcal1	neural visinin-like Ca ²⁺ -binding protein type 3
NP_067589	Tgfb1	transforming growth factor, beta 1
NP_071886	Acreq	activin A receptor type II-like 1
NP_077342	Freq	neuronal calcium sensor-1
NP_077812	Acvr1	activin type I receptor
NP_110476	Bmpr1a	bone morphogenetic protein receptor, type 1A
NP_112393	Tgfb2	transforming growth factor, beta 2
NP_112394	Tgfb2	transforming growth factor-b type II receptor
NP_954700	Acvr1b	activin A receptor, type 1B
XP_217147		similar to mmdj4
XP_217297		activin receptor IIB
XP_217409		similar to Bone morphogenetic protein type II receptor
XP_227759		similar to CFK-43a=bone morphogenetic protein binding ser/thr kinase receptor
XP_232168		similar to DNA replication licensing factor MCM2
XP_342432		activin receptor IIA
XP_342592		bone morphogenetic protein 7
XP_345647		similar to growth/differentiation factor 7

Community 1: Proteasome

NP_001005875	Psmc12	proteasome 26S non-ATPase subunit 12
NP_001008218	Psmc7	proteasome (prosome, macropain) subunit, alpha type 7
NP_001008282	Psmc3	proteasome 26S non-ATPase subunit 3
NP_001008302	Usp14	ubiquitin specific protease 14
NP_058736	Ppp2cb	protein phosphatase 2a, catalytic subunit, beta isoform
NP_058796	Plk1	polo-like kinase 1
NP_058974	Psmc1	proteasome (prosome, macropain) subunit, alpha type 1
NP_058975	Psmc2	proteasome (prosome, macropain) subunit, alpha type 2
NP_058976	Psmc3	proteasome (prosome, macropain) subunit, alpha type 3
NP_058977	Psmc4	proteasome (prosome, macropain) subunit, alpha type 4
NP_058979	Psmc6	proteasome (prosome, macropain) subunit, alpha type 6
NP_058980	Psmc2	proteasome (prosome, macropain) subunit, beta type 2
NP_058981	Psmc3	proteasome (prosome, macropain) subunit, beta type 3
NP_112411	Sug1	proteasomal ATPase (SUG1)
NP_112621	Psmc4	proteasome (prosome, macropain) 26S subunit, non-ATPase, 4
NP_113817	Psmc4	proteasome (prosome, macropain) subunit, beta type 4
NP_150239	Psmc2	proteasome (prosome, macropain) 26S subunit, ATPase 2
NP_476440	Psmc6	proteasome (prosome, macropain) subunit, beta type 6
NP_476463	Psmc4	proteasome 26S ATPase subunit 4
NP_942025	MGC72968	proteasome, 26S, non-ATPase regulatory subunit 6
XP_213456		similar to Protein translation factor SUI1 homolog
XP_215745		similar to 26S proteasome-associated pad1 homolog
XP_218424		similar to nucleotide excision repair protein
XP_220754		similar to 26S proteasome non-ATPase regulatory subunit 11
XP_224534		similar to karyopherin beta 3; Ran_GTP binding protein 5; importin beta-3 subunit
XP_226439		similar to 26S proteasome non-ATPase regulatory subunit 7
XP_233467		similar to cytidine 5-triphosphate synthase
XP_341315		proteasome (prosome, macropain) subunit, beta type 5
XP_342114		similar to Periodic tryptophan protein 2 homolog
XP_344650		similar to Proteasome subunit alpha type 7-like
XP_344977		similar to 26S proteasome subunit p40.5

Community 2: Mitotic spindle checkpoint

NP_001008331	Polr1c	RNA polymerase I subunit
NP_036942	Sycp1	synaptonemal complex protein 1
NP_071582	Rad50	RAD50 homolog
NP_071790	Pmpcb	mitochondrial processing peptidase beta
NP_113771	Cspg6	chondroitin sulfate proteoglycan 6
NP_113871	Smc1l1	SMC1 structural maintenance of chromosomes 1-like 1
NP_114009	Plk2	polo-like kinase 2
NP_703201	Rnf40	ring finger protein 40
NP_955795	Cdk2	cyclin-dependent kinase 2
XP_215573		similar to SMC4 protein
XP_218574		similar to BC013491 protein
XP_222860		similar to NUF2R protein
XP_233108		similar to stromal antigen 2
XP_233337		similar to epidermal growth factor receptor pathway substrate 15
XP_235267		similar to pokeweed agglutinin-binding protein
XP_342838		similar to SMC2 protein

Community 3: Myosin

NP_001009268	Actr2	actin-related protein 2
NP_058936	Myh7	myosin heavy chain, polypeptide 7
NP_068640	Ctnn	cortactin isoform B
NP_112406	Actb	cytoplasm beta-actin
NP_112408	Arpc1a	suppressor of profilin/p41 of actin-related complex 2/3
NP_446372	Trip10	thyroid hormone receptor interactor 10
NP_446438	Myo1b	myosin Ib
NP_476540	Waspip	Wiskott-Aldrich syndrome protein interacting protein
NP_775124	Myo1e	myosin IE
NP_955795	Cdk2	cyclin-dependent kinase 2
XP_217432		similar to actin related protein 2/3 complex subunit 2; ARP2/3 complex subunit 34
XP_218617		similar to nonmuscle myosin heavy chain
XP_228784		similar to Wiskott-Aldrich Syndrome Protein
XP_233945		similar to mKIAA1256 protein
XP_238365		similar to actin related protein 2/3 complex, subunit 4
XP_239604		similar to GluR-delta2 philic-protein
XP_340818		myosin, heavy polypeptide 4
XP_341113		actin-related protein 3 homolog
XP_343046		similar to Sh3yl1

Community 4: Intracellular signaling cascade

NP_001002289	Fut8	fucosyltransferase 8 (alpha (1,6) fucosyltransferase)
NP_001004081	Mpi	mannose phosphate isomerase
NP_001008725	Il6st	interleukin 6 signal transducer
NP_036645	Bdnf	brain derived neurotrophic factor
NP_036742	Ngfr	nerve growth factor receptor, fast
NP_036863	Ntrk2	neurotrophic tyrosine kinase, receptor, type 2
NP_036887	Fyn	fyn proto-oncogene
NP_036921	Dpp4	dipeptidylpeptidase 4
NP_036978	Fgf1	fibroblast growth factor 1
NP_037101	Irs1	insulin receptor substrate 1
NP_037137	Pik3r	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 1
NP_037194	Kdr	kinase insert domain protein receptor
NP_037213	Ptk2	PTK2 protein tyrosine kinase 2
NP_037220	Ptpn11	protein tyrosine phosphatase, non-receptor type 11
NP_037319	Plcg1	phospholipase C, gamma 1
NP_058762	Ptn	pleiotrophin
NP_058767	Insr	insulin receptor
NP_058790	Ghr	growth hormone receptor
NP_058864	Plcg2	phospholipase C, gamma 2
NP_059037	Lipf	lipase, gastric
NP_062121	Ntrk3	neural receptor protein-tyrosine kinase
NP_062178	Fgf2	fibroblast growth factor 2
NP_067600	Ntrk1	trk precursor
NP_071549	Pik3r3	phosphatidylinositol 3-kinase p55 subunit
NP_071963	Arf1	ADP-ribosylation factor 1
NP_110483	Lrrn3	leucine rich repeat protein 3, neuronal
NP_110486	Mdk	midkine
NP_112335	Ntf3	neurotrophin 3
NP_112624	Cdh1	cadherin 1
NP_113702	Jak2	Janus kinase 2

NP_113811	Grb14	growth factor receptor bound protein 14
NP_434694	Igf1r	insulin-like growth factor 1 receptor
NP_445775	Degs	degenerative spermatocyte homolog
NP_445809	Catnb	beta-catenin
NP_445855	Grb7	growth factor receptor binding protein GRB7
NP_476547	Ncr1	lymphocyte antigen 94 (mouse) homolog (activating NK-receptor; NK-p46)
NP_542419	Plcd4	phospholipase C, delta 4
NP_604451	Sh2bpsm1	SH2-B PH domain containing signaling mediator 1
NP_612516	Ptpnc	protein tyrosine phosphatase, receptor type, C
NP_665725	Cyp3a18	cytochrome P450, 3a18
NP_665886	Socs1	suppressor of cytokine signaling 1
NP_849197	Igf1	insulin-like growth factor 1
XP_213997	Ptpn13	similar to protein Tyr phosphatase, non-receptor type 13
XP_214050	Lap3_pred	leucine aminopeptidase 3 (predicted)
XP_217246	Nck1_pred	similar to non-catalytic region of tyrosine kinase adaptor protein 1
XP_217250	Ephb1	PREDICTED: Eph receptor B1
XP_218346	Axl_pred	now NP_001013165. AXL receptor tyrosine kinase (predicted)
XP_221036	LOC303606	now NP_001013996.hypothetical protein
XP_224344	Dok2_pred	similar to docking protein Dok-R
XP_226503	LOC307845	similar to hypothetical protein BC002770
XP_227525	Ngfb	similar to nerve growth factor beta chain precursor - multimammate rat
XP_231137	Abl1	similar to Abl1 protein Proto-oncogene tyrosine-protein kinase ABL1
XP_232763	Lck	lymphocyte-specific protein tyrosine kinase
XP_233522	LOC298528	similar to Ephrin type-A receptor 10
XP_235164	Frs2_pred	similar to fibroblast growth factor receptor substrate 2
XP_236628	Mst1r_pred	similar to hepatocyte growth factor-like protein receptor
XP_341110	Ptpn4	similar to testis-enriched protein tyrosine phosphatase
XP_342283	Shc1	now NP_445969 SHC (Src homology 2 domain-containing) trans-forming protein 1
XP_342864	Tek	similar to TIE-2=receptor-like tyrosine kinase
XP_343062	LOC362737	similar to RIKEN cDNA D930036F22 gene
XP_347256		similar to met proto-onco

Community 5: EGF-like domain containing proteins

NP_058928	Ptgs2	prostaglandin-endoperoxide synthase 2
NP_062020	Jag1	jagged 1
NP_077334	Notch2	notch gene homolog 2
NP_446196	Dlk1	delta-like 1 homolog
NP_620192	Pou3f3	POU domain, class 3, transcription factor 3
NP_942048	Rpl3	ribosomal protein L3
XP_232595		similar to Rbpsi protein
XP_241375		similar to N-terminal acetyltransferase 1
XP_343120		jagged 2

Community 6: Endo/exonuclease

NP_001008383	Cnot8	similar to CCR4-NOT transcription complex, subunit 8 (CAF1-like protein)
NP_001009357	Rqcd1	RCD1 required for cell differentiation1 homolog
XP_216889		similar to CCR4-NOT transcription complex, subunit 2; NOT2 (neg. reg. of transcr)
XP_218062		similar to LATS homolog 1

XP_218187	similar to CCR4-NOT transcription complex, subunit 3
XP_224583	similar to CG31759-PA endo/exonuclease activity
XP_226233	similar to KIAA1007 protein; adrenal gland protein AD-005
XP_231603	similar to potential transcriptional repressor Not4hp
XP_341192	similar to Hypothetical protein 4932442K20Rik

Community 7: Nucleocytoplasm transport

NP_058759	Kpnb1	karyopherin (importin) beta 1
NP_059057	Nup54	nucleoporin 54kDa
NP_445891	Ran	RAN, member RAS oncogene family
NP_446074	Pom121	nucleus pore membrane glycoprotein 121 kD
NP_703206	Krt1-9	keratin complex 1, acidic, gene 9
NP_942021	Kpna1	karyopherin alpha 1 (importin alpha 5)
XP_214639		similar to RIKEN cDNA 2410008G02
XP_218620		similar to nucleus pore glycoprotein p62 (62 kDa nucleoporin)
XP_224534		similar to karyopherin beta 3; Ran_GTP binding protein 5; importin beta-3 subunit
XP_341509		nucleoporin 153kD

Community 8: Cell cycle/cytokinesis

NP_037027	Adk	adenosine kinase
NP_037309	Got2	glutamate oxaloacetate transaminase 2
NP_072138	Sept7	CDC10 (cell division cycle 10, S.cerevisiae, homolog)
NP_076481	Prkaa2	AMP-activated protein kinase alpha 2 catalytic subunit
NP_114025	Sept9	septin 9
NP_446023	Rgpr	regucalcin gene promotor region related protein
NP_446383	Gp1bb	glycoprotein Ib (platelet), beta polypeptide
NP_476489	Sept2	septin 2
NP_620184	Gorasp2	golgi reassembly stacking protein 2
NP_787032	Eef1a1	eukaryotic translation elongation factor 1 alpha 1
XP_213393		similar to CGI-125 protein
XP_213413		similar to H5
XP_213922		similar to Brain protein 44 (0-44 protein)
XP_216170		similar to Hypothetical protein MGC59076
XP_217147		similar to mmDj4
XP_219498		similar to serine/threonine kinase 29
XP_223227		similar to hypothetical protein FIJ10849
XP_231118		similar to kynurenine aminotransferase/glutamine transaminase K
XP_234516		similar to Cyclin K
XP_340794		similar to TBC1 domain, member 8; BUB2-like protein 1; vasc. Rab-GAP/TBC-containing
XP_343156		similar to protein inhibitor of activated STAT gamma
XP_343275		similar to myosin phosphatase targeting subunit 3 MYPT3

Community 9: Nuclear hormone receptors

NP_036787	Sp1	sp1 transcription factor
NP_036804	Thrb	thyroid hormone receptor beta
NP_036937	Rxra	retinoid X receptor alpha
NP_071516	Hnf4a	hepatocyte nucleus factor 4 alpha
NP_077335	Hif1a	hypoxia inducible factor 1, alpha subunit
NP_077364	Nr4a1	nucleus receptor subfamily 4, group A, member 1
NP_112392	Nr2f1	nucleus receptor subfamily 2, group F, member 1
NP_113815	Nr1h3	nucleus receptor subfamily 1, group H, member 3

NP_446294	Mapk1	mitogen activated protein kinase 1
NP_695209	Cops2	COP9 (constitutive photomorphogenic) homolog, subunit 2
XP_216720		similar to Thyroid transcription factor 1 (Thyroid nucleus factor 1) (TTF-1)
XP_226076		similar to putative WDC146
XP_233944		similar to nucleus receptor co-activator

Community 10: Sarcoglycans

NP_001006994	Sgcg	gamma sarcoglycan
NP_036830	Dmd	dystrophin isoform Dp71a
XP_213891		similar to abnormal spindle
XP_220884		similar to alpha-sarcoglycan
XP_223355		similar to beta-sarcoglycan

Community 11: Karyopherins

NP_037347	Akr7a3	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)
NP_058759	Kpnb1	karyopherin (importin) beta 1
NP_058999	Kcnab1	potassium voltage-gated channel, shaker-related subfamily, beta member 1
NP_059000	Kcnab2	potassium voltage-gated channel, shaker-related subfamily, beta member 2
NP_942021	Kpna1	karyopherin alpha 1 (importin alpha 5)

Community 12: Hypoxia inducible factor

NP_077335	Hif1a	hypoxia inducible factor 1, alpha subunit
NP_077338	Arntl	aryl hydrocarbon receptor nucleus translocator-like
XP_234728		similar to Hspca protein
XP_234791		similar to heat shock protein 84 - mouse

Community 13: Peroxisomal proteins

NP_742060	Pex14	peroxisomal membrane anchor protein
XP_218778		similar to peroxisomal PTS2 receptor
XP_223684		similar to Peroxisomal membrane protein PEX13 (Peroxin-13)
XP_232343		similar to Pex5 protein

Community 14: Cell cycle regulation

NP_001009470	MGC108931	similar to cyclin B2
NP_741990	Cdc20	cell division cycle 20 homolog
NP_955795	Cdk2	cyclin-dependent kinase 2
XP_213222		similar to membrane-associated tyrosine-and threonine-spec. cdc2-inhibitory kinase
XP_214152		similar to cyclin-dependent kinase inhibitor 3; CDK2-assoc. dual spec. phosphatase
XP_235722		similar to cell division cycle 2 homolog 2; cell division cycle 2-like 2

Community 15: VEGF

NP_037194	Kdr	kinase insert domain protein receptor
NP_113949	Figf	c-fos induced growth factor
NP_114024	Vegfa	vascular endothelial growth factor
NP_446047	Pgf	placental growth factor
NP_446105	Vegfc	vascular endothelial growth factor C

Community 16: JAK/STAT cascade

NP_058790	Ghr	growth hormone receptor
NP_071775	Stat5b	signal transducer and activator of transcription 5B
NP_113702	Jak2	Janus kinase 2
NP_116001	Stat1	signal transducer and activator of transcription 1

Community 17: Karyopherin docking complex

NP_942021	Kpna1	karyopherin alpha 1 (importin alpha 5)
XP_215550		similar to polymyositis scleroderma overlap syndrome (PM-SCL) antigen 1 a
XP_216949		similar to putative exosome complex exonuclease RRP41
XP_218343		similar to DNA segment, Chr 7, Wayne State University 180, expressed
XP_233673		similar to P100 polymyositis-scleroderma overlap syndrome assoc. autoantigen homolog

Community 18: NF-kappaB regulation

NP_445807	Ikbkb	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
NP_954534	Ikbkg	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
XP_219857		similar to conserved helix-loop-helix ubiquitous kinase
XP_234728		similar to Hspca protein
XP_234791		similar to heat shock protein 84 - mouse
XP_340919		similar to NF-kappaB inducing kinase

Community 19: Calmodulin

NP_036650	Calm3	calmodulin 3
NP_037220	Ptpn11	protein tyrosine phosphatase, non-receptor type 11
XP_213368		similar to Eukaryotic translation initiation factor 5A (eIF-5A) (eIF-4D) (Rev-binding)
XP_214050		similar to leucine aminopeptidase
XP_341110		similar to testis-enriched protein tyrosine phosphatase

Community 20: Actinins

NP_112267	Actn1	actinin, alpha 1
NP_113863	Actn4	alpha actinin 4
NP_653346	Gpsm1	G-protein signalling modulator 1 (AGS3-like, C. elegans)
XP_214499		similar to actinin, alpha 2
XP_216586		similar to Msx-2 interacting nucleus target protein

Community 21: ATP transporter proteins

NP_036686	Eno1	enolase 1, alpha
NP_445954	Abcg1	ATP-binding cassette, sub-family G (WHITE), member 1
XP_214583		similar to grp75
XP_341267		similar to polynucleotide phosphorylase-like protein

Community 22: Tubulin proteins

NP_036942	Sycp1	synaptonemal complex protein 1
NP_112286	Dbn1	drebrin 1
NP_665721	Tubg1	tubulin, gamma 1
NP_954534	Ikbkg	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma

XP_215080		similar to RIKEN cDNA 3230401O13
XP_219470		similar to tubulin, gamma complex associated protein 2
XP_225013		similar to Gamma-tubulin complex 3 (GCP-3) (Spindle pole body Spc98 homol.)

Community 23: Actinin/calmodulin

NP_036650	Calm3	calmodulin 3
NP_059015	Grp58	glucose regulated protein, 58 kDa
NP_112267	Actn1	actinin, alpha 1
NP_113863	Actn4	alpha actinin 4
XP_213891		similar to abnormal spindle
XP_214499		similar to actinin, alpha 2
XP_343034		similar to dystrobrevin B (mDTN-B)

Community 24: Proteasome

NP_001005875	Psmc12	proteasome 26S non-ATPase subunit 12
XP_226439		similar to 26S proteasome non-ATPase regulatory subunit 7 (26S proteasome reg. subunit)
XP_230377		similar to TRAF6
XP_343114		similar to CD40 receptor associated factor 1
XP_344977		similar to 26S proteasome subunit p40.5

Community 25: Serine protease inhibitors

NP_001007619	Rchy1	ring finger and CHY zinc finger domain containing 1
NP_001007733	MGC94010	similar to SPI6
NP_036630	Aldr1	aldehyde reductase 1 (low Km aldose reductase) (5.8 kb PstI fragment)
NP_037215	Hspa5	heat shock 70kD protein 5
NP_059036	Acox1	acyl-Coenzyme A oxidase 1, palmitoyl
NP_067728	Serpinb2	serine (or cysteine) proteinase inhibitor, clade B, member 2
NP_075218	Kcnip1	potassium channel interacting protein 1
NP_113815	Nr1h3	nucleus receptor subfamily 1, group II, member 3
NP_446231	Serpin1	serine (or cysteine) proteinase inhibitor, clade I (neuroserpin), member 1
NP_476449	Serpinb5	serine (or cysteine) proteinase inhibitor, clade B, member 5
NP_543169	Pde11a	phosphodiesterase 11A
NP_596897	Agpat4	1-acylglycerol-3-phosphate O-acyltransferase 4
NP_741984	Spna2	alpha-spectrin 2
NP_742005	Canx	calnexin
NP_742026	Eif2b1	eukaryotic translation initiation factor 2B, subunit 1 alpha
NP_954888	Rela	v-rel reticuloendotheliosis viral oncogene homolog A
XP_213076		similar to GTP-binding protein NGB
XP_233568		similar to Zinc finger protein 436
XP_341644		similar to Polymerase (RNA) II (DNA directed) polypeptide C

Community 26: Myosin

NP_036650	Calm3	calmodulin 3
NP_058936	Myh7	myosin heavy chain, polypeptide 7
XP_218617		similar to nonmuscle myosin heavy chain
XP_340818	Myh4_pred	myosin, heavy polypeptide 4; now: XP_340819, similar to Myh4
XP_343512		similar to leucyl-tRNA synthetase

Community 27: Nucleoporin

NP_036922	Dsp	dentin sialophosphoprotein
XP_218620		similar to nucleus pore glycoprotein p62 (62 kDa nucleoporin)
XP_219265		similar to importin 7
XP_341509		nucleoporin 153kD

Community 28: Laminin

NP_112406	Actb	cytoplasm beta-actin
NP_113708	Myh10	myosin heavy chain 10, non-muscle
XP_218617		similar to nonmuscle myosin heavy chain
XP_225550		similar to RIKEN cDNA 2310068022
XP_228209		similar to Lama4 protein

Community 29: Matrix metalloproteases

NP_037215	Hspa5	heat shock 70kD protein 5
NP_077376	Adamts1	a disintegrin and metalloproteinase with thrombospondin motifs 1
NP_446301	Erp70	protein disulfide isomerase related protein (calcium-binding, intestinal-rel.)
XP_343193		Now NP_001012197 Tra1_predicted tumor rejection antigen gp96 (predicted)

Community 30: Casein kinase

NP_036654	Cbs	cystathionine beta synthase
NP_074046	Csnk1g3	casein kinase 1, gamma 3
XP_213368		similar to Eukaryotic translation initiation factor 5A (eIF-5A, eIF-4D) (Rev-binding)
XP_343107		similar to Vrk1 protein

Community 31: Zinc finger protein

NP_062566	Znf386	zinc finger protein 386 (Kruppel-like)
NP_665887	Lhx1	LIM homeobox protein 1
NP_954525	MGC73008	Unknown (protein for MGC:73008)
XP_342746		hypothetical protein XP_342745

Community 32: Actinin

NP_112267	Actn1	actinin, alpha 1
NP_113863	Actn4	alpha actinin 4
NP_775148	Pdlim7	PDZ and LIM domain 7
XP_214499		similar to actinin, alpha 2
XP_235710		similar to Tenc1 protein

Community 33: Breast cancer anti-estrogen resistance

NP_037063	Bcar1	breast cancer anti-estrogen resistance 1
NP_037213	Ptk2	PTK2 protein tyrosine kinase 2
XP_217250	Ephb1	PREDICTED: Eph receptor B1 (Tyrosine-protein kinase receptor EPH-2)
XP_223781		similar to Vinculin (Metavinculin)
XP_233522		similar to Eph-like receptor tyrosine kinase
XP_341935		transforming growth factor beta 1 induced transcript 1
XP_343143		similar to ring finger protein 41; hypothetical SBBI03 protein

Community 34: Breast cancer anti-estrogen resistance

NP_036890	Syk	spleen tyrosine kinase
-----------	-----	------------------------

NP_037063	Bcar1	breast cancer anti-estrogen resistance 1
NP_067719	ErbB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4
XP_223781		similar to Vinculin (Metavinculin)
XP_341935		transforming growth factor beta 1 induced transcript 1

Community 35: Protein-tyrosine kinases

NP_036887	fyn	fyn proto-oncogene
NP_037137	Pik3r1	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 1
NP_037213	Ptk2	PTK2 protein tyrosine kinase 2
XP_217250		similar to Ephrin type-B receptor 1 precursor (Tyrosine-protein kinase receptor EPH-2)
XP_342283		SHC (Src homology 2 domain-containing) transforming protein 1

Community 36: Fyn proto-oncogene

NP_036887	fyn	fyn proto-oncogene
NP_569089	Khdrbs1	src associated in mitosis, 68 kDa
XP_232763	Lck	lymphocyte-specific protein tyrosine kinase
XP_343333		similar to SMARCD1 protein

Appendix F

Internet resources

URLs for some Internet resources mentioned or used in this work.

3D-JIGSAW Homology Modelling	bmm.cancerresearchuk.org/~3djigsaw
Affymetrix	www.affymetrix.com
ASTRAL	astral.stanford.edu
BIND	www.bind.ca
BioGRID	www.thebiogrid.org
Biomolecular Modelling at CR-UK	bmm.cancerresearchuk.org
BLAST and PSI-BLAST servers	www.ncbi.nlm.nih.gov/BLAST
CATH	www.biochem.ucl.ac.uk/bsm/cath
DIP	dip.doe-mbi.ucla.edu
DomainFishing	bmm.cancerresearchuk.org/~3djigsaw/dom_fish
EBI	www.ebi.ac.uk
Ensembl Genome Browser	www.ensembl.org
Google search tools	www.google.com
HPRD	www.hprd.org
IntAct	www.ebi.ac.uk/intact
NACCESS	wolf.bms.umist.ac.uk/naccess
MINT	cbm.bio.uniroma2.it/mint
MIPS CYGT	mips.gsf.de/proj/yeast
MIPS Mammalian Protein Database	mips.gsf.de/proj/ppi
MySQL	www.mysql.com
NCBI	www.ncbi.nlm.nih.gov
OPHID	ophid.utoronto.ca
PFAM	www.sanger.ac.uk/Software/Pfam
PIP	bmm.cancerresearchuk.org/servers/pip/
POINT	point.bioinformatics.tw
PostgreSQL	www.postgresql.org
Protein Data Bank	www.rcsb.org
SCOP	scop.mrc-lmb.cam.ac.uk/scop
STRING	string.embl.de
Swiss-Prot Protein knowledgebase	expasy.org/sprot

Bibliography

- Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**:1021–1023.
- Ahmad, A. and Hart, I. R. (1997). Mechanisms of metastasis. *Crit Rev Oncol Hematol.*, **26**:163–173.
- Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, **401**:130–131.
- Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, **406**:378–382.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular biology of the cell*. Garland Science, New York, fourth edition.
- Alexander, W. S. and Hilton, D. J. (2004). The role of suppressors of cytokine signaling (SOCS) proteins in regulation of the immune response. *Annu Rev Immunol.*, **22**:503–529.
- Ali, S., Nouhi, Z., Chughtai, N., and Ali, S. (2003). SHP-2 regulates SOCS-1-mediated Janus kinase-2 ubiquitination/degradation downstream of the prolactin receptor. *J Biol Chem.*, **278**:52021–52031.
- All-Ericsson, C., Girnita, L., Seregard, S., Bartolazzi, A., Jager, M. J., and Larsson, O. (2002). Insulin-like growth factor-1 receptor in uveal melanoma: a predictor for metastatic disease and a potential therapeutic target. *Invest Ophthalmol Vis Sci.*, **43**:1–8.
- Aloy, P., Pichaud, M., and Russell, R. B. (2005). Protein complexes: structure prediction challenges for the 21st century. *Curr Opin Struc Biol.*, **15**:15–22.

- Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall, London.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**:3389–3402.
- Amoutzias, G. D., Robertson, D. L., Oliver, S. G., and Bornberg-Bauer, E. (2004). Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, **5**:1–6.
- Andre, F., Janssens, B., Bruyneel, E., van Roy, F., Gespach, C., Mareel, M., and Bracke, M. (2004). Alpha-catenin is required for IGF-I-induced cellular migration but not invasion in human colonic cancer cells. *Oncogene*, **23**:1177–1186.
- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**:223–230.
- Armitage, P. and Berry, G. (1994). *Statistical methods in medical research*. Blackwell Science, Oxford, third edition.
- Armstrong, N. J. and van de Wiel, M. A. (2004). Microarray data analysis: from hypotheses to conclusions using gene expression data. *Cell Oncol.*, **26**:279–290.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**:25–29.
- Atzeni, P., Ceri, S., Paraboschi, S., and Torlone, R. (1999). *Database systems: concepts, languages and architecture*. McGraw-Hill International, Maidenhead, England.
- Bader, G. D., Betel, D., and Hogue, C. W. V. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**:248–250.
- Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol.*, **20**:991–997.

- Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol.*, **22**:78–85.
- Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**:45–48.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**:509–512.
- Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet.*, **5**:101–113.
- Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet.*, **12**:72–77.
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res.*, **32**:D138–D141.
- Bates, P. A., Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. (2001). Enhancement of protein modelling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **45**:39–46.
- Ben-Hur, A. and Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7**:S2.
- Berget, S. M., Moore, C., and Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*, **74**:3171–3175.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**:235–242.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol.*, **112**:535–542.

- Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem.*, **72**:291–336.
- Bock, J. R. and Gough, D. A. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**:455–460.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* (2003). The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**:365–370.
- Bogenrieder, T. and Herlyn, M. (2003). Axis of evil: molecular mechanisms of cancer metastasis. *Oncogene*, **22**:6524–6536.
- Bordner, A. J. and Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins*, **60**:353–366.
- Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol.*, **14**:292–299.
- Bourette, R. P., De Sepulveda, P., Arnaud, S., Dubreuil, P., Rottapel, R., and Mouchiroud, G. (2001). Suppressor of cytokine signaling 1 interacts with the macrophage colony-stimulating factor receptor and negatively regulates its proliferation signal. *J Biol Chem.*, **276**:22133–22139.
- Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**:R35.
- Bradford, J. R. and Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**:1487–1494.
- Brazhnik, P., de la Fuente, A., and Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends Biotechnol.*, **20**:467–472.
- Breiman, L. (2001). Random Forests. *Machine Learning*, **45**:5–32.
- Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, **21**:2076–2082.

- Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., *et al.* (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res.*, **31**:2443–2450.
- Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc.*, **2**:121–167.
- Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**:4364–4375.
- Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. V. (2005). Differential network expression during drug and stress response. *Bioinformatics*, **21**:2898–2905.
- Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**:190–202.
- Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., *et al.* (2005). A network-based analysis of systemic inflammation in humans. *Nature*, **437**:1032–1037.
- Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.*, **3**:285–298.
- Carter, P., Liu, J., and Rost, B. (2003). PEP: Predictions for Entire Proteomes. *Nucleic Acids Res.*, **31**:410–413.
- Cech, T. R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell*, **44**:207–210.
- Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004a). The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**:D189–D192.

- Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004b). The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**:189–192.
- Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5**:823–826.
- Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12**:1–8.
- Clark, F. and Thanaraj, T. A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet.*, **11**:451–464.
- Codd, E. F. (1970). A relational model for large shared data banks. *Commun ACM*, **12**:377–387.
- Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J.-M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome Res.*, **14**:1324–1332.
- Contreras-Moreira, B. and Bates, P. A. (2002). Domain fishing: a first step in protein comparative modelling. *Bioinformatics*, **18**:1141–1142.
- Contreras-Moreira, B., Jonsson, P. F., and Bates, P. A. (2003). Structural context of exons in proteins: implications for protein modelling and design. *J Mol Biol.*, **333**:1045–1059.
- Creighton, T. E. (1993). *Proteins: structures and molecular properties*. W.H. Freeman and Company, New York, second edition.
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**:561–563.
- Cristianini, N. and Shawe-Taylor, J. (2000). *Support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge.
- Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.*, **23**:324–328.

- Dartnell, L., Simeonidis, E., Hubank, M., Tsoka, S., Bogle, I. D. L., and Pappageorgiou, L. G. (2005). Robustness of the p53 network and biological hackers. *FEBS Lett.*, **579**:3037–3042.
- Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, **1**:349–356.
- Deeds, E. J., Ashenberg, O., and Shakhnovich, E. I. (2006). A simple physical model for scaling in protein-protein interaction networks. *P Natl Acad Sci USA*, **103**:311–316.
- Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**:1540–1548.
- Deng, M., Sun, F., and Chen, T. (2003). Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput.*, pages 140–151.
- Di Battista, G., Eades, P., Tamassia, R., and Tollis, I. G. (1994). Algorithms for drawing graphs: an annotated bibliography. *Comp Geom Theory Appl.*, **4**:235–282.
- Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, **301**:827–829.
- Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS Journal*, **272**:5129–5148.
- Dunn, S. E., Ehrlich, M., Sharp, N. J., Reiss, K., Solomon, G., Hawkins, R., Baserga, R., and Barrett, J. C. (1998). A dominant negative mutant of the insulin-like growth factor-I receptor inhibits the adhesion, invasion, and metastasis of breast cancer. *Cancer Res.*, **58**:3353–3361.
- Eddy, S. R. (1996). Hidden Markov models. *Curr Opin Struct Biol.*, **6**:361–365.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA*, **95**:14863–14868.

- Eisenberg, E. and Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Phys Rev Lett.*, **91**:138701.
- Erdős, P. and Rényi, A. (1959). On random graphs. *Publ Math.*, **6**:290–297.
- Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem.*, **269**:1356–1361.
- Fedorova, L. and Fedorov, A. (2003). Introns in gene evolution. *Genetica*, **118**:123–131.
- Ferrara, N., Gerber, H.-P., and LeCouter, J. (2003). The biology of VEGF and its receptors. *Nature Med.*, **9**:669–676.
- Finley, R. L. and Brent, R. (1994). Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc Natl Acad Sci USA*, **91**:12980–12984.
- Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., *et al.* (2005). Protein interaction mapping: a *Drosophila* case study. *Genome Res.*, **15**:376–384.
- Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**:750–752.
- Fraser, H. B., Hirsh, A. E., Wall, D. P., and Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA*, **101**:9033–9038.
- Fraser, H. B., Wall, D. P., and Hirsh, A. E. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol.*, **3**:11.
- Furukawa, M., Raffeld, M., Mateo, C., Sakamoto, A., Moody, T. W., Ito, T., Venzon, D., Serrano, J., and Jensen, R. (2005). Increased expression of insulin-like growth factor I and/or its receptor in gastrinomas is associated with low curability, increased growth, and development of metastases. *Clin Cancer Res.*, **11**:3233–3242.

- Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer*, **4**:177–83.
- Gansner, E. R., Koutsofios, E., North, S. C., and Vo, K. P. (1993). A Technique for drawing directed graphs. *IEEE Trans on Soft Eng.*, **19**:214–230.
- Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., and Cruciat, C. M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**:141–147.
- Ge, H., Walhout, A. J. M., and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, **19**:551–560.
- Gene Ontology Consortium (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**:D322–D326.
- Gerstein, M., Lan, N., and Jansen, R. (2002). Proteomics. Integrating interactomes. *Science*, **295**:284–287.
- Gerstein, M. and Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc Intl Conf Intell Syst Mol Biol.*, **4**:59–67.
- Gfeller, D., Chappelier, J.-C., and De Los Rios, P. (2005). Finding instabilities in the community structure of complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys.*, **72**:056135.
- Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harbor Symp Quant Biol.*, **52**:901–905.
- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al.* (2002). A protein interaction map of *Drosophila melanogaster*. *Science*, **302**:1727–1736.
- Grishin, N. V. and Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.*, **3**:2455–2458.

- Griths, A., Robinson, L. A., and Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *J Document.*, **40**:175–205.
- Guimera, R. and Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, **433**:895–900.
- Guldener, U., Münsterkötter, M., Kastenmüller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., Garcia-Martinez, J., Perez-Ortin, J. E., *et al.* (2005). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**:D364–D368.
- Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J. D., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L. S., *et al.* (2005). Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature*, **436**:861–865.
- Hakes, L., Robertson, D. L., and Oliver, S. G. (2005). Effect of dataset selection on the topological interpretation of protein interaction networks. *BMC Genomics*, **6**:131.
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res.*, **33**:514–517.
- Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol.*, **23**:839–844.
- Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**:3201–3212.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**:29–36.
- Hayles, J. and Nurse, P. (2001). A journey into space. *Nat Rev Mol Cell Biol.*, **2**:647–656.
- Heath, C. W. (2005). Community clusters of childhood leukemia and lymphoma: evidence of infection? *Am J Epidemiol.*, **162**:817–822.

- Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., *et al.* (2004a). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol.*, **22**:177–183.
- Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., *et al.* (2004b). IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**:452–455.
- Hirakawa, S., Kodama, S., Kunstfeld, R., Kajiya, K., Brown, L. F., and Detmar, M. (2005). VEGF-A induces tumor and sentinel lymph node lymphangiogenesis and promotes lymphatic metastasis. *J Exp Med.*, **201**:1089–1099.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., and Boutilier, K. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**:180–183.
- Hofmann, F. and García-Echeverría, C. (2005). Blocking insulin-like growth factor-I receptor as a strategy for targeting cancer. *Drug Discov Today*, **10**:1041–1047.
- Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, **19**:532–538.
- Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V., and Lankelma, J. (2006). Cancer: a systems biology disease. *Biosystems*, **83**:81–90.
- Hoskins, J., Lovell, S., and Blundell, T. L. (2006). An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci.*, **15**:1017–1029.
- Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., and Delisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.*, **33**:W352–W357.
- Huan-Xiang, Z. and Yibing, S. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**:336–343.

- Huang, T.-W., Tien, A.-C., Huang, W.-S., Lee, Y.-C. G., Peng, C.-L., Tseng, H.-H., Kao, C.-Y., and Huang, C.-Y. F. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**:3273–3276.
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *J Molecular Graphics*, **14**:33–38.
- Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**:S233–S240.
- International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**:931–945.
- Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J. S., White, K. P., and Rzhetsky, A. (2004). Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, **20**:1205–1213.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, **98**:4569–4574.
- Jansen, R., Lan, N., Qian, J., and Gerstein, M. (2002a). Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*, **2**:71–81.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2002b). A Bayesian Networks Approach for Predicting Protein-Protein Interactions. *Science*, **302**:449–453.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**:449–453.
- Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**:41–42.

- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**:651–654.
- Johnson, J. M., Castle, J., Garrett-Engle, P., Kan, Z., Loerch, P., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**:2141–2144.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.*, **8**:275–282.
- Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci. USA*, **93**:13–20.
- Jones, S. and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol.*, **272**:121–132.
- Jordan, I. K., Wolf, Y. I., and Koonin, E. V. (2003). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol Biol.*, **3**:1.
- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**:2577–2637.
- Kamal, A., Thao, L., Sensintaffar, J., Zhang, L., Boehm, M. F., Fritz, L. C., and Burrows, F. J. (2003). A high-affinity conformation of Hsp90 confers tumour selectivity on Hsp90 inhibitors. *Nature*, **425**:407–410.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**:D354–D357.
- Karagiannis, T. C. and El-Osta, A. (2005). RNA interference and potential therapeutic applications of short interfering RNAs. *Cancer Gene Ther.*, **12**:787–795.

- Karlin, S. and Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. *J Bacteriol.*, **178**:1881–1894.
- Katona, G., Berglund, G. I., Hajdu, J., Graf, L., and Szilagyi, L. (2002). Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. *J Mol Biol.*, **315**:1209–1218.
- Kaufmann, M. and Wagner, D. (2001). *Drawing graphs : methods and models*. Springer Verlag, Berlin.
- Keil, M., Exner, T. E., and Brickmann, J. (2004). Pattern recognition strategies for molecular surfaces: III. binding site prediction with a neural network. *J Comput Chem.*, **25**:779–789.
- Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**:656–664.
- Kerler, R. and Rabes, H. M. (1994). Rat tumor cytogenetics: a critical evaluation of the literature. *Crit Rev Oncog.*, **5**:271–295.
- Kim, H. and Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.*, **16**:553–560.
- Kohonen, T. (2001). *Self-organizing maps*. Springer-Verlag, New York, third edition.
- Koike, A. and Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel*, **17**:165–173.
- Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J Appl Cryst.*, **24**:946–950.
- LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., *et al.* (2005). A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, **438**:103–107.

- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**:860–921.
- Lee, R. E. C. and Megeney, L. A. (2005). The yeast kinome displays scale free topology with functional hub clusters. *BMC Bioinformatics*, **6**:271.
- Lee, W.-C. and Lee, K. H. (2004). Applications of affinity chromatography in proteomics. *Anal Biochem.*, **324**:1–10.
- Lehner, B. and Fraser, A. G. (2004). A first-draft human protein-interaction map. *Genome Biol.*, **5**:R63.
- LeRoith, D., Werner, H., Beitner-Johnson, D., and Roberts, C. T. (1995). Molecular and cellular aspects of the insulin-like growth factor I receptor. *Endocr Rev.*, **16**:143–163.
- Letunic, I., Copley, R. R., and Bork, P. (2002). Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet.*, **11**:1561–1567.
- Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., *et al.* (2004). A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**:540–543.
- de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, **307**:724–727.
- Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004). Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**:154.
- Liu, Y., Liu, N., and Zhao, H. (2005). Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**:3279–3285.
- Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol.*, **285**:2177–2198.
- Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**:945–953.

- MacQueen, J. (1976). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, Berkeley.
- Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem.*, **70**:437–473.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**:751–753.
- Marlene Belfort, T. C., Mary E. Reaban and Dalgaard, J. Z. (1995). Prokaryotic introns and inteins: a panoply of form and function. *J. Bacteriol.*, **177**:3897–3903.
- Martin, S., Roe, D., and Faulon, J.-L. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**:218–226.
- Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, **296**:910–913.
- Massagué, J. and Czech, M. P. (1982). The Subunit Structures of Two Distinct Receptors for Insulin-like Growth Factors I and I1 and Their Relationship to the Insulin Receptor. *J Biol Chem.*, **257**:5038–5045.
- Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.*, **11**:2120–2126.
- McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci USA*, **97**:4879–4884.
- McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, New York.
- von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and

- predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**:433–437.
- Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**:31–34.
- Modrek, B. and Lee, C. J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, **34**:117–180.
- Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**:2850–2859.
- Müller, P., Kutteneuler, D., Gesellchen, V., Zeidler, M. P., and Boutros, M. (2005). Identification of JAK/STAT signalling components by genome-wide RNA interference. *Nature*, **436**:871–875.
- Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.*, **247**:536–540.
- NCBI Genome Project (2006). Database of sequenced genomes. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>.
- Neverov, A. D., Artamonova, I. I., Nurtdinov, R. N., Frishman, D., Gelfand, M. S., and Mironov, A. A. (2005). Alternative splicing and protein function. *BMC Bioinformatics*, **6**:266.
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc Natl Acad Sci USA*, **98**:404–409.
- Newman, M. E. J. (2004). Detecting community structure in networks. *Eur Phys J B*, **38**:321–330.
- Niehrs, C. and Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature*, **402**:483–487.

- Nooren, I. M. A. and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol.*, **325**:991–1018.
- van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.*, **5**:280–284.
- Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. A. (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21**:993–1001.
- Offman, M. N., Nurtdinov, R. N., Gelfand, M. S., and Frishman, D. (2004). No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. *BMC Bioinformatics*, **5**:41.
- Ofran, Y. and Rost, B. (2003). Analysing six types of protein-protein interfaces. *J Mol Biol.*, **325**:377–387.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**:1093–1108.
- Pacifico, S., Liu, G., Guest, S., Parrish, J. R., Fotouhi, F., and Finley, R. L. J. (2006). A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for *Drosophila*. *BMC Bioinformatics*, **7**:195.
- Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu Rev Biochem.*, **55**:1119–1150.
- Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.-W., *et al.* (2005). The MIPS mammalian protein–protein interaction database. *Bioinformatics*, **21**:832–834.
- Pagel, P., Wong, P., and Frishman, D. (2004). A domain interaction map based on phylogenetic profiling. *J Mol Biol.*, **344**:1331–1346.

- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**:814–818.
- Palmer, J. D. and Logsdon, J. M. (1991). The recent origins of introns. *Curr Opin Genet Dev.*, **1**:470–477.
- Park, E. J., Park, S. Y., Joe, E. H., and Jou, I. (2003). 15d-PGJ2 and rosiglitazone suppress Janus kinase-STAT inflammatory signaling through induction of suppressor of cytokine signaling 1 (SOCS1) and SOCS3 in glia. *J Biol Chem.*, **278**:14747–14752.
- Park, J., Lappe, M., and Teichmann, S. A. (2001). Mapping Protein Family Interactions: Intramolecular and Intermolecular Protein Family Interaction Repertoires in the PDB and Yeast. *J Mol Biol.*, **307**:329–938.
- Parrish, J. R., Gulyas, K. D., and Finley, R. L. (2006). Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol.*, **17**:1–7.
- Pastor-Satorras, R., Smith, E., and Sole, R. V. (2003). Evolving protein interaction networks through gene duplication. *J Theor Biol.*, **222**:199–210.
- Patthy, L. (1987). Intron-dependent evolution: preferred types of exons and introns. *FEBS Letters*, **214**:1–7.
- Patthy, L. (1999a). Genome evolution and the evolution of exon-shuffling—a review. *Gene*, **238**:103–114.
- Patthy, L. (1999b). *Protein evolution*. Blackwell Science, Oxford.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, **96**:4285–4288.
- Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjana, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., *et al.* (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**:2363–2371.

- Persico, M., Ceol, A., Gavrilu, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, **6**:S21.
- Piehlcr, J. (2005). New methodologies for measuring protein interactions in vivo and in vitro. *Curr Opin Struct Biol.*, **15**:4–14.
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**:D501–D504.
- Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**:490–500.
- Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z. (2005). Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput.*, pages 531–542.
- Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., *et al.* (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**:211–215.
- Ramani, A. K., Bunesco, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, **6**:R40.
- Rayward-Smith, V. J., Osman, I. H., Reeves, C. R., and Smith, G. D. (1996). *Modern Heuristic Search Methods*. John Wiley & Sons, New York.
- Reichardt, J. and Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a Potts model. *Phys Rev Lett.*, **93**:218701.
- Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.*, **314**:1041–1052.

- Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol.*, **23**:951–959.
- Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**:R89.
- Rogers, S. and Girolami, M. (2005). A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, **21**:3131–3137.
- Romanucci, M., Marinelli, A., Sarli, G., and Della Salda, L. (2006). Heat shock protein expression in canine malignant mammary tumours. *BMC Cancer*, **6**:171.
- Roy, S. W. (2003). Recent evidence for the exon theory of genes. *Genetica*, **118**:251–266.
- Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**:1173–1178.
- Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**:309–323.
- Saeed, R. and Deane, C. M. (2006). Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*, **7**:128.
- Said, M. R., Begley, T. J., Oppenheim, A. V., Lauffenburger, D. A., and Samson, L. D. (2004). Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*, **101**:18006–18011.
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**:D449–D451.

- Sayle, R. A. and Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci.*, **20**:374–376.
- Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays—a technology review. *Nat Cell Biol.*, **3**:190–195.
- Segal, E., Wang, H., and Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**:i264–i272.
- Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*, **102**:1974–1979.
- Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**:1517–1521.
- de Souza, S. J. (2003). The Emergence of a Synthetic Theory of Intron Evolution. *Genetica*, **118**:117–121.
- de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S., and Gilbert, W. (1998). Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA*, **95**:5094–5099.
- Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol.*, **311**:681–692.
- Sprinzak, E., Sattath, S., and Margalit, H. (2003). How Reliable are Experimental Protein-Protein Interaction Data? *J Mol Biol.*, **327**:919–923.
- Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., and Soreq, H. (2005). Function of alternative splicing. *Gene*, **344**:1–20.
- Stanyon, C. A., Liu, G., Mangiola, B. A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., and Finley, R. L. J. (2004). A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol.*, **5**:R96.

- Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**:535–539.
- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., *et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**:957–968.
- Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M., and Doolittle, W. F. (1994). Testing the exon theory of genes: the evidence from protein structure. *Science*, **265**:202–207.
- Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA*, **102**:4221–4224.
- Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. (2006). A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**:360.
- Szilagyi, A., Grimm, V., Arakaki, A. K., and Skolnick, J. (2005). Prediction of physical protein-protein interactions. *Phys Biol.*, **2**:S1–S16.
- Takahashi, T., Ueno, H., and Shibuya, M. (1999). EGF activates protein kinase C-dependent, but Ras-independent Raf-MEK-MAP kinase pathway for DNA synthesis in primary endothelial cells. *Oncogene*, **18**:2221–2230.
- Teng, S.-C., Chen, Y.-Y., Su, Y.-N., Chou, P.-C., Chiang, Y.-C., Tseng, S.-F., and Wu, K.-J. (2004). Direct activation of HSP90A transcription by c-Myc contributes to c-Myc-induced transformation. *J Biol Chem.*, **279**:14649–14655.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, pages 4673–4680.
- Titz, B., Schlesner, M., and Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics*, **1**:111–121.

- Toby, G. G. and Golemis, E. A. (2001). Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods*, **24**:201–217.
- Tovchigrechko, A., Wells, C. A., and Vakser, I. A. (2002). Docking of protein models. *Protein Sci.*, **11**:1888–1896.
- Uetz, P. (2002). Two-hybrid arrays. *Curr Opin Chem Biol.*, **6**:57–62.
- Uetz, P., Dong, Y.-A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., *et al.* (2006). Herpesviral protein networks and their interaction with the human proteome. *Science*, **311**:239–242.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., and Pochart, P. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**:623–627.
- Ullrich, A., Gray, A., Tam, A. W., Yang-Feng, T., Tsubokawa, M., Collins, C., Henzel, W., Le Bon, T., Kathuria, S., and Chen, E. (1986). Insulin-like growth factor I receptor primary structure: comparison with insulin receptor suggests structural determinants that define functional specificity. *EMBO J.*, **5**:2503–2512.
- Valdar, W. S. and Thornton, J. M. (2001a). Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol.*, **313**:399–416.
- Valdar, W. S. and Thornton, J. M. (2001b). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**:108–24.
- Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struc Biol.*, **12**:368–373.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, Heidelberg.
- Vidal, M. (2005). Interactome modeling. *FEBS Lett.*, **579**:1834–1838.

- Vidalain, P.-O., Boxem, M., Ge, H., Li, S., and Vidal, M. (2004). Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, **32**:363–370.
- Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, **21**:4205–4208.
- Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**:116–122.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**:520–62.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**:440–442.
- Weng, G., Bhalla, U. S., and Iyengar, R. (1999). Complexity in biological signaling systems. *Science*, **284**:92–96.
- Whamond, G. S. and Thornton, J. M. (2006). An analysis of intron positions in relation to nucleotides, amino acids, and protein secondary structure. *J Mol Biol.*, **359**:238–247.
- Williamson, A. R. (2000). Creating a structural genomics consortium. *Nat Struct Biol.*, **S7**:953.
- Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**:D187–D191.
- Wu, C. H., Yeh, L.-S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B. E., *et al.* (2003). The Protein Information Resource. *Nucleic Acids Res.*, **31**:345–347.

- Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network. *Genome Res.*, **14**:1310–1314.
- Wuchty, S. and Almaas, E. (2005). Evolutionary cores of domain co-occurrence networks. *BMC Evol Biol.*, **5**:24.
- Wuchty, S., Oltvai, Z. N., and Barabasi, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet.*, **35**:176–179.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**:303–305.
- Yenush, L. and White, M. F. (1997). The IRS-signalling system during insulin and cytokine action. *Bioessays*, **19**:491–500.
- Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**:1107–1118.
- Yuan, Z., Burrage, K., and Mattick, J. S. (2002). Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**:566–570.
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: a Molecular INTERaction database. *FEBS Lett.*, **513**:135–140.
- Zhao, W., Serpedin, E., and Dougherty, E. R. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, **22**:2129–2135.